Projet de Séries Temporelles Linéaires

Modélisation d'un indice de production industrielle - Extraction de pétrole brut

Antony Albergne | **Carla Lucas**

Mai 2025

Contents

1 Les données

1.1 Description de la série choisie

Dans ce projet, nous étudions l'indice CVS-CJO de la production industrielle, spécifiquement pour l'extraction de pétrole brut (NAF rév. 2, niveau groupe, poste 06.1 [1]). Cette série est publiée sous forme d'indices mensuels corrigés des variations saisonnières et des jours ouvrables. Les données traitées sont de janvier 2000 à février 2025. La série brute initiale est représentée sur la figure 1.

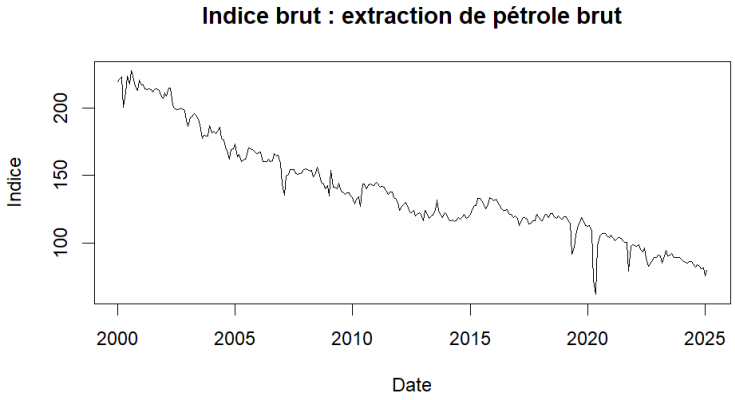

Figure 1: Série brute

1.2 Transformation de la série

La série brute présente une tendance linéaire décroissante, ce qui indique sa non-stationnarité. Pour la rendre stationnaire avant d'y appliquer les modèles ARMA et ARIMA, nous faisons une différenciation d'ordre 1 qui supprime la tendance.

$$X'_t = X_t - X_{t-1}$$

où X_t désigne la valeur de la série au moment t.
La série différenciée est représentée sur la figure 2.

[1] Site de l'Insee : `https://www.insee.fr/fr/statistiques/serie/010767578#Revision`

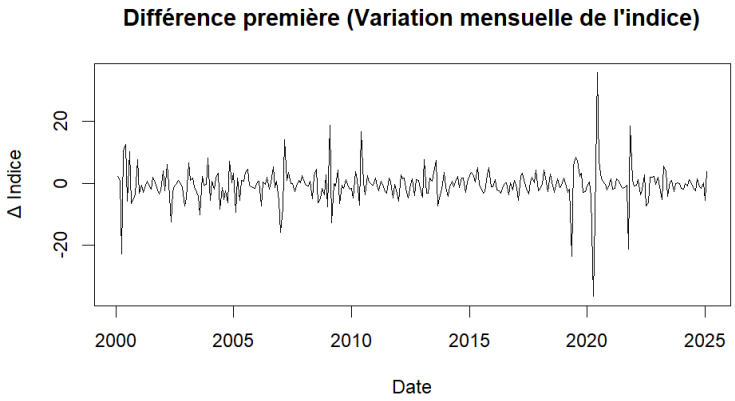

Figure 2: Série différenciée à l'ordre 1

Afin de vérifier la stationnarité de cette nouvelle série, nous faisons le test de Dickey-Fuller Augmenté (ADF) sur la série brute et la série différenciée à l'ordre 1, afin de les comparer.

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 9
  STATISTIC:
    Dickey-Fuller: -2.2718
  P VALUE:
    0.4617
```

Figure 3: ADF sur la série brute

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 5
  STATISTIC:
    Dickey-Fuller: -9.6155
  P VALUE:
    0.01
```

Figure 4: ADF sur la série différenciée

Les résultats 3 et 4 du test ADF montrent que, pour la série brute, la p-valeur très élevée ($= 0,4617$) ne permet pas de rejeter l'hypothèse nulle (présence d'une racine unitaire), ce qui confirme donc que la série brute n'est pas stationnaire. Une fois différenciée, la série est bien stationnaire, avec une p-valeur très faible ($= 0.01$) du test ADF, ce qui nous permet de rejeter l'hypothèse nulle de non-stationnarité de la série.

```
        Phillips-Perron Unit Root Test

data:  dindice
Dickey-Fuller Z(alpha) = -248.2, Truncation lag parameter = 5, p-value = 0.01
alternative hypothesis: stationary
```

Figure 5: Test Phillips-Perron

Nous voyons également sur le test Phillips-Perron 5 que la série différenciée est stationnaire,

avec une p-valeur très faible ($= 0,01$) ce qui nous permet de rejeter l'hypothèse nulle de non-stationnarité au seuil de 5%.
Nous utiliserons donc la série différenciée à l'ordre 1 pour la modélisation ARMA.

