École nationale
de la statistique
et de l'administration
économique

université
PARIS-SACLAY

Séries temporelles linéaires

Compte-rendu

Étude de l'indice de fabrication de malt en France depuis 1990

Théo Lorthios, Mélissa Tamine

9 mai 2022

Table des matières

1 Introduction - Présentation des données

Nous avons choisi de fonder notre projet sur la modélisation et la prévision d'une série temporelle observée particulière : **l'indice brut de production industrielle de malt en France**.

L'indice brut de la production industrielle de malt se définit comme un indicateur de court terme qui permet de mesurer l'évolution de la production des unités industrielles du secteur du malt exerçant sur le territoire national à une période bien définie. Il s'intéresse à l'activité de fabrication de ces unités industrielles et permet de mesurer les quantités physiques de malt produites par ces unités au cours d'une période donnée, de donner l'évolution en volume de la production industrielle de malt et de présenter de façon assez représentative les mouvements observés au sein du tissu industriel lié à ce secteur. Il est produit à un rythme mensuel et son horizon temporel s'étale de janvier 1990 à février 2022.

La période de base choisie pour cette série temporelle est l'année 2015 et la valeur associée est fixée à 100. Cette période de base correspond généralement à une année jugée conforme à la moyenne, qui ne présente pas d'influences particulières sur l'activité économique (c'est-à-dire ni forte croissance, ni faible croissance).

Aucune information ne nous indique que cette série temporelle a subit une transformation logarithmique.

Cette série temporelle est disponible sur le site de l'INSEE au lien suivant :
https://www.insee.fr/fr/statistiques/serie/010537310#Telechargement

On notera dans la suite cette série $(X_t)_{t\in T}$, avec $T = \{1, ..., \bar{t}\}$ l'ensemble contenant toutes les dates où la série est observée (longueur de la série).

En voici une représentation graphique :

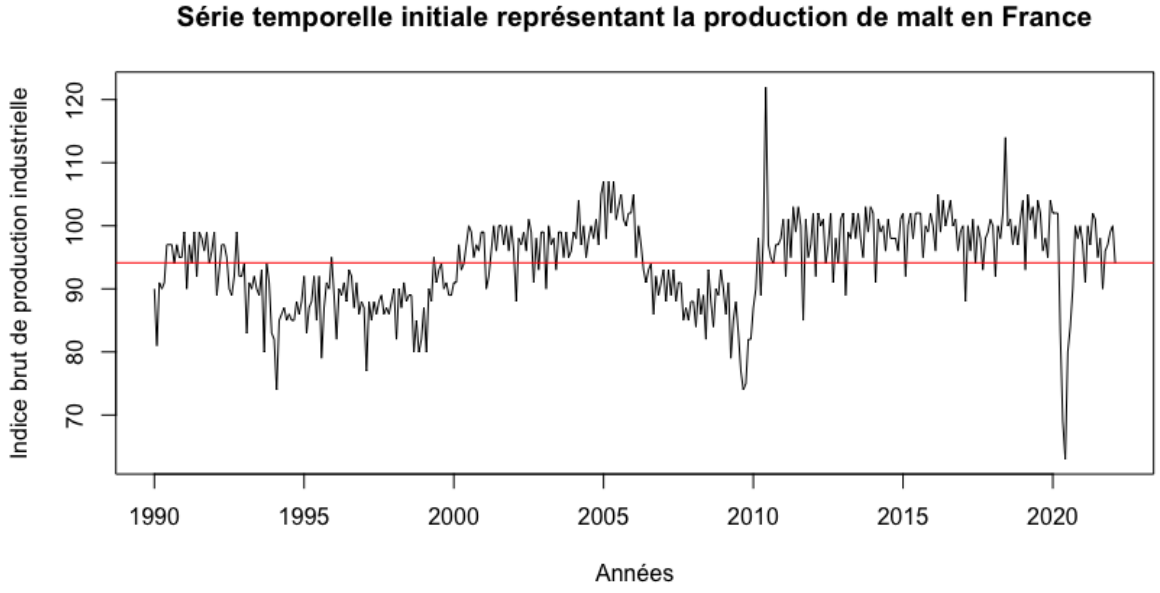


FIGURE 1 – Représentation graphique de la série étudiée - Indice brut de production de malt en France entre janvier 1990 et février 2022

2 Stationnarisation de la série

La série temporelle $(X_t)_{t\in T}$ peut être décomposée selon trois éléments :

1. Une tendance T_t : qui correspond à un comportement croissant ou décroissant de la série au cours du temps. Elle reflète souvent un phénomène de croissance ou décroissance sur le long terme. Elle peut être linéaire, quadratique ou encore exponentielle.
2. Une saisonnalité S_t : qui reflète la présence d'un phénomène périodique qui se répète au long de la série temporelle.
3. Un résidu ou erreur ϵ_t : qui correspond à la partie de la série temporelle que la décomposition ne permet pas d'expliquer, autrement dit sa partie stochastique.

Mathématiquement on peut donc traduire une série temporelle par la formule suivante :

$$X_t = T_t + S_t + \epsilon_t$$

La décomposition cycle-tendance-bruits de la série initiale représentant l'indice brut de production de malt en France entre janvier 1990 et février 2022 disponible sur la figure 2 nous permet d'illustrer ce résulat.

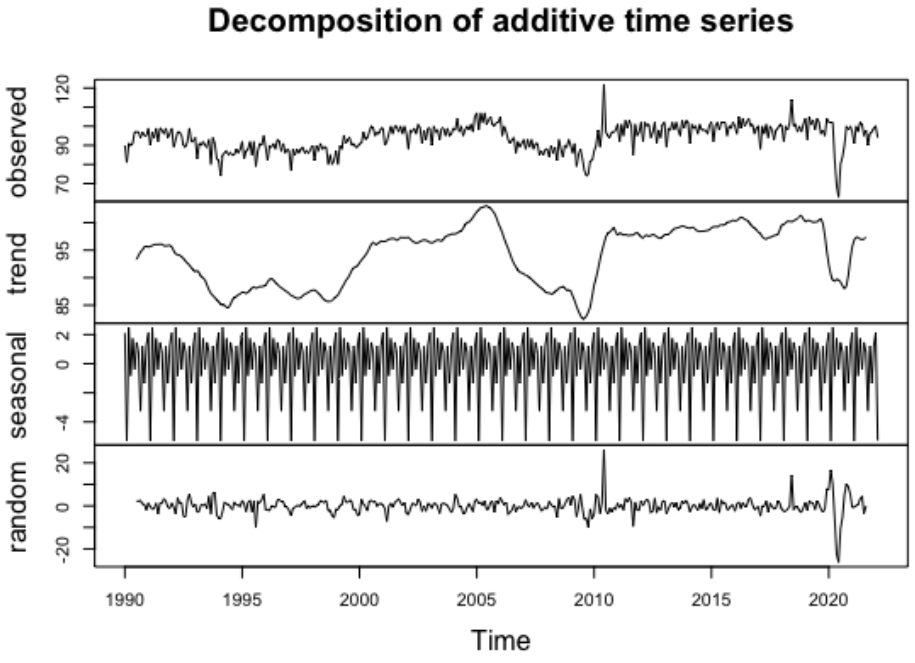


FIGURE 2 – Décomposition cycle - tendance - bruits de la série initiale

L'enjeu est tout d'abord de rendre cette série stationnaire afin de pouvoir l'étudier à l'aide des modèles classiques de séries temporelles, autrement dit de supprimer sa tendance et sa saisonnalité.

