GUYBERT BAPTISTE - BAUD CANDICE

ENSAE

PROJET DE SÉRIES TEMPORELLES

ANNÉE SCOLAIRE 2022-2023

Analyse en séries temporelles de l'indice brut de la production industrielle concernant la fabrication de pesticides et d'autres produits agrochimiques

Mars 2023

Table des matières

1 Partie 1 : Les données

1.1 Question 1

La série sur laquelle nous travaillons est issue des données de l'INSEE [3] et correspond à l'indice brut de la production industrielle de la fabrication de pesticides et autres produits chimiques (NAF rév. 2, niveau groupe, poste 20.2). Ces données sont indicées en base 100 avec pour valeur de référence celle de 2015. D'après le gouvernement français[1], cette classe comprend :

- la fabrication d'insecticides, de rodonticides, de fongicides, d'herbicides, d'acaricides, de molluscicides, de biocides
- la fabrication d'inhibiteurs de germination, de régulateurs de croissance pour plantes
- la fabrication de désinfectants (à usage agricole ou autre)
- la fabrication d'autres produits agrochimiques n.c.a

Cette classe fait partie du secteur de l'industrie chimique et a été corrigée des variations saisonnières et des jours ouvrés.

1.2 Question 2

La série contient 397 observations et s'étend donc sur plus de 30 ans. Nous ne prenons en compte que les 250 valeurs les plus récentes afin de capter les dynamiques les plus récentes de la série. De plus, nous isolons les 2 dernières valeurs de la série qui nous permettront de faire des prédictions *out-of-sample* après avoir analysé et modélisé cette dernière.

La série comportant une tendance déterministe égale à 84.5, nous la retirons afin de centrer la série autour de 0. De plus, on remarque à l'aide d'une représentation ACF (auto-corrélation) et PACF (auto-corrélation partielle) une nette saisonnalité de 12, c'est-à-dire un motif se répétant annuellement.

La tendance déterministe de 84.5 nous indique que l'utilisation de pesticide a, en moyenne, diminué par rapport à celle de 1990. Cela peut s'expliquer aisément par une évolution de la règlementation de l'usage des pesticides plus stricte et par un essor de l'agriculture biologique depuis ces 10 dernières années. A titre d'exemple, depuis 1993, la mise sur le marché d'un nouveau pesticide est encadré par l'Union européenne et la loi Labbé de 2014 et mise à jour en 2017 et en 2022 interdit l'usage de produits phytosanitaires dans les espaces verts publics, les propriétés privées, les lieux fréquentés par le public et les lieux à usage collectif. Ainsi, il n'est pas surprenant d'observer une tendance inférieure à 100, qui indique plutôt une baisse de l'usage des pesticides depuis 1990.

La saisonnalité de notre série semble suivre le cycle des semis et des récoltes. En effet, les valeurs maximales de production de pesticides sont atteintes en janvier-février-mars, soit la période de semis de la plupart des légumes produits en France et les valeurs minimales sont atteintes en juin-juillet-août soit les périodes de récolte. La production de pesticides suit donc le calendrier des semis et suit ainsi la demande variable dans l'année. Ainsi, la saisonnalité de notre série semble parfaitement cohérente avec le cours de l'agriculture.

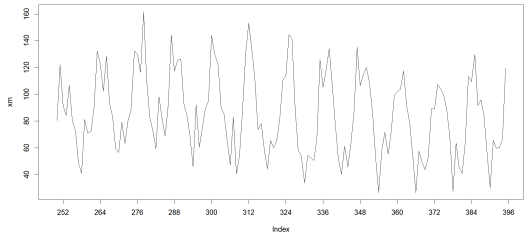
Afin d'obtenir une série sans saisonnalité, nous différencions à l'ordre 12 notre série. C'est-à-dire que nous effectuons l'opération suivante sur notre série X :

$$\Delta_{12} X_t = X_t - X_{t-12}$$

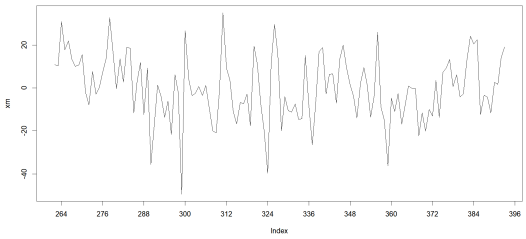
Ces opérations permettent d'obtenir une série stationnaire, sans tendance déterministe ni saisonnalité. En effectuant le test de Pilippe-Perron, on obtient une p-value égale à 0.01 ce qui permet de rejetter l'hypothèse selon laquelle la série n'est pas stationnaire.

1.3 Question 3

Nous représentons ci-dessous la série temporelle avant et après nos opérations. Les nouvelles ACF et PACF sont disponibles en annexe [4].