2 Modèles ARMA

2.1 Sélection du modèle ARMA

Pour sélectionner le modèle ARMA(p, q) approprié, nous examinons les fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF) de la série différenciée.

L'autocorrélation (ACF) théorique $\rho(k)$ et empirique $\hat{\rho(k)}$ au lag k sont définies par :

$$\rho(k) = \frac{\text{Cov}(X_t, X_{t-k})}{\sqrt{\text{Var}(X_t) \cdot \text{Var}(X_{t-k})}} \qquad \hat{\rho}(k) = \frac{\sum_{t=k+1}^{T}(X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_{t=1}^{T}(X_t - \bar{X})^2}$$

où $\bar{X}$ est la moyenne empirique de notre série et T la taille de notre échantillon.

La fonction d'autocorrélation partielle (PACF) mesure la dépendance linéaire entre X_t et X_{t-k} qui n'est pas expliquée par les lags intermédiaires. Pour $k \geq 2$, sa formule théorique est donnée par :

$$r(k) = \text{Corr}(X_t, X_{t-k} \mid X_{t-1}, \ldots, X_{t-k+1})$$

Une définition équivalente est que $r(k)$ est le coefficient de X_{t-k} dans la régression linéaire de X_t sur $\{1, X_{t-1}, \ldots, X_{t-k}\}$, ce qui permet de l'estimer empiriquement.

Les figures 6 et 7 montrent l'ACF et la PACF de notre série différenciée.

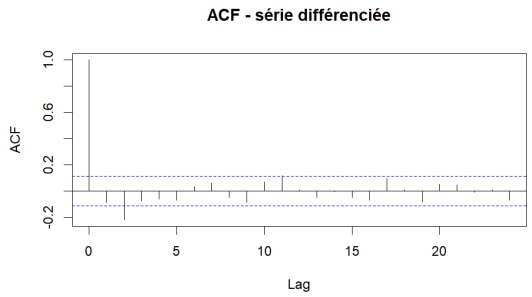

Figure 6: ACF de la série différenciée

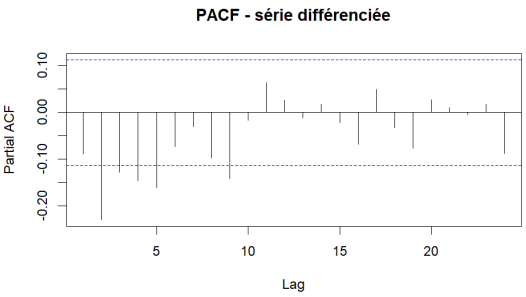

Figure 7: PACF de la série différenciée

L'ACF nous amène à privilégier un modèle MA(2), tandis que la PACF n'apporte pas de conclusion claire sur l'autocorrélation partielle, toutefois un AR(1) semble pertinent à explorer.
Nous estimons donc différents modèles ARMA(p, q) et utilisons les critères d'information d'Akaike (AIC) et bayésien (BIC) afin de les comparer et sélectionner le meilleur modèle. Les tableaux 1 et 2 montrent les valeurs d'AIC et de BIC pour différents modèles ARMA(p, q).

p	q=0	q=1	q=2
0	1890.815	1889.257	1868.437
1	1890.885	1872.132	1867.509
2	1878.146	1867.205	1866.048
3	1876.131	1869.197	1871.152
4	1873.208	1869.809	1866.781
5	1869.562	1870.985	1872.900

Table 1: AIC pour différents ARMA(p, q)

p	q=0	q=1	q=2
0	1894.522	1896.672	1879.558
1	1898.299	1883.253	1882.338
2	1889.268	1882.034	1884.583
3	1890.959	1887.733	1893.394
4	1891.744	1892.051	1892.730
5	1891.805	1896.935	1902.557

Table 2: BIC pour différents ARMA(p, q)

Nous voyons que le modèle qui minimise le critère AIC est le modèle ARMA(2,2) et celui qui minimise le BIC est le modèle MA(2).