2.1 Identification de la tendance

Afin d'identifier la tendance dans la série temporelle initiale, nous régressons X_t sur le temps. Nous obtenons les résultats suivants :

Coefficients	Estimation	Écart-type	t-value	p-value
Constante	89.78207	0.66542	134.926	<2e-16
Temps	0.02248	0.00298	7.544	**3.34e-13**

TABLE 1 – Résultats de la régression linéaire de X_t sur le temps

La p-value associée au test de significativité du coefficient devant la variable temps est plus petite que 1%, le coefficient associé au temps est donc significatif au seuil de 1%. Nous en déduisons que la série initiale possède une tendance.

Cette tendance est représentée sur la figure 3. Nous constatons qu'il s'agit d'une tendance linéaire.

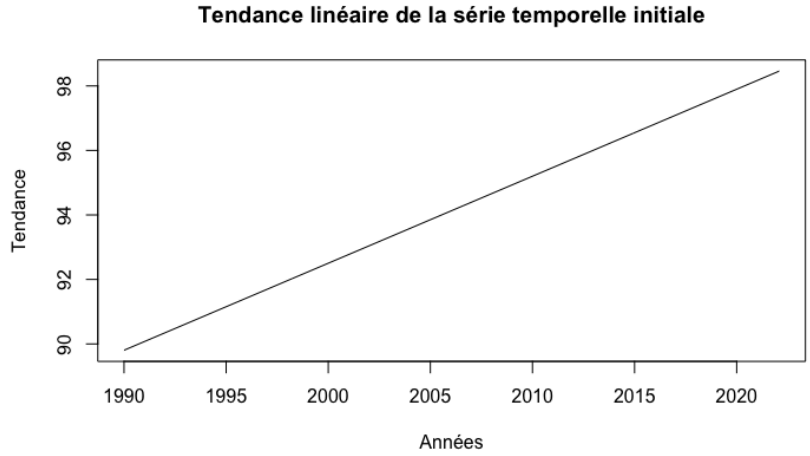


FIGURE 3 – Représentation de la tendance de la série initiale

Nous pouvons alors retirer cette tendance à la série initiale et nous obtenons la série temporelle Y_t suivante (figure 4) qui présente toujours un caractère saisonier.

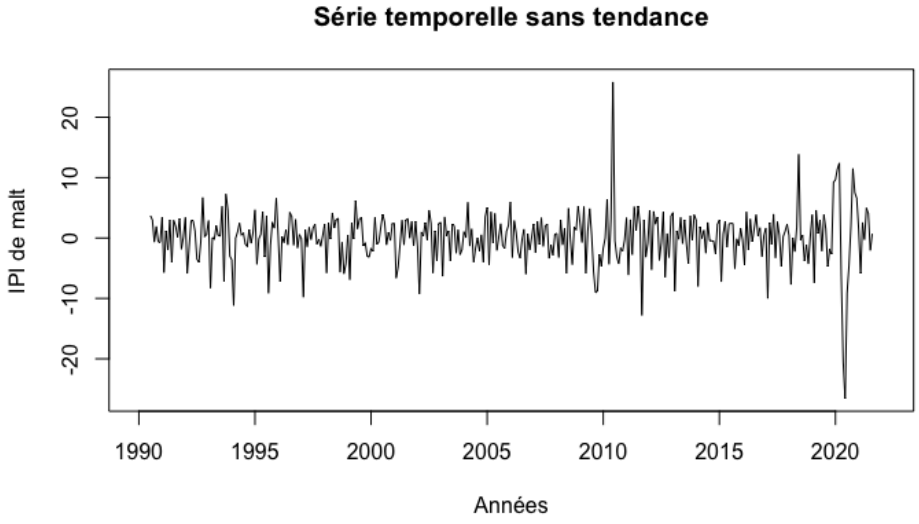


FIGURE 4 – Représentation graphique de la série temporelle sans tendance

2.2 Identification de la saisonnalité

Nous pouvons isoler la saisonnalité de la série initiale et la représenter graphiquement (figure 5) :

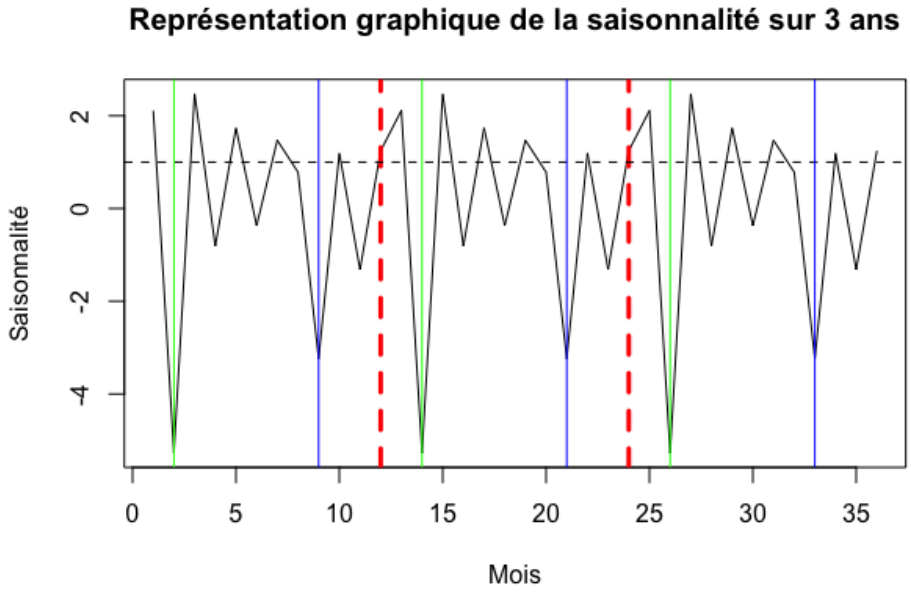


FIGURE 5 – Représentation graphique de la saisonnalité sur 3 ans

Nous constatons que les « pics » significatifs (représentés en vert et bleu) ont une cyclicité très nette tous les 12 mois, ce qui laisse à penser à une saisonnalité à l'ordre 12.

2.3 Transformation de la série initiale

Nous utilisons ensuite la décomposition additive de la série temporelle représentée en figure 2 afin de retirer la saisonnalité de la série temporelle. Nous obtenons la série transformée suivante :

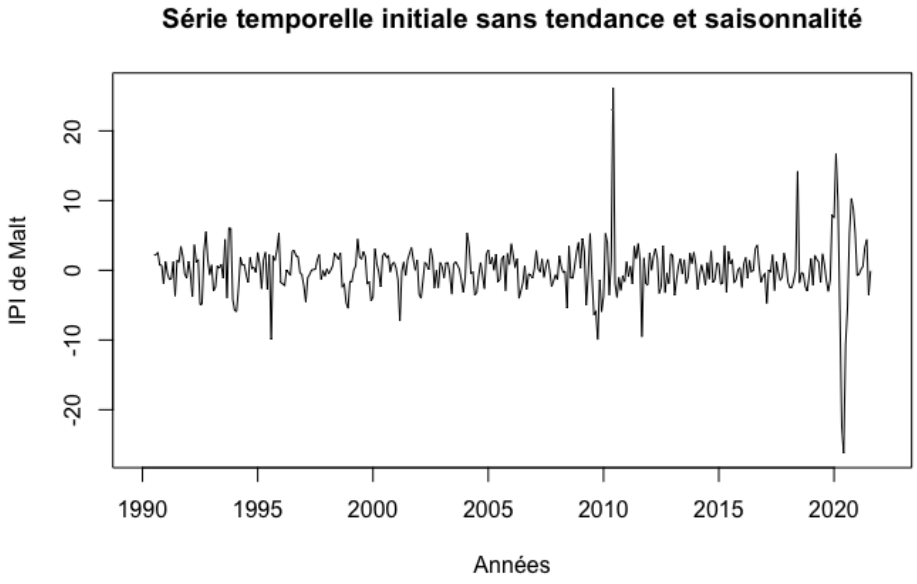


FIGURE 6 – Représentation graphique de la série stationnarisée

2.4 Tests de stationnarité de la série transformée

La série temporelle représentée en figure 6 semble être stationnaire. Il convient maintenant de faire les tests de stationnarité de la série transformée afin de confirmer notre hypothèse. Nous allons effectuer les tests classiques de racine unité : Augmented Dickey-Fuller, Phillips-Perron ainsi que le test de stationnarité KPSS.