(a) Série avant opérations



(b) Série après opérations

FIGURE 1 – Avant et après opérations sur la série temporelle

2 Partie 2 : Modèles ARMA

2.1 Question 4

2.1.1 Méthodologie de Box Jenkins

Pour la modélisation de notre série temporelle, nous suivons la méthodologie de Box-Jenkins. Elle se compose des étapes suivantes :

— **Etape 1. Identification du ou des ordres des modèles** : Il s'agit de choisir les ordres maximums p* et q* de notre modèle ARMA(p,q) en utilisant les fonctions d'autocorrélation et d'autocorrélation partielle de la série

— **Etape 2. Estimation** : On détermine les coefficients de chaque modèle ARMA(p,q) $\forall\ p$ allant de 1 à p^* et $\forall$ q allant de 1 à q^*

— **Etape 3. Validité des modèles estimés** : On teste la significativité des coefficients et l'autocorrélation des résidus afin de sélectionner seulement les modèles valides.

— **Etape 4. Sélection du meilleur modèle parmi ceux ayant passé le test de validité** : A partir de l'AIC et du BIC on choisit le meilleur modèle parmi ceux valides.

2.1.2 Mise en application

Etape 1 : Identification des ordres des modèles. La série étant désormais corrigée, nous identifions à l'aide des auto-corrélations (complètes et partielles) les ordres possibles de notre modèle ARMA.

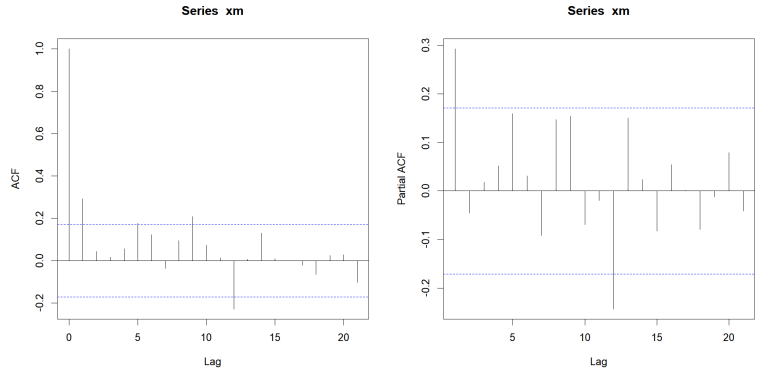


FIGURE 2 – Auto-corrélations de la série

On choisit nos paramètres maxima du modèle ARMA(p,q) en se basant sur la figure [2] :

— L'ACF nous permet d'identifier le paramètre q : on récupère le rang à partir duquel la série reste dans l'intervalle de confiance (représenté par les pointillés bleus) en omettant le premier lag. On choisit donc 11.

— La PACF nous permet d'identifier le paramètre p de manière analogue sans omettre le premier lag. On choisit donc ici 12.

Etape 2 : Estimation des modèles. Nous estimons l'ensemble des coefficients des modèles ARMA(p,q) pour tout p de 0 à 12 et tout q de 0 à 11.

Etape 3 : Validation des modèles. Parmi tous les modèles estimés, nous conservons uniquement ceux étant valides. Pour cela nous calculons la p-value de la t-statistique de chaque coefficient des modèles ARMA et nous considérons comme valides les modèles dont les coefficients sont significatifs au seuil de 5%. Nous effectuons également un test de Ljung Box sur les résidus de chaque modèle ARMA afin de tester leur autocorrélation. De même, nous considérons comme valides les tests significatifs au seuil de 5%.

Etape 4 : Sélection du meilleur modèle. Parmi les modèles valides trouvés à l'étape 3, nous sélectionnons le meilleur selon l'AIC [4](critère d'information d'Akaike) et le BIC [5]. Le modèle sélectionné est donc l'ARMA(5,11), car c'est celui ayant les plus faibles AIC et BIC parmi les modèles valides.

	ARMA(12,2)	ARMA(12,9)	ARMA(5,11)
AIC	1097	1103	1088
BIC	1144	1169	1141

TABLE 1 – Critères de sélection parmi les modèles valides

Nous estimons les coefficients en fittant le modèle ARMA(5,11) et les résumons dans la table [2]. Les coefficients sont significatifs aux niveaux usuels de 5% et même de 1% pour la plupart excepté pour les coefficients estimés du ma7 et ma8. L'intercept n'est pas significatif, ce qui correspond au fait que notre série est centrée et a donc une moyenne nulle.

	ar1	ar2	ar3	ar4	ar5	ma1	ma2	ma3	ma4	ma5	ma6
coef	-0.84	-0.77	-0.58	-0.56	-0.38	1.20	1.23	1.05	1.01	0.97	0.58
se	0.14	0.15	0.16	0.15	0.12	0.13	0.17	0.21	0.23	0.24	0.24
pval	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01

	ma7	ma8	ma9	ma10	ma11	intercept
coef	0.37	0.39	0.53	0.61	0.58	0.43
se	0.24	0.23	0.21	0.16	0.10	2.34
pval	0.12	0.10	0.01	0.00	0.00	0.86

TABLE 2 – Estimations du modèle

Les diagnostics du modèle sont disponibles en annexe [5] et montrent des résidus gaussiens, non corrélés, qui s'apparentent à un bruit blanc. Cette hypothèse n'est pas rejetée par le Q-test (Ljungbox) pour les ordres 1 à 24. L'hypothèse nulle de ce test est la présence d'un bruit blanc fort.

