

SERIES TEMPORELLES LINEAIRES
ÉVOLUTION DU CHIFFRE D’AFFAIRES GENERE PAR LA
FABRICATION D’ARMES ET DE MUNITIONS EN FRANCE

Table des matières

1	Données	1
1.1	Présentation de la série	1
1.2	Stationnarisation par différenciation	1
2	Modélisation ARIMA	2
2.1	Choix d'un modèle $ARMA(p, q)$ pour la série transformée (X_t)	2
2.1.1	Modèles $ARMA(p, q)$ candidats	2
2.1.2	Comparaison de la performance des modèles candidats	3
2.1.3	Étude de la qualité d'ajustement du modèle $MA(1)$	3
2.1.4	Étude de la validité du modèle $MA(1)$	4
2.2	Choix d'un modèle $ARIMA(p, d, q)$ pour la série non transformée (Y_t)	4
3	Prédiction	4
3.1	Région de confiance de niveau α sur les valeurs futures X_{t+1} et X_{t+2}	5
3.1.1	Méthode 1 : Région bivariée	5
3.1.2	Méthode 2 : Région univariée	5
3.2	Hypothèses sous-jacentes	5
3.3	Visualisation de la région de confiance à 95%	6
3.4	Améliorer la prédiction de X_{T+1} à l'aide d'une autre variable disponible plus rapidement	6
A	Annexes	7
A.1	Annexe 1 - Autocorrélogrammes simple et partiel de la série non transformée (Y_t)	7
A.2	Annexe 2 - Test de racine unitaire sur la série non transformée (Y_t)	8
A.3	Annexe 3 - Tests de racine unitaire sur la série transformée (X_t)	9
A.3.1	Test ADF sur la série transformée (X_t)	9
A.3.2	Test KPSS sur la série transformée (X_t)	9
A.3.3	Test Philips-Perron sur la série transformée (X_t)	10
A.4	Annexe 4 - Transformation logarithmique de la série avant différenciation (non retenue)	11
A.5	Annexe 5 - ACF des résidus du modèle $MA(1)$ appliqué à la série différenciée	12
A.6	Annexe 6 - Normalité des résidus du modèle $MA(1)$ appliqué à la série différenciée	13
B	Code R	14

1 Données

1.1 Présentation de la série

L'objectif de ce projet est d'étudier l'évolution du chiffre d'affaires mensuel généré par la fabrication d'armes et de munitions en France, destinées au marché intérieur ou à l'export. Pour cela, nous disposons d'une série publiée par l'INSEE¹, qui comporte 253 données entre janvier 1999 et janvier 2020. Il s'agit d'un indice de chiffre d'affaires (ICA) qui permet de mesurer mensuellement, et de façon précoce, l'évolution de l'activité des entreprises en valeur. Les données sont des estimations réalisées à partir des déclarations TVA déposées par les entreprises chaque mois.

Selon la nomenclature d'activités française, cette série comprend le chiffre d'affaires généré par :

- la fabrication d'armes lourdes, légères, de chasse, de tir sportif, de défense ;
- la fabrication de munitions de guerre telles que les bombes, les mines, les roquettes ou les obus.

Les données mises à notre disposition sont en base 2015, c'est-à-dire qu'elles ont été normalisées de manière à ce que la moyenne des observations de l'année 2015 soit 100. Ainsi, par exemple, la dernière valeur disponible (179,98 en janvier 2020) signifie que, en janvier 2020, le chiffre d'affaires généré par ce secteur d'activité a été supérieur de 79,98% au chiffre d'affaires mensuel moyen produit par ce même secteur en 2015. Par ailleurs, les données utilisées ont été corrigées des variations saisonnières et des jours ouvrés (CVS-CJO).

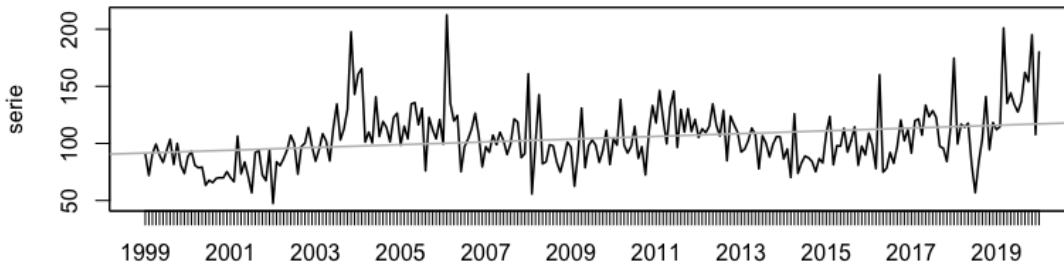


FIGURE 1 – Évolution de la série non transformée, notée (Y_t) , et droite de tendance

Ce graphique permet de visualiser la série avant transformation. Dans la suite, nous la noterons $(Y_t)_{t \in \mathcal{T}}$, avec \mathcal{T} l'ensemble des mois entre janvier 1999 et janvier 2020. Rappelons que cette série est corrigée de toute saisonnalité par l'INSEE, ce que semble confirmer son ACF en annexe 1, sur laquelle ne figure aucun pic proche de 1.

1.2 Stationnarisation par différenciation

Visuellement, la série ci-dessus semble présenter une certaine persistance, et son espérance ne semble pas constante. Ceci nous conduit à douter de sa stationnarité. Nous souhaitons tester l'existence d'une racine unitaire. Nous choisissons pour cela d'utiliser un test de Dickey-Fuller augmenté (ADF).

- Étape 1 : une régression linéaire de (Y_t) sur le temps nous indique qu'il existerait une constante et une tendance significatives (*cf.* droite grise sur la Fig. 1). Nous utiliserons donc un test ADF dans le cas avec constante et avec tendance. Nous rappelons que celui-ci consiste à tester l'hypothèse nulle $H_0 : \beta = 0$ dans la régression

suivante : $\Delta Y_t = c + bt + \beta Y_{t-1} + \sum_{l=1}^k \phi_l \Delta Y_{t-l} + \varepsilon_t$ avec k le nombre de retards nécessaires à considérer pour rendre les résidus non autocorrélés.

- Étape 2 : nous utilisons une fonction, programmée en amont, qui ajoute des retards (ce qui revient à incrémenter k dans la formule ci-dessus) jusqu'à rendre les résidus non autocorrélés d'après des tests de Ljung-Box². Celle-ci nous conduit à utiliser un test ADF avec $k = 6$ retards pour que celui-ci soit valide.
- Étape 3 : interprétation. Ce test ADF valide renvoie une p-valeur de 0,37 (*cf.* annexe 2) ce qui ne nous permet pas de rejeter l'hypothèse nulle d'existence d'une racine unitaire au seuil de confiance usuel de 95%.

Nous en concluons que (Y_t) n'est pas stationnaire, elle est donc au moins $\mathcal{I}(1)$. Nous choisissons de la différencier afin d'essayer d'y remédier : dans la suite, nous considérons donc $X_t = Y_t - Y_{t-1} = (1 - B)Y_t$ avec B l'opérateur retard de Y_t . Le graphique ci-dessous permet de visualiser la série transformée. À première vue, il semble déjà plus plausible d'affirmer que celle-ci serait stationnaire : elle évolue autour d'une droite horizontale et avec une variance relativement stable.