Nous rappelons que pour les deux premiers tests, l'hypothèse nulle est la présence de racine unité et donc la non stationnarité de la série alors que pour le test KPSS, l'hypothèse nulle est l'hypothèse de stationnarité de la série.

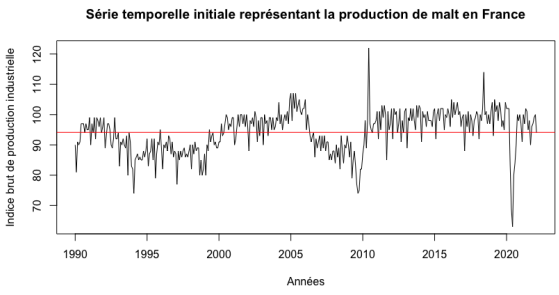
Type de test	Lag order	t-value	p-value
Augmented Dickey-Fuller	30	-5.16	< 0.01
Phillips-Perron	30	-12.8	< 0.01

Type de test	Lag order	KPSS level	p-value
KPSS	4	0.00644	> 0.1

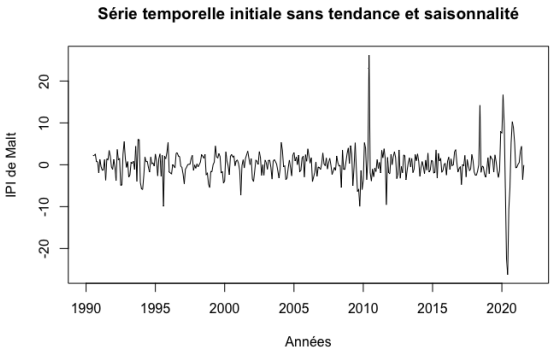
Les deux tests de racine unitaire (ADF et PP) nous ont donné une p-value inférieure à 0.01, ce qui nous permet de rejeter H_0 au seuil de 1% en faveur de l'hypothèse alternative de stationnarité de la série. Concernant le test KPSS, nous ne parvenons pas à rejeter l'hypothèse nulle au niveau 0.1, nous décidons donc de ne pas la rejeter (p-value supérieure à 0.1), ce qui est également en faveur de l'hypothèse de stationnarité de la série. Les trois tests confirment ainsi que notre série n'est pas une marche aléatoire. Nous pouvons donc la modéliser telle qu'elle sans l'intégrer ou la différencier.

2.5 Comparaison de la série initiale et de la série transformée

Finalement, nous pouvons comparer nos deux séries : la série brute initiale avec tendance linéaire et saisonnalité ainsi que la série stationnaire. Nous les avons représentées dans la figure ci-après.



(a) Série brute initiale $(X_t)_{t\in T}$



(b) Série transformée $(Y_t)_{t\in T}$

3 Modélisation de la série stationarisée

Dans cette partie, nous allons tenter de trouver le modèle qui correspond le mieux aux données de la série temporelle. Dans le cadre de notre étude, la modélisation qui semble la plus adaptée étant donnée la saisonnalité est la modélisation SARIMA, cette dernière permettant de générer une composante saisonnière de nature aléatoire, c'est-à-dire qui ne

se répète pas à l'identique d'un cycle à l'autre.

Nous pouvons rappeller qu'une série temporelle X_t est un processus $\text{SARIMA}_s[(p,d,q),(P,D,Q)]$ de période s si X_t est tel que le processus $Z_t = (I-L)^d \times (I-L^s)^D \times X_t$ est un $\text{ARMA}(p,q)$ causal tel que $\phi(L)\Phi(L^s) \times Z_t = \lambda(L)\Lambda(L^s)\epsilon_t$ avec ϕ, Φ, λ et Λ des polynômes de degrés respectivement p, P, q, Q.

Nous avons déjà identifié l'ordre s correspondant à la saisonnalité (pour rappel $s = 12$). De plus, comme nous n'avons pas différencié la série temporelle initiale afin de la stationnariser, nous avons $d = D = 0$. L'enjeu à présent est donc d'identifier les ordres p, P, q, Q. Les étapes que nous allons suivre sont les suivantes :

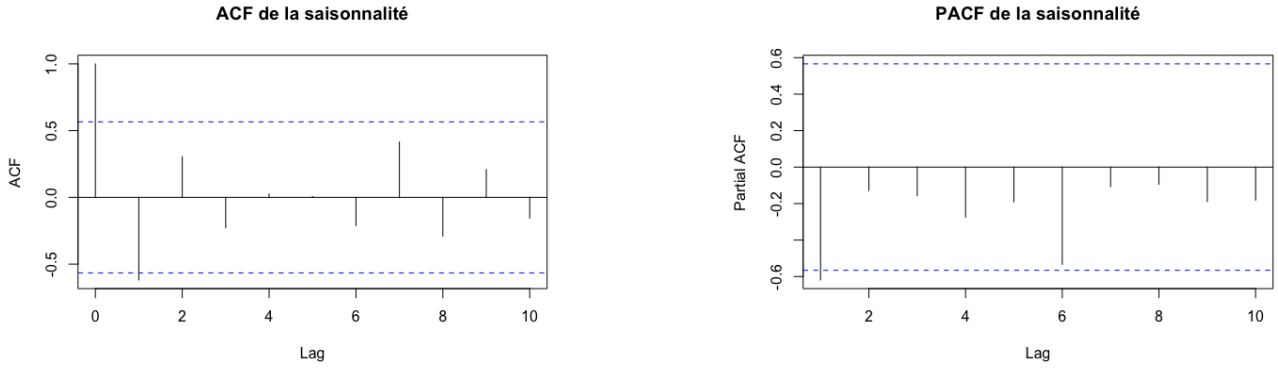
1. Déterminer les ordres P et Q du modèle SARIMA en analysant l'ACF et la PACF de la saisonnalité.

2. Identifier les ordres maximaux $p_{\max}$ et $q_{\max}$ du modèle ARMA censé modéliser la série stationnarisée $(Y_t)_{t\in T}$ puis trouver le modèle $\text{ARMA}(p'_{\max}, q'_{\max})$ qui minimise les critères AIC et BIC en testant tout couple $(i,j) \in [[0, p_{\max}]] \times [[0, q_{\max}]]$ (tout en vérifiant la blancheur et la normalité des résidus du modèle optimal obtenu). **L'objectif est d'affiner la plage des ordres potentiels pour le modèle SARIMA que nous cherchons à déterminer**.

3. Réintroduire la saisonnalité dans $(Y_t)_{t\in T}$ (pour retrouver $(X_t)_{t\in T}$) et trouver les ordres p et q de la modélisation SARIMA de $(X_t)_{t\in T}$ qui minimise les critères AIC et BIC en testant tout couple $(i,j) \in [[0, p'_{\max}]] \times [[0, q'_{\max}]]$ (tout en vérifiant la blancheur et la normalité des résidus du modèle optimal obtenu).