2.2 Question 5

Notre série différenciée à l'ordre 12 suit un modèle ARMA(p,q) avec p = 5, q = 11. En remplaçant avec les coefficients estimés, l'équation vérifiée par Z notre série différenciée est :

$$Z_t + 0.84Z_{t-1} + 0.77Z_{t-2} + 0.58Z_{t-3} + 0.56Z_{t-4} + 0.38Z_{t-5} = \epsilon_t - 1.20\epsilon_{t-1} - 1.23\epsilon_{t-2} - 1.05\epsilon_{t-3} - 1.01\epsilon_{t-4} - 0.97\epsilon_{t-5} - 0.58\epsilon_{t-6} - 0.37\epsilon_{t-7} - 0.39\epsilon_{t-8} - 0.53\epsilon_{t-9} - 0.61\epsilon_{t-10} - 0.58\epsilon_{t-11}$$

En notant Ψ et Φ les polynômes associés aux coefficients tels que

$$Z_t = \sum_{i=1}^{5} \phi_i X_{t-i} + \epsilon_t + \sum_{i=1}^{11} \psi_i \epsilon_{t-i} = \Psi(X_t) + \Phi(\epsilon_t)$$

où les ϕ_i et ψ_i sont estimés dans la table [2], et où $Z_t = \Delta_{12}(X_t - \mathbb{E}(X_t))$, on obtient que

$$(1 - B^{12})(X_t - \mathbb{E}(X_t)) = \Psi(X_t) + \Phi(\epsilon_t)$$

Notre série X_t suit donc un $SARIMA((5,0,11)(0,1,0)_{12})$.

De plus, en utilisant *polyroot*, il est possible d'identifier les racines du polynôme Φ associé à notre modèle ARMA(5,11) sur la série modifiée. En utilisant le cours, un modèle est causal si les racines de la partie AR en module sont toutes en dehors du cercle unité. Les racines obtenues [3] sont en dehors du cercle unité ce qui permet de conclure que le modèle est bien **causal**.

3 Partie 3 : Prévision

3.1 Question 6

Notons $(X_t)_{t\in\mathbb{R}}$ notre série temporelle différenciée à l'ordre 12 qui suit un ARMA(5,11) d'équation $X_t = \sum_{i=1}^{5} \phi_i X_{t-i} + \epsilon_t + \sum_{i=1}^{11} \psi_i \epsilon_{t-i}$, alors, les meilleures prédictions de X aux temps T+1 et T+2 sachant les X précédents sont

$$\hat{X}_{T+1|T} = \sum_{i=1}^{5} \phi_i X_{T+1-i} + \sum_{i=1}^{11} \psi_i \epsilon_{T+1-i}$$

$$\hat{X}_{T+2|T} = \phi_1 \hat{X}_{T+1|T} + \sum_{i=2}^{5} \phi_i X_{T+2-i} + \sum_{i=2}^{11} \psi_i \epsilon_{T+2-i}$$

En calculant [4.1] les différences et en écrivant sous forme matricielle on obtient :

$$\hat{\mathbf{X}} = \begin{pmatrix} \hat{X}_{T+1|T} \\ \hat{X}_{T+2|T} \end{pmatrix} et \mathbf{X} = \begin{pmatrix} X_{T+1} \\ X_{T+2} \end{pmatrix}$$

$$\text{alors, } \mathbf{X} - \hat{\mathbf{X}} = \begin{pmatrix} \tilde{X}_{T+1} \\ \tilde{X}_{T+2} \end{pmatrix} = \begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} \\ X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} = \begin{pmatrix} \epsilon_{T+1} \\ \epsilon_{T+2} + (\psi_1 + \phi_1)\epsilon_{T+1} \end{pmatrix}$$

On peut alors calculer la variance comme suit en utilisant le fait que $\forall t, \mathbb{V}(\epsilon_t) = \sigma^2$ et que les ϵ_t ne sont pas corrélés entre eux (c'est-à dire que ce sont des bruits blancs [5])

$$\mathbb{V}(X_{T+1} - \hat{X}_{T+1|T}) = \mathbb{V}(\epsilon_{T+1}) = \sigma^2$$

$$\mathbb{V}(X_{T+2} - \hat{X}_{T+2|T}) = \mathbb{V}(\epsilon_{T+2} + (\psi_1 + \phi_1)\epsilon_{T+1}) = \sigma^2(1 + (\psi_1 + \phi_1)^2)$$