Avant de s'intéresser aux coefficients de ces modèles, il faut vérifier leurs validités, c'est-à-dire que leurs résidus ne sont pas autocorrélés.
Pour cela nous menons un test de Ljung-Box sur les résidus des modèles jusqu'au lag 24 (correspondant à deux années).

```
tests d'absence d'autocorrélation des résidus :
     lag  pval lag  pval lag  pval lag  pval
[1,]   1    NA   7 0.088  13 0.126  19 0.189
[2,]   2    NA   8 0.127  14 0.161  20 0.215
[3,]   3 0.052   9 0.151  15 0.168  21 0.252
[4,]   4 0.049  10 0.135  16 0.136  22 0.297
[5,]   5 0.023  11 0.071  17 0.143  23 0.348
[6,]   6 0.049  12 0.099  18 0.185  24 0.253
```

Figure 8: Test de Ljung-Box - MA(2)

	MA(1)	MA(2)
Coefficient	-0.186	-0.293
Erreur standard	0.055	0.055
p-value	0.001	0.000

Figure 9: Coefficients - MA(2)

On remarque alors sur les résultats 8 et 9 que les coefficients du MA(2) sont significatifs, mais que les résidus du modèle sont autocorrélés. Le MA(2) n'est donc pas valide.

```
tests d'absence d'autocorrélation des résidus :
     lag  pval lag  pval lag  pval lag  pval
[1,]   1    NA   7 0.408  13 0.217  19 0.106
[2,]   2    NA   8 0.485  14 0.284  20 0.141
[3,]   3    NA   9 0.493  15 0.188  21 0.180
[4,]   4    NA  10 0.426  16 0.182  22 0.177
[5,]   5 0.093  11 0.235  17 0.145  23 0.216
[6,]   6 0.235  12 0.320  18 0.186  24 0.248
```

Figure 10: Test de Ljung-Box - ARMA(2,2)

	AR(1)	AR(2)	MA(1)	MA(2)
Coefficient	0.0419	-0.065	-0.0552	-0.177
Erreur standard	0.317	0.222	0.314	0.273
p-value	0.186	0.770	0.079	0.518

Figure 11: Coefficients - ARMA(2,2)

De même, on voit sur les résultats 10 et 11 que, pour le modèle ARMA(2,2), l'hypothèse de non-autocorrélation des résidus est rejetée au seuil de 10% au lag 5.

Il convient alors d'explorer les autres modèles ARMA par ordre de préférence en fonction des critères d'information.

```
tests d'absence d'autocorrélation des résidus :
     lag  pval lag  pval lag  pval lag  pval
[1,]   1    NA   7 0.493  13 0.251  19 0.105
[2,]   2    NA   8 0.545  14 0.324  20 0.139
[3,]   3    NA   9 0.529  15 0.219  21 0.177
[4,]   4 0.389  10 0.472  16 0.201  22 0.165
[5,]   5 0.197  11 0.270  17 0.158  23 0.202
[6,]   6 0.334  12 0.355  18 0.199  24 0.237
```

Figure 12: Test de Ljung-Box - ARMA(2,1)

	AR(1)	AR(2)	MA(1)
Coefficient	0.615	-0.193	-0.754
Erreur standard	0.099	0.070	0.084
p-value	0.000	0.006	0.000

Figure 13: Coefficients - ARMA(2,1)

Finalement, après une analyse exploratoire des modèles ARMA alternatifs via des tests de Ljung-Box et tests de nullité des coefficients, nous avont retenu le modèle ARMA(2,1), dont les résultats sont en figures 12 et 13. En effet, ce modèle est valide et a un AIC et un BIC satisfaisants.

2.2 Modèle ARIMA pour la série choisie

D'après l'analyse précédente, le modèle sélectionné est un ARMA(2,1). Comme nous avons appliqué une différenciation d'ordre 1 pour obtenir une série stationnaire, le modèle ARIMA qui convient pour la série brute est un ARIMA(2,1,1).

Le modèle ARIMA(2,1,1) peut être exprimé comme suit :

$$(1 - \phi_1 B - \phi_2 B^2)(1 - B)X_t = (1 + \psi_1)\epsilon_t$$

où :

- X_t est l'indice de la production de l'industrie manufacturière au moment t,
- $\phi_1 = 0.615$ et $\phi_2 = -0.193$ sont les coefficients autorégressifs,
- $\psi_1 = -0.754$ est le coefficient de la moyenne mobile,
- B est l'opérateur de décalage, tel que $BX_t = X_{t-1}$,
- ϵ_t est un bruit blanc.

