Projet de Séries Temporelles Linéaires

ENSAE 2A 2021-2022

Modélisation de la fabrication française de produits azotés et d'engrais par ARIMA

KERROS

Victor

9 mai 2022

Table des matières

1 Les données

1.1 Présentation

Pour ce projet, j'ai décidé d'étudier la fabrication de produits azotés et d'engrais en France (lien).
Cette série chronologique compilée par l'INSEE représente l'indice mensuel de production industrielle depuis janvier 1990 jusqu'à mars 2022 avec une base 100 en 2015. Les données sont corrigées des variations saisonnières et des jours ouvrés (CVS-CJO). Aucune autre transformation (logarithmique, par exemple) n'a été effectuée.
J'ai donc à disposition 387 observations de la série temporelle représentée ci-dessous (Figure 1).

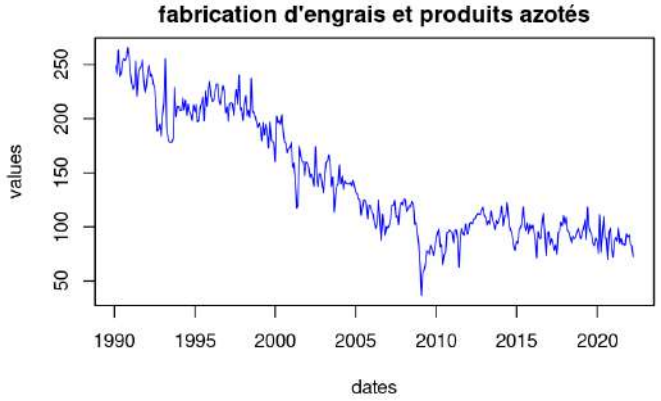

FIGURE 1 – Représentation de la série temporelle en niveau

On constate sur la Figure 1 une tendance négative important. On note également une chute de la production avec la crise de 2009 sur cette industrie. On ne remarque pas d'hétéroscédasticité ni de saisonnalité. On le confirme avec la décomposition de la série en tendance - saisonnalité - aléas en annexe (5) : tendance marquée, faible saisonnalité. En outre, l'aléas ne dépendant pas du temps. Il n'y a donc *a priori* pas besoin de faire une transformation logarithmique pour lisser l'hétéroscédasticité.

Par ailleurs, j'ai mis de côté les six dernières valeurs de la série afin de pouvoir tester les performances de nos différents modèles sur un semestre (en particulier avec le score RMSE - *root-mean-square error* ou erreur quadratique moyenne).

1.2 Stationnarisation

Le but de cette section est de vérifier si la série est stationnaire. Si elle ne l'est pas, il s'agira de la rendre stationnaire.
Pour rappel, la stationnarité d'une série traduit la constance dans le temps de certaines de ses propriétés statistiques : sa moyenne, sa variance et sa fonction d'autocorrélation. Cela permet de ensuite de modéliser la série par un processus ARMA.

1.2.1 Série en niveau

En premier lieu, j'ai vérifié si la série en niveau était stationnaire. Pour cela, j'ai réalisé un test de racine unitaire Dickey Fuller dont l'hypothèse H1 est la stationnarité de la série.

Ceci étant, il convient de vérifier au préalable par une régression linéaire les présences significatives ou non (p-valeur < 0.05) d'une tendance linéaire négative dans nos données et d'une constante afin de paramétrer le test ADF.

On effectue donc bien le test Dickey Fuller dans le cas avec constante et tendance d'après les résultats de la régression en annexe (6). Il en ressort une statistique de test - 5.5666 d'où une p-valeur de 0.01. Ainsi, le test est accepté : la série est stationnaire.

Cependant, le test Dickey Fuller n'est valide que si les résidus ne sont pas auto-corrélés. Or, l'absence d'autocorrélation des résidus est rejetée, le test ADF avec aucun retard n'est donc pas valide. On ajoute alors des retards jusqu'à ce que les résidus ne soient plus autocorrélés avant d'appliquer le test de Dickey Fuller Augmenté (ADF pour *Augmented Dickey Fuller*).