En définissant le vecteur $\tilde{\mathbf{X}} = (\tilde{X}_{T+1}, \tilde{X}_{T+2})'$, celui-ci suit une loi $\mathcal{N}(0, \Sigma)$ où

$$\Sigma = \begin{bmatrix} \sigma^2 & (\psi_1 + \phi_1)\sigma^2 \\ (\psi_1 + \phi_1)\sigma^2 & (1 + (\psi_1 + \phi_1)^2)\sigma^2 \end{bmatrix}$$

Alors, la matrice est inversible si et seulement si $\det(\Sigma) \neq 0$ [1], et dans ce cas,

$$\tilde{\mathbf{X}}^T \Sigma^{-1} \tilde{\mathbf{X}} \sim \chi(2)$$

La région de confiance au niveau α associée est : $\left\{ X \in \mathbb{R}^2 | \tilde{\mathbf{X}}^T \Sigma^{-1} \tilde{\mathbf{X}} \leq q^{1-\alpha}_{\chi(2)} \right\}$

Autrement dit, l'équation pour la région de confiance de niveau α sur la valeur X_{T+1} est :

$$\hat{X}_{T+1|T} \pm Z_{\alpha/2} \times \hat{\sigma}^2$$

Et l'équation pour la région de confiance univariée de niveau α sur la valeur future X_{T+2} est :

$$\hat{X}_{T+2} \pm Z_{\alpha/2} \times \hat{\sigma}^2 \times \sqrt{(1 + (\hat{\psi}_1 + \hat{\phi}_1)^2)}$$

Où $Z_{\alpha/2}$ est le quantile de la distribution normale standard pour un niveau de confiance $\alpha/2$.

3.2 Question 7

Les hypothèses nécessaires sont :

1. Ce qui est le cas si $\sigma^2 \neq 0$ car $det(\Sigma) = \sigma^2(\psi_1^2 + \phi_1^2 + 2)$

— La stationnarité de la série, au moins au second ordre (moyenne et autocovariance de la série ne dépendant pas du temps),

— Le processus des innovations est gaussien i.i.d. (ie le bruit blanc ϵ_t suit une loi $\mathcal{N}(0, \sigma^2)$ où $\sigma^2 \neq 0$),

— Le fait que notre modèle est parfaitement identifié, c'est-à-dire que les coefficients trouvés lors des premières parties sont les coefficients réels (ou du moins que les estimateurs sont convergents).

Si la variance du résidu est inconnue, on doit l'estimer, et l'intervalle de confiance dé- pendra alors de la répartition d'une loi de Student, avec des queues plus larges et donc un intervalle moins précis.

3.3 Question 8

Nous représentons en figure [3] notre série réelle en noir, les valeurs prédites en rouge, et l'intervalle de confiance à 95% en zone grisée.

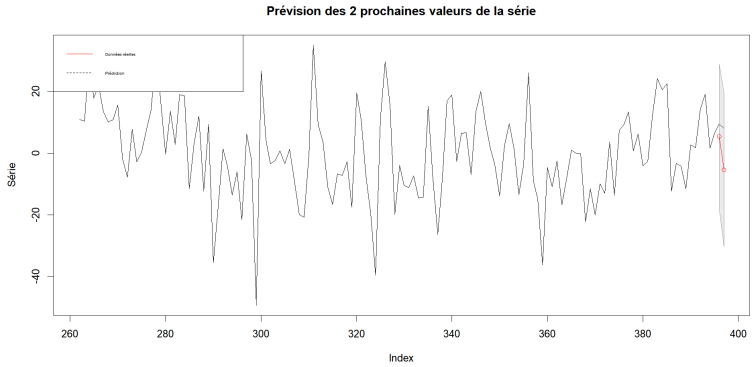


FIGURE 3 – Prévision des 2 prochaines valeurs de la série

3.4 Question 9

Si l'on dispose de l'observation Y_{T+1} d'une série stationnaire Y_t avant de disposer de l'observation X_{T+1} de la série temporelle X_t que l'on souhaite prédire, il est possible d'utiliser cette information pour améliorer la prédiction de X_{T+1}.

Plus précisément, l'utilisation de l'observation Y_{T+1} permettrait d'améliorer la prédiction de X_{T+1} si : (Y_t) cause (X_t) instantanément au sens de Granger. En effet, la causalité de Granger de X_t sur Y_t est explicitement définie dans le cours [2] comme le fait que X_t soit utile à la prédiction de Y_t. C'est à dire si :

$$\hat{Y}_{t+h|X_u,Y_u,u\leq t} \neq \hat{Y}_{t+h|Y_u,u\leq t}$$

pour au moins un h strictement positif et un t entier relatif.

Afin de tester cette condition, il est possible de faire un test de significativité globale des coefficients associés aux valeurs passées de la variable causale dans l'équation de la variable causée.

Références

[1] Gouvernement FRANÇAIS. *Nomenclature d'Activités Française de l'Artisanat - Révision 2*. 2008. URL : https://www.entreprises.gouv.fr/files/files/directions_services/secteurs-professionnels/artisanat/nafa/nomenclature_nafa_rev2.pdf.