Ainsi, le modèle ARIMA(2,1,1) pour notre série est :

$$(1 - 0.615B + 0.193B^2)(1 - B)X_t = (1 - 0.754B)\epsilon_t$$

Ce modèle capture à la fois la structure autorégressive et la structure de moyenne mobile de la série, tout en tenant compte de la différenciation nécessaire pour atteindre la stationnarité.

3 Prévision

Le modèle ARMA(2,1) pour la série différenciée s'écrit :

$$X_t = 0.615X_{t-1} - 0.193X_{t-2} + \epsilon_t - 0.754\epsilon_{t-1}$$

Par ailleurs, pour la suite de ce projet, nous noterons T la longueur de la série différenciée X_t. De plus, certaines hypothèses sont nécessaires, nous supposerons que :

- La série est stationnaire après transformation.
- Le modèle est connu.
- Les coefficients estimés sont les vrais coefficients de notre modèle.
- Le bruit blanc est gaussien et *iid*, $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$.
- $\sigma_\epsilon^2 > 0$, la variance du résidu est connue et non-nulle.

On peut déterminer directement la forme de $\hat{X}_{T+1|T}$ et $\hat{X}_{T+2|T}$ justement grâce au fait qu'ϵ_t soit une innovation linéaire. En effet, la forme canonique du modèle ARMA implique que les racines soient à l'extérieur du cercle unité et qu'il n'y ait pas de racines communes entre les polynômes auto-régressifs.

3.1 Région de confiance de niveau α

Sachant que $E[\epsilon_{T+h}|X_T, X_{T-1}, \ldots] = 0$ pour tout $h > 0$, par le cours, nous savons que les prévisions optimales en T sont données par :

$$\begin{cases} \hat{X}_{T+1|T} = 0.615X_T - 0.193X_{T-1} - 0.754\epsilon_T \\ \hat{X}_{T+2|T} = 0.615\hat{X}_{T+1|T} - 0.193X_T \end{cases}$$

Calculons alors les erreurs de prédiction $X_{T+1} - \hat{X}_{T+1|T}$ et $X_{T+2} - \hat{X}_{T+2|T}$. On a :

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

Ainsi :

$$X - \hat{X} = \begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} \\ X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} = \begin{pmatrix} \epsilon_{T+1} \\ \epsilon_{T+2} + 0.754\epsilon_{T+1} \end{pmatrix}$$

Nous pouvons ensuite calculer la variance des erreurs de prédiction :

$$\begin{cases} V(X_{T+1} - \hat{X}_{T+1|T}) = V(\epsilon_{T+1}) = \sigma_\epsilon^2 \\ V(X_{T+2} - \hat{X}_{T+2|T}) = V(\epsilon_{T+2} + 0.754\epsilon_{T+1}) = \sigma_\epsilon^2(1 + 0.754^2) \end{cases}$$

$X - \hat{X}$ étant un vecteur dont les composantes sont des combinaisons linéaires de variables gaussiennes indépendantes, il s'agit donc d'un vecteur gaussien qui suit une loi normale multivariée de paramètres $\mu = 0$ et Σ, c'est-à-dire $X - \hat{X} \sim \mathcal{N}(0, \Sigma)$, où Σ est la matrice de variance-covariance telle que :

$$\Sigma = \sigma_\epsilon^2 \begin{pmatrix} 1 & 0.754 \\ 0.754 & 1 + 0.754^2 \end{pmatrix}$$

Comme $\text{Det}(\Sigma) = \sigma_\epsilon^2$, la matrice de variance-covariance est inversible si et seulement si $\sigma_\epsilon^2 > 0$, ce que nous avons supposé vrai.
Par le cours, nous avons finalement ${}^t(X - \hat{X})\Sigma^{-1}(X - \hat{X}) \sim \chi^2(2)$. Ce qui nous permet d'en déduire directement la région de confiance de niveau α. On a ainsi $\forall \alpha \in [0, 1]$:

$$\{X \in R^2 \mid {}^t(X - \hat{X})\Sigma^{-1}(X - \hat{X}) \leq q_{\chi^2(2)}^{1-\alpha}\}$$

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

3.2 Représentation graphique

Nous pouvons maintenant représenter graphiquement cette région pour $\alpha = 5\%$. Nous pouvons voir en gris l'intervalle de confiance à 95% (figure 14). Les points bleus correspondent à la prévision.

