UNIVERSITÉ SORBONNE PARIS NORD

SAE S3.03 :
Description et prévision de données temporelles

Réalisé par :
Ibrahim **BENKHERFELLAH**
Mohammed **BOUKHRISS**
Cherif Oumar **ABATCHA**

BUT SD - 2025/2026

RAPPORT D'ANALYSE
TOURISME AUX MALDIVES

Contexte : Analyse des séries chronologiques
Méthodologie : Approches Déterministes & Stochastiques
Modélisation : Désaisonnalisation & Méthode de Box-Jenkins
Outils : Google-Sheets en 1er puis Python
Rédaction : LaTeX

Date de rendu : 25 Janvier 2026
Professeur référent : Laurent LAVAL

IUT Villetaneuse
Université Sorbonne Paris Nord

Table des matières

Introduction

L'analyse des séries chronologiques permet d'étudier l'évolution d'un phénomène au cours du temps et de réaliser des prévisions à partir de données passées. Elle repose notamment sur l'identification de composantes telles que la tendance, la saisonnalité et les fluctuations aléatoires, dont la prise en compte est essentielle pour obtenir des modèles pertinents.

Ce projet est structuré en deux parties. La première partie consiste à analyser une série chronologique réelle présentant une composante saisonnière, issue d'une base de données officielle. Après l'étude de la série brute, une correction des variations saisonnières est effectuée afin d'obtenir une série CVS. Celle-ci est ensuite exploitée pour établir des prévisions à l'aide de deux méthodes distinctes : l'ajustement de la tendance par la méthode des moindres carrés et le lissage exponentiel. Une analyse de la saisonnalité complète cette étude.

La seconde partie est consacrée à l'étude des modèles ARIMA. À partir des données fournies, deux modèles, ARIMA(1,1,1) et ARIMA(1,1,2), sont estimés et comparés. Les prévisions obtenues sont confrontées aux valeurs observées afin d'évaluer la qualité des modèles et de mieux comprendre leur comportement prédictif.

1 Partie 1

1.1 Choix de la série chronologique

La série chronologique retenue pour cette étude est issue de la base de données officielle de la *Maldives Monetary Authority (MMA) Statistics Database*, en collaboration avec le *Ministry of Tourism and Environment*. Elle correspond au **nombre mensuel d'arrivées de touristes en provenance de la France aux Maldives**, sur la période allant de **janvier 1988 à novembre 2025**.

Cette série est particulièrement adaptée à l'analyse de la saisonnalité pour plusieurs raisons. Tout d'abord, il s'agit d'une série mensuelle, ce qui permet d'observer des variations régulières au cours de l'année. Or, l'activité touristique est fortement dépendante des saisons, notamment des conditions climatiques, des périodes de vacances scolaires et des habitudes de voyage, ce qui laisse naturellement présager l'existence d'une composante saisonnière marquée.

L'examen de la série brute met en évidence des fluctuations récurrentes avec **des pics et des creux qui se répètent d'une année sur l'autre**, traduisant une périodicité annuelle. Ces variations saisonnières sont cohérentes avec les périodes de forte et de faible fréquentation touristique. Par ailleurs, la série présente également des évolutions de long terme, liées à des facteurs économiques, géopolitiques ou sanitaires, justifiant la présence d'une tendance.

Ainsi, cette série constitue un support pertinent pour étudier la décomposition d'une série chronologique en tendance, saisonnalité et résidu, établir une série corrigée des variations saisonnières (CVS) et mettre en œuvre différentes méthodes de prévision.

Les données utilisées sont disponibles à l'adresse suivante : https://database.mma.gov.mv/viya/series/141

1.2 Nettoyage et préparation des données

Le nettoyage des données a été une étape rapide car le jeu de données initial était de bonne qualité. Afin de préparer la série chronologique pour l'analyse, nous avons procédé à une restructuration de l'axe temporel : la colonne *Date* a été convertie pour obtenir une vision trimestrielle (Année et Trimestre).

NB : Il a fallu procéder à une **agrégation temporelle** des données. Nous avons regroupé les relevés mensuels initiaux pour constituer une série trimestrielle.

Exemple :

TABLE 1 – Comparaison des jeux de données (Avant/Après)

Dataset Avant		Dataset Après		
Date	**Amount**	**Year**	**Quarter**	**Amount**
31/01/1988	751	1988	T1	2394
29/02/1988	807	...	...	...
31/03/1988	826			
...	...			

1.3 Modèle Additif ou Multiplicatif ?

Avant de procéder aux tests statistiques, une première observation empirique de la série chronologique a été réalisée. Celle-ci laisse apparaître une tendance générale à la hausse sur le long terme. La question principale est alors de déterminer comment les fluctuations saisonnières se comportent vis-à-vis de cette tendance : restent-elles constantes (modèle additif) ou s'amplifient-elles avec le temps (modèle multiplicatif) ?

Pour répondre à cette question de manière rigoureuse, nous avons exploité l'intégralité des données disponibles et croisé deux méthodes distinctes : l'analyse quantitative de Buys-Ballot et l'analyse graphique de la méthode de la bande.

1.3.1 Méthode de Buys-Ballot

Cette méthode consiste à étudier la corrélation entre les moyennes ($\overline{X_t}$) et les écarts-types ($\sigma_{\overline{X_t}}$) calculés annuellement.

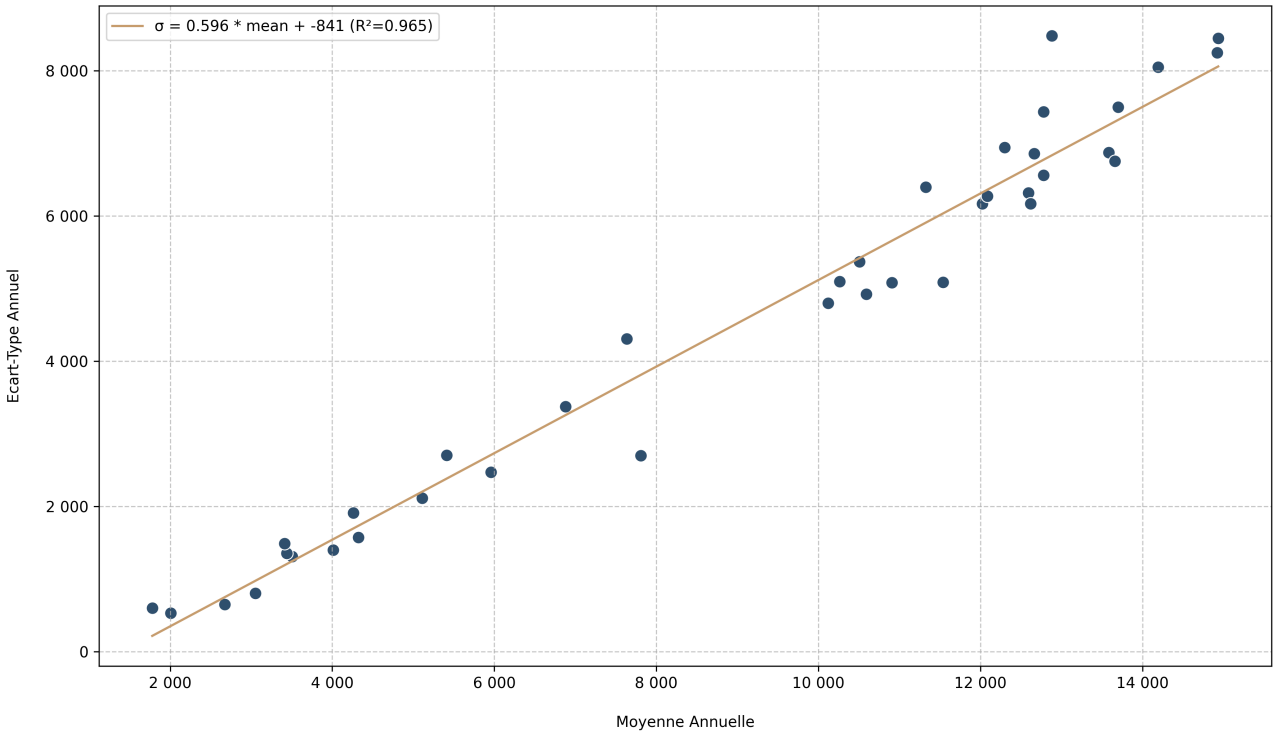

FIGURE 1 – Méthode de Buys-Ballot

La régression linéaire effectuée sur le nuage de points fournit l'équation d'ajustement suivante :

$$\sigma = 0,596\overline{X_t} - 841$$

Le coefficient directeur α est égal à **0,596**. Ce coefficient étant significativement différent de 0, nous constatons que l'écart-type n'est pas constant mais dépend fortement du niveau de la moyenne. Cela signifie que l'amplitude des variations saisonnières augmente proportionnellement à la tendance. Conformément aux règles de décision de la méthode de Buys-Ballot, ce résultat impose le choix d'un **modèle multiplicatif**.

1.3.2 Méthode de la bande

En complément, nous avons tracé l'enveloppe de la série chronologique en reliant d'une part les points extrêmes minima (creux saisonniers) et d'autre part les points extrêmes maxima (pics saisonniers).

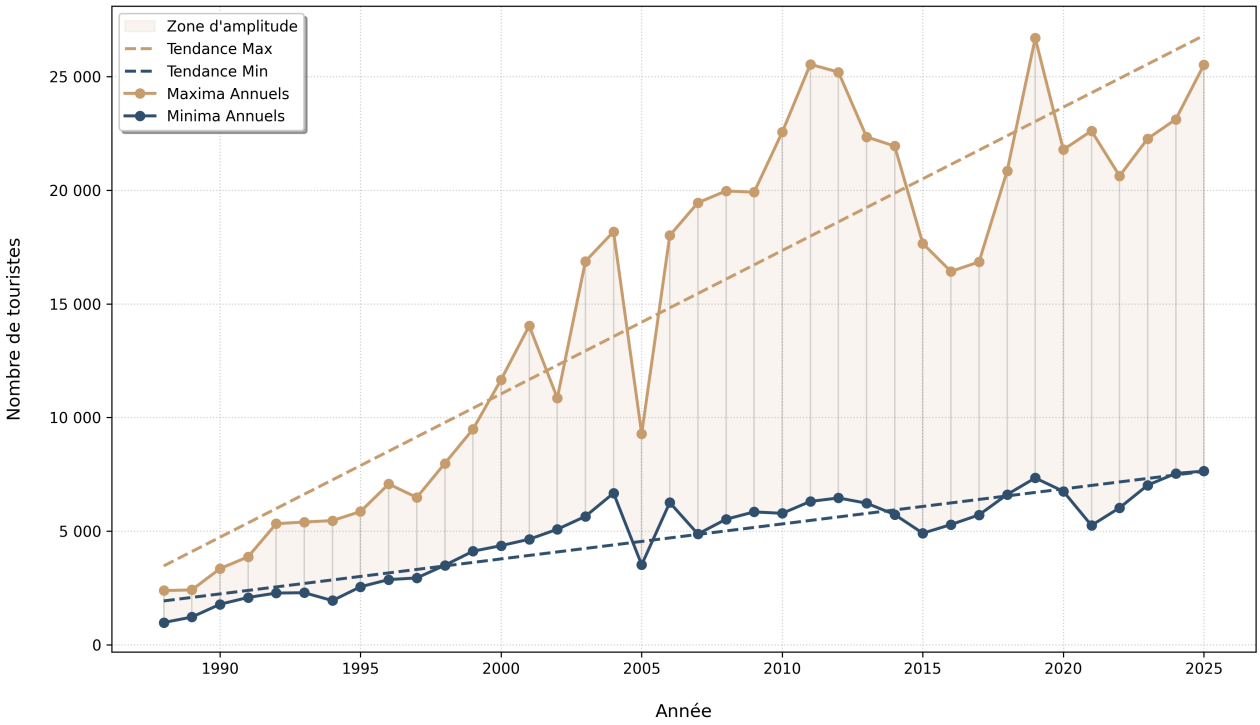

FIGURE 2 – Méthode de la Bande : visualisation de la Divergence

En appliquant cette méthode, on constate visuellement que les droites de tendance passant par les minima et les maxima ne sont pas parallèles mais **divergentes**. L'écart entre la courbe basse et la courbe haute s'agrandit au cours du temps (forme d'entonnoir) : l'amplitude des variations saisonnières n'est donc pas constante, elle croît avec la tendance générale. Cette géométrie valide, elle aussi, le choix du **modèle multiplicatif**.