[2] Christian FRANCQ. *Linear Times series*. Chapitre 5 : Var and Cointegration. 2023.

[3] INSEE. *Indice brut de la production industrielle (base 100 en 2015) - Fabrication de pesticides et d'autres produits agrochimiques (NAF rév. 2, niveau groupe, poste 20.2)*. 2023. URL : https://www.insee.fr/fr/statistiques/serie/010537439.

4 Annexes

4.1 Calculs

Ci-dessous sont présentés les calculs pour obtenir les valeurs pour $X_{T+1} - \hat{X}_{T+1|T}$ et $X_{T+2} - \hat{X}_{T+2|T}$.

D'une part, $X_{T+1} - \hat{X}_{T+1|T} = \sum_{i=1}^{5} \phi_i X_{T+1-i} + \epsilon_{T+1} + \sum_{i=1}^{11} \psi_i \epsilon_{T+1-i} - (\sum_{i=1}^{5} \phi_i X_{T+1-i} + \sum_{i=1}^{11} \psi_i \epsilon_{T+1-i}) = \epsilon_{T+1}$

D'autre part,

$X_{T+2} - \hat{X}_{T+2|T} =$

$\sum_{i=1}^{5} \phi_i X_{T+2-i} + \epsilon_{T+2} + \sum_{i=1}^{11} \psi_i \epsilon_{T+2-i} - (\phi_1 \hat{X}_{T+1|T} + \sum_{i=2}^{5} \phi_i X_{T+2-i} + \sum_{i=2}^{11} \psi_i \epsilon_{T+2-i}) =$

$\sum_{i=1}^{5} \phi_i X_{T+2-i} + \epsilon_{T+2} + \sum_{i=1}^{11} \psi_i \epsilon_{T+2-i} - (\phi_1 (X_{T+1} - \epsilon_{T+1}) + \sum_{i=2}^{5} \phi_i X_{T+2-i} + \sum_{i=2}^{11} \psi_i \epsilon_{T+2-i}) =$

$\phi_1 X_{T+1} + \sum_{i=2}^{5} \phi_i X_{T+2-i} + \epsilon_{T+2} + \sum_{i=1}^{11} \psi_i \epsilon_{T+2-i} - (\phi_1 X_{T+1} - \phi_1 \epsilon_{T+1} + \sum_{i=2}^{5} \phi_i X_{T+2-i} + \sum_{i=2}^{11} \psi_i \epsilon_{T+2-i}) =$

$\epsilon_{T+2} + \sum_{i=2}^{11} \psi_i \epsilon_{T+2-i} + \psi_1 \epsilon_{T+1} + \phi_1 \epsilon_{T+1} - \sum_{i=2}^{11} \psi_i \epsilon_{T+2-i}) = \epsilon_{T+2} + \epsilon_{T+1} \phi_1 + \epsilon_{T+1} \psi_1$

d'où les résultats affichés.

Les termes colorés sont ceux qui se simplifient entre eux.

