ENSAE Paris

Projet de séries temporelles linéaires

Fabrication de produits laitiers

Matthieu Bricaire Yseult Masson

Avril - Mai 2022

Sommaire

1 Les données

1.1 Série choisie

Nous allons étudier dans ce projet la fabrication de produits laitiers, à travers l'indice de la production industrielle (IPI) associé.

Le site de l'INSEE indique que "les indices de la production industrielle permettent de suivre l'évolution mensuelle de l'activité industrielle de la France et de la construction". Ils sont calculés en utilisant les données de la France entière en ce qui concerne l'industrie (champ dont fait partie la fabrication de produits laitiers). Ces indices sont calculés mensuellement par l'INSEE, et ont pour année de référence 2015, ce qui signifie qu'ils ont pour moyenne 100 en 2015.

Nous avons à disposition une base de données contenant l'IPI de la fabrication de produits laitiers, entre janvier 1990 et février 2022. Cette série est corrigée des variations saisonnières et des jours ouvrés (CVS-CJO), et est disponible sur le site de l'INSEE à l'adresse https://www.insee.fr/fr/statistiques/serie/010537262.

La série est tracée sur la figure 1.

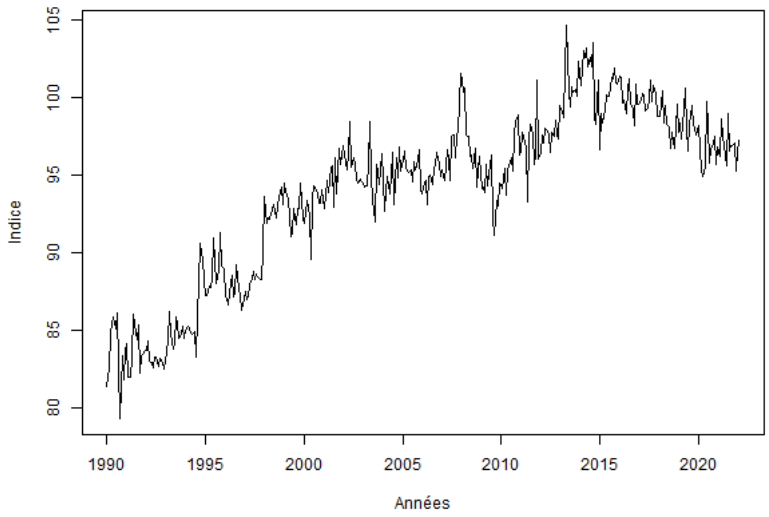


FIGURE 1 – Série initiale de l'indice de la production industrielle de produits laitiers

1.2 Transformation de la série

Comme on peut le voir sur la figure 1, la série ne semble pas stationnaire (il semble y avoir une tendance déterministe croissante). Nous avons effectué une régression linéaire de l'IPI sur les dates : le coefficient vaut 1.39×10^{-3}, et est significatif à tous les niveaux usuels (p-valeur inférieure à 2×10^{-16}), ce qui témoigne de l'existence d'une tendance déterministe croissante.

Nous avons vérifié la non-stationnarité à l'aide des tests ADF et KPSS. Les résultats de ces tests sont répertoriés dans la table 1.

Test	Lag order	Test statistic	p-value
ADF	7	-1.6955	0.7057
KPSS	5	5.3237	< 0.01

TABLE 1 – Résultats des tests de stationnarité pour la série initiale

La p-valeur du test ADF vaut 0.7057, donc on ne rejette pas l'hypothèse nulle d'existence de racine unitaire (donc de non stationnarité). D'autre part, pour le test KPSS, l'hypothèse nulle de stationnarité est rejetée à tous les niveaux usuels (p-valeur inférieure à 0.01). Ainsi, ces deux tests montrent que la série initiale de l'IPI de fabrication de produits laitiers n'est pas stationnaire.