1. Série disponible à l'adresse <https://www.insee.fr/fr/statistiques/serie/010540139>, poste 25.4 de la NAF rév. 2

2. L'implémentation des tests de Ljung-Box d'absence d'autocorrélation sera détaillée dans la section 2.1.4.

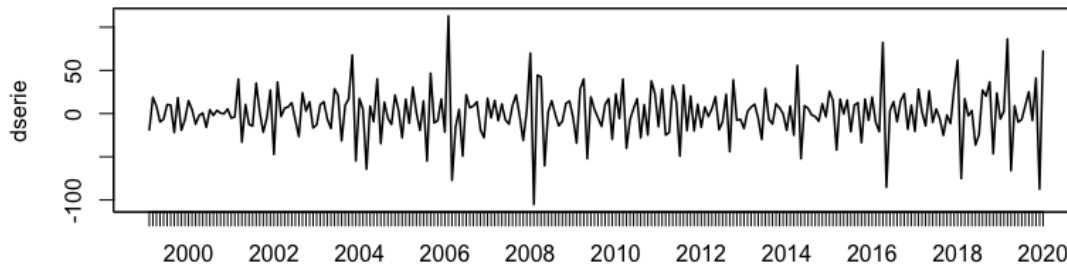


FIGURE 2 – Évolution de la série différenciée, notée (X_t)

Nous souhaitons de nouveau tester l'existence d'une racine unitaire. Comme précédemment, nous réalisons un test de Dickey-Fuller augmenté.

- Étape 1 : une régression sur le temps nous pousse à considérer un test sans tendance ni constante.
- Étape 2 : la fonction programmée en amont nous incite à considérer un test ADF avec 5 retards ce qui permet de ne jamais rejeter l'absence d'autocorrélation des résidus dans les tests de Ljung-Box³.
- Étape 3 : interprétation. Ce test ADF valide renvoie une p-valeur de 0,01 (cf. annexe 3) ce qui nous permet de rejeter l'hypothèse nulle d'existence d'une racine unitaire au seuil de confiance usuel de 95%.

Par acquit de conscience, nous réalisons également deux tests complémentaires qui livrent des conclusions similaires : le test Philips-Perron (p-valeur 0,01 - cf. annexe 3) nous incite à rejeter l'hypothèse nulle de non-stationnarité tandis que le test KPSS (p-valeur 0,10 - cf. annexe 3) ne rejette pas l'hypothèse nulle de stationnarité.

Enfin, en ce qui concerne l'hypothèse d'homoscédasticité, nécessaire à la stationnarité, l'allure graphique de la série (X_t) ci-dessus nous laisse penser que la variance est bien stable au fil du temps. Nous essayons néanmoins par précaution de recourir à une transformation logarithmique de la série originale avant de la différencier, afin de voir si celle-ci présente un intérêt éventuel. L'allure de cette transformation (cf. annexe 4), nous pousse à ne pas la retenir.

En conclusion : les tests de racine unitaire réalisés ci-dessus, conjugués à une allure graphique très favorable, nous conduisent à affirmer avec une certitude suffisante que la série différenciée (X_t) est bien stationnaire. On conclut que la série non transformée (Y_t) est $\mathcal{I}(1)$.

2 Modélisation ARIMA

2.1 Choix d'un modèle $ARMA(p, q)$ pour la série transformée (X_t)

Comme vu précédemment, il semble raisonnable de supposer que la série (X_t) est stationnaire. On souhaite déterminer le modèle le mieux ajusté et délivrant la meilleure qualité d'information parmi les différents modèles $ARMA(p, q)$ valides. Nous emploierons pour cela la méthodologie de Box-Jenkins⁴. Contrairement à la procédure originale décrite par les auteurs, nous commençons par sélectionner le meilleur modèle au sens de l'AIC et du BIC, puis nous vérifions son bon ajustement et sa validité. Si besoin, nous réitérons ce procédé jusqu'à obtenir un modèle délivrant la meilleure information possible tout en étant bien ajusté et valide. Cette approche, inspirée du TD5, est équivalente à la procédure de Box et Jenkins, qui suggèrent quant à eux d'étudier l'ajustement et la validité de tous les modèles candidats avant de comparer leur AIC et leur BIC. Elle peut s'avérer moins coûteuse en temps de calcul car le coût computationnel des critères d'information est généralement inférieur à celui des tests de validité.

2.1.1 Modèles $ARMA(p, q)$ candidats

La première étape consiste à déterminer les ordres maxima p^* et q^* vraisemblables pour la série transformée (X_t) . Il faut pour cela analyser les autocorrélogrammes simple et partiel de la série et observer quel est l'ordre significatif le plus élevé. En effet, si l'on note ρ et r les fonctions d'autocorrélation respectivement simple (ACF) et partielle (PACF), on sait que :

- Si (X_t) suit un $MA(q)$, alors $\forall h > q, \rho(h) = 0$
- Si (X_t) suit un $AR(p)$, alors $\forall h > p, r(h) = 0$

Ces autocorrélogrammes nous permettent de poser $p^* = 6$ (ordre maximum significatif de la PACF) et $q^* = 1$

3. L'implémentation des tests de Portmanteau d'absence d'autocorrélation sera détaillée dans la section 2.1.4.

4. *Time series analysis : forecasting and control*, George E. P. Box, Gwilym M. Jenkins, 1976

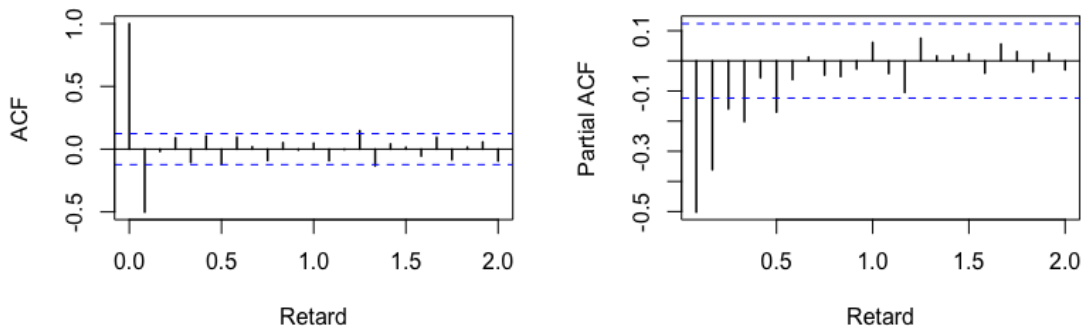


FIGURE 3 – Autocorrélogrammes simple et partiel de la série transformée (X_t)

(ordre maximum significatif de l'ACF)⁵. Les modèles envisagés à ce stade sont donc tous les $ARMA(p, q)$ avec $p \in \llbracket 0, 6 \rrbracket$ et $q \in \llbracket 0, 1 \rrbracket$, autrement dit, les modèles : $AR(1)$, $AR(2)$, $AR(3)$, $AR(4)$, $AR(5)$, $AR(6)$, $MA(1)$, $ARMA(1, 1)$, $ARMA(2, 1)$, $ARMA(3, 1)$, $ARMA(4, 1)$, $ARMA(5, 1)$, $ARMA(6, 1)$.

2.1.2 Comparaison de la performance des modèles candidats

On commence par calculer les AIC⁶ et BIC⁷ pour chacun des modèles candidats : ces critères permettent de comparer la qualité de l'information délivrée par les différents modèles statistiques. Un modèle est d'autant plus qualitatif que son critère d'information est faible.

	p=0	p=1	p=2	p=3	p=4	p=5	p=6
q=0		2 329	2 296	2 292	2 284	2 285	2 278
q=1	2 272	2 274	2 276	2 277	2 278	2 280	2 279

TABLE 1 – AIC des modèles $ARMA(p, q)$ candidats

	p=0	p=1	p=2	p=3	p=4	p=5	p=6
q=0		2 336	2 307	2 306	2 301	2 306	2 303
q=1	2 279	2 284	2 290	2 294	2 299	2 304	2 307

TABLE 2 – BIC des modèles $ARMA(p, q)$ candidats

On constate que le modèle $ARMA(0, 1)$ a le bon goût de minimiser à la fois l'AIC et le BIC. On choisit donc de focaliser notre attention sur ce modèle spécifique, dont on va vérifier le bon ajustement et la validité ci-après.