4.2 Annexes

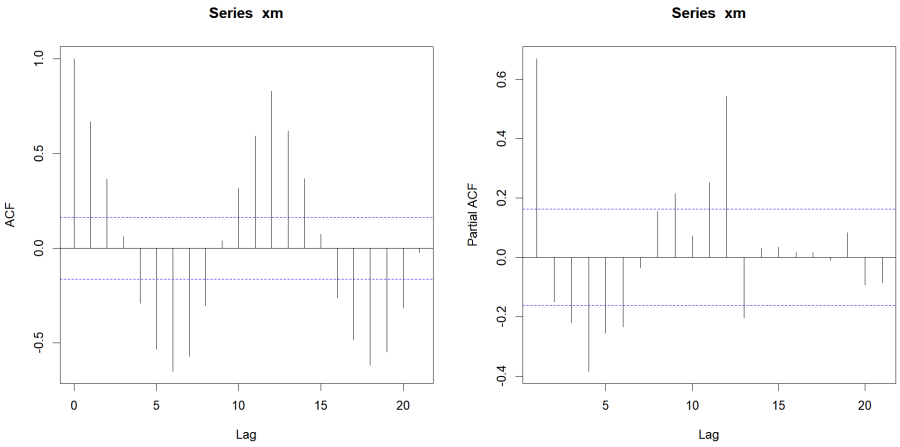


FIGURE 4 – Auto-corrélations de la série avant traitement

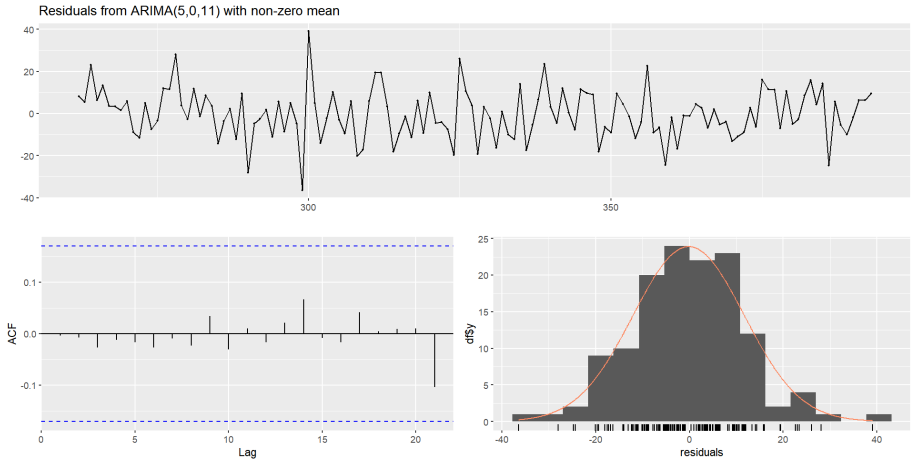


FIGURE 5 – Résidus du modèle

Racines :	1.2	1.26	1.26	1.2

TABLE 3 – Module des racines du polynôme Φ

	q=1	q=2	q=3	q=4	q=5	q=6	q=7	q=8	q=9
p=1	1099.06	1099.52	1101.34	1102.75	1101.67	1101.81	1103.26	1099.73	1102.12
p=2	1100.39	1101.42	1103.46	1103.74	1103.54	1100.70	1105.23	1101.60	1101.68
p=3	1100.83	1101.73	1103.47	1098.62	1098.42	1099.64	1101.96	1100.72	1094.32
p=4	1102.01	1101.48	1101.29	1096.97	1098.94	1100.55	1102.43	1104.27	1096.16
p=5	1102.74	1104.74	1099.25	1098.94	1096.87	1099.54	1100.96	1102.71	1097.35
p=6	1104.04	1100.06	1102.54	1100.58	1098.61	1098.22	1099.75	1101.75	1097.43
p=7	1105.13	1101.91	1102.15	1102.54	1100.45	1099.75	1100.72	1102.51	1099.30
p=8	1103.08	1100.29	1102.21	1104.91	1104.08	1101.33	1102.60	1102.92	1099.56
p=9	1099.78	1101.08	1100.80	1097.00	1098.92	1100.91	1104.26	1103.41	1099.59
p=10	1103.82	1102.12	1101.98	1098.92	1100.98	1102.76	1106.25	1108.17	1103.17
p=11	1102.89	1102.75	1104.66	1100.90	1102.67	1102.29	1108.26	1107.29	1103.19
p=12	1098.78	1097.81	1096.20	1098.17	1098.30	1106.07	1105.51	1105.53	1103.14

	q=8	q=9	q=10	q=11
p=1	1099.73	1102.12	1103.92	1097.55
p=2	1101.60	1101.68	1104.19	1105.83
p=3	1100.72	1094.32	1105.68	1097.61
p=4	1104.27	1096.16	1098.14	1096.17
p=5	1102.71	1097.35	1098.41	1088.94
p=6	1101.75	1097.43	1099.38	1090.88
p=7	1102.51	1099.30	1101.08	1092.88
p=8	1102.92	1099.56	1101.27	1094.53
p=9	1103.41	1099.59	1101.57	1103.30
p=10	1108.17	1103.17	1103.60	1098.33
p=11	1107.29	1103.19	1105.23	1103.02
p=12	1105.53	1103.14	1100.12	1106.50

Table 4 – AIC des différents modèles

	q=1	q=2	q=3	q=4	q=5	q=6	q=7	q=8	q=9
1	1110.65	1114.01	1118.73	1123.04	1124.85	1127.89	1132.24	1131.60	1136.89
2	1114.88	1118.80	1123.75	1126.93	1129.62	1129.67	1137.11	1136.38	1139.36
3	1118.22	1122.02	1126.65	1124.70	1127.40	1131.51	1136.73	1138.39	1134.89
4	1122.30	1124.66	1127.37	1125.95	1130.81	1135.32	1140.10	1144.84	1139.63
5	1125.93	1130.82	1128.23	1130.82	1131.65	1137.21	1141.53	1146.18	1143.71
6	1130.12	1129.03	1134.42	1135.35	1136.28	1138.79	1143.22	1148.11	1146.69
7	1134.10	1133.79	1136.92	1140.21	1141.02	1143.22	1147.08	1151.77	1151.46
8	1134.96	1135.06	1139.88	1145.48	1147.54	1147.69	1151.86	1155.08	1154.62
9	1134.55	1138.75	1141.37	1140.47	1145.28	1150.18	1156.42	1158.47	1157.55
10	1141.49	1142.69	1145.45	1145.29	1150.24	1154.92	1161.31	1166.12	1164.03
11	1143.46	1146.22	1151.03	1150.17	1154.83	1157.35	1166.22	1168.15	1166.94
12	1142.25	1144.17	1145.46	1150.33	1153.36	1164.03	1166.37	1169.28	1169.79