1.3.3 Conclusion sur le choix du modèle

Les deux méthodes aboutissant à la même conclusion, nous retenons définitivement un **modèle multiplicatif** pour la décomposition de cette série chronologique. Le modèle s'écrira sous la forme suivante :

$$X_t = C_t \times (1 + S_t) \times (1 + \epsilon_t)$$

où X_t est la série brute, C_t la tendance, S_t la composante saisonnière et ϵ_t (ou C_R) la composante résiduelle.

1.4 Périmètre d'étude et préparation de la série

Délimitation de la période d'étude

Même si la base couvre une longue période, il est important d'adapter la fenêtre d'analyse aux objectifs de l'étude. Ici, l'enjeu principal est d'identifier la saisonnalité trimestrielle, d'estimer une tendance récente et de produire des prévisions à court terme. Dans cette logique, on retient la période 2015–2025 : elle est suffisamment longue pour mettre en évidence des variations saisonnières et calculer des coefficients saisonniers représentatifs, tout en restant centrée sur une dynamique **actuelle** du tourisme.

Traitement de la rupture structurelle (COVID-19)

Avant de procéder au calcul de la série CVS, une analyse critique de la série brute est nécessaire. L'observation des données sur la période 2020-2021 révèle une rupture structurelle importante liée à la pandémie de COVID-19 (fermeture des frontières, arrêt du trafic aérien).
Cette anomalie ne constitue pas une variation aléatoire classique mais une aberration qui, si elle est conservée telle quelle, introduirait deux gros biais dans notre modèle :

- **Biais de Tendance** : l'effondrement des valeurs tirerait artificiellement la moyenne mobile vers le bas.
- **Biais de Saisonnalité** : le calcul des coefficients saisonniers serait faussé par des valeurs quasi-nulles, détruisant la structure saisonnière moyenne.

Conformément à la méthodologie du cours sur la ***Mise en forme préparatoire des données***, il est nécessaire de procéder à un nettoyage des données avant analyse. Le support de cours préconise explicitement d'***éliminer les aberrations*** et valide la ***création de données artificielles pour combler des manques***. Nous avons opté pour une méthode de ***reconstitution par années témoins fixes***. Cette approche consiste à remplacer les valeurs aberrantes (identifiées du T2-2020 au T2-2021) par la moyenne des valeurs observées sur des années de référence saines, situées avant et après la crise.

Les années témoins retenues sont :

- **Avant-COVID :** 2018 et 2019
- **Post-COVID :** 2022 et 2023

Soit $x_{A,T}$ la valeur brute pour l'année A et le trimestre T. Soit $\mathcal{A} = \{2018, 2019, 2022, 2023\}$, l'ensemble des années témoins. La valeur rectifiée est calculée ainsi :

$$x_{A,T}^{corr} = \frac{1}{|\mathcal{A}|} \sum_{k \in \mathcal{A}} x_{k,T}$$

Cette méthode présente l'avantage de **préserver le profil saisonnier** tout en ajustant le niveau de la tendance à une valeur théorique "normale". Cela garantit que les Coefficients Saisonniers calculés par la suite refléteront la saisonnalité structurelle du tourisme aux Maldives, et non l'accident conjoncturel de la crise sanitaire. De plus, nous pensons que ces valeurs estimées représentent celles qui auraient dû être observées durant cette période. C'est sur cette base de données rectifiée que nous établirons la série CVS et les modèles de prédictions.

1.5 Établissement de la série CVS via la MMC

1. Extraction de la tendance lissée (via la MMC)

Afin d'éliminer les variations saisonnières et accidentelles, nous calculons une moyenne mobile centrée d'ordre 4 (trimestres). Pour centrer correctement la moyenne sur le temps t, nous utilisons donc une moyenne pondérée sur 5 termes via la formule suivante :

Nous retenons un **ordre 4** car la série est **trimestrielle** : une année est composée de **4 trimestres**. Utiliser une moyenne mobile centrée d'ordre 4 permet donc de lisser la série sur un cycle complet, ce qui atténue naturellement les fluctuations saisonnières et met mieux en évidence la tendance.

$$\begin{aligned}\widehat{C_t} = MMC_4 &= \frac{1}{4}\left(\frac{x_{t-2}}{2} + \sum_{i=-1}^{1} x_{t+i} + \frac{x_{t+2}}{2}\right) \\ &= \frac{1}{4}\left(\frac{x_{t-2}}{2} + x_{t-1} + x_t + x_{t+1} + \frac{x_{t+2}}{2}\right)\end{aligned}$$

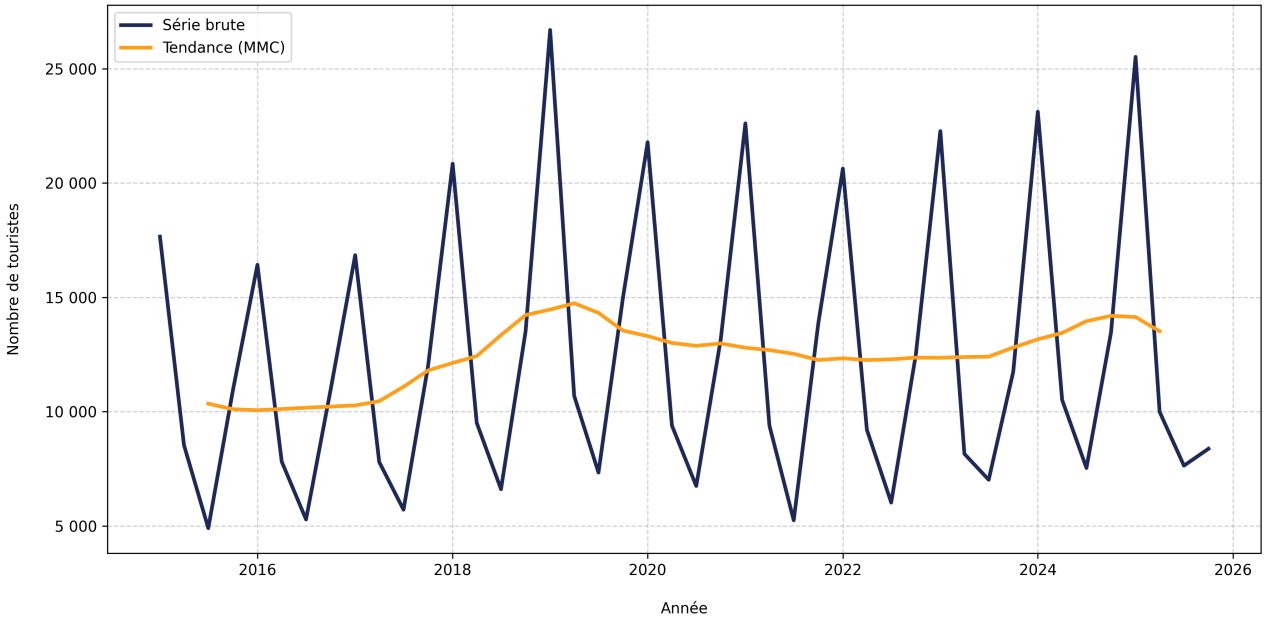

FIGURE 3 – Comparaison : Série brute vs Composante Tendancielle Ĉt

La **MMC** est très utile ici pour visualiser la tendance, car elle atténue fortement les variations trimestrielles qui rendent la série brute difficile à lire. En lissant les valeurs sur plusieurs périodes, elle met davantage en évidence l'évolution de fond : on distingue alors plus clairement les phases de hausse, de stabilisation ou de reprise, sans être "parasité" par la saisonnalité.

2. Calcul des Coefficients Saisonniers

La détermination de la composante saisonnière repose sur la méthode du *Rapport à la Moyenne Mobile*. Cette approche s'effectue en trois étapes successives : le calcul des rapports bruts, leur moyenne par trimestre, et enfin leur normalisation.

Calcul des Rapports Bruts (r_t)

Dans un modèle multiplicatif en divisant la série brute par la tendance lissée (estimée par MMC_t), on isole le produit de la saisonnalité et de l'aléa :

$$r_t = \frac{X_t}{MMC_t} = \frac{C_t \times (1 + S_t) \times (1 + \epsilon_t)}{C_t} \approx (1 + S_t) \times (1 + \epsilon_t)$$

Ces rapports quantifient, pour chaque trimestre, l'écart relatif par rapport à la tendance générale.

Estimation des Coefficients Saisonniers

Pour éliminer la composante aléatoire ϵ_t, nous calculons la moyenne arithmétique des rapports r_t pour chaque trimestre i (T1, T2, T3, T4).

$$1 + S_i^{brut} = \frac{1}{n_t} \sum_k r_{i,k}$$

Avec :

- i : l'indice du trimestre (fixe pour le calcul) ;
- k : l'indice représentant l'année (variable de sommation) ;
- n_i : le nombre total d'années disponibles pour le trimestre i.

Normalisation des Coefficients

Dans un modèle multiplicatif trimestriel, la somme des coefficients saisonniers doit être strictement égale à 4 (pour que leur moyenne soit 1). Si ce n'est pas le cas, un facteur de correction est appliqué :

$$1 + S_i = (1 + S_i^{brut}) \times \frac{4}{\sum (1 + S_j^{brut})}$$

Dans notre cas, on a :

$$\sum_{i=1}^{4} (1 + \widehat{S}_i) = 1,73 + 0,74 + 0,51 + 1,02 = 4,00 \quad \Longleftrightarrow \quad \sum_{i=1}^{4} \widehat{S}_i = 4 - 1 \times 4 = 0$$

Donc pas besoin d'appliquer de facteur de correction. Les résultats obtenus pour le tourisme aux Maldives sont présentés dans le tableau ci-dessous :

Table 2 – Coefficients Saisonniers Normalisés

Trimestre	**T1**	**T2**	**T3**	**T4**
Coefficient $(1 + \widehat{S}_i)$	1,73	0,74	0,51	1,02
Interprétation	Pic annuel (+73%)	Creux (-26%)	Basse saison (-49%)	Reprise (+2%)

Graphique à la page suivante.

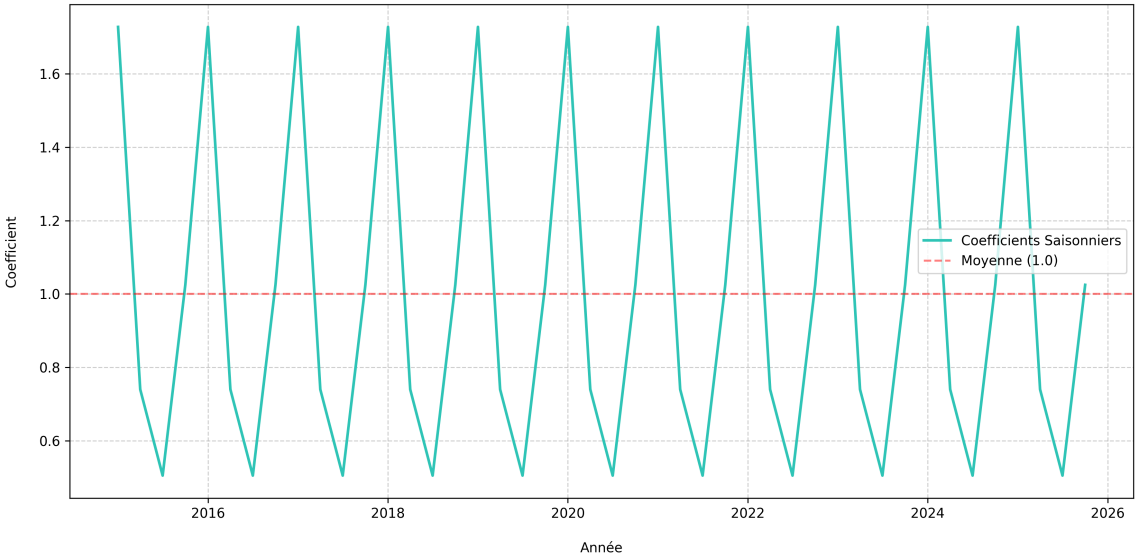

FIGURE 4 – Profil Saisonnier du Tourisme aux Maldives

Analyse Graphique

La visualisation de ces coefficients confirme la forte saisonnalité du tourisme. On observe un pic marqué au premier trimestre (haute saison hivernale européenne) et un creux significatif en milieu d'année (saison des pluies / été européen).

