Projet Séries Temporelles

Evolution mensuelle de de la fabrication d'instruments et de fournitures à usage médical et dentaire de janvier 1990 à janvier 2020

Khairaldin Ahmed, RAZIG Amine

ENSAE Paris
Palaiseau, France
Année universitaire 2023-2024

April 24, 2025

1 Partie I: Les données

1.1 Que représente la série choisie ?

Dans le cadre de ce projet, nous allons étudier l'évolution mensuelle de la fabrication d'instruments et de fournitures à usage médical et dentaire de janvier 1990 à janvier 2020. Cette série représente l'indice de production industrielle en base 100 en 2021, c'est-à-dire qu'un indice supérieur à 100 indique une production supérieure à l'année de référence (2021). De plus, la série est corrigée aux variations saisonnières et des jours ouvrables (CVS-CJO).

La série brute initiale, qu'on note X_t, est tracée sur la Figure 1:

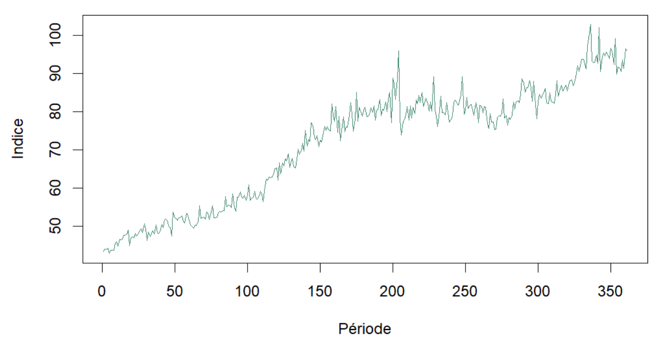


Figure 1: Série Brute Initiale

1.2 Transformation de la série

Graphiquement, la série semble avoir une tendance croissante. Cette tendance peut être due à une composante linéaire en t (déterministe) ou par un processus stochastique non linéaire. Pour corriger ce problème, nous différencierons plus tard notre série. Tout d'abord, vérifions que la série n'est pas stationnaire.

Pour ce faire, nous réalisons un test Augmented Dickey-Fuller dont l'hypothèse nulle est la présence d'une racine unité et donc la non-stationnarité de la série. L'hypothèse alternative est la stationnairté de la série. Ce test consiste en la régression de suivante (avec constante et tendance : cas général):

$$\Delta X_t = c + bt + \beta X_{t-1} + \sum_{\ell=1}^{k} \phi_\ell \Delta X_{t-\ell} + \varepsilon_t \tag{1}$$

où k représente le nombre de retards de la différence de la série temporelle inclus dans le modèle.

Ce paramètre k permet de déterminer le nombre de retards qu'il faudrait inclure dans la régression pour que le test soit valide, c'est-à-dire pour que les résidus ne soient pas autocorrélés. Nous vérifions alors pour chaque $k \geq 0$ l'autocorrélation des résidus [1].

Pour supprimer l'autocorrélation des résidus, il a fallu considérer 5 retards au test ADF dont les résultats sont en annexe. On ne rejette pas l'hypothèse nulle à un seuil de 95% : la série n'est donc pas stationnaire.

Nous différencions alors la série : $Y_t = X_t - X_{t-1}$. Pour tester la stationnarité de cette série, nous effectuons les mêmes étapes pour notre série initiale. Il a également fallu 5 retards au test ADF pour supprimer l'autocorrélation des résidus. Le test ADF correspondant donne une p-value inférieure à 0.01.

Par ailleurs, nous effectuons un test de Philipps-Perron sur la série différenciée dont l'hypothèse nulle est identique au test ADF. Les hypothèses du test PP sont :

- H_0 : La série a une racine unitaire (non-stationnaire).
- H_A : La série est stationnaire.

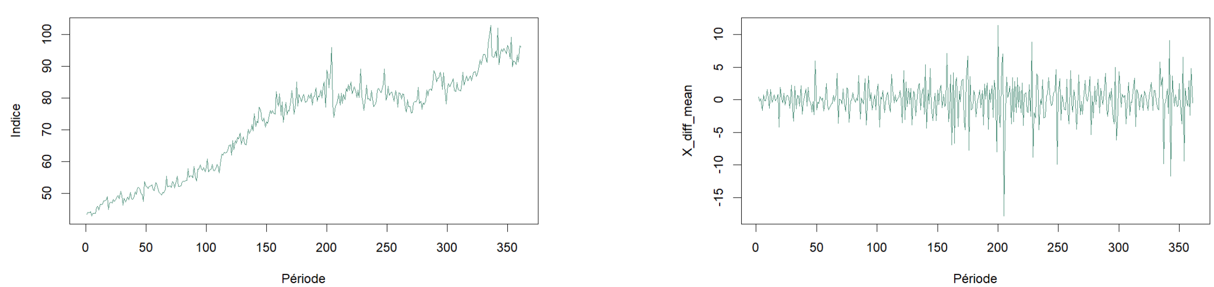
Résultats du test de Phillips-Perron:

Statistique de test	-418.98
Paramètre Lag	5
Valeur p	0.01
Hypothèse alternative	Stationary

Table 1: Résultats du test de Phillips-Perron pour la stationnarité

On rejette l'hypothèse nulle au seuil de 1% en faveur de l'hypothèse alternative de stationnarité. Nous considérons par la suite la série Y_t stationnaire (La série X_t est intégrée d'ordre 1).