	q=8	q=9	q=10	q=11
1	1131.60	1136.89	1141.59	1138.12
2	1136.38	1139.36	1144.76	1149.29
3	1138.39	1134.89	1149.15	1143.98
4	1144.84	1139.63	1144.51	1145.43
5	1146.18	1143.71	1147.67	1141.10
6	1148.11	1146.69	1151.54	1145.94
7	1151.77	1151.46	1156.14	1150.84
8	1155.08	1154.62	1159.23	1155.38
9	1158.47	1157.55	1162.43	1167.05
10	1166.12	1164.03	1167.36	1164.98
11	1168.15	1166.94	1171.88	1172.57
12	1169.28	1169.79	1169.67	1178.95

Table 5 – BIC des différents modèles

5 Code

```
require(zoo)
require(tseries)
library(dplyr)

#fonctions utilisees dans le code

#fonction de test des significations individuelles des coefficients
signif <- function(estim){
  coef <- estim$coef
  se <- sqrt(diag(estim$var.coef))
  t <- coef/se
  pval <- (1-pnorm(abs(t)))*2
  return(rbind(coef,se,pval))
}

#fonction de test de Ljung-Box pour tester l'hypothese d'autocorrelation nulle
Qtests <- function(series, k, fitdf=0) {
  pvals <- apply(matrix(1:k), 1, FUN=function(l) {
    pval <- if (l<=fitdf) NA else Box.test(series, lag=l, type="Ljung-Box", fitdf=fitdf)$p.value
    return(c("lag"=l,"pval"=pval))
  })
  return(t(pvals))
}

#fonction qui applique nos deux fonctions signif et Qtest :
#le test de significations individuelles des coefficients et le test d'absence d'autocorrelation des residus
arimafit <- function(estim){
  adjust <- round(signif(estim),3)
  pvals <- Qtests(estim$residuals,24,length(estim$coef)-1)
  pvals <- matrix(apply(matrix(1:24,nrow=6),2,function(c) round(pvals[c,],3)),nrow=6)
  colnames(pvals) <- rep(c("lag", "pval"),4)
  cat("tests de nullite des coefficients :\n")
  print(adjust)
  cat("\n tests d'absence d'autocorrelation des residus : \n")
  print(pvals)
}

#fonction qui selectionne les modeles valides,
#c'est-a-dire ceux avec des coefficients significatifs et des residus non correles
modelchoice <- function(p,q,data=xm, k=24){
  estim <- try(arima(data, c(p,0,q),optim.control=list(maxit=20000)))
```

```
  if (class(estim)=="try-error") return(c("p"=p,"q"=q,"arsignif"=NA,"masignif"=NA,"resnocorr"=NA, "ok"=NA))
  arsignif <- if (p==0) NA else signif(estim)[3,p]<=0.05
  masignif <- if (q==0) NA else signif(estim)[3,p+q]<=0.05
  resnocorr <- sum(Qtests(estim$residuals,24,length(estim$coef)-1)[,2]<=0.05,na.rm=T)==0
  checks <- c(arsignif,masignif,resnocorr)
  ok <- as.numeric(sum(checks,na.rm=T)==(3-sum(is.na(checks))))
  return(c("p"=p,"q"=q,"arsignif"=arsignif,"masignif"=masignif,"resnocorr"=resnocorr,"ok"=ok))
}

#fonction qui ressort les modeles qui ont ete valides par la fonction modelchoice definie ci-dessus
armamodelchoice <- function(pmax,qmax){
  pqs <- expand.grid(0:pmax,0:qmax)
  t(apply(matrix(1:dim(pqs)[1]),1,function(row) {
    p <- pqs[row,1]; q <- pqs[row,2]
    cat(paste0("Computing ARMA(",p,",",q,") \n"))
    modelchoice(p,q)
  }))
}



#base de donnees
setwd("C:/Users/candi/Desktop/ETUDES/ENSAE2A/semestre 2/s ries temporelles/series temp/series_temp")
data <- read.csv('valeurs_mensuelles_pesticides.csv', sep=";")
data <- data[2]

data <- as.data.frame(data[-(1:3),])#on enleve les premieres lignes qui ne sont pas des donnees
indice <- as.data.frame(as.numeric(unlist(data)))

xm.source <- zoo(indice[[1]]) #convertit le premier element de data en serie temporelle de type "zoo"
T <- length(xm.source)
test <- tail(xm.source, n=2) #pour comparer nos previsions avec les vraies donnees
xm <- xm.source[(250):(T-2)] #pour le modele

mean(xm.source)
plot(xm, xaxt="n") #plot des donn es
axis(side=1,at=seq(0,400,12)) #pour mettre l'axe x

par(mfrow=c(1,2))
acf(xm)
pacf(xm) #saisonnalite apparente : saisonnalite de 12 donc annuelle, avec ete et hiver differencies (corr pos et neg)