Pour rendre la série stationnaire, nous l'avons différenciée. On peut voir sur la figure 2 que la série différenciée semble stationnaire :

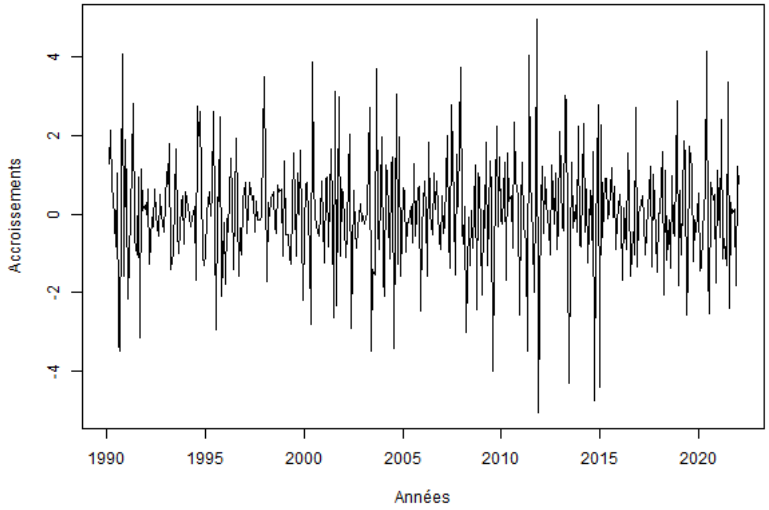


FIGURE 2 – Série différenciée de l'indice de la production industrielle de produits laitiers

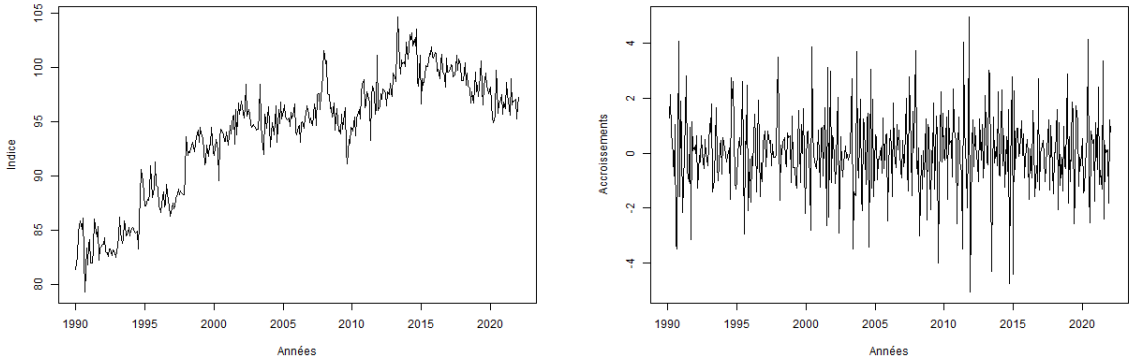
On effectue à nouveau les tests ADF et KPSS. Comme on peut le voir dans la table 2, la p-valeur du test ADF est inférieure à 0.01, donc l'hypothèse de non-stationnarité est rejetée à tous les niveaux usuels. De plus, le test KPSS ne rejette l'hypothèse de stationnarité à aucun des niveaux usuels (1%, 5% et 10%), car sa p-valeur est supérieure à 0.1. Ainsi, on peut considérer dans la suite que la série différenciée est stationnaire.

Test	Lag order	Test statistic	p-value
ADF	7	-9.2073	< 0.01
KPSS	5	0.10019	> 0.1

TABLE 2 – Résultats des tests de stationnarité pour la série differenciée

1.3 Représentation graphique

Les représentations graphiques des séries initiale et différenciée (figure 3) semblaient montrer que la série n'était initialement pas stationnaire, mais qu'une différentiation permettait d'obtenir la stationnarité, ce que nous avons vérifié ci-dessus avec les tests ADF et KPSS.