Avec six retards, les résidus ne sont plus autocorrélés. On constate alors que la série en niveau n'est pas stationnaire : le test ADF ne rejette pas l'hypothèse de racine unitaire (p-valeur $= 0.5003 > 0.05$) i.e. de non stationnarité de la série.

1.2.2 Série intégrée d'ordre 1

On passe à la série en différence première ou intégrée d'ordre 1 représentée sur la Figure 2.

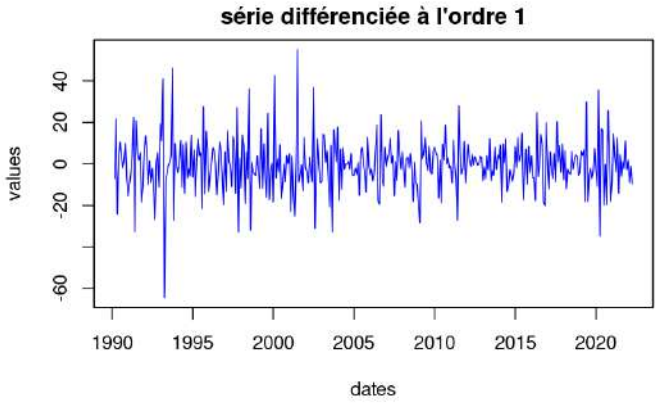

FIGURE 2 – Représentation de la série temporelle en différence première

On vérifie d'abord avec une régression linéaire que la série ne présente pas de tendance ni de constante.

Le test ADF sans tendance ni constante est valide avec cinq retards pour la série intégrée d'ordre 1. On retient son hypothèse alternative avec une p-valeur de 0.01 (statistique de test de -11.8503) : la série est stationnaire. On détermine désormais pmax et qmax pour évaluer le modèle ARIMA(p,1,q) adéquat.

1.3 Représentation graphique avant et après transformation

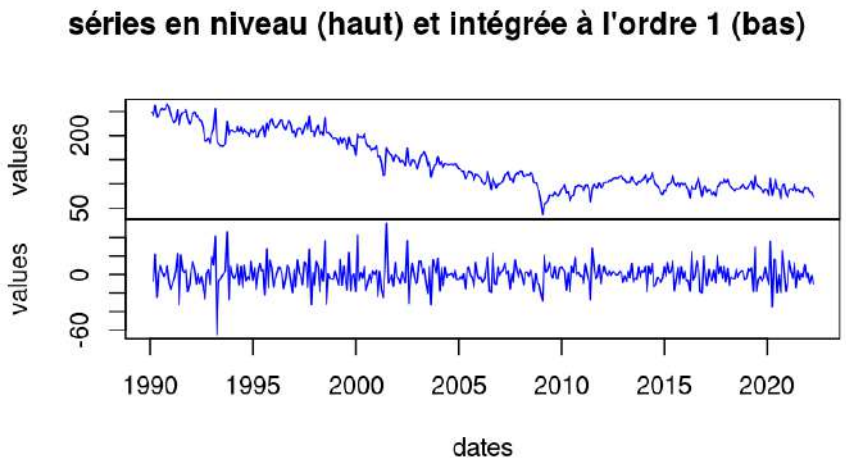

FIGURE 3 – Représentation de la série en niveau (haut) et intégrée d'ordre 1 (bas)

2 Modèles ARMA

2.1 Modèle $ARIMA(p, 1, q)$

2.1.1 Détermination de p_{max} et q_{max}

Pour modéliser la série à l'aide d'un modèle ARIMA(p,1,q), nous avons recours à la méthodologie Box-Jenkins décrite en annexe.

Il faut d'abord déterminer les ordres p_{max} et q_{max}. Pour cela, j'identifie les derniers pics significatifs des autocorrélogrammes de la série différenciée.
Sur l'autocorrélogramme en annexe (7), le dernier pic significatif est celui avec sept retards.
Sur l'autocorrélogramme partiel en annexe (8), le dernier pic significatif est celui avec six retards.
Je choisis donc $q_{max} = 1$ et $p_{max} = 6$.