1.3 Représentation de la série avant et après transformation



(a) Série initiale

(b) Série différenciée

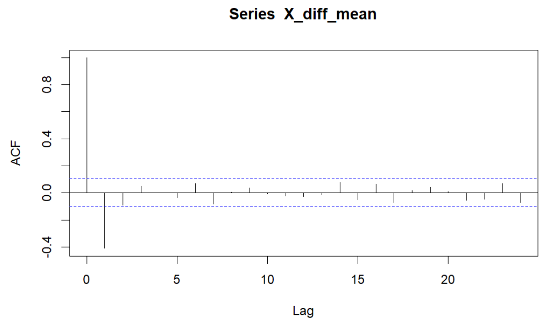
Figure 2: Série avant et après transformation

2 Partie II: Modèles ARMA

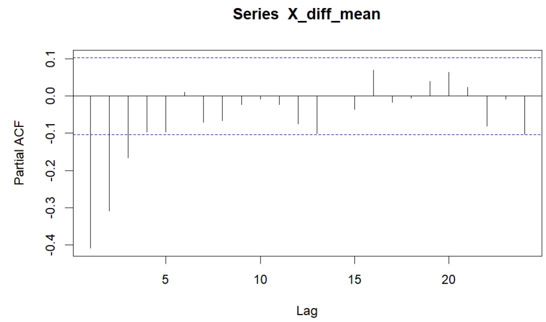
2.1 Modèle ARMA de la série différenciée

Afin de déterminer le modèle ARMA(p,q) approprié à la série différenciée Y_t, nous nous basons dans un premier temps sur les fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF). Elles sont représentées dans la Figure 3:

[1] Comme la série est mensuelle, nous testons l'autocorrélation des résidus jusqu'à l'ordre 24 (2 ans) grâce à des tests de Ljung-Box.



(a) ACF



(b) PACF

Figure 3: ACF et PACF de la série différenciée

Nous déduisons des deux figures les degrés maximaux : $p_{max} = 3$ et $q_{max} = 1$. Les modèles possibles sont alors tous les ARMA(p,q) tels que $p \leq 3$ et $q \leq 1$. Nous cherchons un modèle :

- bien ajusté, c'est-à-dire que les coefficients sont statistiquement significatifs.
- valide : les résidus ne sont pas autocorrélés.

Dans cette optique, nous effectuons pour chaque modèle possible un test de significativité des coefficients estimés et un test d'autocorrélations des résidus. Pour ce dernier, nous optons pour un test de Ljung-Box vérifiant l'absence jointe d'autocorrélations des résidus jusqu'à un ordre k donné. Comme dans la partie précédente, nous choisissons ici $k = 24$.

A titre d'exemple, les modèles AR(2) et ARMA(2,1) ne valident pas les deux tests (voir annexe). Le premier est bien ajusté mais n'est pas valide: les résidus sont autocorrélés. Tandis que le modèle ARMA(2,1) n'est pas valide, mais est bien ajusté: les coefficients correspondants aux plus hauts degrés sont significatifs.

Parmi tous les modèles possibles, les seuls modèles valides et bien ajustés sont les modèles AR(3) et MA(1) (voir annexe). Afin de déterminer lequel est le meilleur parmi les deux, nous calculons les AIC et BIC correspondants. Nous obtenons les valeurs suivantes des critères AIC et BIC:

	AR(3)	**MA(1)**
AIC	1708.614	1698.607
BIC	1728.044	1710.265

Table 2: Comparaison des modèles retenus

C'est le modèle MA(1) qui minimise les deux critères. On modélise alors notre série différenciée Y_t par un ARMA(0,1).

2.2 Modèle ARIMA pour la série initiale

Comme nous avons différencié la série initiale à l'ordre 1, nous la modélisons par un ARIMA(0,1,1).