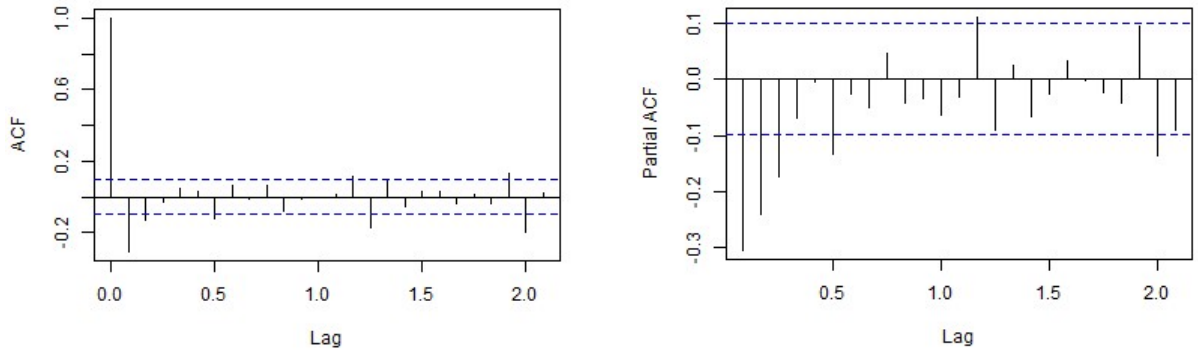
(a) Série initiale (b) Série différenciée

FIGURE 3 – Comparaison des séries initiale et différenciée

2 Modèle ARMA

2.1 Choix du modèle et validité

On veut à présent modéliser la série différenciée par un modèle ARMA(p, q). Pour choisir les paramètres p et q, on regarde ses graphes d'autocorrélation (pour q) et d'autocorrélation partielle (pour p), donnés dans la figure 4 :



(a) Autocorrélogramme (b) Autocorrélogramme partiel

FIGURE 4 – ACF et PACF de la série différenciée

Pour la partie MA du modèle, on étudie les autocorrélations (figure 4(a)). On considère qu'elles ne sont plus significatives à partir de q = 2 (on voit aussi un pic à q = 6, ainsi qu'à quelques ordres plus élevés, qui sortent de l'intervalle de confiance, mais on fait de choix de les ignorer afin de ne pas avoir des modèles avec trop de paramètres). En ce qui concerne la partie AR, on voit sur la figure 4(b) que les autocorrélations partielles sont significatives jusqu'à l'ordre 3 (de même, on ignore le pic à p = 6 et aux ordres supérieurs).

Ainsi, on cherche à modéliser la série différenciée par un ARMA(p, q) avec $p \in [0, 3]$ et $q \in [0, 2]$. Pour choisir le modèle le plus adapté, on utilise les critères AIC et BIC, le but étant de minimiser ceux-ci. Nous avons appliqué à la série les modélisations ARMA(p, q) pour $p \in [0, 3]$ et $q \in [0, 2]$, et calculé les critères AIC et BIC pour chacune d'entre elles. Les résultats sont dans les tables 3 et 4.

p \ q	0	1	2
0		1342.456	**1334.319**
1	1373.487	1334.720	1336.250
2	1351.599	1336.243	1338.233
3	1341.295	1338.183	1340.242

TABLE 3 – AIC

p \ q	0	1	2
0		1354.315	**1350.132**
1	1385.347	1350.533	1356.017
2	1367.412	1356.010	1361.953
3	1361.061	1361.903	1367.915