J'ignore volontairement les éventuels autres pics légèrement significatifs au-delà du douzième retard dans l'objectif d'obtenir un modèle raisonnablement simple (de toute façon, la sélection par AIC et BIC effectuée ci-après pénalise les modèles insuffisamment parcimonieux).

2.1.2 Evaluation des modèles $ARMA(p, q)$ possibles pour la série différenciée

On cherche ensuite tous les modèles valides et justifiés i.e. dont les coefficients sont significatifs et dont les résidus ne sont pas autocorrélés.

On obtient alors cinq modèles valides (résidus non corrélés : colonne resnocorr) et justifiés (coefficients significatifs : colonnes arsignif et masignif) dont on évalue la qualité avec les critères AIC et BIC.

Les critères AIC et BIC sont obtenus à partir des formules suivantes où L est le maximum de la fonction de vraisemblance du modèle $ARMA(p, q)$:

$$AIC(p, q) = 2(p + q) - 2ln(L)$$

$$BIC(p, q) = (p + q)ln(n) - 2ln(L)$$

Après analyse des résultats disponibles en annexe (**??**), on retient les modèles arma(6,7) (minimise l'AIC) et arma(2,1) (minimise le BIC).

2.1.3 Choix du modèle $ARIMA(p, 1, q)$ pour la série en niveau

Pour choisir entre ces deux modèles, on effectue des prévisions pour les derniers six mois qui avaient initialement été retirés de la série. Ces prévisions sont disponibles en annexes (10).

Ensuite, on compare les performances prédictives des modèles avec le score RMSE où : - n représente le nombre d'observation (6 ici) - obs_i et $pred_i$ représentent respectivement l'observation et la prédiction à la

date i.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(obs_i - pred_i\Big)^2}$$

On sélectionne donc le modèle $ARIMA(2,1,1)$ qui a le score RMSE le plus faible.

	arima(6,1,7)	arima(2,1,1)
score RMSE	0.151	**0.139**

2.1.4 Expression du modèle $ARIMA(p,1,q)$ retenu

Soit, X_t la série différenciée qui suit un modèle $ARMA(2,1)$, alors on a l'expression suivante :

$$X_t = 0.3865X_{t-1} + 0.1101X_{t-2} + \epsilon_t + 0.8274\epsilon_{t-1} \tag{1}$$

ou, avec L l'opérateur retard,

$$(1 - 0.3865L - 0.1101L^2)X_t = (1 + 0.8274L)\epsilon_t \tag{2}$$

Y_t la série en niveau suit un modèle $ARIMA(2,1,1)$ et $Y_t = (1-L)X_t$ alors elle s'exprime comme suit :

$$(1 - 0.3865L - 0.1101L^2)(1 - L)Y_t = (1 + 0.8274L)\epsilon_t \tag{3}$$

3 Prévision et région de confiance

3.1 Equation de la région de confiance

On conserve les mêmes notations et on suppose que les résidus sont gaussiens. On réécrit alors X_{T+1} (où T est le nombre d'observations) à l'aide de l'équation 1.

$$X_{T+1} = \epsilon_{T+1} + \sum_{i=1}^{2} \phi_i X_{T+1-i} - \psi_1 \epsilon_T \tag{4}$$

On note $\hat{X}_{T+1}$ (respectivement $\hat{X}_{T+2}$) la meilleure prévision linéaire de X_{T+1} (respectivement X_{T+2}) à partir des valeurs connues en T (respectivement T+1).
Comme les résidus sont l'innovation linéaire (orthogonaux à toute fonction du passé linéaire), $EL(\epsilon_{T+1}|X_T, ...) = 0$. Par suite, on obtient d'après 4 l'équation suivante :

$$\hat{X}_{T+1} = EL(X_{T+1}|X_T, ...) = \sum_{i=1}^{2} \phi_i X_{T+1-i} - \psi_1 \epsilon_T = X_{T+1} - \epsilon_{T+1} \tag{5}$$