3.1 Identification des ordres P et Q

Nous commençons par déterminer les ordres P et Q du modèle $SARIMA_{12}[(p,0,q),(P,0,Q)]$ que nous cherchons à déterminer. Pour cela, nous analysons l'autocorrélation ainsi que l'autocorrélation partielle de la saisonnalité de la série temporelle $(X_t)_{t\in T}$ appelée S_t. Pour P, nous regarderons le graphique de la PACF tandis que pour Q nous regarderons la partie ACF. La figure 8 nous permet de constater un unique pic significatif dans chacun des graphiques. Nous en déduisons ainsi que $P = 1$ et $Q = 1$.



(a) ACF de la saisonnalité S_t (b) PACF de la saisonnalité S_t

FIGURE 8 – ACF et PACF de la saisonnalité S_t sur 12 mois

3.2 Identification des ordres $p'_{\max}$ et $q'_{\max}$

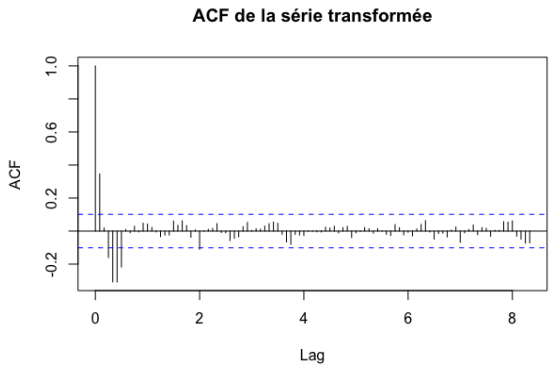
3.2.1 Identification des ordres $p_{\max}$ et $q_{\max}$

Afin de sélectionner le modèle approprié, et en particulier les ordres p et q du modèle SARIMA, nous allons tout d'abord regarder la fonction d'autocorrélation (ACF) et la fonction d'autocorrélation partielle (PACF) de la série transformée $(Y_t)_{t\in T}$ (série supposée stationnaire). Cela va nous permettre de trouver les ordres p et q maximaux du modèle $ARMA(p,q)$ censé modéliser cette série transformée $(Y_t)_{t\in T}$.

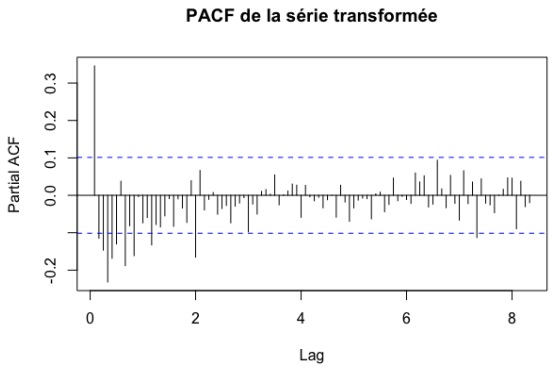
Pour p, nous regarderons la PACF (partie AR) tandis que nous regarderons l'ACF pour q (partie MA). **Nous réintroduirons ensuite la saisonalité dans notre étude**.

Nous pouvons voir dans la figure 9 que l'ACF et la PACF ne présentent plus de « pics » significativement différents de zéro au delà de 6 retards dans les deux cas (nous faisons le choix d'ignorer délibérément les pics pour des retards supérieurs à 6, en particulier dans la PACF, l'objectif étant d'obtenir des modèles parcimonieux). Ainsi, nous estimons tous les modèles $ARMA(p,q)$ avec $0 \leq p \leq 6$ et $0 \leq q \leq 6$.

De plus, on peut remarquer des chocs significatifs à des ordres bien supérieures sur la PACF. Ces chocs sont des changements temporaires de la structure, on explique celui de 2020 avec la crise Covid qui a bloqué toutes les commandes de bière car les bars et restaurants étaient fermés. L'autre choc significatif est en 2018, il peut être expliqué par la coupe du monde de foot qui a montré des chiffres exceptionnelles de consommation de bière. On peut aussi constaté que l'année 2018 a été une année de canicule impressionnante avec des temperatures chocs.



(a) ACF de la série transformée $(Y_t)_{t\in T}$



(b) PACF de la série transformée $(Y_t)_{t\in T}$

FIGURE 9 – ACF et PACF de $(Y_t)_{t\in T}$ (série stationnarisée)

3.2.2 Critères AIC et BIC

Afin d'obtenir le meilleur modèle censé modéliser la série transformée $(Y_t)_{t\in T}$, nous allons utiliser les critères d'informations d'Akaike (AIC) ainsi que le critère critère d'information bayésien (BIC). Nous choisirons le modèle qui minimisent ces deux critères. Afin de trouver un tel modèle, nous allons ajuster un modèle ARMA(i,j) pour tout couple $(i,j) \in [[0, p_{\max} = 6]] \times [[0, q_{\max} = 6]]$.

q \ p	0	1	2	3	4	5	6
0	1915.2	1917.2	1914.7	1915.6	1916.3	1916.7	1918.5
1	2002.8	1996.7	1978.1	1969.4	1965.0	2051.3	2005.7
2	2004.3	1938.3	2004.7	1934.2	1911.3	2004.2	2004.7
3	1919.1	1921.3	1912.3	1913.9	1915.8	2004.5	2006.3
4	1920.8	1912.4	1913.9	1919.7	1917.3	2006.3	1944.2
5	1938.6	1913.9	1915.8	1913.0	1916.9	1949.2	1934.2
6	1921.3	1915.7	1917.8	1913.4	1914.9	1938.3	1932.7

TABLE 2 – AIC pour les différents modèles ARMA(p, q)

q \ p	0	1	2	3	4	5	6
0	1950.5	1956.4	1949.9	1962.6	1964.1	1971.7	1949.9
1	2018.5	2016.3	2001.6	1996.8	1996.4	2059.1	2017.5
2	1944.2	1944.1	1947.3	1952.7	1941.3	2015.9	2020.4
3	1942.6	1949.0	1943.7	1949.2	1955.1	2019.9	2025.9
4	1948.3	1943.8	1949.2	1958.9	1960.4	2025.9	1967.7
5	1970.0	1949.2	1955.1	1956.2	1963.0	1972.8	1961.7
6	1956.6	1955.0	1961.0	1960.5	1965.9	1965.7	1964.1

TABLE 3 – BIC pour les différents modèles ARMA(p, q)

Dans notre cas, nous constatons que le modèle qui minimise ces deux critères est identique et correspond à un ARMA$(4, 2)$. Nous pouvons ensuite effectuer un test de significativité des coefficients de notre modèle ARMA$(4, 2)$ par un test de Student. Ce test a pour hypothèse nulle la nullité des coefficients. Nous avons renseigné dans le tableau ci-après les résultats de ce test.

	Coefficients	Écart-types	p-value
AR 1	0.20	0.06	2.6e-4
AR 2	0.71	0.058	0.00
AR 3	-3.6e-1	5.4e-1	9.5e-12
AR 4	-0.18	0.052	6.9e-4
MA 1	-0.03	0.003	0.02
MA 2	-0.96	0.03	0.00
Intercept	-7.93e-5	4.5e-3	9.9e-1

TABLE 4 – Coefficients du modèle ARMA$(4, 2)$

Les p-values associées à tous les coefficients nous permettent de rejeter l'hypothèse nulle au seuil de 5%. Les coefficients sont donc bien significatifs.

3.2.3 Blancheur et normalité des résidus

Nous pouvons également vérifier la blancheur des résidus du modèle ARMA$(4, 2)$ afin d'être certains qu'ils correspondent bien à un bruit blanc gaussien. Pour cela, nous pouvons réaliser le test du Portemanteau/Ljung-box. L'hypothèse nulle H_0 de ce test est la présence d'un bruit blanc fort.