La présence d'une saisonnalité est justifiée par le fait que le même schéma se répète d'une année sur l'autre : à trimestre identique, les niveaux observés restent systématiquement plus élevés ou plus faibles que la moyenne annuelle. Autrement dit, la variation n'est pas aléatoire : elle suit un cycle régulier de période 4 (données trimestrielles), ce que confirment les coefficients saisonniers normalisés. En particulier, les valeurs $(1+\widehat{S}_i)$ sont durablement **supérieures à 1** en T1 (pic récurrent) et **inférieures à 1** en T2-T3 (creux récurrent), tandis que T4 se situe proche de 1 (reprise). Le profil saisonnier obtenu étant stable et répétitif, on conclut donc à l'existence d'une saisonnalité marquée sur la période étudiée.

1.5.1 Validation Statistique et Analyse Graphique

Vérification de la Composante Résiduelle

Contrairement aux coefficients saisonniers qui se vérifient sur une période ($p = 4$), la validité de la composante résiduelle s'analyse sur l'ensemble de la série temporelle (n observations). Dans notre modèle, le résidu multiplicatif est exprimé sous la forme $(1 + \epsilon_t)$. Pour valider le modèle, la moyenne de ce facteur doit être égale à 1, ce qui implique que la somme des écarts résiduels C_R doit être nulle. Dans notre cas, le calcul sur l'ensemble de la colonne donne un résultat exact :

$$\sum_{t=1}^{n} \epsilon_t = 0$$

L'absence totale d'écart confirme que la moyenne des résidus $(1+\epsilon_t)$ est strictement égale à 1,00. Cela valide mathématiquement que les composantes C_t et S_t ont été correctement et intégralement extraites par le modèle.

Obtention de la Série Corrigée CVS

L'objectif final de la désaisonnalisation est d'obtenir une série temporelle épurée de ses fluctuations saisonnières, ne laissant apparaître que la tendance de fond et les variations cycliques ou accidentelles. Dans un modèle multiplicatif, pour chaque observation à l'instant t correspondant au trimestre i, la valeur corrigée est donnée par :

$$CVS_t = \frac{X_t}{1 + S_i}$$

Analyse Graphique

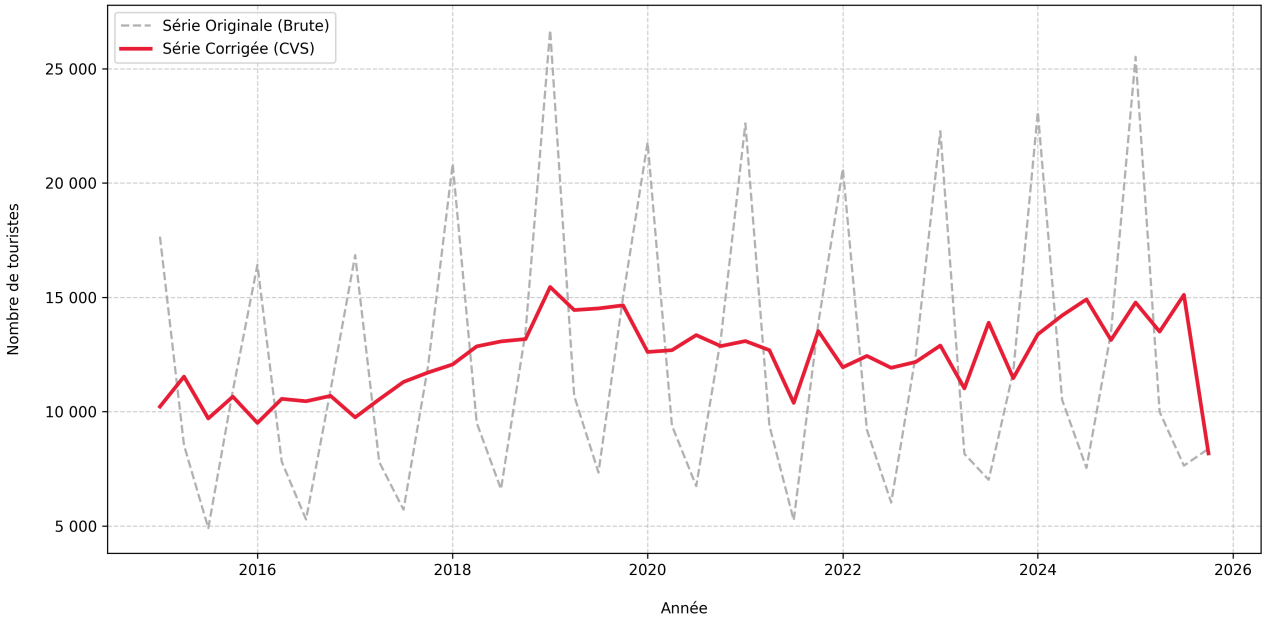

FIGURE 5 – Comparaison : Série Brute vs Série CVS

L'observation graphique permet de valider la qualité du traitement :

- La courbe rouge est nettement plus **lisse** que la courbe grise.
- Les pics récurrents du premier trimestre ont disparu, tout comme les creux systématiques du troisième trimestre.
- La courbe présente une tendance haussière modérée, non uniforme dans le temps.

1.5.2 Conclusion

Cette première étude nous a permis de décomposer la série chronologique X_t en identifiant ses différentes composantes déterministes. Après avoir validé graphiquement et statistiquement le choix d'un modèle **multiplicatif**, nous avons pu isoler la saisonnalité S_t grâce à la méthode du rapport à la moyenne mobile. Le calcul de la série CVS nous offre maintenant une vision plus claire de l'évolution réelle du tourisme, débarrassée des fluctuations périodiques. Bien que le traitement des données aberrantes liées au COVID-19 introduise une part d'approximation, la série corrigée semble suffisamment propre pour modéliser la tendance C_t. La prochaine étape consistera donc à estimer mathématiquement cette tendance par une **régression au sens des moindres carrés**, ce qui nous permettra de prolonger les courbes et de proposer une première prévision pour les années à venir.

1.6 Prévision par la Méthode des Moindres Carrés

Afin d'anticiper la fréquentation touristique pour l'année 2026, nous avons opté pour une modélisation analytique de la tendance. Cette approche consiste à ajuster une droite de régression sur la série désaisonnalisée puis à réintroduire la saisonnalité.

1.6.1 Modélisation et Estimation de la Tendance

Le Modèle Linéaire

Nous faisons l'hypothèse que la tendance de fond (C_t) évolue de manière linéaire en fonction du temps. Le modèle s'écrit donc :

$$CVS_t = a \cdot t + b + \epsilon_t$$

Où :
- t est le rang de la période (de $t = 1$ pour T1 2015 à $t = 44$ pour T4 2025) ;
- a est la pente ;
- b est l'ordonnée à l'origine.

Estimation et Calculs

Pour estimer les paramètres $\hat{a}$ et $\hat{b}$, nous avons appliqué la méthode des **Moindres Carrés** sur la série CVS. Cette méthode minimise la somme des carrés des écarts entre la droite et les points observés. Les calculs effectués sur nos 44 observations nous donnent l'équation de prévision suivante :

$$\hat{C}_t = \hat{a} \cdot t + \hat{b}$$

Cette droite représente la direction "moyenne" prise par le tourisme aux Maldives sur la décennie, abstraction faite des variations saisonnières et des chocs ponctuels (traités en amont). Pour obtenir la prévision finale ($\hat{X}_{2026}$), nous projetons cette tendance sur les rangs futurs ($t = 45$ à 48) et nous la remultiplions par les coefficients saisonniers correspondants :

$$\hat{X}_t = (\hat{a} \cdot t + \hat{b}) \times (1 + S_i)$$

Pour la suite, on retient donc un horizon de prévision d'une année, soit 4 trimestres.

1.6.2 Validation du Modèle et Prévisions

Validation de l'Hypothèse de Linéarité

Avant d'utiliser ce modèle pour prédire le futur, il est important de vérifier sa validité sur les données passées. Notre hypothèse de départ suppose que **la tendance est linéaire sur la période 2015-2025**. Pour ce faire, nous avons reconstitué une série théorique en appliquant notre formule aux rangs historiques ($t = 1$ à 44) et nous l'avons superposée aux données réelles.

Graphique à la page suivante.

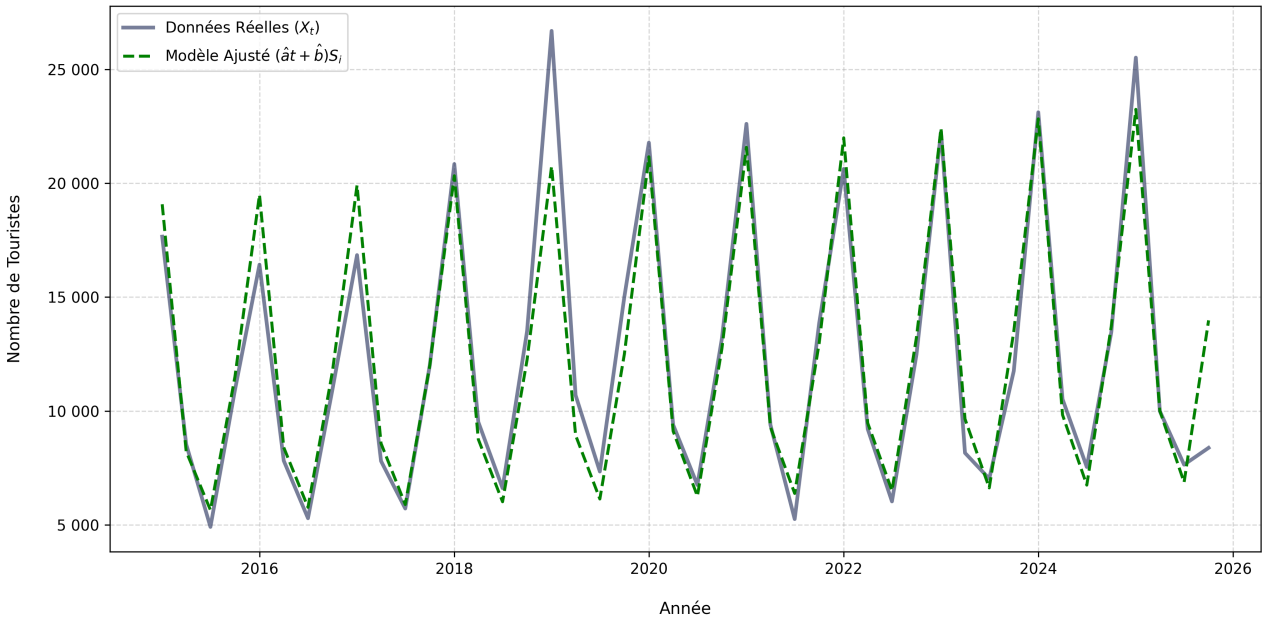

FIGURE 6 – Test d'adéquation : Série Réelle vs Modèle Ajusté

Interprétation : Comme le montre la Figure 6, la courbe verte épouse fidèlement la courbe bleue. Les écarts semblent visuellement limités et sans dérive majeur. Cela confirme que **l'hypothèse de linéarité est vérifiée** : la combinaison d'une droite de tendance et de nos coefficients saisonniers suffit à expliquer la dynamique passée du tourisme. Nous pouvons donc utiliser ce modèle pour la prévision avec un degré de confiance satisfaisant.