On réitère pour X_{T+2} et avec 5 on obtient le système suivant des erreurs de prédiction :

$$\begin{cases} \hat{X}_{T+1} - X_{T+1} = -\epsilon_{T+1} \\ \hat{X}_{T+2} - X_{T+2} = -\epsilon_{T+2} + (\psi_1 - \phi_1)\epsilon_{T+1} \end{cases} \tag{6}$$

Comme les résidus sont gaussiens (indépendants, de variance σ^2), on en déduit d'après 6 :

$$\begin{pmatrix} \hat{X}_{T+1} - X_{T+1} \\ \hat{X}_{T+2} - X_{T+2} \end{pmatrix} \to \mathcal{N}(0, E) \text{ avec } E = \begin{pmatrix} \sigma^2 & \sigma^2(\phi_1 - \psi_1) \\ \sigma^2(\phi_1 - \psi_1) & \sigma^2(1 + (1 + \psi_1 - \phi_1)^2) \end{pmatrix} \tag{7}$$

Le déterminant de E vaut $\sigma^4 > 0$ donc la matrice est inversible et par normalité des résidus :

$$\begin{pmatrix} \hat{X}_{T+1} - X_{T+1} \\ \hat{X}_{T+2} - X_{T+2} \end{pmatrix}' E^{-1} \begin{pmatrix} \hat{X}_{T+1} - X_{T+1} \\ \hat{X}_{T+2} - X_{T+2} \end{pmatrix} \to \chi^2(2) \tag{8}$$

D'où la région de confiance $RC(1-\alpha)$ avec $q_{1-\alpha}^{\chi^2(2)}$ quantile d'ordre $1-\alpha$ de la loi du $\chi^2(2)$:

$$\left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \begin{pmatrix} x_1 - \hat{X}_{T+1} \\ x_2 - \hat{X}_{T+2} \end{pmatrix}' E^{-1} \begin{pmatrix} x_1 - \hat{X}_{T+1} \\ x_2 - \hat{X}_{T+2} \end{pmatrix} \leq q_{1-\alpha}^{\chi^2(2)} \right\} \tag{9}$$

On peut aussi obtenir des intervalles de confiance (voir annexe 11).

3.2 Hypothèses de la région de confiance

On a fait deux hypothèses principales pour obtenir la région de confiance :
1) les résidus sont un bruit blanc gaussiens (*white noise*) ;
2) notre modèle $ARMA(2,1)$ est sous forme canonique.

3.3 Représentation des prédictions et de la région de confiance

J'ai réalisé les prévisions pour les mois d'octobre et novembre 2021. Les zones en gris clair représentent les intervalles de confiance à 95% pour $\hat{X}_{T+1}$ et $\hat{X}_{T+2}$ et les points rouges les prédictions.

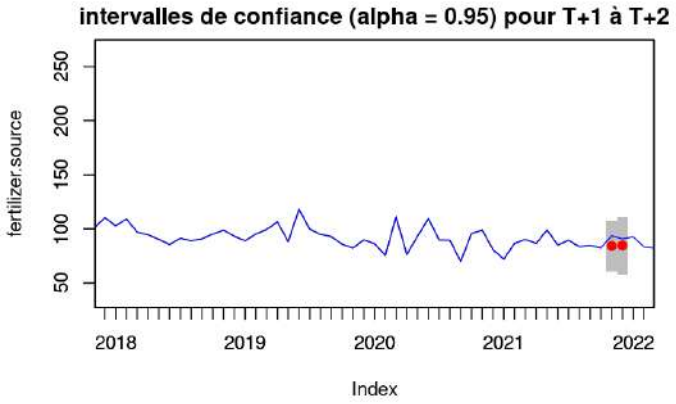

FIGURE 4 – Prédictions et région de confiance

Les prédictions sur la Figure 4 ne sont pas excellentes mais ont le mérite de suivre une tendance haussière conformément à la réalité. En revanche, on se rend compte sur la Figure 11 en annexe que ce modèle ARIMA(2,1,1) ne parvient pas à rendre compte de la variance de la série.