3 Partie III: Prévision

3.1 Région de confiance de niveau α sur les valeurs futures

Notre série différenciée (non centrée) suit un ARMA(1,1) qu'on note :

$$Y_t = \mu + \epsilon_t - \theta\epsilon_{t-1} \tag{2}$$

Ainsi, comme on a que $Y_t = X_t - X_{t-1}$, notre série initiale s'écrirait:

$$X_t = X_{t-1} + Y_t = \mu + X_{t-1} + \epsilon_t - \theta\epsilon_{t-1} \tag{3}$$

Les prédictions sur les valeurs futures (X_{T+1}, X_{T+2}) sont données par:

$$\begin{cases} \hat{X}_{T+1|T} = \mu + X_T - \theta\epsilon_T \\ \hat{X}_{T+2|T} = \mu + \hat{X}_{T+1|T} - \theta\epsilon_{T+1} \end{cases}$$

car $\mathbb{E}[\epsilon_{T+h} \mid X_T, X_{T-1}, \ldots] = 0$

On remplace $\hat{X}_{T+1|T}$ dans la seconde équation, nous obtenons :

$$\begin{cases} \hat{X}_{T+1|T} = \mu + X_T - \theta\epsilon_T \\ \hat{X}_{T+2|T} = \mu + (\mu + X_T - \theta\epsilon_T) - \theta\epsilon_{T+1} = 2\mu + X_T - \theta\epsilon_T - \theta\epsilon_{T+1} \end{cases}$$

Les erreurs de prédiction sont données alors par:

$$\begin{pmatrix} e_{T+1} \\ e_{T+2} \end{pmatrix} = \mathbf{X} - \hat{\mathbf{X}} = \begin{pmatrix} \epsilon_{T+1} \\ \epsilon_{T+2} + (1-\theta)\epsilon_{T+1} \end{pmatrix}$$

avec :

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

En supposant que nos résidus suivent une loi gaussienne de variance σ_ϵ^2 alors, le vecteur d'erreurs de prévision est un vecteur gaussien de moyenne nulle et de matrice de variance-covariance :

$$\Sigma = \sigma_\epsilon^2 \begin{pmatrix} 1 & 1-\theta \\ 1-\theta & 1+(1-\theta)^2 \end{pmatrix}$$

Puisque θ est dans le cercle unité (la série Y_t étant stationnaire), la matrice Σ est inversible si et seulement si $\sigma_\epsilon > 0$, ce que nous supposons vrai. Le rang de la matrice est donc égal à 2. D'où, $\|\Sigma^{-1/2}(\mathbf{X}-\hat{\mathbf{X}})\| \sim \chi^2(2)$ avec $\|\cdot\|$ la norme euclidienne. La région de confiance de niveau α est alors donnée par:

$$R_\alpha = \{x \in \mathbb{R}^2 \mid \|\Sigma^{-1/2}(x - \hat{\mathbf{X}})\|^2 \leq \chi^2_{1-\alpha}(2)\}$$

où $\chi^2_{1-\alpha}(2)$ est le quantile $1-\alpha$ de la loi χ^2 à 2 degrés de liberté.

3.2 Hypothèses utilisées

Pour obtenir les résultats précédents, nous avons tout d'abord supposé que le modèle est parfaitement connu et que les coefficients obtenus dans la partie précédente sont les véritables coefficients de notre modèle.

Par ailleurs, comme mentionné au début de cette partie, nous avons supposé la normalité des résidus et le fait que leur variance est strictement positive dans le calcul de la matrice de variance-covariance. La variance qu'on utilise est un estimateur de la variance qu'on suppose égal à la variance théorique des rédidus (méthode plug-in).

3.3 Représentation graphique de la région de confiance

L'ellipse de confiance représente la région où nous nous attendons à trouver les valeurs futures X_{T+1} et X_{T+2} avec une probabilité de 95%. Les prédictions se trouvent au centre de l'ellipse. La taille et la forme de l'ellipse sont déterminées par la variance des résidus et les coefficients de notre modèle.

On remarque que l'ellipse est légèrement allongée et orientée, ce qui suggère qu'il y a une certaine corrélation entre les valeurs prévues à T+1 et T+2. Géneralement, si l'ellipse était plus circulaire, cela indiquerait une moindre corrélation.