Prévision pour 2026

Fort de cette validation, nous appliquons le modèle aux rangs futurs correspondant à l'année 2026.

En appliquant notre modèle aux quatre trimestres de l'année 2026, nous obtenons les estimations suivantes. On remarque l'impact direct des coeff. saisonniers qui modulent fortement la prévision finale :

TABLE 3 – Prévisions des arrivées touristiques (Année 2026)

Année	**Trimestre**	**Rang** (t)	**Coeff.** ($1 + \widehat{S}_i$)	**Prévision** ($\hat{X}_t$)
2026	T1	45	1,73	23 662,09
2026	T2	46	0,74	10 183,62
2026	T3	47	0,51	6 987,08
2026	T4	48	1,02	14 225,94

Ces résultats chiffrés confirment la forte disparité attendue entre la haute saison (T1) et la basse saison (T3), avec un volume de touristes plus de trois fois supérieur au premier trimestre.

1.6.3 Interprétation et Conclusion

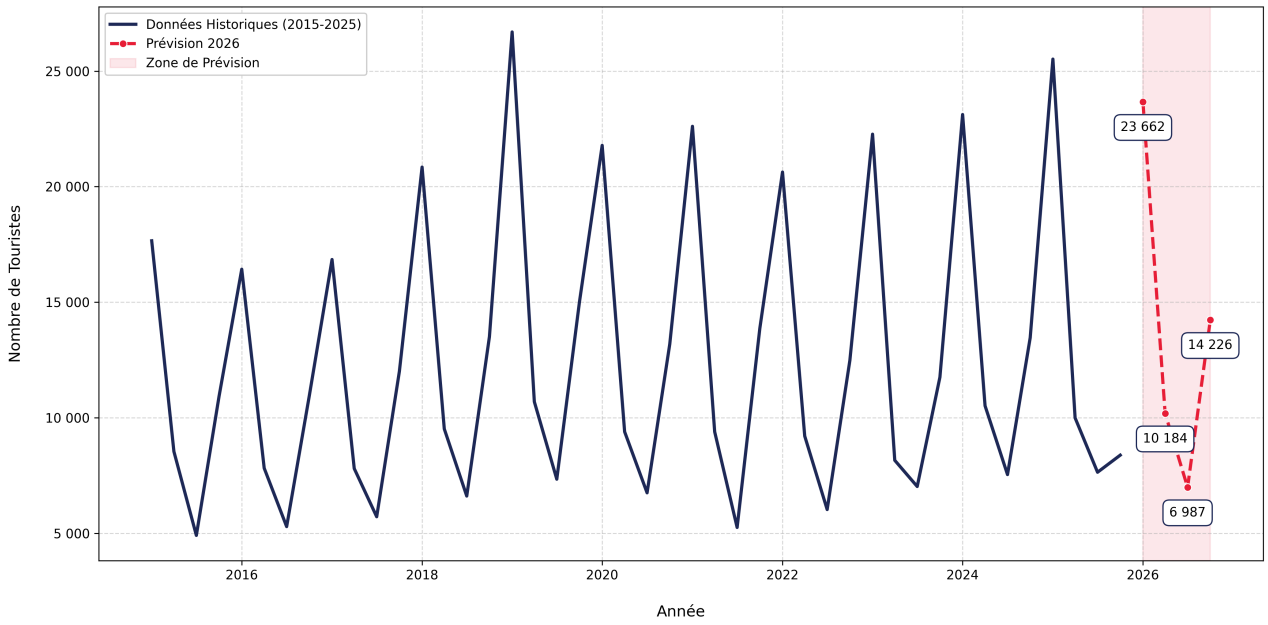

FIGURE 7 – Projection des arrivées touristiques pour 2026 (Modèle Ajusté)

Interprétation des Résultats

Graphiquement, nous observons une jonction cohérente entre la fin de la série brute (2025) et le début de la prévision (2026). Le modèle capture bien la dynamique de reprise :

- La tendance haussière est prolongée, reflétant l'attractivité croissante de la destination.
- Le profil saisonnier est parfaitement reproduit : on anticipe un pic marqué au premier trimestre 2026, suivi du creux habituel de milieu d'année.

Ces estimations fournissent un scénario réaliste, **sous l'hypothèse que la croissance structurelle observée ces dix dernières années se poursuive au même rythme**.

Conclusion et Ouverture

La méthode de décomposition (MMC) couplée à une régression linéaire nous a permis de construire une prévision robuste à court terme. Ce modèle présente l'avantage de la simplicité et de l'explicabilité : chaque composante (Tendance, Saison) est clairement identifiée. Cependant, ce modèle présente une limite non négligeable : il suppose que la tendance est constante sur toute la période. Si la dynamique du tourisme venait à s'accélérer ou à ralentir brutalement, la droite de régression s'ajusterait trop lentement. Pour pallier cette rigidité, il serait pertinent d'explorer la méthode du **Lissage Exponentiel**. Contrairement aux moindres carrés qui traitent tout le passé avec le même poids, le lissage exponentiel permet de donner de l'importance aux observations selon que l'on souhaite privilégiez les données récente ou lointaine, permettant au modèle d'être plus réactif aux changements de tendance de dernière minute.

1.7 Établissement de la série CVS via le LES

Dans une démarche alternative à la MMC, nous estimons ici la tendance à l'aide d'un LES. Cette approche présente l'avantage technique de conserver l'intégralité des observations, évitant la perte de données aux extrémités de la série, ce qui est important pour une analyse sur des données récentes.

1.7.1 Estimation de la tendance et choix du paramètre

Le modèle de lissage exponentiel simple repose sur la **formule récursive de mise à jour** des prévisions, telle que définie dans le support de cours. L'objectif est de calculer la nouvelle prévision $\hat{x}_{t+h}$ (à l'horizon h)en corrigeant la précédente via la nouvelle observation x_t.

La formule s'écrit :

$$\begin{aligned}\hat{x}_{t+h} &= \beta\hat{x}_{(t-1)+h} + (1-\beta)x_t \\ &= \hat{x}_{(t-1)+h} + (1-\beta)(x_t - \hat{x}_{(t-1)+h})\end{aligned}$$

Où :

- $\hat{x}_{t+h}$ est la nouvelle prévision calculée à l'instant t ;
- $\hat{x}_{(t-1)+h}$ est la prévision précédente ;
- x_t est la nouvelle observation réelle ;
- β est le facteur de mémoire (poids du passé).

Choix du paramètre β

Dans le cadre du **LES**, la mise à jour suit la relation :

$$\hat{x}_{t+1} = \beta\,\hat{x}_t + (1-\beta)\,x_t$$

où β représente le **poids du passé** et $(1-\beta)$ celui de la valeur courante. Afin de retenir une valeur de β cohérente avec les données, on compare plusieurs valeurs possibles et on mesure l'écart entre les observations x_t et les valeurs lissées $\hat{x}_t$. Dans ce travail, l'écart est résumé par une erreur quadratique moyenne (puis racine), ce qui revient à calculer une **RMSE** (ou EQM) basée sur les résidus $e_t = x_t - \hat{x}_t$:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

Le tableau ci-dessous présente la RMSE obtenue pour chaque β testé.

β	RMSE
0.9	6550.97
0.8	6658.53
0.7	6864.59
0.6	7107.92
0.5	7364.46
0.4	7617.30
0.3	7852.35
0.2	8057.91
0.1	8226.16
Meilleur β	0.9

TABLE 4 – Comparaison des RMSE selon β pour le LES.

On obtient ainsi un minimum pour $\beta = 0.9$. Ce résultat est aussi cohérent avec le choix réalisé *a priori* : une valeur élevée de β implique un lissage plus marqué (forte mémoire), ce qui aide à dégager une tendance générale sans réagir excessivement aux fluctuations trimestrielles.

NB : attention, une RMSE minimale $\not\Rightarrow$ un β optimal. Il s'agit d'un indicateur utile, mais ce n'est pas une condition suffisante : l'interprétation des courbes (stabilité du lissage, cohérence des prévisions et comportement des résidus) doit aussi être prise en compte.

À noté également, cette stabilité peut être un inconvénient : en accordant un poids important au passé, le **LES** s'adapte plus lentement lorsque la série évolue rapidement.

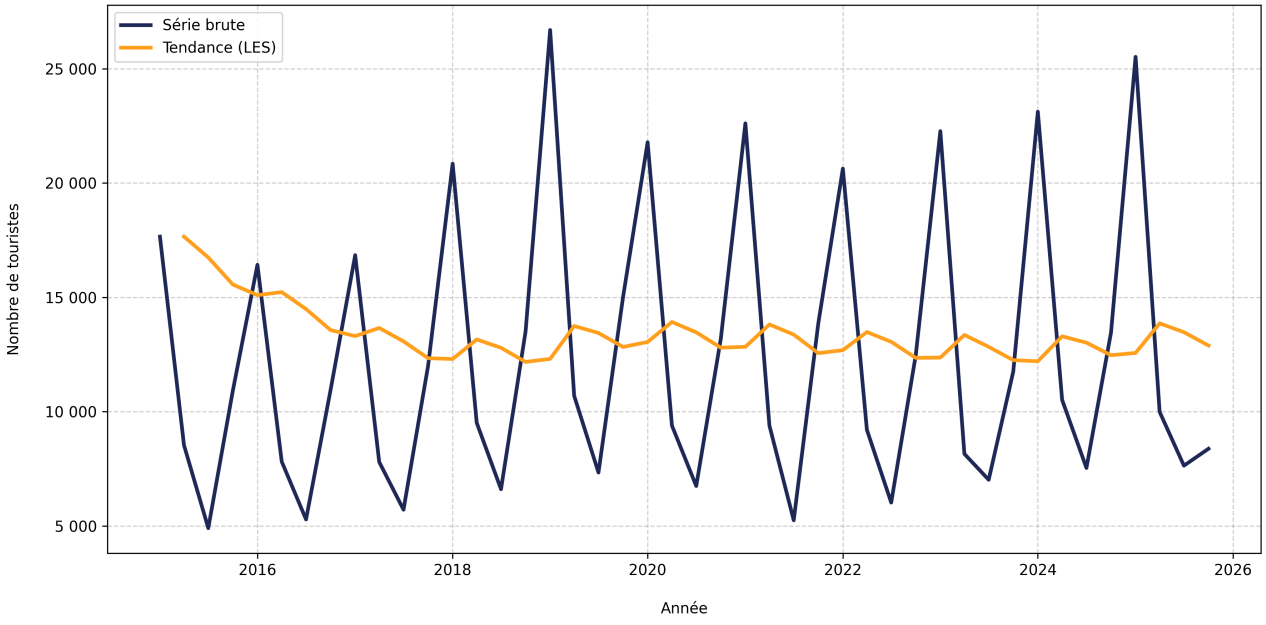

FIGURE 8 – Analyse du LES et choix du paramètre optimal ($\beta = 0,9$)

Ci-dessus, la tendance reste relativement lisse et ne suit pas immédiatement les variations récentes : elle amortit fortement les pics et les creux, ce qui est l'objectif du lissage, mais elle peut aussi retarder la prise en compte d'une hausse en fin de période.

Graphiquement, on observe que la courbe de tendance se situe souvent en dessous des pics de la série brute (notamment sur les trimestres de forte fréquentation), et qu'elle progresse plus lentement lorsque

les niveaux augmentent. Cela se traduit par une légère **sous-estimation** lors des phases de reprise, cohérente avec un modèle à forte mémoire.

1.7.2 Extraction de la saisonnalité et correction

Sur la base de cette tendance estimée, nous appliquons le principe du rapport à la tendance pour isoler les mouvements saisonniers. Nous calculons d'abord les rapports bruts $r_t = \frac{X_t}{L_t}$, puis nous en déduisons les coefficients saisonniers en faisant la moyenne de ces rapports pour chaque trimestre :

$$1 + S_i = \frac{1}{k} \sum r_t \quad \text{(pour le trimestre } i\text{)} \tag{1}$$

Ces coefficients sont ensuite normalisés afin que leur somme sur une période soit strictement égale à 4. En effet, en faisant le calcul, nous **n'obtenons pas zéro**. Ainsi pour ce faire, nous avons procédé **une normalisation des coefficients** afin que leur somme sur une période soit strictement égale à 4 (et donc $\widehat{S}_i = 0$).