3.4 Question ouverte

Soit Y_t est une série stationnaire disponible de $t = 1$ à T. De plus, Y_{T+1} est disponible plus rapidement que X_{T+1}. D'après le cours (chapitre 5), Y_{T+1} améliore la prévision de X_{T+1} si (Y_t) cause (X_t) au sens de Granger i.e :

$$\hat{X}_{t+1|\{X_u,Y_u,u\leq t\}\cup\{Y_{t+1}\}} \neq \hat{X}_{t+1|\{X_u,Y_u,u\leq t\}} \tag{10}$$

où $\hat{X}_{t+1|\{X_u,Y_u,u\leq t\}}$ est la prévision linéaire de X_t sachant $X_u, Y_u, \forall u \leq t$. On peut tester cette hypothèse avec un test de Wald et ainsi éventuellement utiliser Y_{T+1} pour améliorer la prévision de X_{T+1}.

4 Annexes

4.1 Les données

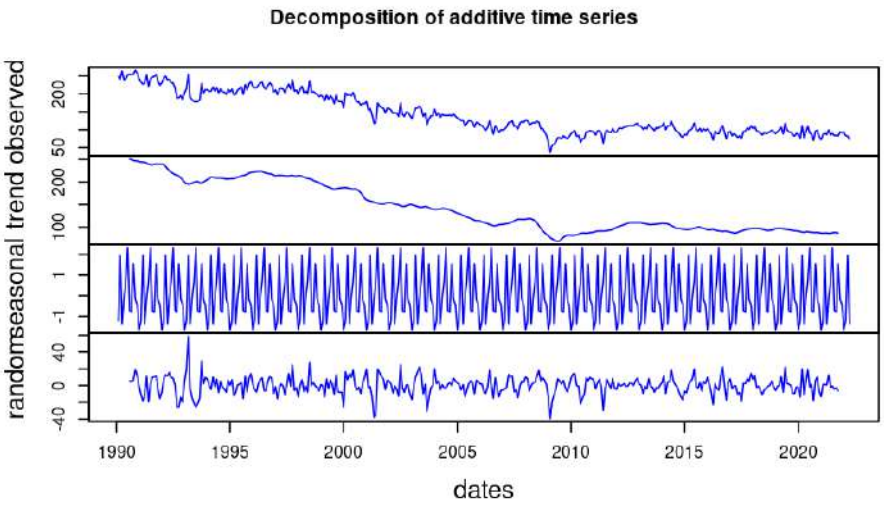

FIGURE 5 – Décomposition de la série temporelle en niveau

Coefficient	Valeur	p-value
intercept	11009.7244	$<$ 2e-16
dates	-5.4168	$<$ 2e-16

FIGURE 6 – Résultats de la régression linéaire de la série en niveau sur les dates

4.2 Modèles ARMA

4.2.1 Méthodologie Box-Jenkins

La méthodologie Box-Jenkins pour modéliser une série par ARMA(p,q) (ou ARIMA) se fait en quatre étapes :
- détermination des ordre p et q en utilisant les autocorrélogrammes (ACF, PACF) ;
- estimation des paramètres par méthode des moindres carrés ou maximum de vraisemblance ;
- vérifier que les modèles sont ajustés (les coefficients des ordres AR et MA les plus élevés sont statistiquement siginificatifs) et valides (les résidus ne sont pas autocorrélés) avec des tests (Portmanteau) ;
- choix du modèle (AIC, BIC, RMSE).