TABLE 4 – BIC

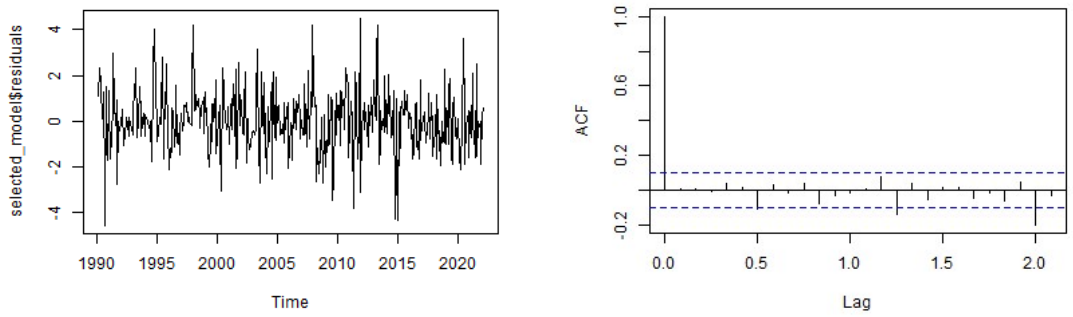
Les deux critères sont ici minimisés pour p=0 et q=2. On choisit donc le modèle ARMA(0,2) ou MA(2). On cherche maintenant à savoir si ce modèle est bien valide, en regardant la significativité des coefficients MA1 et MA2 et en vérifiant que les résidus ne sont pas autocorrélés.

Tout d'abord, nous avons testé la significativité des coefficients, à l'aide de tests de Student. D'après la table 5, les deux coefficients MA1 et MA2 sont significatifs (p-valeurs inférieures à 0.1, donc l'hypothèse nulle est rejetée à tous les seuils usuels). La constante n'est pas significative, donc peut être égale à 0.

	ma1	ma2	intercept
coeff	-0.45463255	-0.164659723	0.03588801
se	0.05000696	0.050678587	0.02641369
t-value	-9.09138537	-3.249098554	1.35868982
p-value	0.00000000	0.001157714	0.17424490

TABLE 5 – Coefficients du modèle MA(2)

D'autre part, pour que le modèle soit valide, il faut aussi que les résidus correspondent à un bruit blanc. Ces résidus et leurs autocorrélations sont tracés sur la figure 5 et semblent en effet se comporter comme un bruit blanc.



(a) Résidus (b) Autocorrélogramme des résidus

FIGURE 5 – Résidus du modèle MA(2)

Nous avons testé l'absence d'autocorrélation des résidus à l'aide de tests de Ljung-Box et de Box-Pierce. Les résultats, disponibles en annexe (tables 6 et 7), montrent que les résidus ne sont pas autocorrélés, pour des retards entre 3 et 20 (on considère qu'il n'y a pas non plus d'autocorrélation pour des retards supérieurs).

Ainsi, le modèle MA(2) est valide, et on peut l'appliquer à la série différenciée X_t. Avec l'estimation des coefficients vus plus haut, on a :

$$X_t = 0.04 + \epsilon_t - 0.45\epsilon_{t-1} - 0.16\epsilon_{t-2}$$

2.2 Modèle ARIMA

La série différenciée suit un modèle MA(2), donc la série initiale de l'IPI de fabrication de produits laitiers, qu'on note Y_t, suit un modèle ARIMA(0, 1, 2). Comme $X_t = Y_t - Y_{t-1}$, on obtient :

$$Y_t = 0.04 + Y_{t-1} + \epsilon_t - 0.45\epsilon_{t-1} - 0.16\epsilon_{t-2}$$

3 Prévision

On note T la longueur de la série. On suppose que les résidus de la série sont Gaussiens, c'est à dire que $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

3.1 Région de confiance de niveau α

On étudie un modèle MA(2) défini par :

$$X_t = m + \epsilon_t - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2}$$

Comme $\mathbb{E}[\hat{\epsilon}_{T+1}|X_T, X_{T-1}, ...] = \mathbb{E}[\hat{\epsilon}_{T+2}|X_T, X_{T-1}, ...] = 0$, les meilleures prévisions de X_{T+1} et X_{T+2} sachant le passé sont :

$$\begin{cases} \hat{X}_{T+1|T} &= m - \theta_1\epsilon_T - \theta_2\epsilon_{T-1} \\ \hat{X}_{T+2|T} &= m - \theta_2\epsilon_T \end{cases}$$