2.1.3 Étude de la qualité d'ajustement du modèle $MA(1)$

On commence par évaluer ce modèle, on choisit de ne pas inclure de constante dans la régression car celle-ci n'est pas significativement non nulle (cf. annexe 3).

```
Call:
arima(x = dserie, order = c(0, 0, 1), include.mean = F)
```

```
Coefficients:
      ma1
    -0.7987
s.e.    0.0378
```

5. En toute rigueur, il faudrait poser $q^* = 15$ car on remarque un pic significativement non nul sur l'ACF à cet ordre. On décide de l'ignorer. D'une part, prendre en compte ce pic alourdirait fortement notre étude : il nous faudrait alors évaluer $(6+1)*(15+1)-1 = 111$ modèles ARIMA possibles au lieu de $(6+1)*(1+1)-1 = 13$ puisque le précédent pic significatif est en $q=1$. D'autre part, l'autocorrélation d'ordre 15 dépasse tout juste le seuil de significativité : cela pourrait très bien résulter d'une aberration statistique.

6. Akaike information criterion, calculé à partir de la formule $AIC(p, q) = \log \hat{\sigma}_\varepsilon^2 + \frac{2(p+q)}{n}$

7. Bayesian information criterion, calculé à partir de la formule $BIC(p, q) = \log \hat{\sigma}_\varepsilon^2 + (p+q) \frac{\log n}{n}$

sigma^2 estimated as 472.9: log likelihood = -1134.09, aic = 2272.17

Pour vérifier la qualité d'ajustement d'un modèle $ARMA(p, q)$ dans le cas général, il convient de s'assurer que les coefficients associés aux ordres maxima p de la relation AR et q de la relation MA sont bien significatifs. Dans le cas d'un modèle $MA(1)$, cela revient donc à tester la significativité du seul coefficient. On calcule pour cela une statistique de test à partir de la sortie ci-dessus, en divisant le coefficient par son écart-type : $\frac{-0.7987}{0.0378} \approx -21, 1$. Cette valeur est nettement supérieure, en valeur absolue, à la valeur seuil de 1,96 (quantile à 97,5% d'une loi $\mathcal{N}(0, 1)$) ce qui nous permet de rejeter l'hypothèse nulle de nullité du coefficient $MA(1)$ au seuil de confiance usuel de 95%.

On conclut que le coefficient associé au $MA(1)$ est bien significatif. Ce modèle est donc bien ajusté et nous estimons, sous réserve de validité, que la série transformée (X_t) peut être modélisée par la relation suivante :

$$X_t = \varepsilon_t - 0,7987\varepsilon_{t-1} \tag{1}$$

2.1.4 Étude de la validité du modèle $MA(1)$

On teste l'absence d'autocorrélation des résidus en utilisant le test de Portmanteau (utilisant la statistique de Ljung-Box) dont l'hypothèse nulle est la "nullité jointe des autocorrélations" jusqu'à un ordre k donné, et dont la statistique de test $Q(k) = n(n+2) \sum_{l=1}^k \frac{\hat{\rho}_l^2}{n-l}$ suit une loi du χ_2 à $k-1$ degrés de liberté⁸, où n est le nombre d'observations, k est le nombre de retards à tester, $\hat{\rho}_l$ est l'autocorrélation empirique d'ordre l .

Retard	1	2	3	4	5	6	7	8	9	10	11	12
P-valeur	NA	0,51	0,51	0,52	0,61	0,56	0,48	0,58	0,54	0,62	0,71	0,79

Retard	13	14	15	16	17	18	19	20	21	22	23	24
P-valeur	0,69	0,76	0,51	0,44	0,51	0,58	0,64	0,62	0,64	0,69	0,73	0,74

TABLE 3 – P-valeurs retournées par les tests Portmanteau d'absence d'autocorrélation des résidus du modèle $MA(1)$

Pour nous assurer de l'absence d'autocorrélation, nous testons l'hypothèse nulle sur deux ans, soit 24 retards ici. Les p-valeurs, toutes largement supérieures à 0,05, nous conduisent à ne jamais rejeter l'hypothèse nulle d'absence d'autocorrélation des résidus au seuil de confiance usuel de 95%. En complément, nous traçons également l'ACF des résidus (*cf.* annexe 5) : celle-ci ne montre aucun pic significatif, ce qui est cohérent. On conclut que les résidus du modèle $MA(1)$ ne sont pas autocorrélés. Enfin, on vérifie à l'aide d'un test de Student que la moyenne des résidus est bien significativement nulle. Celui-ci renvoie une p-valeur de 0,36 qui nous conduit à ne pas rejeter cette hypothèse nulle au seuil de confiance usuel de 95%.

Ces éléments nous permettent d'affirmer que le modèle $MA(1)$ est bien valide.

2.2 Choix d'un modèle $ARIMA(p, d, q)$ pour la série non transformée (Y_t)

On a modélisé (X_t) avec un $MA(1)$ tel que $X_t = (1 - 0,7987B)\varepsilon_t$. Par définition, ce modèle est évidemment causal. La seule racine du polynôme MA est en dehors du disque unité puisqu'elle vaut $\frac{1}{0,7987} > 1$. Par définition, le processus (X_t) est donc également inversible⁹.

Par ailleurs, comme $X_t = (1 - B)Y_t$, on en déduit par définition que Y_t peut être modélisé par un $ARIMA(0, 1, 1)$ et on a, en remplaçant ci-dessus :

$$(1 - B)Y_t = \varepsilon_t - 0,7987\varepsilon_{t-1} \tag{2}$$

3 Prédiction

À partir du modèle $MA(1)$ que nous avons estimé ci-dessus et dont nous avons vérifié la pertinence, la validité et la qualité d'ajustement, nous souhaitons prédire les deux prochaines valeurs de la série, c'est-à-dire celles qui seront publiées pour les mois de février et mars 2020.

8. $k - p - q$ degrés de liberté dans le cas général pour un modèle $ARMA(p, q)$

9. On peut aisément l'inverser en remarquant que $\frac{1}{1-0,7987B} = \sum_{k=0}^{+\infty} 0,7987^k B^k$ et donc $\varepsilon_t = \sum_{k=0}^{+\infty} 0,7987^k X_{t-k}$

Nous posons ci-après $X_t = \varepsilon_t - \psi\varepsilon_{t-1}$ et nous faisons l'hypothèse que les résidus (ε_t) suivent tous une loi $\mathcal{N}(0, \sigma_\varepsilon^2)$. L'annexe 6 montre en effet que cette hypothèse est crédible compte tenu de notre jeu de données. Nous supposons en outre que ces résidus ne sont pas autocorrélés (hypothèse vérifiée ci-dessus), ce qui, dans le cas gaussien, équivaut à l'indépendance. On a donc $\varepsilon_t \underset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$.

3.1 Région de confiance de niveau α sur les valeurs futures X_{t+1} et X_{t+2}

Par propriété, ce modèle étant causal et inversible (cf. 2.2), et puisque $\forall h > 0, \mathbb{E}[\varepsilon_{T+h}|X_T, X_{T-1}, \dots] = 0$, alors la meilleure prévision de $X = \begin{pmatrix} X_{T+1} \\ X_{T+2} \end{pmatrix}$ au temps T est donnée par le vecteur $\hat{X} = \begin{pmatrix} \hat{X}_{T+1|T} \\ \hat{X}_{T+2|T} \end{pmatrix} = \begin{pmatrix} -\psi\varepsilon_T \\ 0 \end{pmatrix}$.