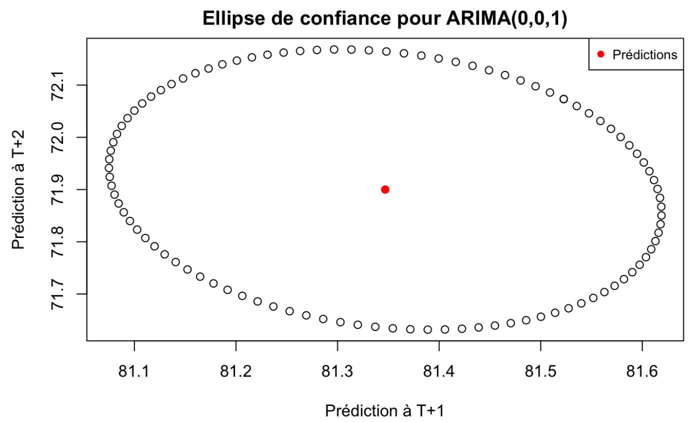


Figure 4: Ellipse de confiance

3.4 Question ouverte

Dans le cadre où Y_{T+1} est disponible plus rapidement que X_{T+1}, nous pouvons utiliser cette information pour améliorer la prédiction de X_{T+1} si Y cause X instantanément au sens de Granger. La causalité instantanée au sens de Granger signifie que, conditionnellement aux valeurs passées de Y_t et X_t, la valeur présente de Y_{T+1} améliore la prédiction de X_{T+1}. Plus formellement, on a :

$$\hat{Y}_{t+1|\{X_u,Y_u,u\leq t\}} \cup \{X_{t+1}\} \neq \hat{Y}_{t+1|\{X_u,Y_u,u\leq t\}} \tag{4}$$

Pour tester cette causalité, nous pouvons effectuer un test de causalité instantanée au sens de Granger pour lequel l'hypothèse nulle est : est la non-causalité instatanée de Y_t et X_t (dernière slide du cours). S'il y avait uniquement causalité au sens de Granger (non instantanée) :$\hat{Y}_{t+1|\{X_u,Y_u,u\leq t\}} \neq \hat{Y}_{t+1|\{X_u,Y_u,u\leq t\}}$, nous pourrions effectuer un test de Wald dont l'hypothèse nulle est la non causalité de Y_t et X_t.

4 Annexe

	Série Initiale	**Série Différenciée**
Lag Order	5	5
Dickey-Fuller	-2.0284	-10.0594
P-Valeur	0.5649	0.01

Table 3: Résultats des tests Augmented Dickey-Fuller sur les séries initiale et différenciée

4.1 Modèles valides et bien ajustés

Coefficient	**P-Value**
ma1	0.000
intercept	0.928

Table 4: P-Values pour la nullité des coefficients du modèle ARIMA(0,0,1)

Lag	**P-Value**	**Lag**	**P-Value**
1	NA	13	0.776
2	0.171	14	0.636
3	0.360	15	0.700
4	0.564	16	0.670
5	0.681	17	0.714
6	0.795	18	0.753
7	0.554	19	0.737
8	0.643	20	0.789
9	0.738	21	0.643
10	0.791	22	0.543
11	0.736	23	0.600
12	0.779	24	0.632

Table 5: P-Values pour les différents lags des tests d'absence d'autocorrélation des résidus

Coefficient	**P-Value**
ar1	0.000
ar2	0.000
ar3	0.001
intercept	0.965

Table 6: P-Values pour la nullité des coefficients du modèle ARIMA(3,0,0)

Lag	**P-Value**	**Lag**	**P-Value**
1	NA	13	0.313
2	NA	14	0.195
3	NA	15	0.254
4	0.015	16	0.233
5	0.046	17	0.275
6	0.090	18	0.312
7	0.070	19	0.320
8	0.122	20	0.382
9	0.184	21	0.269
10	0.240	22	0.174
11	0.246	23	0.212
12	0.238	24	0.247

Table 7: P-Values pour les différents lags des tests d'absence d'autocorrélation des résidus, ARIMA(3,0,0)

4.2 Exemples : modèles non retenus

Lag	**P-Value**	**Lag**	**P-Value**
1	NA	13	0.044
2	NA	14	0.021
3	0.000	15	0.032
4	0.001	16	0.037
5	0.002	17	0.040
6	0.005	18	0.048
7	0.005	19	0.047
8	0.009	20	0.065
9	0.016	21	0.047
10	0.028	22	0.028
11	0.027	23	0.037
12	0.029	24	0.049

Table 8: P-Values pour les différents lags des tests d'absence d'autocorrélation, AR(2)

Coefficient	**P-Value**
ar1	0.000
ar2	0.000
intercept	0.978

Table 9: P-Values pour la nullité des coefficients du modèle ARIMA(2,0,0)

Lag	**P-Value**	**Lag**	**P-Value**
1	NA	13	0.806
2	NA	14	0.649
3	NA	15	0.725
4	0.576	16	0.663
5	0.730	17	0.710
6	0.855	18	0.749
7	0.500	19	0.751
8	0.620	20	0.803
9	0.740	21	0.682
10	0.793	22	0.581
11	0.746	23	0.636
12	0.730	24	0.659

Table 10: P-Values pour les différents lags des tests d'absence d'autocorrélation (ARIMA(2,0,1))

Coefficient	**P-Value**
ar1	0.641
ar2	0.520
ma1	0.000
intercept	0.921

Table 11: P-Values pour la nullité des coefficients du modèle ARIMA(2,0,1)