Les erreurs de prévision sont donc, en notant $X = (X_{T+1}, X_{T+2})^T$ et $\hat{X} = (\hat{X}_{T+1|T}, \hat{X}_{T+2|T})^T$:

$$\begin{aligned} \tilde{X} &= X - \hat{X} \\ &= \begin{pmatrix} (m + \epsilon_{T+1} - \theta_1\epsilon_T - \theta_2\epsilon_{T-1}) - (m - \theta_1\epsilon_T - \theta_2\epsilon_{T-1}) \\ (m + \epsilon_{T+2} - \theta_1\epsilon_{T+1} - \theta_2\epsilon_T) - (m - \theta_2\epsilon_T) \end{pmatrix} \\ &= \begin{pmatrix} \epsilon_{T+1} \\ \epsilon_{T+2} - \theta_1\epsilon_{T+1} \end{pmatrix} \end{aligned}$$

Les résidus sont supposés Gaussiens, de variance σ^2, donc :

$$\begin{aligned} &V(X_{T+1} - \hat{X}_{T+1|T}) = \sigma^2 \\ &V(X_{T+2} - \hat{X}_{T+2|T}) = \sigma^2 + \theta_1^2\sigma^2 = (1 + \theta_1^2)\sigma^2 \\ &Cov(X_{T+1} - \hat{X}_{T+1|T}, X_{T+2} - \hat{X}_{T+2|T}) = Cov(\epsilon_{T+1}, \epsilon_{T+2} - \theta_1\epsilon_{T+1}) = -\theta_1\sigma^2 \end{aligned}$$

Ainsi, le vecteur $\tilde{X}$ est gaussien, centré, et admet pour matrice de variance-covariance :

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & -\theta_1 \\ -\theta_1 & 1+\theta_1^2 \end{pmatrix}$$

Le déterminant de Σ vaut $\sigma^4 > 0$, donc Σ est inversible.

On en déduit que $\tilde{X}^T \Sigma^{-1} \tilde{X} \sim \chi^2(2)$. Cela nous permet d'établir une région de confiance de niveau α sur les valeurs futures de $X = (X_{T+1}, X_{T+2})^T$:

$$R_\alpha = \{ \, X \in \mathbb{R}^2 \;\; | \;\; \tilde{X}^T \Sigma^{-1} \tilde{X} \leq q_{1-\alpha} \, \}$$

où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(2)$.

3.2 Hypothèses nécessaires

Le calcul de la région de confiance de niveau α réalisé ci-dessus nécessite que les hypothèses suivantes soient vérifiées :

1. Le modèle théorique est connu, et les coefficients estimés obtenus en partie 2.1 sont égaux aux coefficients théoriques.
2. Les résidus sont un bruit blanc Gaussien suivant la loi $\mathcal{N}(0, \sigma^2)$.
3. La variance est connue et strictement positive ($\hat{\sigma^2} = \sigma^2 > 0$).

3.3 Représentation graphique

Nous avons tracé les prévisions de X_{T+1} et de X_{T+2} sur la figure 6. Les points bleus représentent les prévisions, et les zones grises les régions de confiance à 95% associées.

Les prévisions annoncent des variations négatives de l'IPI de fabrication de produits laitiers entre février et mars 2022 et entre mars et avril 2022 (les valeurs exactes sont en annexe, table 8). Cependant, les intervalles de confiance à 95% contiennent à la fois des valeurs positives et négatives, donc on ne peut pas conclure à une diminution certaine de l'IPI.

Enfin, selon les intervalles de confiance, il y a plus de 95% de chance que les variations de cet indice entre février et mars 2022, et entre mars et avril 2022, soient comprises dans l'intervalle [-3, 3].