Le vecteur des erreurs de prédiction s'écrit donc :

$$X - \hat{X} = \begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} \\ X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} = \begin{pmatrix} \varepsilon_{T+1} \\ \varepsilon_{T+2} - \psi\varepsilon_{T+1} \end{pmatrix} \quad (3)$$

Puisque les résidus sont normaux par hypothèse, on en déduit que : $X - \hat{X} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right)$, où Σ est la matrice de variance-covariance telle que $\Sigma = \sigma_\varepsilon^2 \begin{pmatrix} 1 & -\psi \\ -\psi & 1 + \psi^2 \end{pmatrix}$. En effet, le vecteur $X - \hat{X}$ est une transformation affine du vecteur $\begin{pmatrix} \varepsilon_{T+1} \\ \varepsilon_{T+2} \end{pmatrix}$ qui est lui-même gaussien, puisque composé de variables normales indépendantes.

3.1.1 Méthode 1 : Région bivariée

Afin de déterminer une région de confiance, on utilise la propriété selon laquelle, si Z est un vecteur gaussien centré à n dimensions ($n \in \mathbb{N}^*$) ayant pour matrice de variance-covariance V , alors ${}^t Z V^{-1} Z \sim \chi^2(n)$. L'application directe¹⁰ de ce résultat nous permet d'écrire que :

$$\begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} & X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} \Sigma^{-1} \begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} \\ X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} \sim \chi^2(2) \quad (4)$$

On conclut que la région de confiance de niveau α sur les valeurs futures $\begin{pmatrix} X_{T+1} \\ X_{T+2} \end{pmatrix}$ s'écrit comme suit, en notant $q_{1-\alpha}(2)$ le quantile d'ordre $1 - \alpha$ d'une loi $\chi^2(2)$:

$$R_\alpha = \left\{ \begin{pmatrix} X_{T+1} \\ X_{T+2} \end{pmatrix} \in \mathbb{R}^2 \mid \begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} & X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} \Sigma^{-1} \begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} \\ X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} \leq q_{1-\alpha}(2) \right\} \quad (5)$$

3.1.2 Méthode 2 : Région univariée

De manière alternative, on peut déduire de ce qui précède que $\begin{cases} X_{T+1} - \hat{X}_{T+1|T} \sim \mathcal{N}(0, \sigma_\varepsilon^2) \\ X_{T+2} - \hat{X}_{T+2|T} \sim \mathcal{N}(0, \sigma_\varepsilon^2(1 + \psi^2)) \end{cases}$

Ainsi, la région de confiance de niveau α s'écrit, en notant $q_{1-\alpha}$ le quantile d'ordre $1 - \alpha$ d'une $\mathcal{N}(0, 1)$:

- Pour X_{T+1} : $R_{T+1, \alpha} = \{X_{T+1} \in \mathbb{R} \mid \hat{X}_{T+1|T} - q_{1-\frac{\alpha}{2}}\sigma_\varepsilon \leq X_{T+1} \leq \hat{X}_{T+1|T} + q_{1-\frac{\alpha}{2}}\sigma_\varepsilon\}$
- Pour X_{T+2} : $R_{T+2, \alpha} = \{X_{T+2} \in \mathbb{R} \mid \hat{X}_{T+2|T} - q_{1-\frac{\alpha}{2}}\sigma_\varepsilon\sqrt{1 + \psi^2} \leq X_{T+2} \leq \hat{X}_{T+2|T} + q_{1-\frac{\alpha}{2}}\sigma_\varepsilon\sqrt{1 + \psi^2}\}$

3.2 Hypothèses sous-jacentes

Les résultats énoncés ci-dessus reposent sur plusieurs hypothèses :

- Les résidus sont supposés gaussiens : l'annexe 6 de cette étude montre que cette hypothèse est plausible au vu de notre jeu de données.
- Les résidus sont supposés d'espérance nulle : le test de Student mené dans la section 2.1.4 nous confirme que cette hypothèse est compatible avec nos observations.

10. En toute rigueur, cette affirmation n'est licite que si la matrice Σ est bien inversible. Cette condition revient à supposer que $\det(\Sigma) = \sigma_\varepsilon^2 \neq 0$, ce qui est évident puisque la série (X_t) serait constante dans le cas contraire.

- Les résidus sont supposés de variance connue et constante : l'évolution temporelle décrite dans l'annexe 6 indique que cette dernière hypothèse semble réaliste.
- La série différenciée (X_t) est supposée suivre un modèle $MA(1)$: nous supposons que la modélisation faite en section 2 est exacte et qu'il n'y a pas d'erreur d'estimation.
- La série différenciée (X_t) est supposée inversible et causale, ce qui est compatible avec les estimations réalisées (cf. section 2.2).

3.3 Visualisation de la région de confiance à 95%

On utilise la méthode 2 ci-dessus permettant d'établir une région de confiance pour chaque prédiction. On estime, d'après la modélisation réalisée en amont, que $\psi \approx -0.7987$ et $\sigma_\varepsilon^2 \approx 21.75239$. On prédit les valeurs suivantes en février et en mars 2020 : $\hat{X}_{T+1} \approx -29.18403$ et $\hat{X}_{T+2} = 0$. On en déduit les intervalles représentés ci-dessous.

Le graphique de gauche présente les prédictions pour la série différenciée (X_t). Le graphique de droite présente celles pour la série non différenciée (Y_t), qui découlent immédiatement de ce qui précède, en faisant l'opération réciproque à celle de différenciation, ce qui revient à estimer un $ARIMA(0, 1, 1)$ pour (Y_t).

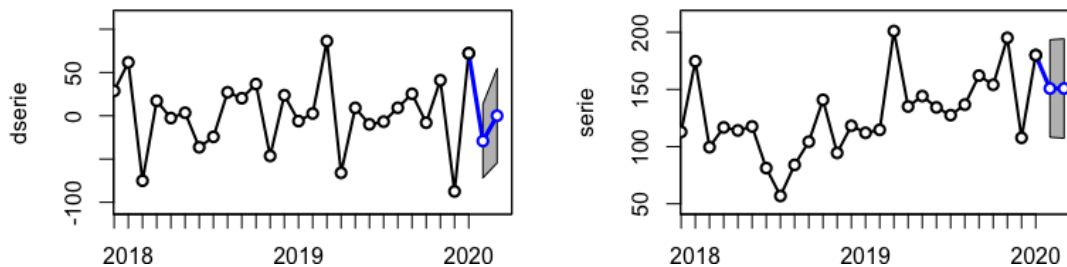


FIGURE 4 – Prédications (en bleu) et IC à 95% pour la série différenciée (gauche) et la série d'origine (droite)

Pour des raisons de lisibilité, ces graphiques reprennent les valeurs historiques des deux dernières années uniquement. Quelques commentaires :

- La prédiction est peu précise : pour la série différenciée, l'amplitude des intervalles de confiance est de 85 points en février et de 109 points en mars, ce qui est énorme. En outre, les intervalles de confiance contiennent des valeurs positives et négatives pour les deux prédictions. On est donc incapable de prédire avec un niveau de confiance satisfaisant si le chiffre d'affaires augmentera ou diminuera dans les deux prochains mois.
- De manière intuitive, l'amplitude des intervalles de confiance est plus grande en $T + 2$ qu'en $T + 1$: on a plus de difficultés à prédire des valeurs lointaines.
- Le modèle $MA(1)$ nous conduit à prédire une valeur nulle pour la série différenciée pour toutes les dates à partir de $T + 2$: cela revient à estimer que la série non transformée sera constante à partir de $T + 2$, ce qui est évidemment peu réaliste.