Modèle	df	χ^2	p-value
ARMA(4, 2)	17	24.241	0.11

TABLE 5 – Test du Portemanteau/Ljung-box

On obtient ainsi une p-value de 0.11. L'hypothèse nulle n'est pas rejetée au seuil de 5% ce qui semble bien confirmer la blancheur des résidus du modèle.

Enfin, afin d'obtenir la prédiction la plus fiable possible, il est intéressant d'avoir des résidus normaux (donc gaussiens). Pour tester la normalité de nos résidus, nous pouvons représenter graphiquement leur densité.

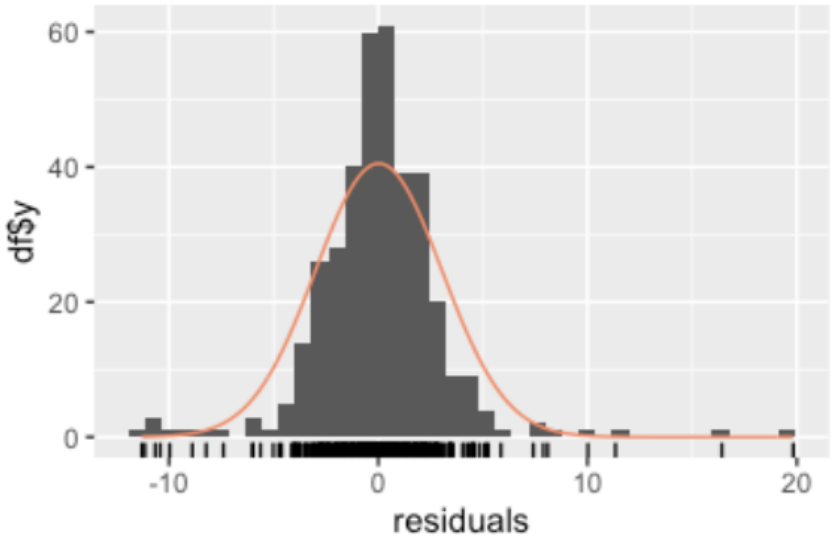


FIGURE 10 – Densité des résidus

Sur l'histogramme de la figure 10, nous pouvons constater que les résidus semblent normaux.

3.3 Détermination du modèle SARIMA$_s[(p, d, q), (P, D, Q)]$

Nous avons déterminé $P = Q = 1$, $D = d = 0$, $p'_{\max} = 4$ et $q'_{\max} = 2$ dans les parties précédentes. À présent, nous cherchons à déterminer la modélisation SARIMA$_{12}[(p, 0, q), (1, 0, 1)]$ de $(X_t)_{t\in T}$ qui minimise les critères AIC et BIC en testant tout couple $(p, q) \in [[0, p'_{\max} = 4]] \times [[0, q'_{\max} = 2]]$ (tout en vérifiant la blancheur et la normalité des résidus du modèle optimal obtenu).

3.3.1 Critères AIC et BIC

Afin d'obtenir le meilleur modèle SARIMA$_{12}[(p, 0, q), (1, 0, 1)]$ censé modéliser la série $(X_t)_{t\in T}$, nous allons utiliser les critères d'informations d'Akaike (AIC) ainsi que le critère critère d'information bayésien (BIC). Nous choisirons le modèle qui minimisent ces deux critères. Afin de trouver ce modèle qui minimise les deux critères, nous allons ajuster un modèle SARIMA$_{12}[(i, 0, j), (1, 0, 1)]$ pour tout couple $(i, j) \in [[0, p'_{\max} = 4]] \times [[0, q'_{\max} = 2]]$.

q \ p	0	1	2	3	4
0	6.51	5.84	5.64	5.67	5.74
1	6.12	5.74	5.73	5.71	5.79
2	5.91	5.78	5.86	5.72	6.11

TABLE 6 – AIC pour les différents modèles $\text{SARIMA}_{12}[(i,0,j),(1,0,1)]$

q \ p	0	1	2	3	4
0	6.52	5.81	5.8	5.72	5.82
1	6.24	5.80	5.8	5.83	5.91
2	6.0	5.84	5.73	5.82	6.19

TABLE 7 – BIC pour les différents modèles $\text{SARIMA}_{12}[(i,0,j),(1,0,1)]$

Dans notre cas, nous constatons que le modèle qui minimise ces deux critères est identique et correspond à un $\text{SARIMA}_{12}[(3,0,1),(1,0,1)]$.

3.3.2 Blancheur et normalité des résidus

Nous devons également vérifier la blancheur des résidus du modèle $\text{SARIMA}_{12}[(3,0,1),(1,0,1)]$ afin d'être certain qu'ils correspondent bien à un bruit blanc gaussien. Pour cela, nous pouvons réaliser le test du Portemanteau/Ljung-box. L'hypothèse nulle H_0 de ce test est la présence d'un bruit blanc fort.

Modèle	df	χ^2	p-value
$\text{SARIMA}_{12}[(3,0,1),(1,0,1)]$	17	24.241	0.11

TABLE 8 – Test du Portemanteau/Ljung-box

On obtient ainsi une p-value de 0.11. L'hypothèse nulle n'est pas rejetée au seuil de 5% ce qui semble bien confirmer la blancheur des résidus du modèle.

Enfin, afin d'obtenir la prédiction la plus fiable possible, il est intéressant d'avoir des résidus normaux (donc gaussiens). Pour tester la normalité de nos résidus, nous pouvons réaliser un QQplot. Ainsi, si les points sont alignés par rapport à la loi normale sur le QQplot, alors les résidus sont normaux.

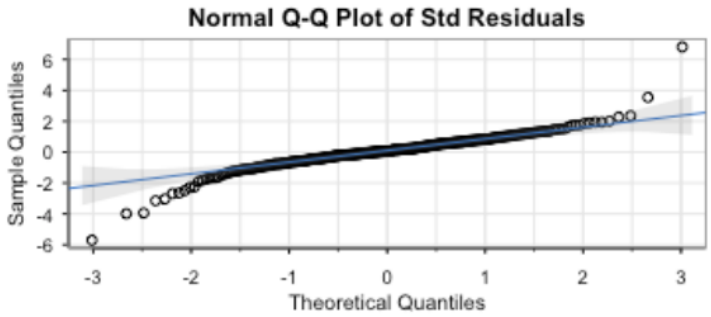


FIGURE 11 – QQplot des résidus

Sur le QQplot de la figure 11, nous pouvons voir que nos résidus semblent normaux. Nous constatons néanmoins que les probabilité des quantiles de la loi normale ont du mal sur

les queues de distribution, nous l'expliquons facilement avec les outliers de crises.

On a déjà remarqué des chocs temporaires comme la crise covid ou la coupe du monde de 2018. Il est possible de gérer ce genre chocs avec des modèles de discontinuité se modifie en cas de changement stucturel de l'économie.

3.3.3 Écriture du modèle

En conclusion de cette partie, nous avons sélectionné un modèle $\text{SARIMA}_{12}[(3,0,1),(1,0,1)]$ avec des résidus normaux pour notre série $(X_t)_{t\in T}$. Nous avons donc avec l'estimation des coefficients suivante :

Coefficient	Estimation	Écart-type	t-value	p-value
ϕ_1	-0.27	0.06	-4.15	0
ϕ_2	0.70	0.05	13.26	0
ϕ_3	0.23	0.05	4.57	0
λ_1	0.92	0.04	20.05	0
Φ_1	0.99	0.004	229.9	0
Λ_1	-0.93	0.04	-24.14	0

TABLE 9 – Résultats de la spécification $\text{SARIMA}_{12}[(3,0,1),(1,0,1)]$

Nous notons $\phi_j, j \in \{1,2,3\}$ les coefficients de la partie AR du SARIMA, λ_1 le coefficient de la partie MA du SARIMA, Λ_1 le coefficient de la partie SMA et Φ_1 le coefficient de la partie SAR.