3.4 Question ouverte

Supposons que l'on dispose d'une série stationnaire Y_t, disponible de t = 1 à T, et que Y_{T+1} est disponible plus rapidement que X_{T+1}. Cela peut permettre d'améliorer la prévision de X_{T+1} si et seulement si (Y_t) cause instantanément (X_t) au sens de Granger, soit

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

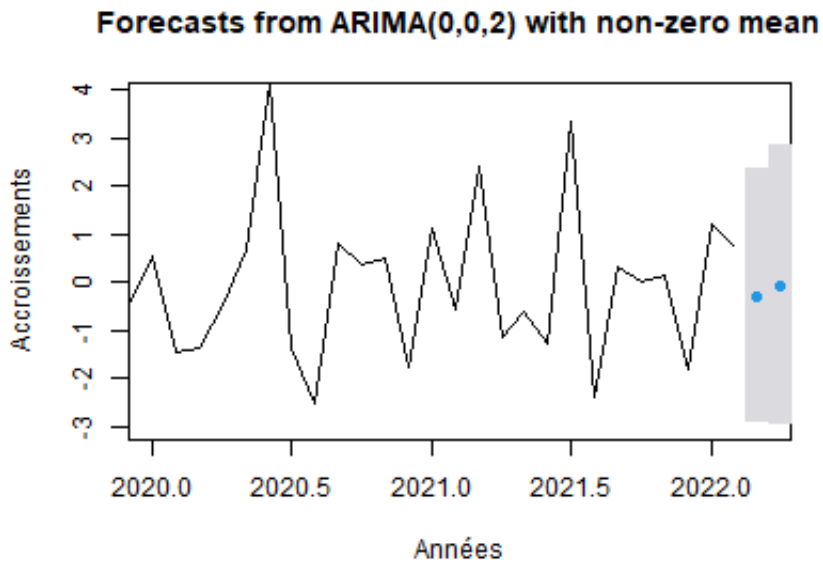


FIGURE 6 – Prévisions : mars et avril 2022

Cette condition peut être testée à l'aide d'un test de Wald, dont l'hypothèse nulle est H_0 : « Y ne cause pas X ». On peut notamment utiliser la fonction `causality` de R, qui renvoie le résultat du test de causalité et de causalité instantanée (ce qui nous intéresse ici) au sens de Granger.

4 Annexe

χ^2	df	p-value
0.072243	1	0.7881
0.57843	2	0.7489
0.67021	3	0.8802
4.9066	4	0.297
5.3768	5	0.3716
5.4587	6	0.4865
6.0487	7	0.5341
8.4234	8	0.3932
8.7583	9	0.4599
8.8201	10	0.5493
8.8535	11	0.6354
11.143	12	0.5167
18.79	13	0.1298
19.346	14	0.1521
20.432	15	0.156
20.534	16	0.1971
20.599	17	0.2447
21.43	18	0.2583
21.49	19	0.3104
22.879	20	0.2948

TABLE 6 – Test de Box-Pierce

χ^2	df	p-value
0.073037	1	0.787
0.5872	2	0.7456
0.68067	3	0.8777
5.0064	4	0.2866
5.4879	5	0.3593
5.572	6	0.4728
6.1792	7	0.519
8.6299	8	0.3745
8.9764	9	0.4395
9.0405	10	0.5283
9.0753	11	0.6149
11.464	12	0.4897
19.462	13	0.1095
20.045	14	0.1287
21.187	15	0.131
21.295	16	0.1675
21.364	17	0.2104
22.245	18	0.2213
22.308	19	0.2692
23.789	20	0.2517

TABLE 7 – Test de Ljung-Box

	Point Forecast	Lo 95	Hi 95
Mar 2022	-0.27520957	-2.928903	2.378484
Apr 2022	-0.05835238	-2.973420	2.856716

TABLE 8 – Prévisions aux horizons T+1 et T+2