3.4 Améliorer la prédiction de X_{T+1} à l'aide d'une autre variable disponible plus rapidement

Soit (Z_t) une série stationnaire dont la $(T + 1)$ -ième valeur est disponible plus rapidement que X_{T+1} . Sous quelles conditions permet-elle d'améliorer la prédiction de X_{T+1} ?

Si Z_{T+1} est disponible plus rapidement que X_{T+1} , alors on peut utiliser l'information de Z_{T+1} afin d'améliorer la prédiction de X_{T+1} si et seulement si la variable (Z_t) cause instantanément, au sens de Granger, la variable (X_t). On rappelle qu'une série (Z_t) cause au sens de Granger une autre série (X_t) si, conditionnée aux valeurs passées de (Z_t), l'erreur quadratique moyenne de prédiction de X_{T+1} est inférieure par rapport à celle où les informations relatives aux valeurs passées de (Z_t) seraient omises. Mathématiquement :

$$\mathbb{E} [(X_{T+1} - \mathbb{E}(X_{T+1}|\cdot))^2 | X_T, X_{T-1}, \dots; Z_T, Z_{T-1}, \dots] \leq \mathbb{E} [(X_{T+1} - \mathbb{E}(X_{T+1}|\cdot))^2 | X_T, X_{T-1}, \dots]$$

Concrètement, dans notre étude : on a vu que l'erreur de prévision de la $(T + 1)$ -ième donnée pouvait s'écrire $X_{T+1} - \hat{X}_{T+1|T} = \varepsilon_{T+1}$, erreur que l'on souhaiterait réduire en valeur absolue. Une solution consisterait à disposer d'une série (Z_t) corrélée au résidu (ε_t). En régressant, on pourrait alors écrire que $\varepsilon_t = \alpha + \beta Z_t + \varepsilon'_t$ avec $\beta \neq 0$ et $\mathbb{E}[\varepsilon'_t] = 0$. On aurait ainsi $X_{T+1} = \alpha + \beta Z_{T+1} + \varepsilon'_{T+1} - \psi \varepsilon_T$, expression que l'on serait capable d'estimer au temps $T + 1$ puisque Z_{T+1} est connue de manière anticipée. Cette solution présente un intérêt sous l'hypothèse que $\mathbb{V}[\varepsilon'_t] < \mathbb{V}[\varepsilon_t]$.

A Annexes

A.1 Annexe 1 - Autocorrélogrammes simple et partiel de la série non transformée (Y_t)

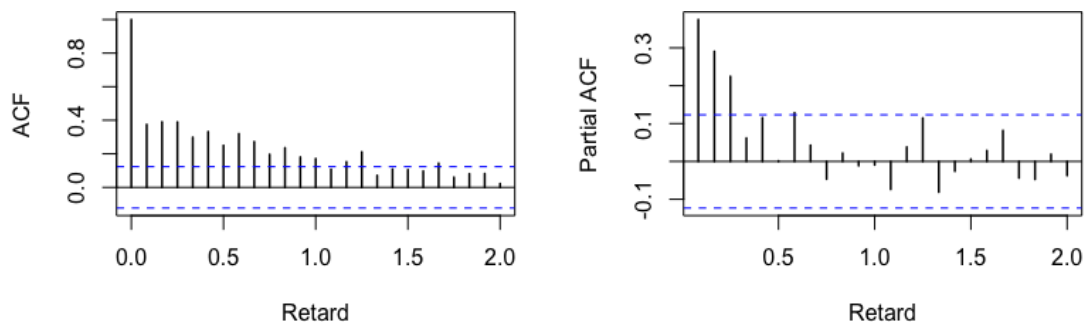


FIGURE 5 – ACF et PACF de la série non transformée (Y_t)

L'ACF, avec ses nombreuses autocorrélations significativement positives, illustre le phénomène de persistance observé sur la série non transformée, qui nous a poussé à la différencier.

A.2 Annexe 2 - Test de racine unitaire sur la série non transformée (Y_t)

Au préalable, on montre que la série non transformée (Y_t) présente une constante et une tendance significatives en la régressant sur le temps. Cette régression renvoie deux coefficients significatifs.

Call:

```
lm(formula = serie ~ index(serie))
```

Residuals:

```
   Min       1Q   Median       3Q      Max
-58.37 -16.10  -1.91  11.79 112.27
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2332.5268   511.0992  -4.564 7.87e-06 ***
index(serie)   1.2127     0.2543   4.768 3.16e-06 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.62 on 251 degrees of freedom

Multiple R-squared: 0.08305, Adjusted R-squared: 0.0794

F-statistic: 22.73 on 1 and 251 DF, p-value: 3.157e-06

On effectue un test ADF (avec constante et tendance) sur la série non transformée en incluant 6 retards dans la régression. L'algorithme a en effet permis de montrer que cela était le nombre minimum de retards à ajouter pour obtenir un test valide.

Title:

Augmented Dickey-Fuller Test

Test Results:

PARAMETER:

Lag Order: 6

STATISTIC:

Dickey-Fuller: -2.4789

P VALUE:

0.3742

On vérifie que les résidus de la régression effectuée par le test ADF ne sont pas autocorrélés à l'aide de tests Portmanteau.

Retard	1	2	3	4	5	6	7	8	9	10	11	12
P-valeur	NA	NA	NA	NA	NA	NA	NA	NA	NA	0,47	0,61	0,80

Retard	13	14	15	16	17	18	19	20	21	22	23	24
P-valeur	0,69	0,78	0,36	0,37	0,47	0,57	0,66	0,63	0,68	0,75	0,80	0,84

TABLE 4 – P-valeurs retournées par les tests Portmanteau de Ljung-Box

Aucun des tests Portmanteau (utilisant la statistique de Ljung-Box) ne rejette l'absence d'autocorrélation des résidus, le test ADF ci-dessus est donc bien valide et on peut l'interpréter. On ne rejette pas l'hypothèse nulle d'existence d'une racine unitaire.

A.3 Annexe 3 - Tests de racine unitaire sur la série transformée (X_t)

A.3.1 Test ADF sur la série transformée (X_t)

Au préalable, on montre que la série transformée (X_t) ne présente ni constante ni tendance significatives en la régressant sur le temps. Cette régression renvoie deux coefficients qui ne sont pas significatifs.

Call:

```
lm(formula = dserie ~ index(dserie))
```

Residuals:

	Min	1Q	Median	3Q	Max
	-105.451	-14.345	1.549	14.918	112.900

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-141.55287	592.88893	-0.239	0.811
index(dserie)	0.07062	0.29504	0.239	0.811

Residual standard error: 28.39 on 250 degrees of freedom

Multiple R-squared: 0.0002291, Adjusted R-squared: -0.00377

F-statistic: 0.05729 on 1 and 250 DF, p-value: 0.811

On effectue un test ADF (sans constante ni tendance) sur la série transformée en incluant 5 retards dans la régression. L'algorithme a en effet permis de montrer que cela était le nombre minimum de retards à ajouter pour obtenir un test valide.

Title:

Augmented Dickey-Fuller Test

Test Results:

PARAMETER:

Lag Order: 5

STATISTIC:

Dickey-Fuller: -10.4316

P VALUE:

0.01

On vérifie que les résidus de la régression effectuée par le test ADF ne sont pas autocorrélés à l'aide de tests Portmanteau (utilisant la statistique de Ljung-Box). Aucun des tests Portmanteau (utilisant la statistique de Ljung-Box) ne rejette

Retard	1	2	3	4	5	6	7	8	9	10	11	12
P-valeur	NA	NA	NA	NA	NA	NA	0,18	0,41	0,51	0,68	0,77	0,85
Retard	13	14	15	16	17	18	19	20	21	22	23	24
P-valeur	0,74	0,82	0,52	0,50	0,59	0,67	0,74	0,73	0,75	0,81	0,85	0,88

TABLE 5 – P-valeurs retournées par les tests Portmanteau de Ljung-Box

l'absence d'autocorrélation des résidus, le test ADF ci-dessus est donc bien valide et on peut l'interpréter. On rejette l'hypothèse nulle d'existence d'une racine unitaire. Les deux tests ci-dessous sont là pour confirmer ce résultat.

A.3.2 Test KPSS sur la série transformée (X_t)

KPSS Test for Level Stationarity

data: dserie

KPSS Level = 0.049064, Truncation lag parameter = 5, p-value = 0.1

Warning message:

```
In kpss.test(dserie) : p-value greater than printed p-value
```

On ne rejette pas l'hypothèse nulle de stationnarité au seuil de confiance usuel de 95%.

A.3.3 Test Philips-Perron sur la série transformée (X_t)

Phillips-Perron Unit Root Test

```
data: dserie
```

```
Dickey-Fuller Z(alpha) = -316.35, Truncation lag parameter = 5, p-value = 0.01
```

```
alternative hypothesis: stationary
```

On rejette l'hypothèse nulle d'absence de stationnarité au seuil de confiance usuel de 95%.

A.4 Annexe 4 - Transformation logarithmique de la série avant différenciation (non retenue)

Il est courant de passer en logarithme des séries temporelles pour "lisser les pics" et mieux contrôler une variance instable qui pourrait remettre en cause l'hypothèse d'homoscédasticité et de stationnarité. Nous comparons ici la série différenciée (haut) et la transformation logarithmique de la série différenciée (bas).

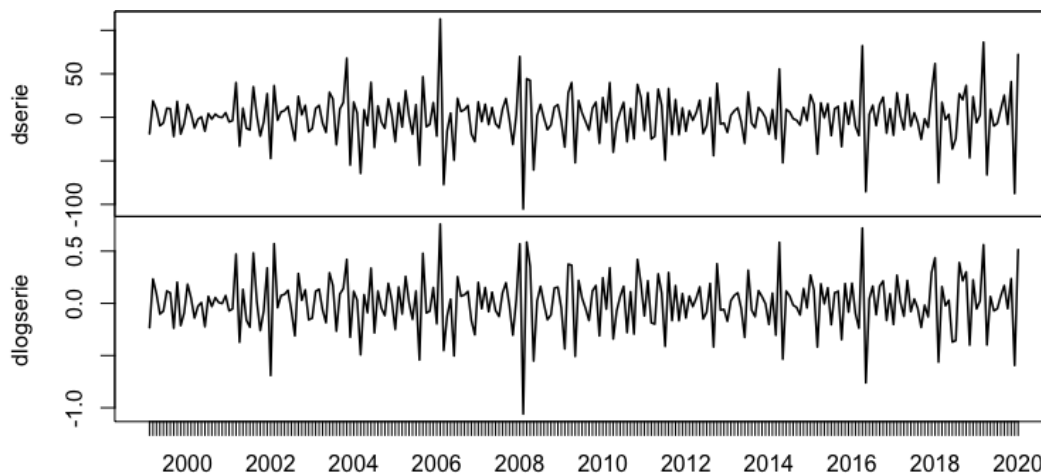


FIGURE 6 – Comparaison série différenciée vs. log(série) différenciée

On ne constate pas d'amélioration flagrante de la stabilité de la variance, et il ne nous semble pas pertinent de retenir cette transformation. On se concentrera donc sur la série en niveau différenciée, notée (X_t) .

A.5 Annexe 5 - ACF des résidus du modèle $MA(1)$ appliqué à la série différenciée

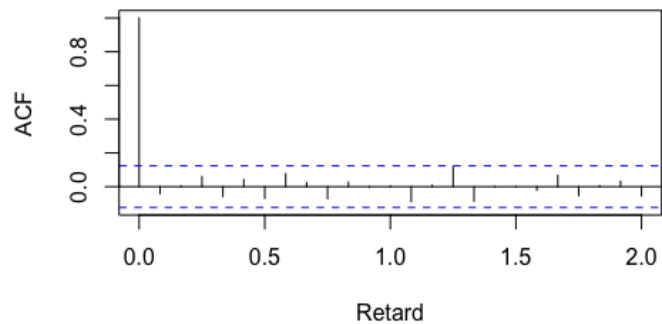


FIGURE 7 – ACF des résidus associés au modèle $MA(1)$ de la série différenciée (X_t)

Cet autocorrélogramme total ne signale aucune corrélation significative entre les résidus du modèle $MA(1)$ et leurs retards. Ce constat visuel est cohérent avec les tests Portmanteau menés dans notre étude, qui indiquent l'absence d'autocorrélation des résidus et la validité du modèle.

A.6 Annexe 6 - Normalité des résidus du modèle $MA(1)$ appliqué à la série différenciée

On s'intéresse ici aux résidus du modèle $MA(1)$ appliqué à la série différenciée, représentés ci-dessous.

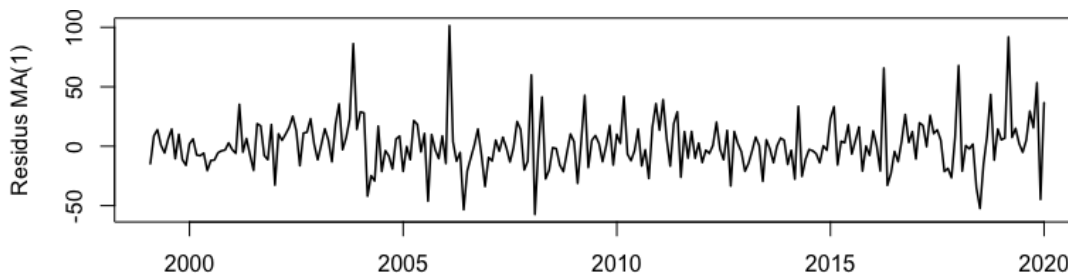


FIGURE 8 – Évolution temporelle des résidus du modèle $MA(1)$ appliqué à la série différenciée

Nous montrons ci-après que l'hypothèse de normalité des résidus semble crédible au vu de notre jeu de données.

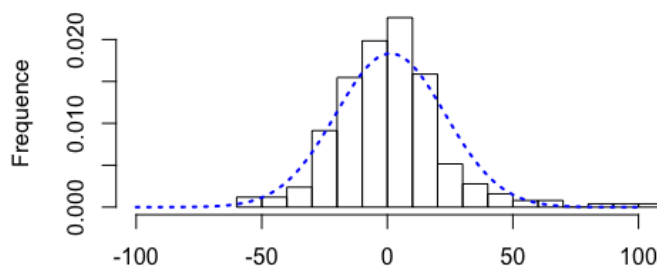


FIGURE 9 – Distribution des résidus du modèle $MA(1)$ vs. distribution théorique d'une loi normale (en bleu)

Cette courbe en cloche laisse penser que les résidus pourraient effectivement être normaux, sans certitude toutefois car celle-ci semble trop pointue et ses queues semblent un peu trop lourdes pour une gaussienne.

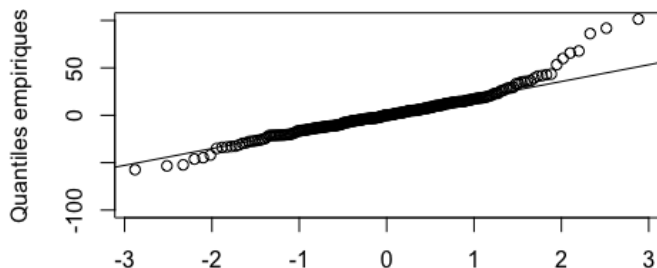


FIGURE 10 – QQ-plot comparant les quantiles de la distribution des résidus avec ceux d'une loi normale

Même conclusion : l'alignement des quantiles le long d'une droite laisse penser que la distribution pourrait être normale, sans certitude toutefois car les quantiles les plus extrêmes s'éloignent fortement de la distribution théorique. Cela pourrait être dû à des valeurs aberrantes mais la normalité des résidus est à nouveau remise en cause.