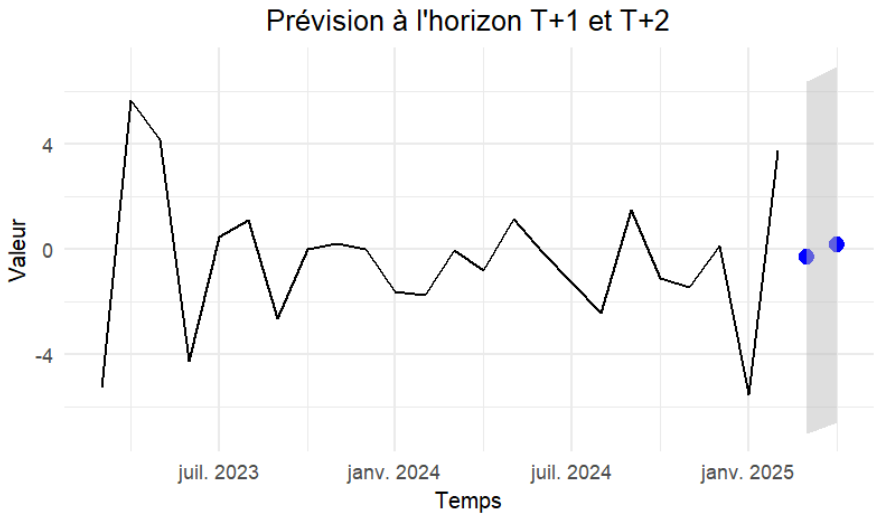

Figure 14: Prévision à l'horizon $T+1$ et $T+2$

3.3 Question ouverte

La question ouverte est la suivante : soit Y_t une série stationnaire disponible de $t = 1$ à T. On suppose que Y_{T+1} est disponible plus rapidement que X_{T+1}. Sous quelles conditions cette information permet-elle d'améliorer la prévision de X_{T+1} ? Comment les tester ?

Si l'on suppose que Y_{T+1} est disponible plus rapidement que X_{T+1}, alors on peut utiliser l'information de Y_{T+1} afin d'améliorer la prédiction de X_{T+1} si et seulement si :

- Les séries X_t et Y_t sont corrélées.
  C'est-à-dire $Corr(X_t, Y_t) \neq 0$.
  On peut tester la corrélation de X_t et Y_t en menant le test suivant :

$$\begin{aligned} H_0 &: \rho(X_t, Y_t) = 0 \quad \text{(absence de corrélation)} \\ H_1 &: \rho(X_t, Y_t) \neq 0 \quad \text{(corrélation)} \end{aligned}$$

  Statistique de test :

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim \mathcal{T}(n-2)$$

  Avec $r = \frac{\sum_{t=1}^{n}(X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{n}(X_t - \bar{X})^2 \sum_{t=1}^{n}(Y_t - \bar{Y})^2}}$

  On rejetterait alors H_0 si $|t| > t_{\alpha/2}(n-2)$.

- La série Y_t cause la série X_t au sens de Granger.

  Cette causalité au sens de Granger signifie que l'information passée de Y_t améliore significativement la prévision de X_t, c'est-à-dire que l'erreur quadratique moyenne de la prévision de X_{T+1}, en prenant en compte Y_t, est inférieure à l'erreur quadratique moyenne de prévision de X_{T+1} si nous n'avions pas pris en compte Y_t.

  Pour tester cette causalité, on estimerait deux modèles : un modèle 1 prenant en compte Y_t (avec γ_t les coefficients de Y_t dans ce modèle) et un modèle 2 ne prenant pas en compte Y_t.

  Nous testerions ensuite :

$$\begin{aligned} &H_0 : \gamma_1 = \gamma_2 \cdots = 0 \quad (Y \text{ ne cause pas } X) \\ &H_1 : \exists j : \gamma_j \neq 0 \end{aligned}$$

  En utilisant une statistique de test avec les erreurs quadratiques moyennes des modèles 1 et 2, nous rejetterions H_0 si $F > F_\alpha(q, n-p-q-1)$.

- Les séries X_t et Y_t sont dynamiquement corrélées.

  C'est-à-dire $Corr(X_{t+1}, Y_t) \neq 0$, donc que la série Y_t a un lien anticipé avec la série X_t.

  Pour tester la corrélation à différents décalages, nous calculerions :

$$\rho_{XY}(k) = \frac{\text{Cov}(X_t, Y_{t-k})}{\sqrt{\text{Var}(X_t)\text{Var}(Y_t)}}$$

  Nous testerions ensuite :

$$\text{Sous } H_0 : \rho_{XY}(k) = 0, \quad \hat{\rho}_{XY}(k) \sim \mathcal{N}\left(0, \frac{1}{n}\right)$$