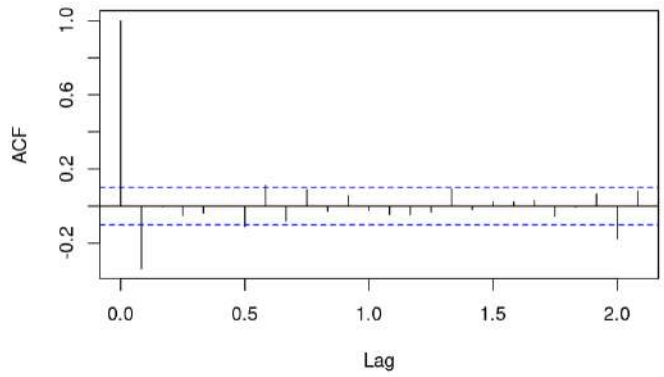

FIGURE 7 – Autocorrélogramme de la série différenciée à l'ordre 1

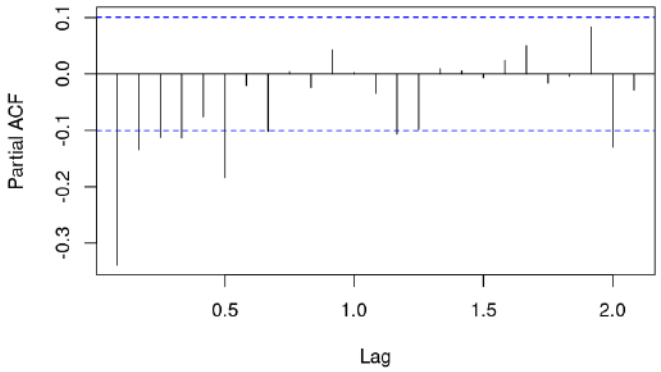

FIGURE 8 – Autocorrélogramme partiel de la série différenciée à l'ordre 1

Modèle	**AIC**	**BIC**
arma(6,0)	2970.048	3001.569
arma(2,1)	2966.686	2986.387
arma(0,3)	2968.878	2988.579
arma(3,6)	2970.190	3013.531
arma(6,7)	2965.223	3024.326

FIGURE 9 – AIC et BIC des différents modèles validés

mois	**obs**	**arima(6,1,7)**	**arima(2,1,1)**
nov. 2021	93.57	83.88629	84.21822
déc. 2021	90.88	86.69630	84.65183
janv. 2022	92.62	84.62050	85.01410
févr. 2022	83.39	88.58849	85.20186
mars 2022	82.21	84.87560	85.31432
avril 2022	72.59	87.29928	85.37846

FIGURE 10 – Observations et prédictions de la série temporelle en niveau

4.3 Prévision et région de confiance

Intervalle de confiance :

$$\begin{cases} \hat{X}_{T+1} - X_{T+1} \to \mathcal{N}(0, \sigma^2) \\ \hat{X}_{T+2} - X_{T+2} \to \mathcal{N}(0, \sigma^2(1 + (1 + \psi_1 - \phi_1)^2) \end{cases} \tag{11}$$

On obtient donc d'après 11 les intervalles de confiance au niveau α respectivement pour $\hat{X}_{T+1}$ et $\hat{X}_{T+2}$:

$$\begin{cases} [X^T_{T+1} + / - \sigma q_{1-\frac{\alpha}{2}}] \\ [X^T_{T+2} + / - \sigma \sqrt{(1 + (1 + \psi_1 - \phi_1)^2)} q_{1-\frac{\alpha}{2}}] \end{cases} \tag{12}$$

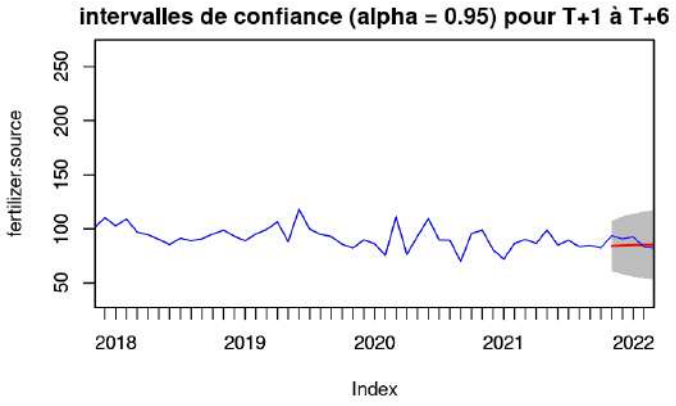

FIGURE 11 – Prédictions et région de confiance de T+1 à T+6