Ci-dessous la représentation graphiques des coefficients saisonniers :

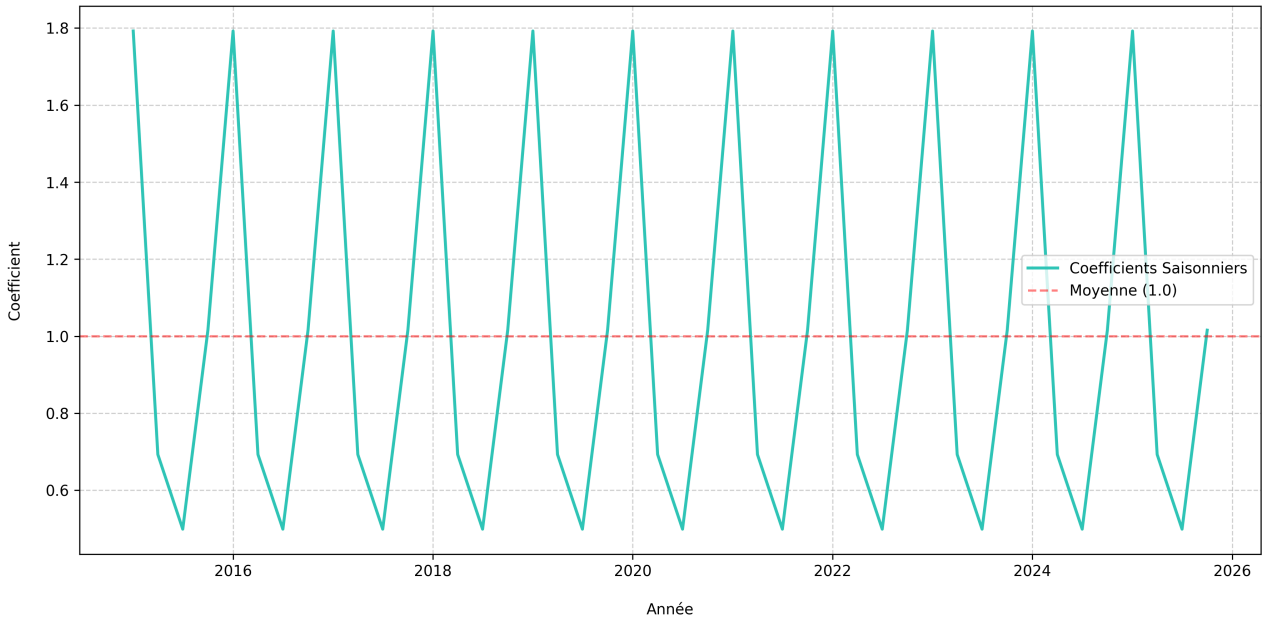

FIGURE 9 – Coefficients saisonniers issus du LES

L'examen des résultats met en évidence un profil saisonnier stable, et l'on retrouve globalement le même schéma que dans la méthode précédente : un pic récurrent en T1, un creux marqué en T2-T3, puis une reprise en T4. Cette répétition d'un cycle de période 4 permet ainsi de confirmer la présence d'une **saisonnalité** marquée sur la série.

L'application de ces coefficients permet d'obtenir la nouvelle Série CVS selon la formule classique du modèle multiplicatif :

$$CVS_t = \frac{X_t}{(1 + S_i)_{\text{corrigé}}} \tag{2}$$

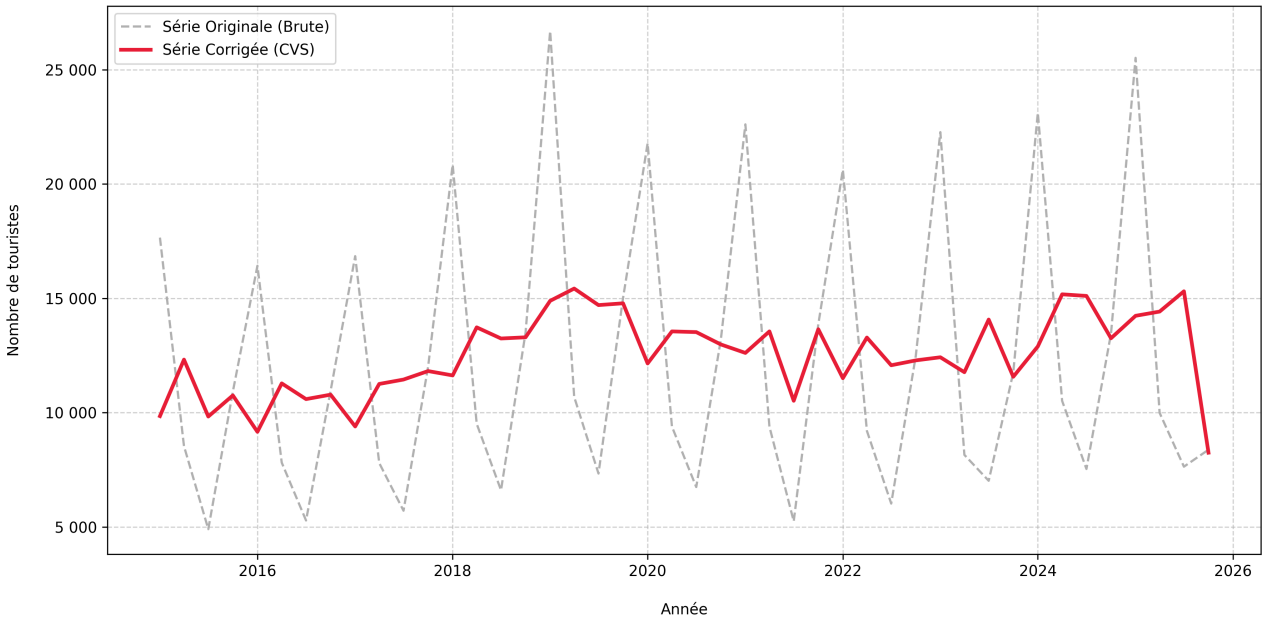

FIGURE 10 – Série CVS obtenue par Lissage Exponentiel Simple ($\beta = 0,9$)

L'analyse graphique confirme la disparition des cycles périodiques, ne laissant apparaître que la dynamique de fond et les fluctuations irrégulières.

1.8 Prévisions par lissage exponentiel double

Conformément aux consignes du projet, on complète l'approche par ajustement aux moindres carrés par une seconde méthode de prévision : le **lissage exponentiel double**.

Principe général

Le **LED** (lissage exponentiel double) permet de produire des prévisions en tenant compte d'une **tendance** à court terme. La méthode repose sur deux lissages exponentiels successifs, notés $LES1$ et $LES2$. Comme pour le LES, le paramètre $\beta \in]0,1[$ représente le **poids du passé** (mémoire).

Construction des deux lissages

On initialise à la première observation :

$$LES1_1 = x_1 \qquad \text{et} \qquad LES2_1 = x_1$$

Puis, pour tout $t \geq 2$, on applique les récurrences suivantes (même logique que le tableur) :

$$LES1_t = \beta\, LES1_{t-1} + (1-\beta)\, x_{t-1}$$

$$LES2_t = \beta\, LES2_{t-1} + (1-\beta)\, LES1_{t-1}$$

où x_t désigne la valeur observée au trimestre t.

Estimation de la pente et du niveau

À partir de $LES1$ et $LES2$, on estime les paramètres de tendance au temps t :

$$\hat{a}_t = \frac{1-\beta}{\beta}\left(LES1_{t-1} - LES2_t\right) \qquad \text{et} \qquad \hat{b}_t = 2\,LES1_{t-1} - LES2_t$$

$\hat{a}_t$ représente la **pente** et $\hat{b}_t$ l' **estimation de niveau**.

Choix du paramètre β pour le LED

Comme pour le LES, on compare plusieurs valeurs de β et on évalue les écarts entre les observations et les valeurs prévues à court terme. L'erreur est à nouveau résumée par une RMSE construite à partir des résidus.

Le tableau suivant présente la RMSE obtenue selon β.

β	RMSE
0.8	5920.756980
0.7	6086.325275
0.9	6117.444549
0.6	6737.311341
0.5	7871.302008
0.4	9395.704691
0.3	11183.049615
0.2	13099.278494
0.1	15019.197532
Meilleur β	0.8

TABLE 5 – Comparaison des RMSE selon β pour le LED.

Le minimum est obtenu pour $\beta = 0.8$. Là encore, ce choix est cohérent avec une sélection intuitive : une valeur de β relativement élevée assure un lissage suffisant pour extraire la tendance. Nous utiliserons donc cette valeur pour la suite de nos analyses. **Attention**, nous ne prenons pas seulement cette valeur pour sa RMSE minimal, elle permet également d'obtenir un bon lissage et des prédictions cohérentes.

Horizon de prévision

On note t le dernier trimestre observé (ici 2025–T4) et h l'**horizon de prévision**, c'est-à-dire le nombre de trimestres dans le futur. Prévoir l'année 2026 revient à considérer :

$$h \in \{1,2,3,4\}, \qquad \text{avec} \quad h=1 \quad \text{pour 2026–T1}, \quad h=2 \quad \text{pour 2026-T2, etc...}$$

Prévision LED

En notant $\hat{a}$ et $\hat{b}$ les dernières valeurs disponibles, la prévision LED à horizon h est :

$$\widehat{x}_{t+h} = \hat{b} + \hat{a}\,h$$

Cette quantité correspond à une prévision de la **dynamique de fond** (tendance) prolongée à court terme.

Réintroduction de la saisonnalité

Le modèle retenu étant **multiplicatif**, on réintroduit la saisonnalité à l'aide des coefficients trimestriels estimés précédemment, notés $(1 + \widehat{S})$. Pour le trimestre $i \in \{T1, T2, T3, T4\}$, la prévision finale est donc :

$$\widehat{X}_{t+h} = \widehat{x}_{t+h} \times (1 + \widehat{S}_i)_{\text{corrigé}}$$

Concrètement, chaque trimestre de 2026 est multiplié par son coefficient saisonnier correspondant.

Visualisation des prévisions

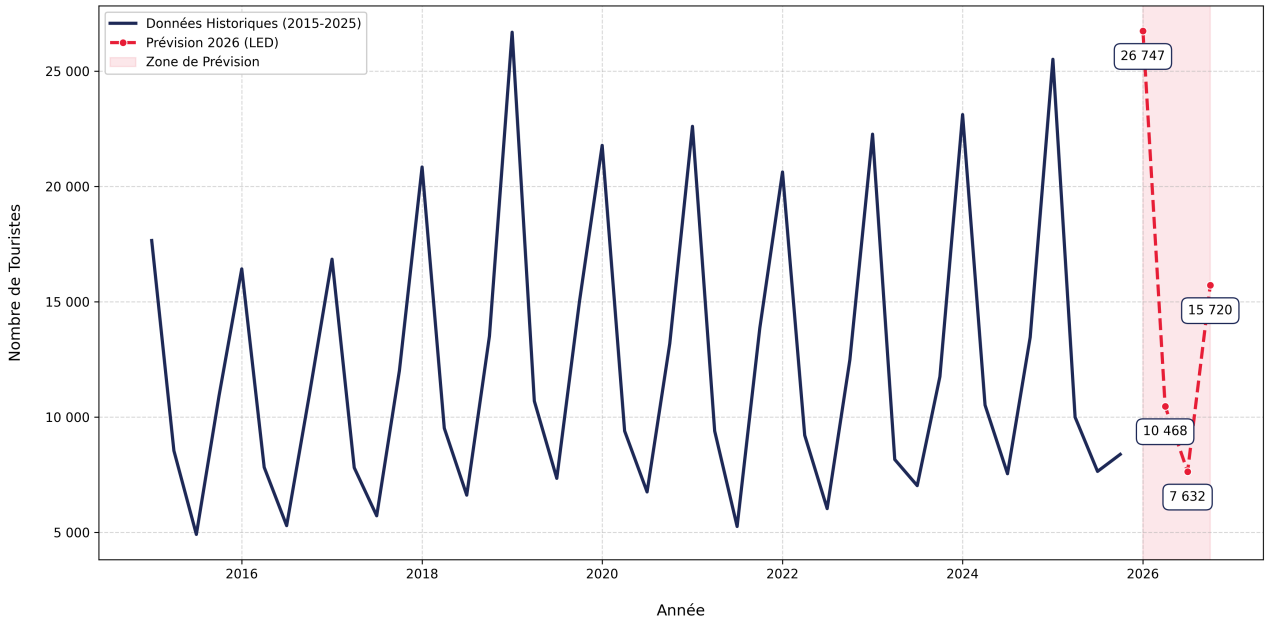

FIGURE 11 – Projection des arrivées touristiques pour 2026 (LED)

Lecture des résultats

La Figure 11 met en évidence une continuité globale entre 2025 et le début des prévisions (2026), ce qui suggère que la composante de tendance estimée en fin de période reste cohérente à court terme. On retrouve également un profil trimestriel très marqué, déjà observé sur la série brute. Cela est logique car la resaisonnalisation applique directement les coefficients $(1 + \widehat{S}_i)$ calculés précédemment.