4 Prévision

4.1 Région de confiance de niveau α

Nous notons toujours T la longueur de la série. Nous avons vu que la série $(X_t)_{t\in T}$ suivait un processus $\text{SARIMA}_{12}[(3,0,1),(1,0,1)]$ avec des résidus normaux (c'est-à-dire que $\epsilon_t \sim \mathcal{N}(0,\sigma_\epsilon^2)$).

De manière équivalente, la série $(X_t)_{t\in T}$ suit un processus $\text{ARMA}(p+sP, q+sQ) = \text{ARMA}(3+12\times 1, 1+12\times 1) = \text{ARMA}(15,13)$. Plus précisément :

$$(1-\sum_{j=1}^{3}\phi_j L^j)(1-\Phi_1 L^{12})X_t = (1-\lambda_1 L)(1-\Lambda_1 L^{12})\epsilon_t$$

$$\Leftrightarrow X_t = \sum_{j=1}^{3}\phi_j X_{t-j} + \Phi_1 X_{t-12} - \Phi_1\sum_{j=1}^{3}\phi_j X_{t-12-j} + \epsilon_t - \lambda_1\epsilon_{t-1} - \Lambda_1\epsilon_{t-12} + \lambda_1\Lambda_1\epsilon_{t-13}$$

Nous nous plaçons à la date T^* et nous cherchons la prévision linéaire optimale des valeurs futures (X_{T^*+1}, X_{T^*+2}), notées $(\hat{X}^{T^*}_{T^*+1}, \hat{X}^{T^*}_{T^*+2})$. De manière classique, nous avons :

$$\begin{cases} \hat{X}^{T^*}_{T^*+1} = \sum_{j=1}^{3}\phi_j X_{T^*+1-j} + \Phi_1 X_{T^*+1-12} - \Phi_1\sum_{j=1}^{3}\phi_j X_{T^*+1-12-j} - \lambda_1\epsilon_{T^*+1-1} - \Lambda_1\epsilon_{T^*+1-12} \\ +\lambda_1\Lambda_1\epsilon_{T^*+1-13} \\ \hat{X}^{T^*}_{T^*+2} = \phi_1\hat{X}^{T^*}_{T^*+1} + \sum_{j=1}^{3}\phi_j X_{T^*+2-j} + \Phi_1 X_{T^*+2-12} - \Phi_1\sum_{j=1}^{3}\phi_j X_{T^*+2-12-j} - \lambda_1\epsilon_{T^*+2-1} \\ -\Lambda_1\epsilon_{T^*+2-12} + \lambda_1\Lambda_1\epsilon_{T^*+2-13} \end{cases}$$

$$\begin{cases} \hat{X}^{T^*}_{T^*+1} = X_{T^*+1} - \epsilon_{T^*+1} \\ \hat{X}^{T^*}_{T^*+2} = X_{T^*+2} + \phi_1\hat{X}^{T^*}_{T^*+1} - \phi_1 X_{T^*+1} - \epsilon_{T^*+2} = -\phi_1\epsilon_{T^*+1} - \epsilon_{T^*+2} \end{cases}$$

Nous en déduisons ainsi que : $\begin{cases} X_{T^*+1} - \hat{X}^{T^*}_{T^*+1} = \epsilon_{T^*+1} \\ X_{T^*+2} - \hat{X}^{T^*}_{T^*+2} = \phi_1 \epsilon_{T^*+1} + \epsilon_{T^*+2} \end{cases}$

Ainsi, nous obtenons aisément la matrice variance-covariance Σ des erreurs de prévision. En outre, si nous faisons l'hypothèse que le processus des innovations suit un processus Gaussien et que les erreurs sont i.i.d, alors par linéarité et stabilité de la loi normale on a : $\mathrm{X} = \begin{pmatrix} X_{T^*+1} - \hat{X}^{T^*}_{T^*+1} \\ X_{T^*+2} - \hat{X}^{T^*}_{T^*+2} \end{pmatrix} \sim \mathrm{N}(0,\Sigma)$ avec $\Sigma = \begin{pmatrix} \sigma_\epsilon^2 & \sigma_\epsilon^2 \phi_1 \\ \sigma_\epsilon^2 \phi_1 & \sigma_\epsilon^2 (1+\phi_1^2) \end{pmatrix}$

En supposant que Σ est bien inversible, nous pouvons tester au seuil de $\alpha \in]0,1[$, le jeu d'hypothèse : $\begin{cases} H_0 : X = 0_{R^2} \\ H_1 : X \neq 0_{R^2} \end{cases}$

La statistique de test vérifie :

$$X^T \Sigma^{-1} X \sim_{H_0} \chi_2(2)$$

La région de rejet de l'hypothèse H_0 est donnée par :

$$R_\alpha = \{ \hat{X}^T \hat{\Sigma}^{-1} \hat{X} \geq q_{1-\alpha}^{\chi^2(2)} \}$$

où $q_{1-\alpha}^{\chi^2(2)}$ est le quantile d'ordre $1-\alpha$ d'une loi du chi-2 à deux degrés de liberté.

Une fois les estimations des coefficients et des écarts types réalisées, nous obtenons aisément les intervalles de confiance suivants pour les deux composantes du vecteur X :

$$\begin{cases} IC_{95\%}(X_{T^*+1}) = [\hat{X}^{T^*}_{T^*+1} - 1,96 \times \hat{\sigma}_\epsilon ; \hat{X}^{T^*}_{T^*+1} + 1,96 \times \hat{\sigma}_\epsilon] \\ IC_{95\%}(X_{T^*+2}) = [\hat{X}^{T^*}_{T^*+2} - 1,96 \times \hat{\sigma}_\epsilon \sqrt{(1+\hat{\phi_1}^2)} ; \hat{X}^{T^*}_{T^*+2} + 1,96 \times \hat{\sigma}_\epsilon \sqrt{(1+\hat{\phi_1}^2)}] \end{cases}$$

4.2 Hypothèses sous-jacentes

Comme nous l'avons dit dans ce qui précède, les résultats pour les bornes de l'intervalle de confiance à 95 % et 99 % nécessitent de faire l'hypothèse que le processus des innovations est gaussiens i.i.d. et 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).