dev.off()

pp.test(xm) #test de philippe perron, on rejette a 1% l'hypothese que la serie n est pas stationnaire
```

```r
#on retire la moyenne de xm
xm <- xm - mean(xm)

#on enleve la saisonnalite apparente
xm <- diff(xm, lag = 12)
par(mfrow=c(1,2))
acf(xm)
pacf(xm) #la saisonnalite a bien disparu

#on identifie avec l'acf et la pacf les ordres maximums a tester
pmax = 12
qmax = 11

#modele
#estimation de tous les modeles et selection des modeles valides
arma_valid <- armamodelchoice(12,11)
selec <- arma_valid[arma_valid[,"ok"]==1&!is.na(arma_valid[,"ok"]),]

#les modeles possibles sont donnes par
selec
#on peut donc choisir p=12, q=2 ou p=12, q=9, ou p=5, q=11

#on fit les trois modeles et on calcule les aic
arma_12_2 <- arima(xm, c(12,0,2))
arma_12_9 <- arima(xm, c(12,0,9))
arma_5_11 <- arima(xm, c(5,0,11))

#aic
arma_12_2$aic
arma_12_9$aic
arma_5_11$aic #modele ayant le plus petit aic

#bic
BIC(arma_12_2)
BIC(arma_12_9)
BIC(arma_5_11) #modele ayant le plus petit bic


#valeurs selectionnees pour notre modele
p = 5
q = 11

#fit du modele
arma_fit <- arima(xm, c(5,0,11))
```

```
arma_fit

#residus
plot(arma_fit$residuals)
acf(arma_fit$residuals)
pacf(arma_fit$residuals)

hist(arma_fit$residuals)
library(forecast)
checkresiduals(arma_fit)

#Q test
#test
Qtests(arma_fit$residuals, 24, 5) #tests de LB pour les ordres 1 a 24
#on rejette le fait que les residus soient correles

signific <- as.data.frame(signif(arma_fit))

#causalite
roots <- polyroot(sort(arma_fit$coef[c('ar1', 'ar2', 'ar3', 'ar4', 'ar5')]))
modulus_roots <- Mod(roots)
modulus_roots #les coefficients sont bien plus grands que 1 donc le modele est causal

#prevision
model_pred <- predict(arma_fit, n.ahead=2)
serie_pred <- zoo(c(xm, model_pred$pred))

#graphiques
xm_all <- xm.source[250:T] - mean(xm.source[250:(T-2)])
xm_all <- diff(xm_all, lag = 12)

dev.off()

plot(xm_all, col = 'black', ylab = 'S rie', main = 'Pr vision des 2 prochaines valeurs de la s rie')
#lines(xm_all, col = 'black', type = 'p') pour avoir des ronds a chaque valeur de la serie temporelle
U = model_pred$pred + 1.96*model_pred$se
L = model_pred$pred - 1.96*model_pred$se
xx = c(time (U), rev (time (U)))
yy = c(L, rev(U))
polygon(xx, yy, border = 8, col = gray (0.6, alpha=0.2))
lines(model_pred$pred, type = "p", col = "red")
lines(model_pred$pred, type = 'l', col = 'red')
legend("topleft", legend=c("Donn es r elles", "Pr diction"), col=c("red", "black"), lty=1:2, cex=0.4)
```

```
#export de la table de significativite des modeles pour le document latex
library(xtable)
xtable(signific)

xtable(signific %>% select(ar1, ar2, ar3, ar4, ar5, ma1, ma2, ma3, ma4, ma5, ma6))

xtable(signific %>% select(ma7, ma8, ma9, ma10, ma11, intercept))



#bonus: table avec les aic et bic de tous les modeles
selection_print_aic <- function(pmax, qmax){
  res_aic <- matrix(nrow=pmax, ncol=qmax)
  for (p in 1:pmax){
    for (q in 1:qmax){
      model <- arima(xm, c(p,0, q))
      res_aic[p,q] <- model$aic
      print(c(p,q))
    }}
  return(res_aic)
}
selec_aic <- selection_print_aic(pmax,qmax)

xtable((as.data.frame(selec_aic))%>%select(V1, V2, V3, V4, V5, V6, V7, V8, V9))

xtable((as.data.frame(selec_aic))%>%select(V8, V9, V10, V11))

selection_print_bic <- function(pmax, qmax){
  res_bic <- matrix(nrow=pmax, ncol=qmax)
  for (p in 1:pmax){
    for (q in 1:qmax){
      model <- arima(xm, c(p,0, q))
      res_bic[p,q] <- BIC(model)
      print(c(p,q))
    }}
  return(res_bic)
}
selec_bic <- selection_print_bic(pmax,qmax)

xtable((as.data.frame(selec_bic))%>%select(V1, V2, V3, V4, V5, V6, V7, V8, V9))

xtable((as.data.frame(selec_bic))%>%select(V8, V9, V10, V11))
```