On rappelle que le paramètre β joue un rôle important : plus β est élevé, plus le modèle donne de poids au passé et devient **inerte**, ce qui stabilise le lissage mais peut retarder l'adaptation en cas de reprise récente. À l'inverse, un β plus faible rend la méthode plus réactive aux dernières observations, mais au prix d'un lissage moins stable, susceptible de reproduire davantage les fluctuations de court terme.

On remarque enfin que les prévisions issues du LED sont ici légèrement plus élevées que celles obtenues par régression ; ce qui paraît cohérent car, avec $\beta = 0{,}8$, la tendance, ici, est davantage pilotée par le lissage de fin de période, alors que la régression résume l'ensemble de la période par une tendance moyenne.

1.9 Conclusion générale

Cette première partie a permis de construire une démarche complète d'analyse déterministe.

Les résultats obtenus par via les deux méthodes sont globalement compatibles : dans les deux cas, la prévision repose sur une tendance estimée. La régression présente l'avantage d'offrir une lecture très directe via une unique tendance linéaire, tandis que le LED est conçu pour accorder un poids différent au passé et peut donc, selon β, mieux refléter la situation de fin de période.

Sans affirmer une supériorité, on peut considérer que, pour un objectif de **prévision à court terme** sur 4 trimestres, le LED constitue une alternative intéressante lorsque l'on souhaite s'appuyer davantage sur l'information récente, tandis que la régression reste un choix solide lorsque l'on privilégie la simplicité et l'interprétation directe de la tendance. Cette comparaison motive la suite du rapport, où l'on cherchera à modéliser plus finement la dépendance temporelle par une approche stochastique (ARIMA).

2 Partie 2

Dans cette partie, nous présentons le processus de construction d'un modèle ARIMA (Auto Regressive Integrated Moving Average) appliqué à la série temporelle fournie dans le fichier `dataset.txt`, que vous, Monsieur LAVAL, nous avez imposé pour ce projet. On ne va pas s'en plaindre cette fois, surtout que dans la partie 1, nous avions eu le choix !

Ce fichier contient des observations mensuelles de valeurs numériques allant de juillet 1991 à juin 2008, soit près de 17 ans de données, avec une seule variable d'intérêt, *value*.

L'objectif de cette étude n'est pas seulement de reproduire les analyses vues sur le site Kaggle, mais de **démontrer une compréhension approfondie des méthodes de modélisation de séries temporelles et de justifier les choix effectués à chaque étape**. Pour cela, nous suivons une démarche méthodique qui se décompose en plusieurs étapes.

Nous commençons par vérifier la stationnarité de la série et appliquer les différenciations nécessaires pour la stabiliser. Ensuite, nous déterminons les ordres AR et MA en analysant les graphiques PACF et ACF de la série différenciée. Nous présentons également les ajustements possibles lorsque la série est légèrement sous- ou sur-différenciée, afin de corriger les tendances résiduelles ou les fluctuations trop importantes des résidus. Enfin, nous construisons et évaluons les modèles ARIMA pour comparer leurs performances et vérifier leur adéquation à la série.

Cette démarche illustre la **combinaison de méthodes statistiques et d'analyses visuelles** pour la modélisation de séries temporelles. Les graphiques de la série originale, des différenciations, ainsi que des fonctions ACF et PACF permettent de justifier chaque choix de paramètre, tandis que l'analyse des résidus assure que le modèle n'a pas laissé de structure non modélisée.

Ainsi, l'étude ne se limite pas à l'exécution du code elle **montre une maîtrise des concepts mathématiques et statistiques**, une compréhension du comportement de la série temporelle, et une capacité à interpréter correctement les résultats pour construire un **modèle ARIMA fiable et efficace** sur des données réelles. Elle met également en évidence l'importance de combiner o**bservation graphique, tests statistiques et ajustements de modèle** pour obtenir des prévisions robustes et justifiables.

2.1 Détermination de l'ordre de différenciation (d)

L'objectif principal de cette étape est de déterminer le nombre de différenciations nécessaires pour rendre la série stationnaire, condition essentielle pour qu'un modèle ARIMA fonctionne correctement.

Pour rappel, un processus stochastique est dit **stationnaire** si ses propriétés statistiques fondamentales ne varient pas avec le temps. Concrètement, cela signifie que son espérance et sa variance sont constantes, et que l'autocovariance ne dépend que du décalage temporel (lag) et non de l'instant t. Cette stabilité est une condition *sine qua non* pour que le modèle puisse identifier des motifs prédictifs fiables.

Pour tester la stationnarité, nous avons utilisé le test **Augmented Dickey-Fuller (ADF)**, disponible dans le package `statsmodels`. Le test repose sur les hypothèses suivantes :

- H_0 : la série est non stationnaire
- H_1 : la série est stationnaire

La statistique du test ADF se base sur le modèle de régression suivant :

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta Y_{t-i} + \varepsilon_t$$

où ΔY_t désigne la première différence de la série : $\Delta Y_t = Y_t - Y_{t-1}$

Dans cette formule, ΔY_t représente la première différence de la série, c'est-à-dire la variation entre une valeur et la valeur précédente. Le terme α correspond à la constante du modèle, tandis que βt capture une éventuelle tendance linéaire dans la série. Le coefficient γ mesure l'influence de la valeur précédente de la série sur la valeur actuelle, et les coefficients δ_i sont associés aux différences retardées qui permettent de corriger l'autocorrélation dans les résidus. Enfin, ϵ_t désigne le terme d'erreur aléatoire qui reste une fois que toutes les influences précédentes ont été prises en compte.

Le test renvoie une p-value qui permet de décider si la série est stationnaire. Si la p-value est inférieure au seuil de signification (5%), on rejette l'hypothèse nulle H_0 et on considère que la série est stationnaire.

Dans notre cas, la série originale `df.value` a donné une p-value de 1,00, indiquant qu'elle n'est pas stationnaire. Nous avons donc appliqué successivement :

Première différenciation

```
df.value.diff()
```

Cette ligne de code calcule la différence entre chaque valeur et la valeur précédente, éliminant les tendances linéaires de la série. Si la première différenciation ne suffit pas à rendre la série stationnaire, il est nécessaire de procéder à une deuxième différenciation.

Deuxième différenciation

```
df.value.diff().diff()
```

Mathématiquement, la différenciation d'ordre 2 s'écrit :

$$\Delta^2 Y_t = Y_t - 2Y_{t-1} + Y_{t-2}$$

Cette étape est nécessaire si la première différenciation ne suffit pas à rendre la série stationnaire. Cette opération élimine les tendances quadratiques ou les cycles persistants dans la série, rendant la série plus proche de la stationnarité.

Pour visualiser l'effet de la différenciation et décider du bon ordre d, nous avons tracé à la fois la série différenciée et sa fonction d'autocorrélation (ACF) à l'aide de :

```
from statsmodels.graphics.tsaplots import plot_acf
plot_acf(df.value.diff().dropna(), ax=axes[1, 1])
```

Le graphique de la série différenciée permet d'observer si la série oscille autour d'une moyenne constante, tandis que le graphique ACF montre comment les autocorrélations se comportent après différenciation. Si les autocorrélations chutent rapidement vers zéro, cela indique que la série est devenue stationnaire.

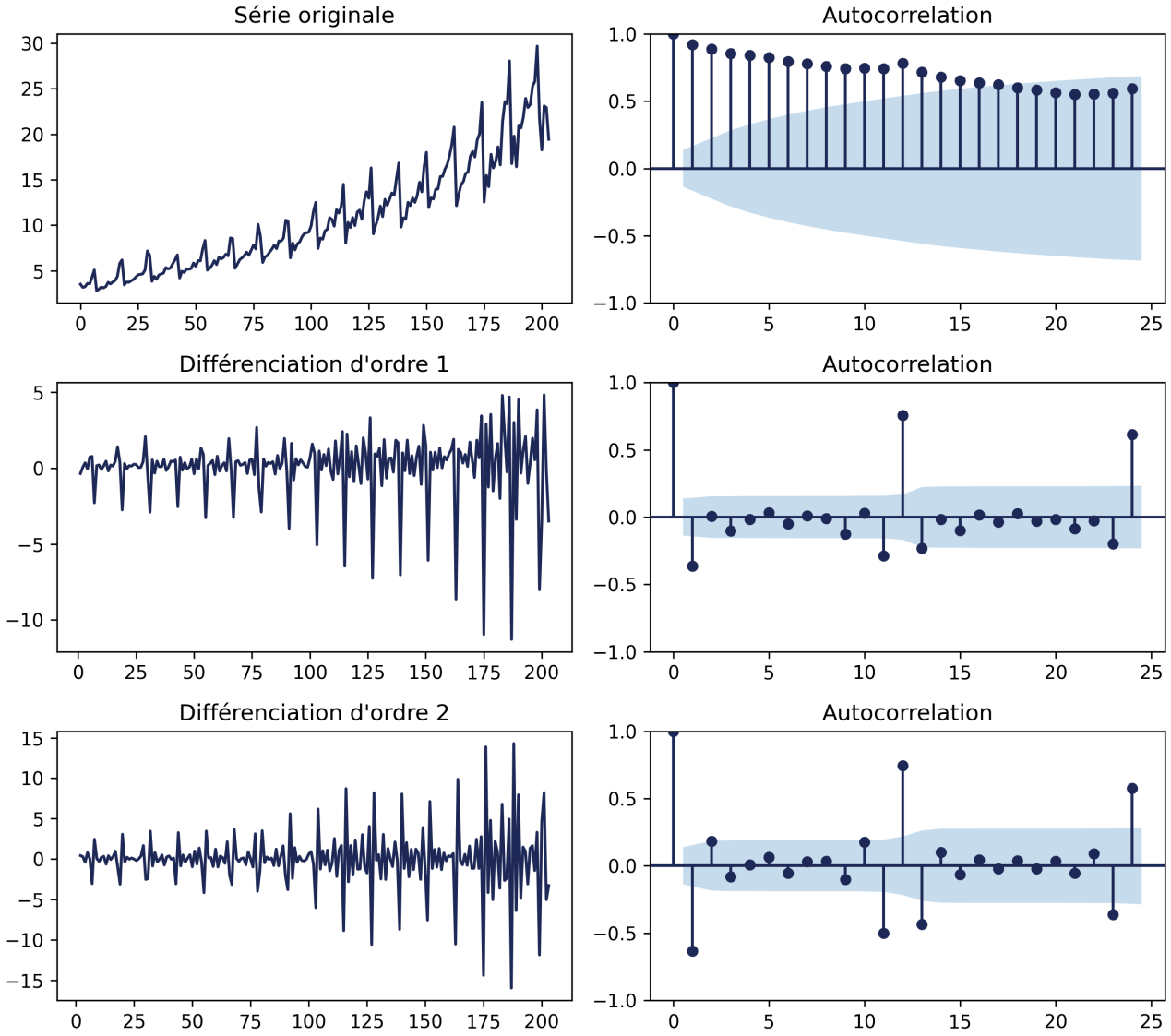

FIGURE 12 – Analyse de la stationnarité : Série originale vs Différenciations

Dans notre cas, l'ACF de la première différenciation devient rapidement proche de zéro, hormis quelques pics significatifs par exemple les lags 12 et 24 liés à une composante saisonnière persistante. Cela indique que la série est globalement stationnaire après une différenciation d'ordre 1. En revanche, la deuxième

différenciation n'apporte pas d'amélioration significative de la structure d'autocorrélation et accroît même la variabilité de la série. Ainsi, le choix d = 1 est le plus approprié pour stabiliser la série.