4.3 Visualisation de la région critique pour $\alpha = 95\%$

Nous pouvons maintenant représenter graphiquement cette région pour $\alpha = 95\%$. Nous pouvons voir en gris clair l'intervalle de confiance à 95% sur la figure 12. Les points rouges correspondent à la prédiction :

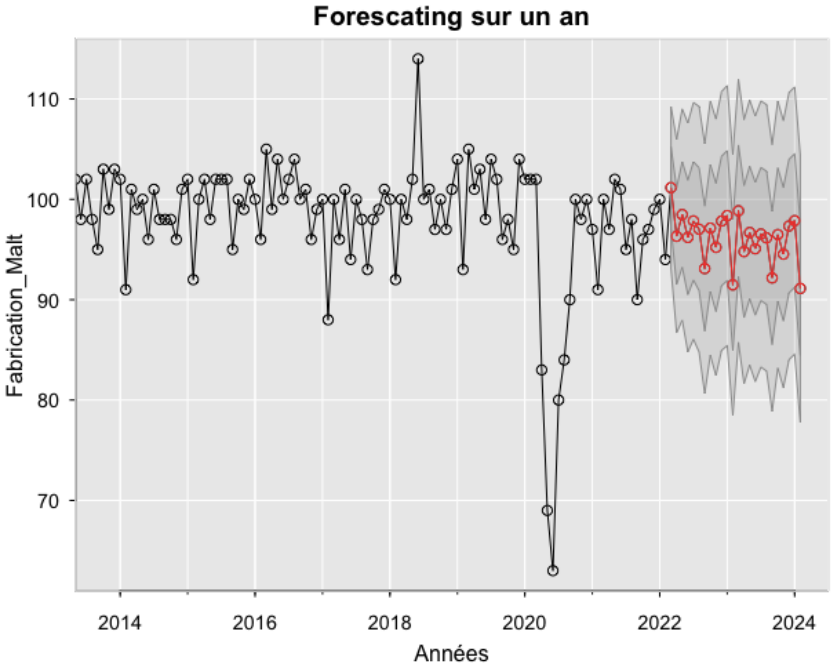


FIGURE 12 – Prédiction à l'horizon T+1 et T+2

Ici, on a fait un prévision jusqu'à 12 périodes, les valeurs des prédictions de $T+1$ et $T+2$ sont des les deux premiers points rouges.

5 Question ouverte

Dans cette question, l'objectif est de trouver un lien entre X_t et Y_t. L'idée est de trouver la combinaison linéaire reliant les deux séries entre elles : c'est le principe de variables co-intégrées.

Pour rappel, on dit que deux variables sont co-intégrées si elles possèdent toutes deux une tendance stochastique commune, en d'autres termes, elles admettent une relation de long terme.

Dans le cadre de notre question, nous souhaitons prédire X_{t+1} avec le plus de précision possible, en sachant que l'on possède Y_{t+1} avant X_{t+1}. Pour cela, il suffit de modéliser X_t avec son passé et la valeur présente de Y_t. On pourrait penser à un modèle de type VAR/VECM mais on peut aller plus loin en augmentant l'équation avec Y_t car on sait que Y_t est stationnaire.

Le modèle VAR(p) s'écrirait de la forme suivante :

$$\begin{cases} X_t = c + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{i=0}^{p} \omega_i Y_{t-i} + \epsilon_{1,t} \\ Y_t = b + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{p} \zeta_i X_{t-i} + \epsilon_{2,t} \end{cases} \tag{1}$$

Afin de tester la pertinence de l'ajout de Y_t dans la valeur de X_t, on testera le degrès de co-intégration entre ces deux variables et notamment la significativité de ω_0.

$$\begin{cases} H0 : \omega_0 = 0 \\ H1 : \omega_0 \neq 0 \end{cases}$$

Le concept que l'on souhaite tester est la causalité au sens de Granger(1969). Pour rappel, on dit qu'une variable X_t "cause" Y_t si X_t est utile à la prédiction de Y_t. On peut écrire

cette relation telle que :

$$C^{(h)}_{X \Leftrightarrow Y} :\equiv \hat{Y}_{t+h|\{X_u, Y_u, u \leq t\}} \neq \hat{Y}_{t+h|\{Y_u, u \leq t\}}, \forall h > 0. \tag{2}$$

Il faut tout de même tester la causalité instantanée au sens de Granger, c'est-à-dire, l'utilité de X_{t+1} pour prédire Y_{t+1} au temps t :

$$\hat{Y}_{t+1|\{X_u, Y_u, u \leq t\} \cup \{X_{t+1}\}} \neq \hat{Y}_{t+1|\{Y_u, u \leq t\}} \tag{3}$$

Pour cela, on peut regarder la corrélation des erreurs du système :
$\epsilon_t = \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} = \begin{pmatrix} X_t - \hat{X}_{t|\{X_u, Y_u, u \leq t-1} \\ Y_t - \hat{Y}_{t|\{X_u, Y_u, u \leq t-1} \end{pmatrix}$
On peut dire que (X_t) ne cause pas (Y_t) instannément si $\epsilon_{1,t}$ n'est pas corrélé à $\epsilon_{2,t}$.

Le test en découlant est un test de Wald qui suit une loi de khi deux de la forme qui admet comme hypothèse :

$$\begin{cases} H0 : R\vec{A} = 0 \\ H1 : R\vec{A} \neq 0 \end{cases} \tag{4}$$

Avec R est la matrice des contraintes composé de 0 et de 1 qui permet de sélectionner les coefficients de la régression linéaire à tester et $\vec{A}$ le vecteur des coefficients.

Si on rejette l'hypothèse nulle, on peut n'accepte pas la causalité au sens de Ganger, la série Y_t n'a donc aucune utilité pour la prédiction de X_t au temps t.

Si les tests s'avèrent significatifs, on peut écrire la prédiction de X_t avec une information supplémentaire et utile. Cette prédiction sera d'une variance d'erreur plus faible et donc plus efficace qu'un modèle VAR classique.

Pratiquement, on pourrait penser à la consommation de bière en France stationnarisé par intégration si nécessaire ou la demande mensuelle d'houblon en France par les brasseries industrielles.

Sans stationnarisé, on peut voir la relation proche entre la Fabrication de bière en France et celle du Malt :

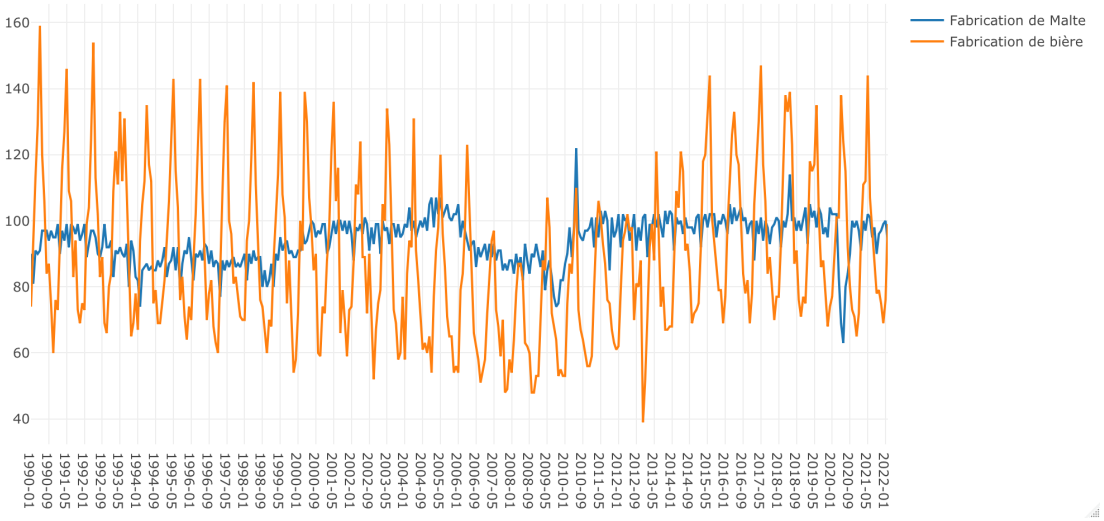


FIGURE 13 – Fabrication de la bière et du Malte en France depuis 1990

Les prochaines étapes sont la stationnarisation de la série Y_t : La fabrication de bière en France depuis 1990, puis les test de causalité au sens de Granger. On a récupéré cette série sur le site de l'Insee tout comme la série étudiée.