```
> ks.test(residus, "pnorm", mean=moy_residus, sd=sd_residus)
```

One-sample Kolmogorov-Smirnov test

```
data: residus
D = 0.082364, p-value = 0.06548
alternative hypothesis: two-sided
```

Le test de Kolmogorov-Smirnov d'ajustement à une loi continue ne rejette pas l'hypothèse nulle selon laquelle les résidus suivraient une loi normale au seuil usuel de 5%. Cette hypothèse reste donc plausible bien qu'incertaine.

B Code R

```
#####
# PROJET DE SERIES TEMPORELLES LINEAIRES - MODELISATION ARIMA #####
# PIERRE-ALEXANDRE ARANEGA - ENSAE PARIS - MAI 2020 #####
# FABRICATION D'ARMES ET DE MUNITIONS - SOURCE INSEE - 010540139 #####
# INDICE DE CHIFFRE D'AFFAIRES - MARCHE INTERIEUR ET EXPORT #####
# SERIE MENSUELLE CVS-CJO - FRANCE - BASE 100 EN 2015 #####
#####

#####
##### PARTIE 0 - IMPORT DE PACKAGES / DEFINITION DE FONCTIONS #####
#####

# setwd("your_directory")
require(zoo)
require(tseries)
require(fUnitRoots)

Qtests <- function(series, k, fitdf=0) {
  # Cette fonction, inspiree du TD5, realise des tests de Ljung-Box d'absence
  # d'autocorrelation des residus pour les lags entre 1 et k.
  # Elle renvoie les p-valeurs des tests realises pour chaque lag.
  FUN <- function(lag) {
    pval <- if (lag<=fitdf) NA else Box.test(series, lag, "Ljung-Box", fitdf)$p.value
    return(c("lag"=lag,"pval"=pval))
  }
  pvals <- apply(as.matrix(1:k), 1, FUN)
  return(t(pvals))
}

adfTest_valid <- function(series, kmax, adftype) {
  # Cette fonction, inspiree du TD5, realise des tests ADF successifs en ajoutant
  # des lags jusqu'a obtenir des residus qui ne soient pas autocorrees.
  # Elle renvoie le premier test ADF retenu qui soit valide.
  k <- 0 # nb de lags
  noautocorr <- FALSE
  while (noautocorr == FALSE) {
    cat(paste0('ADF avec ', k, ' lags: residus non autocorrees ? '))
    adf <- adfTest(series, lags=k, type=adftype)
    fitdf <- length(adf@test$lm$coefficients)
    pvals <- Qtests(adf@test$lm$residuals, 24, fitdf=fitdf)[,2]
    if (sum(pvals<0.05, na.rm=T)==0) {
      # Toutes les pvals sont >.05 donc on ne rejette pas l'absence d'autocor
      noautocorr <- TRUE
      cat('OUI - tous les LB tests sont accept s ! \n') # La boucle s'arr te alors
    } else cat('NON - au moins un LB test est rejet . \n')
    k <- k+1 # On ajoute un lag pour passer la boucle suivante.
  }
  return(adf) # Le dernier test fait
}

#####
##### PARTIE 1 - IMPORT ET DESCRIPTION DES DONNEES #####
#####

# IMPORT DES DONNEES
data <- read.csv("010540139.csv", sep=";", skip=3,
                colClasses = c("character", NA, "NULL"),
                col.names = c("Date", "Valeur", "Skip"))
serie <- zoo(data$Valeur, order.by = as.yearmon(data$Date))
dserie <- diff(serie, 1) # Calcul de la serie differenciee
```



```

61 # VISUALISATION DES DONNEES
62
63 titre_graph = "Evolution mensuelle de la serie non transformee"
64 plot(serie, main=titre_graph, xlab="", lwd=1.5)
65 abline(lm(serie ~ index(serie)), col="grey", lwd=1.5)
66 # * serie ne semble pas stationnaire : elle n'evolue pas le long
67 # d'une droite, son esperance semble varier avec le temps, et
68 # sa variance ne semble pas constante.
69 titre_graph = "Evolution mensuelle de la serie transformee"
70 plot(dserie, main=titre_graph, xlab="", lwd=1.5)
71 # * dserie semble stationnaire : elle evolue le long de l'axe
72 # des abscisses (cste=0) et sa variance semble plus stable.
73
74 # MONTRONS QUE LA SERIE N'EST PAS I(0)
75 # Testons existence d'une constante et d'une tendance significatives.
76 summary(lm(serie ~ index(serie))) # Les deux coefs sont significatifs a 95%
77 # Test ADF de racine unitaire dans le cas avec constante et tendance (ct).
78 adf <- adfTest_valid(serie, 24, adftype='ct')
79 # Il a ete necessaire d'ajouter 6 lags pour obtenir un test ADF
80 # valide (sans autocorrelation des residus).
81 fitdf <- length(adf@test$lm$coefficients)
82 Qtests(adf@test$lm$residuals, 24, fitdf=fitdf)
83 # Petite verification : en effet, avec 6 lags (+ 3 parametres c, t, beta),
84 # on ne rejette jamais l'absence d'autocor des residus.
85 # -> On peut desormais interpreter ce test ADF valide.
86 adf
87 # H0 (existence d'une racine unitaire) n'est pas rejetee au seuil de 95%
88 # (p-val > .05) -> la serie presente une racine unitaire.
89 # -> elle n'est pas I(0). Diff rencions -la.
90
91 # MONTRONS QUE LA SERIE EST I(1)
92 summary(lm(dserie ~ index(dserie))) # Aucun coef significatif a 95%.
93 # On va donc realiser un test ACF sans constante ni tendance (type "nc").
94 adf <- adfTest_valid(dserie, 24, adftype='nc')
95 # Il faut ajouter 5 lags pour obtenir un test ADF valide
96 # (ie. sans autocorrelation des residus).
97 fitdf <- length(adf@test$lm$coefficients)
98 Qtests(adf@test$lm$residuals, 24, fitdf=fitdf) # Petite verification
99 # En effet, avec 5 lags (+ 1 parametre beta), on ne rejette jamais
100 # l'absence d'autocorrelation des residus.
101 adf # H0 (existence d'une racine unitaire) est rejetee au seuil de 95% (p-val <.05).
102 # En compl ment : on r alise un test KPSS et Philips-Perron.
103 kpss.test(dserie) # On ne rejette pas l'hypothese H0 "dserie est stationnaire".
104 pp.test(dserie) # On rejette l'hypothese H0 "dserie n'est pas stationnaire"
105
106 # CONCLUSION : LA SERIE DIFFERENCIEE N'A NI CONSTANTE NI TENDANCE.
107 # LES TESTS ADF ET KPSS SEMBLent INDIQUER QUE dserie EST STATIONNAIRE,
108 # IE. serie EST INTEGREE D'ORDRE 1.
109
110 #//////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////#
111 ##### PARTIE 2 - MODELISATION ARIMA #####
112 #//////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////#
113
114 # DETERMINONS LES MODELES CANDIDATS
115 par(mfrow=c(1,2))
116 acf(dserie, 24, main = "ACF de la serie differenciee", xlab="Retard", lwd = 1.5)
117 pacf(dserie, 24, main = "PACF de la serie differenciee", xlab="Retard", lwd = 1.5)
118 pmax = 6 ; qmax = 1 # Il y a donc 13 mod les possibles (7*2-1)
119
120 # COMPARONS LES MODELES CANDIDATS A L'AIDE DES CRITERES AIC ET BIC
121 pqs <- expand.grid(0:pmax, 0:qmax) # Combinaisons possibles de p<=p* et q<=q*
122 mat <- matrix(NA, nrow=pmax+1, ncol=qmax+1) # matrice qui contiendra les AIC et BIC
123

```