NB : une série qui a besoin d'être différencier pour atteindre la stationnarité est considérée comme une version **intégrée** d'une série stationnaire. **Ce qui est le cas ici !**

2.2 Détermination de l'ordre AR (p)

Après avoir déterminé l'ordre de différenciation d nécessaire pour rendre la série stationnaire, l'étape suivante dans la construction du modèle ARIMA consiste à identifier l'ordre autorégressif p.

Le paramètre p correspond au nombre de valeurs passées de la série qui influencent directement la valeur actuelle. En d'autres termes, il indique dans quelle mesure la valeur Y_t dépend des valeurs précédentes $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$. L'objectif est donc de déterminer jusqu'à quel retard les valeurs précédentes de la série contribuent de manière significative à l'explication de la dynamique temporelle.

Un modèle autorégressif d'ordre p, noté $AR(p)$, peut être représenté mathématiquement par l'équation suivante :

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t$$

Dans cette équation, Y_t représente la valeur actuelle de la série, μ est une constante, ϕ_i sont les coefficients autorégressifs qui mesurent l'influence du retard i sur la valeur actuelle, et ϵ_t est le terme d'erreur aléatoire autrement dit la composante aléatoire.

NB : un processus auto-régressifs n'est stable (convergent) que si les **coéfficients d'AR** (ϕ_i) sont définis dans l'intervalle $]-1,1[$.

L'objectif est de déterminer la valeur de p qui capture correctement la dépendance temporelle sans introduire de sur-ajustement.

Afin de déterminer la valeur appropriée de p, nous avons analysé la fonction d'autocorrélation partielle (PACF) de la série différenciée. Contrairement à la fonction d'autocorrélation simple (ACF), la PACF mesure la corrélation entre la série et un retard donné en éliminant l'effet des retards intermédiaires. Elle permet ainsi d'isoler l'influence pure de chaque retard sur la série et d'identifier précisément quels retards doivent être inclus dans la partie autorégressive du modèle.

Le principe est le suivant : si le coefficient partiel d'un certain retard k est significatif, cela signifie que la valeur Y_{t-k} contribue directement à expliquer Y_t et doit donc être incluse dans la partie AR du modèle. Les retards dont les coefficients partiels ne dépassent pas l'intervalle de confiance peuvent être ignorés, car leur influence est jugée négligeable. Dans le code, la PACF de la série différenciée est tracée avec :

```
from statsmodels.graphics.tsaplots import plot_pacf
plot_pacf(df.value.diff().dropna(), ax=axes[1,1])
```

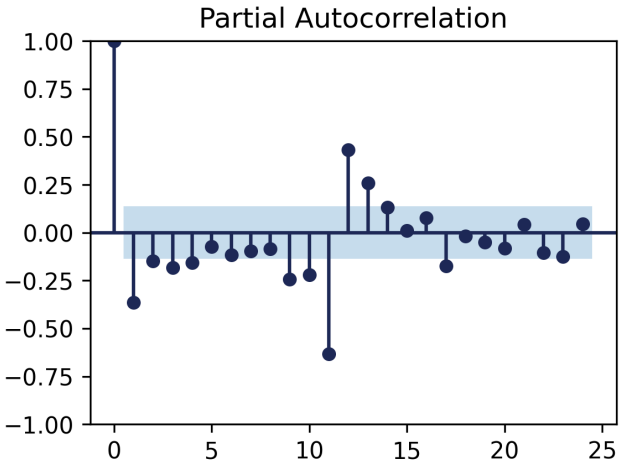

FIGURE 13 – Autocorrélation Partielle (PACF)

Le graphique obtenu permet d'observer la significativité statistique des différents retards. Chaque barre du graphique correspond au coefficient d'autocorrélation partielle d'un retard donné, tandis que les lignes horizontales représentent les bornes de l'intervalle de confiance. Lorsqu'une barre dépasse ces bornes, le retard associé est considéré comme statistiquement significatif.

L'analyse du graphique de la PACF, appliquée à la série préalablement différenciée afin d'assurer sa stationnarité, met clairement en évidence le rôle prépondérant du premier retard dans la dynamique temporelle de la série. En effet, le coefficient partiel associé au retard 1 se distingue nettement des autres retards par son amplitude et dépasse de manière franche les bornes de l'intervalle de confiance statistique.

Ce dépassement indique l'existence d'une corrélation partielle statistiquement significative entre la valeur actuelle de la série Y_t et sa valeur à l'instant précédent Y_{t-1}. Autrement dit, même après avoir éliminé les effets des retards intermédiaires et appliqué une différenciation pour rendre la série stationnaire, la valeur observée à l'instant $t-1$ conserve une influence directe et significative sur la valeur observée à l'instant t.

À l'inverse, l'examen des coefficients partiels correspondant aux retards d'ordre supérieur, à partir de $k > 12$, ne révèle pas de structure autorégressive autant marquée. Les valeurs associées à ces retards restent globalement proches de zéro et demeurent majoritairement à l'intérieur des bornes de l'intervalle de confiance. Bien que certains coefficients puissent ponctuellement s'en approcher, ces variations restent de faible amplitude, ne présentent pas de régularité particulière et alternent fréquemment de signe. Ce comportement est caractéristique de fluctuations aléatoires imputables au bruit statistique plutôt qu'à une dépendance temporelle significative.

En conséquence, une fois l'effet du premier retard pris en compte, les retards supplémentaires n'apportent pas d'information explicative pertinente sur l'évolution de la série. Ainsi, la PACF suggère que la dépendance temporelle de la série est essentiellement concentrée sur le premier retard. **L'ajout de termes autorégressifs supplémentaires conduirait à complexifier inutilement le modèle** sans amélioration notable de sa capacité à représenter la structure sous-jacente des données.

Conformément au principe de parcimonie, qui consiste à privilégier le modèle le plus simple capable de décrire correctement la dynamique observée, il apparaît donc pertinent de retenir un **modèle autorégressif d'ordre 1**. Nous fixons ainsi le paramètre autorégressif à :

$$p = 1$$

Ce choix permet de modéliser efficacement la structure temporelle de la série tout en limitant le nombre de paramètres à estimer, ce qui **réduit le risque de sur-ajustement** et **favorise la stabilité** ainsi que la robustesse des prévisions. À ce stade, l'ordre autorégressif du modèle ARIMA est donc déterminé de manière cohérente au regard de l'analyse statistique de la PACF.

2.3 Détermination de l'ordre MA (q)

Après avoir identifié l'ordre de différenciation d nécessaire pour rendre la série stationnaire ainsi que l'ordre autorégressif p, l'étape suivante consiste à déterminer l'ordre de la partie moyenne mobile, noté q, du modèle ARIMA.

Le paramètre q représente le nombre d'erreurs passées qui influencent directement la valeur courante de la série. Contrairement à la composante autorégressive, qui modélise l'influence des valeurs passées de la série elle-même, la composante moyenne mobile permet de capturer l'impact des chocs aléatoires antérieurs, c'est-à-dire des erreurs de prédiction commises aux instants précédents.

Un modèle de moyenne mobile d'ordre q, noté $MA(q)$, peut s'écrire sous la forme suivante :

$$Y_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Dans cette expression, Y_t désigne la valeur de la série à l'instant t, μ est une constante, ϵ_t correspond au terme d'erreur aléatoire à l'instant t (supposé de moyenne nulle et de variance constante), et θ_i sont les coefficients de moyenne mobile associés aux erreurs passées. Chaque coefficient θ_i mesure l'influence directe de l'erreur commise à l'instant $t - i$ sur la valeur observée à l'instant t.

Afin d'identifier la valeur appropriée de q, on s'appuie classiquement sur l'analyse l'ACF de la série différenciée. Contrairement à la fonction d'autocorrélation partielle, utilisée pour déterminer l'ordre autorégressif, l'ACF mesure la corrélation globale entre la série et ses retards successifs, en incluant à la fois les effets directs et indirects. Cette propriété rend l'ACF particulièrement adaptée à l'identification de la composante MA d'un modèle ARIMA.

Dans le code, la fonction d'autocorrélation est tracée à l'aide de la commande suivante :

```
from statsmodels.graphics.tsaplots import plot_acf
plot_acf(df.value.diff().dropna(), ax=axes[1, 0])
```

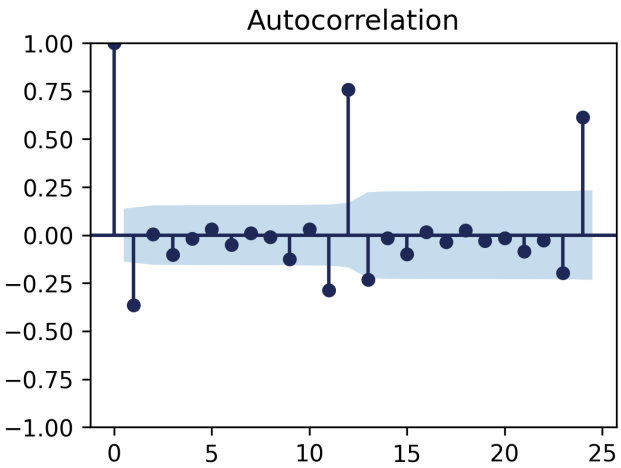

FIGURE 14 – Fonction d'Autocorrélation (ACF)

Le graphique obtenu présente, pour chaque retard, un coefficient d'autocorrélation ainsi que les bornes de l'intervalle de confiance statistique. Lorsqu'un coefficient dépasse ces bornes, le retard associé est considéré comme statistiquement significatif. L'analyse de ce graphique permet ainsi d'identifier jusqu'à quel ordre les erreurs passées influencent de manière significative la valeur actuelle de la série.

Dans notre cas, l'ACF met en évidence un coefficient significatif au retard 1, qui dépasse les bornes de l'intervalle de confiance. Cela indique que l'erreur commise à l'instant $t-1$ exerce une influence directe sur la valeur observée à l'instant t.

En revanche, pour les retards d'ordre supérieur, les coefficients d'autocorrélation décroissent rapidement mais ne restent pas tous à l'intérieur de l'intervalle de confiance ; on observe une absence de structure persistante ou de décroissance lente. Ce comportement est caractéristique d'un **processus de moyenne mobile de faible ordre**, pour lequel l'influence des erreurs passées est limitée aux premiers retards.

Ainsi, l'ACF suggère que la dynamique de la série peut être correctement capturée par un unique terme de moyenne mobile. **L'introduction de termes supplémentaires n'apporterait pas d'amélioration significative à la modélisation**, mais augmenterait inutilement la complexité du modèle.

Conformément au principe de parcimonie, il est donc pertinent de retenir :

$$q = 1$$

Ce choix permet de prendre en compte l'effet des chocs aléatoires récents tout en limitant le nombre de paramètres à estimer, ce qui contribue à améliorer la stabilité du modèle et la robustesse des prévisions. À ce stade, les trois paramètres fondamentaux du modèle ARIMA, à savoir p, d et q, sont désormais déterminés.

Finalement, en assemblant les différentes parties étudiées, on obtient la formule générale du modèle complet ARIMA(p, d, q) :

$$\widehat{Y}_t = \mu + \underbrace{\phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p}}_{\text{Composante AR}} + \underbrace{\theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}}_{\text{Composante MA}} + \epsilon_t$$

Détail des composantes :

- $\widehat{Y}_t$: La valeur stationnarisée de la série à l'instant t (après différenciation d'ordre d).
- μ : La constante du modèle.
- ϕ_i : Les coefficients d'auto-régression.
- θ_i : Les coefficients de moyenne mobile.
- ϵ_t : Le terme d'erreur aléatoire à l'instant t.