6 Script R

```
#########################################################
###### Projet - S ries temporelles lin aires #########
################### ENSAE ############################
##Auteurs : Th o Lorthios, M lissa Tamine ############
#########################################################


# Installation et chargement des packages utiles
#install.packages("tidyverse")
library(tidyverse)
#install.packages("plyr")
#install.packages("dplyr")
library(plyr)
#install.packages("hablar")
library(hablar)
#install.packages("tibble")
#install.packages("tseries")
library(tseries)
#install.packages("forecast")
library(forecast)
#install.packages("TSA")
library(TSA)
#install.packages("FitARMA")
library(FitARMA)
#install.packages("aTSA")
library(aTSA)
#install.packages('LSTS')
library(LSTS)
#install.packages('astsa')
library(astsa)
#install.packages('fUnitRoots')
library(fUnitRoots)
library(timeDate)
#install.packages('mvmeta')
library(mvmeta)
#install.packages('fpp2')
library(fpp2)
#install.packages('equatiomatic')
library(equatiomatic)

#########################################################
#                 Partie 1 : Les donn es                #
#########################################################

# Chargement et traitement des donn es

path = '/Users/melissa/Desktop/Scolaire/ENSAE/2A/Semestre 2/Time Series/Proje
df <- read.csv(path,sep=';', header =T)
names(df) <- c('Time','Y','Code')
date  <- df[,1]
Tmax = length(date)
df <- df[4:Tmax,1:2]
df <- arrange(df,Time)
```

```r
df <- df %>% convert(int(Y))
summary(df$Y)

# D finition de tsY : un objet de traval pratique (plus l ger et qui assoc
tsY <-ts(df$Y,start = c(1990,1), end = c(2022,2), frequency = 12, names = c(

# Visualisation de la s rie temporelle initiale
plot(tsY,type="l")
abline(h=mean(df$Y),col="red")

#Quelques commentaires :
# - Les chocs semblent  tre  stochastique (pas de changement structurel de la
# - Le choc de 2018 peut  tre  expliqu  par la coupe du monde et les fortes
# - Le choc de 2020 peut  tre  expliqu  par la crise du covid (confinement
# - Il ne semble pas y avoir de tendance apparante, la saisonnalit  m'a pas


#Nous affichons les ACF/PACF de la s rie initiale :

acf(tsY,lag.max = 200)
pacf(tsY,lag.max = 50)

# Puis nous affichons les d compositions additive et multiplicative de la s
decomposedadd <- decompose(tsY, type="additive")
plot(decomposedadd)
decomposedmu <- decompose(tsY, type="multiplicative")
plot(decomposedmu)

# Afin d'identifier la tendance dans la s rie temporelle initiale, nous r g
fita <- tslm(tsY ~ trend + season)
summary(fita)

# Nous supprimons la tendance et nous l'isolons pour l'afficher
dtY <- tsY-decomposedadd$trend
plot (dtY)
fitdt <- tslm(dtY ~ season)
summary(fitdt)
temps <-c(1:length(tsY))
reg <- lm(tsY~temps)
summary(reg)
res <- residuals(reg)
trendl <-tsY-res
plot(trendl) #Nous constatons que la tendance est lin aire

# Nous affichons ensuite la saisonnalit  sur une p riode de 3 ans
s <- decomposedadd$seasonal
plot(s[1:36],type='l')
abline(v=c(12,24),col="red",lwd=3, lty=2)
abline(h=1,lty=2,lwd=1,color='purple')
abline(v=c(2,14,26),col="green")
abline(v=c(9,21,33),col="blue") #Nous observons des piques significatifs en

#S rie initiale sans seasonalit
ssY <- tsY - decomposedadd$seasonal
plot(ssY,type="l")
```

```
#S rie initiale ans tendance et saisonnalit  (semble stationnaire)
sdY <- tsY - decomposedadd$seasonal-decomposedadd$trend
plot(sdY,type="l")

#3 tests de v rification de la stationnarit  de la s rie transform e
adf <- adf.test(sdY,nlag = 30)
pp.test(sdY, type = "Z_tau", lag.short = TRUE, output = TRUE)
kpss.test(sdY)

# - Les trois tests confirment que notre s rie n'est pas une marche al atoi
# - Nous pouvons donc mod liser la s rie telle qu'elle sans int grer ou d

#########################################################
#                 Partie 2 : Mod lisation               #
#########################################################


# Nous affichons l'ACF et la PACF de la saisonnalit  pour d terminer les o
acf(s[1:12]) #P = 1
pacf(s[1:12]) #Q = 1

#D termination des ordres p_max et q_max gr ce    l'ACF et la PACF de la s
sdYn <- na.omit(sdY)
acf(sdYn,lag.max = 100 )
pacf(sdYn,lag.max = 100)


#- PACF nous pr sente un MA(6) et ACF un AR(6) (les chocs lointain sont des
#- Nous construisons l'ensemble des mod les possibles et nous regardons pou

model <- list()
AIC <- list()
BIC <- list()
for (p in 0:6){
  for (q in 0:6){
    model <- paste('AR',(as.character(p)),'_','MA',as.character(q))
    fit <- Arima(sdY,order=c(p,0,q))
    bic <- BIC(fit)
    aic <- AIC(fit)
    models <- append(models,model)
    BIC <- append(BIC,bic)
    AIC <- append(AIC,aic)
  }
}

paste((models),(AIC),(BIC))

# Le mod le (ARIMA(4,0,2) minimse les crit res BIC et AIC

# Nous regardons ensuite les r sidus et les tests de Ljung-Box/Portmanteau
#H0 = r sidus ind pendant dans le temps (pas d'autocorr lation) => White
#H1 = r sidus d pendant => white noise rejet
#On a besoin d'une P-value > 5 % pour "valider" notre mod le
```

```
fit <- Arima(sdY, order=c(4,i,2))
checkresiduals(fit) #La p-value    11 % => white noise ok    5 %

# Nous pouvons ensuite v rifier la significativit  des coefficients avec u
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))
}

signif(fit) #ici le 1 er coefficient du MA(1) = 0

#Nous avons    pr sent les ordres maximums p,q pour notre mod le SARIMA =>
#Nous prenons le mod le le plus large et nous testons les combinaisons possi

#Nous test donc SARIMA(p,0,q,1,0,1)[12] avec p = 0,..,4 et q = 0,..,2
models <- list()
AIC <- list()
BIC <- list()

for (p in 0:4){
  for (q in 0:2){
    model <- paste('AR',(as.character(p)),'_','MA',as.character(q))
    fit <- sarima(tsY,p,0,q,1,0,1,12)
    bic <- fit$BIC
    aic <- fit$AIC
    models <- append(models,model)
    BIC <- append(BIC,bic)
    AIC <- append(AIC,aic)
  }}

paste((models),(AIC),(BIC))

#Le mod le le plus parcimonieux avec comme r sidus un bruit blanc qui mini

fit <- sarima(tsY,3,0,1,1,0,1,12)

signif(fit)

# Remarques : nous constatons que les probabilit  des quantiles de la loi n

#########################################################
#                  Partie 3 : Pr vision                 #
#########################################################

# Nous lan ons le forecasting :
forecasting <- sarima.for(tsY,24,3,0,1,1,0,1,12, gg=TRUE, main='Forescating␣

#Question ouverte

library(plotly)

dfFB <- read.csv('/Users/theoalegretti/Desktop/serie_010537308_09052022/valeu
```

```
names(dfFB) <- c('Time','X','Code')
date  <- dfFB[,1]
Tmax = length(date)
dfFB <- dfFB[4:Tmax-1,1:2]
dfFB <- arrange(dfFB,Time)
dfFB <- dfFB %>% convert(int(X))


fig <- plot_ly()
fig <- fig %>% add_trace(x = df$Time, y = df$Y, name = "Fabrication de Malte
fig <- fig %>% add_trace(x = df$Time, y = dfFB$X[1:length(df$Y)], name = "Fa
fig
```