```

rownames(mat) <- paste0("p=", 0:pmax)
colnames(mat) <- paste0("q=", 0:qmax)
AICs <- mat # matrice ou assigner les AICs
BICs <- mat
for (row in 1:dim(pqs)[1]) {
  # Pour chq combinaison (p,q) on va calculer l'ARIMA correspondant puis AIC et BIC
  p <- pqs[row, 1]
  q <- pqs[row, 2]
  estim <- try(arima(dserie, c(p, 0, q), include.mean=F)) # on a vu que pas de cste
  # try() permet de ne pas bloquer si erreur, notamment pour la 1e occurrence (0,0,0)
  AICs[p+1, q+1] <- if (class(estim) == "try-error") NA else estim$aic
  BICs[p+1, q+1] <- if (class(estim) == "try-error") NA else BIC(estim)
}
AICs == min(AICs) # min en p=0, q=1
BICs == min(BICs) # min en p=0, q=1
# Le modele ARIMA(0,1,1) minimise les criteres AIC et BIC

# VERIFIONS LA QUALITE D'AJUSTEMENT DE CE MODELE
darima001 <- arima(dserie, c(0, 0, 1), include.mean = F) # pas de cste signif (cf supra)
darima001$coef / sqrt(darima001$var.coef) # Statistique de test > 1.96 en val absolue.
# -> On ne peut pas ameliorer la qualite d'ajustement du modele.

# VERIFIONS LA VALIDITE DU MODELE (RESIDUS NON AUTOCORRELES)
dresidus = darima001$residuals
Qtests(dresidus, 24, fitdf=length(darima001$coef)) # fitdf=1-> 1 deg de liberte car 1
  param.
# Toutes les p-val des tests de Ljung Box sont >.05 donc on ne rejette jamais H0
# (absence d'autocor) -> les residus sont bien autocorrees et le modele est donc valide
# Confirmation graphique :
dev.off() ; acf(dresidus, main="ACF des residus du MA(1) applique a dserie",
  xlab="Retard")
t.test(dresidus, mu=0)
# Le test de Student indique qu'on ne rejette pas la nullite de la moyenne des residus
# Conclusion : ces r sultats confirment la blancheur des r sidus.

##### PARTIE 3 - PREVISIONS A PARTIR DU MODELE ARIMA #####
#####

# OBSERVONS QUE LES RESIDUS SUIVENT UNE DISTRIBUTION NORMALE
plot(dresidus, main="Residus du modele MA(1)", ylab="Residus MA(1)", lwd=1.5)
moy_residus = mean(dresidus)
sd_residus = sd(dresidus)
titre_graph = "Distrib des residus ARIMA \n vs. distrib theorique d'une loi normale"
hist(dresidus, breaks=20, xlim=c(-100,100), prob=T, xlab="", ylab="Frequence",
  main = titre_graph)
plot(function(x){dnorm(x, mean=moy_residus, sd=sd_residus)},
  -100, 100, add=T, lwd=2, lty=3, col="blue")
titre_graph = "Quantiles des residus ARIMA vs. quantiles theoriques d'une loi normale"
qqnorm(dresidus, ylim=c(-100,100), main=titre_graph,
  xlab="Quantiles theoriques", ylab="Quantiles empiriques") ; qqline(dresidus)
ks.test(dresidus, "pnorm", mean=moy_residus, sd=sd_residus)
# Le test de Kolmogorov-Smirnov ne rejette pas la normalite des residus au seuil de 95%.

# PREDICTION DES 2 PROCHAINES VALEURS DE dserie A L'AIDE D'UN MA(1)
dprediction <- predict(darima001, 2)
dserie_pred <- zoo(dprediction$pred, frequency=12)
dserie_se <- zoo(dprediction$se, frequency=12)
print(dserie_pred)

# INTERVALLES DE CONFIANCE EN UTILISANT LES ECARTS TYPES CALCULES AUTOMATIQUEMENT
IC_sup <- dserie_pred + qnorm(0.975) * dserie_se
IC_inf <- dserie_pred - qnorm(0.975) * dserie_se

```

```

# REMARQUONS QUE L'ON OBTIENT LES MEMES VALEURS AVEC LA FORMULE THEORIQUE (cf Rapport PDF)
dserie_pred[1] - qnorm(0.975) * sd_residus # Borne inf T+1
dserie_pred[1] + qnorm(0.975) * sd_residus # Borne sup T+1
dserie_pred[2] - qnorm(0.975) * sd_residus * sqrt(1+darima001$coef^2) # Borne inf T+2
dserie_pred[2] + qnorm(0.975) * sd_residus * sqrt(1+darima001$coef^2) # Borne sup T+2

# REPRESENTATION GRAPHIQUE
dates_prev <- index(dserie_pred)
plot(dserie, xlim=c(2018, max(dates_prev)), type="o", pch=21, bg='white', lwd=2, xlab="",
     main = "Prediction des 2 prochaines valeurs de la serie differenciee avec IC a 95%")
polygon(x=c(dates_prev[1], dates_prev[2], dates_prev[2], dates_prev[1]),
       y=c(as.numeric(IC_inf[1]), as.numeric(IC_inf[2]), as.numeric(IC_sup[2]),
          as.numeric(IC_sup[1])),
       col="grey")
lines(c(tail(dserie,1), dserie_pred), col="blue", lwd=3)
points(dserie_pred, col="blue", lwd=2, type="o", pch=21, bg='white')
points(tail(dserie, 1), type="o", pch=21, bg='white', lwd=2)

# ON FAIT DE MEME POUR LA SERIE NON DIFFERENCIEE
arima011 <- arima(serie, c(0, 1, 1), include.mean = F)
prediction <- predict(arima011, 2)
serie_pred <- zoo(prediction$pred, frequency=12)
serie_se <- zoo(prediction$se, frequency=12)
print(serie_pred)
IC_sup <- serie_pred + 1.96*serie_se
IC_inf <- serie_pred - 1.96*serie_se
dates_prev <- index(serie_pred)
plot(serie, xlim=c(2018, max(dates_prev)), type="o", pch=21, bg='white', lwd=2, xlab="",
     main = "Prediction des 2 prochaines valeurs de la serie originale avec IC a 95%")
polygon(x=c(dates_prev[1], dates_prev[2], dates_prev[2], dates_prev[1]),
       y=c(as.numeric(IC_inf[1]), as.numeric(IC_inf[2]), as.numeric(IC_sup[2]),
          as.numeric(IC_sup[1])),
       col="grey")
lines(c(tail(serie,1), serie_pred), col="blue", lwd=3)
points(serie_pred, col="blue", lwd=2, type="o", pch=21, bg='white')
points(tail(serie, 1), type="o", pch=21, bg='white', lwd=2)

# REMARQUONS QUE L'ON OBTIENT LES MEMES VALEURS EN CALCULANT MANUELLEMENT DEPUIS dserie
prev1 <- as.numeric(dserie_pred[1]) + as.numeric(tail(serie, 1))
prev2 <- as.numeric(dserie_pred[2]) + prev1
zoo(c(prev1, prev2), order.by=index(dserie_pred))

```