Par exemple pour le modèle **ARIMA(1,1,2)**, nous obtenons :

$$\widehat{Y}_t = \mu + \underbrace{\phi_1 \widehat{Y}_{t-1}}_{\text{AR(1)}} + \underbrace{\theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}}_{\text{MA(2)}} + \epsilon_t$$

La section suivante sera consacrée aux ajustements en cas de sous-différenciation ou sur-différenciation.

2.4 Ajustements en cas de sous-différenciation ou sur-différenciation

La détermination de l'ordre de différenciation d constitue une étape clé dans la construction d'un modèle ARIMA, car une mauvaise différenciation peut fortement dégrader les performances du modèle. En pratique, il est possible que la série soit insuffisamment différenciée, on parle alors de **sous-différenciation**, soit excessivement différenciée, ce que l'on appelle la **sur-différenciation**. Dans les deux cas, des ajustements sont nécessaires afin de garantir la validité des hypothèses du modèle et la qualité des prévisions.

La sous-différenciation

Elle se produit lorsque le **nombre de différenciations appliquées à la série est insuffisant pour la rendre stationnaire**. Dans ce cas, la série conserve une tendance ou une structure de dépendance de long terme, ce qui viole l'hypothèse fondamentale de stationnarité requise par les modèles ARIMA.

D'un point de vue graphique, une série sous-différenciée présente généralement une **moyenne non constante dans le temps** et des **autocorrélations qui décroissent lentement**. Sur les graphiques de l'ACF, ce phénomène se traduit par des corrélations significatives persistantes sur de nombreux retards, indiquant que l'information passée continue d'influencer fortement la série sur le long terme.

D'un point de vue statistique, la sous-différenciation peut également être détectée à l'aide du **test de Dickey-Fuller augmenté**. Une p-value élevée, supérieure au seuil de significativité de 5%, indique que l'hypothèse nulle de non-stationnarité ne peut pas être rejetée. Dans ce cas, il est nécessaire d'augmenter l'ordre de différenciation en appliquant une différenciation supplémentaire, afin d'éliminer les tendances résiduelles et de stabiliser la série. Mathématiquement, cela revient à augmenter l'ordre d du modèle ARIMA.

La sur-différenciation

À l'inverse, elle correspond à l'application d'un **nombre de différenciations trop élevé**. Bien que la série soit alors stationnaire, elle peut perdre une partie de sa structure temporelle utile. Une série sur-différenciée présente souvent une variance artificiellement accrue et un comportement marqué par de fortes fluctuations aléatoires, traduisant une perte de structure temporelle.

Sur les graphiques de l'ACF, la sur-différenciation se manifeste généralement par des **autocorrélations fortement négatives** au premier retard, suivies de valeurs proches de zéro pour les retards suivants. Ce type de structure est un indicateur clair que la différenciation appliquée est trop importante.

D'un point de vue pratique, la sur-différenciation entraîne des **modèles moins interprétables** et **des prévisions moins fiables**, car le signal utile de la série a été en partie supprimé. Dans ce cas, il convient de réduire l'ordre de différenciation, en revenant à une différenciation d'ordre inférieur. L'objectif est de trouver un compromis permettant de rendre la série stationnaire tout en conservant l'essentiel de l'information temporelle nécessaire à la modélisation.

Validation de notre choix

Dans notre étude, l'analyse conjointe des graphiques de la série différenciée, de l'ACF, de la PACF ainsi que les résultats du test ADF ont permis de vérifier que la différenciation d'ordre $d = 1$ permettait d'obtenir une série "stationnaire" sans introduire de comportements caractéristiques d'une sur-différenciation. Les autocorrélations chutent rapidement vers zéro et la série oscille autour d'une

moyenne stable, ce qui confirme la pertinence de ce choix. Aucun ajustement supplémentaire n'a donc été nécessaire à ce stade.

Dans notre cas, par exemple, si nous avions choisi $d = 2$ alors nous aurions eu une **sur-différenciation**.

Ainsi, la vérification des phénomènes de sous-différenciation et de sur-différenciation constitue une étape de validation essentielle dans la construction d'un modèle ARIMA. Elle permet de s'assurer que le paramètre d retenu est optimal et que le modèle repose sur une série stationnaire tout en conservant une structure temporelle exploitable pour la prévision.

2.5 Construction et évaluation du modèle ARIMA

Après avoir déterminé les paramètres p, d et q du modèle ARIMA, l'étape suivante consiste à construire concrètement le modèle et à évaluer sa capacité à représenter la dynamique de la série temporelle ainsi qu'à produire des prévisions pertinentes. Cette phase permet de valider empiriquement les choix effectués lors des étapes précédentes et de vérifier que le modèle est à la fois statistiquement cohérent et performant du point de vue prédictif.

La construction du modèle ARIMA repose sur l'estimation des paramètres autorégressifs et de moyenne mobile à partir des données observées. Dans notre cas, deux modèles ont été ajustés afin de comparer leurs performances : un modèle $ARIMA(1, 1, 1)$ et un modèle $ARIMA(1, 1, 2)$.

Le paramètre $d = 1$ correspond à l'ordre de différenciation retenu pour rendre la série stationnaire, tandis que les paramètres p et q ont été déterminés à partir de l'analyse de la PACF et de l'ACF. Ces deux modèles permettent d'évaluer l'impact de l'ajout d'un terme de moyenne mobile supplémentaire sur la qualité de l'ajustement et des prévisions.

L'ajustement du modèle est réalisé à l'aide de la classe `ARIMA` du module `statsmodels.tsa.arima.model`. Cette classe estime les paramètres du modèle par la méthode du maximum de vraisemblance, qui consiste à déterminer les valeurs des coefficients maximisant la probabilité d'observer les données compte tenu du modèle. Une fois le modèle ajusté, la méthode `fit()` fournit un objet contenant l'ensemble des résultats de l'estimation, notamment les coefficients estimés, leurs écarts-types, les statistiques de test ainsi que plusieurs critères d'évaluation globaux du modèle.

Suite à la page suivante.

SARIMAX Results

Dep. Variable:	value	No. Observations:	204
Model:	ARIMA(1, 1, 2)	Log Likelihood	-424.570
Date:	Sun, 25 Jan 2026	AIC	857.140
Time:	10:15:20	BIC	870.393
Sample:	0	HQIC	862.502
	- 204		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4178	0.356	1.174	0.240	-0.280	1.115
ma.L1	-0.9546	0.377	-2.531	0.011	-1.694	-0.215
ma.L2	0.0969	0.272	0.356	0.722	-0.437	0.631
sigma2	3.8259	0.269	14.209	0.000	3.298	4.354

Ljung-Box (L1) (Q):	0.46	Jarque-Bera (JB):	135.61
Prob(Q):	0.50	Prob(JB):	0.00
Heteroskedasticity (H):	9.82	Skew:	-0.80
Prob(H) (two-sided):	0.00	Kurtosis:	6.67

FIGURE 15 – Résultats du modèle ARIMA(1,1,2).

SARIMAX Results

Dep. Variable:	value	No. Observations:	204
Model:	ARIMA(1, 1, 1)	Log Likelihood	-424.762
Date:	Sun, 25 Jan 2026	AIC	855.524
Time:	10:15:20	BIC	865.463
Sample:	0	HQIC	859.545
	- 204		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3009	0.094	3.195	0.001	0.116	0.485
ma.L1	-0.8300	0.048	-17.204	0.000	-0.925	-0.735
sigma2	3.8327	0.259	14.790	0.000	3.325	4.341

Ljung-Box (L1) (Q):	0.72	Jarque-Bera (JB):	130.26
Prob(Q):	0.40	Prob(JB):	0.00
Heteroskedasticity (H):	9.98	Skew:	-0.75
Prob(H) (two-sided):	0.00	Kurtosis:	6.63

FIGURE 16 – Résultats du modèle ARIMA(1,1,1).

L'analyse du résumé du modèle permet de vérifier la significativité statistique des paramètres estimés. Les coefficients associés aux composantes autorégressives et de moyenne mobile doivent idéalement être significatifs, c'est-à-dire associés à des p-values inférieures au seuil de 5%. Cela indique que les termes inclus dans le modèle contribuent effectivement à expliquer la dynamique de la série.

Le résumé fournit également des critères d'information tels que l'AIC (Akaike Information Criterion) et le BIC (Bayesian Information Criterion), qui permettent de comparer plusieurs modèles entre eux. Des valeurs plus faibles de ces critères indiquent un meilleur compromis entre qualité d'ajustement et complexité du modèle.

Une fois le modèle ajusté, une étape essentielle consiste à analyser les résidus, c'est-à-dire les différences entre les valeurs observées et les valeurs estimées par le modèle. Les résidus doivent se comporter comme un bruit blanc, c'est-à-dire avoir une moyenne nulle, une variance constante et ne présenter aucune autocorrélation significative. Cette vérification est fondamentale, car la présence de structure dans les résidus indiquerait que le modèle n'a pas capturé l'ensemble de l'information contenue dans la série.

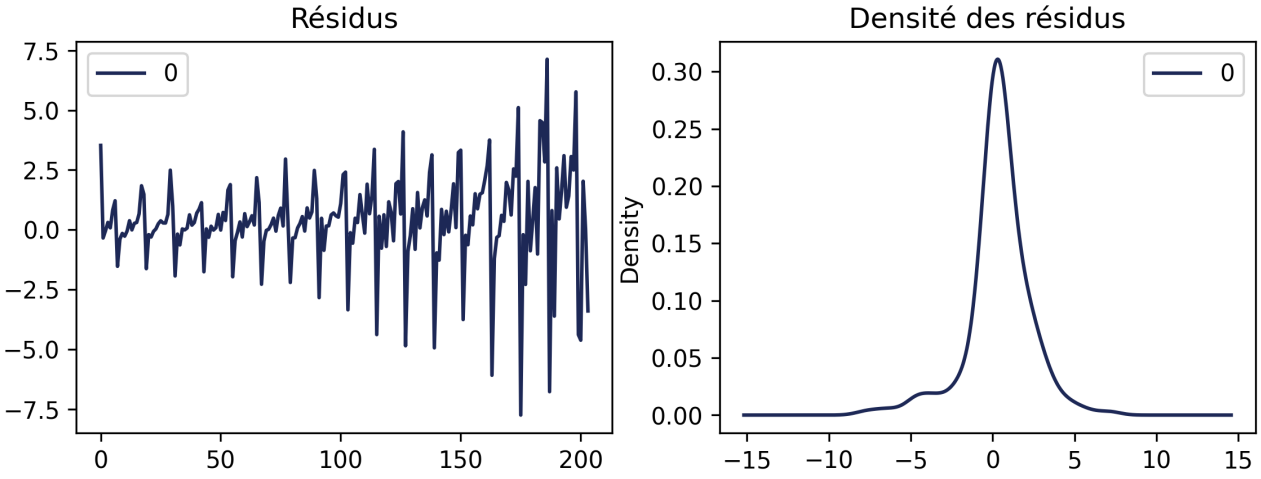

FIGURE 17 – Analyse des résidus : Évolution temporelle et Densité

L'examen graphique des résidus permet de vérifier visuellement l'absence de tendance et de structure temporelle résiduelle. Le graphique de densité permet quant à lui d'évaluer si la distribution des résidus est approximativement symétrique et centrée autour de zéro, ce qui est cohérent avec l'hypothèse d'erreurs aléatoires.

Enfin, l'évaluation du modèle passe par la comparaison entre les valeurs réelles de la série et les valeurs prédites par le modèle. Les prédictions sont calculées à partir des modèles ARIMA ajustés, puis représentées graphiquement sur une même figure afin de comparer directement les performances des modèles $ARIMA(1,1,1)$ et $ARIMA(1,1,2)$. Cette comparaison visuelle permet d'apprécier la capacité de chaque modèle à suivre la dynamique observée de la série, notamment lors des variations importantes.

Graphique à la page suivante.

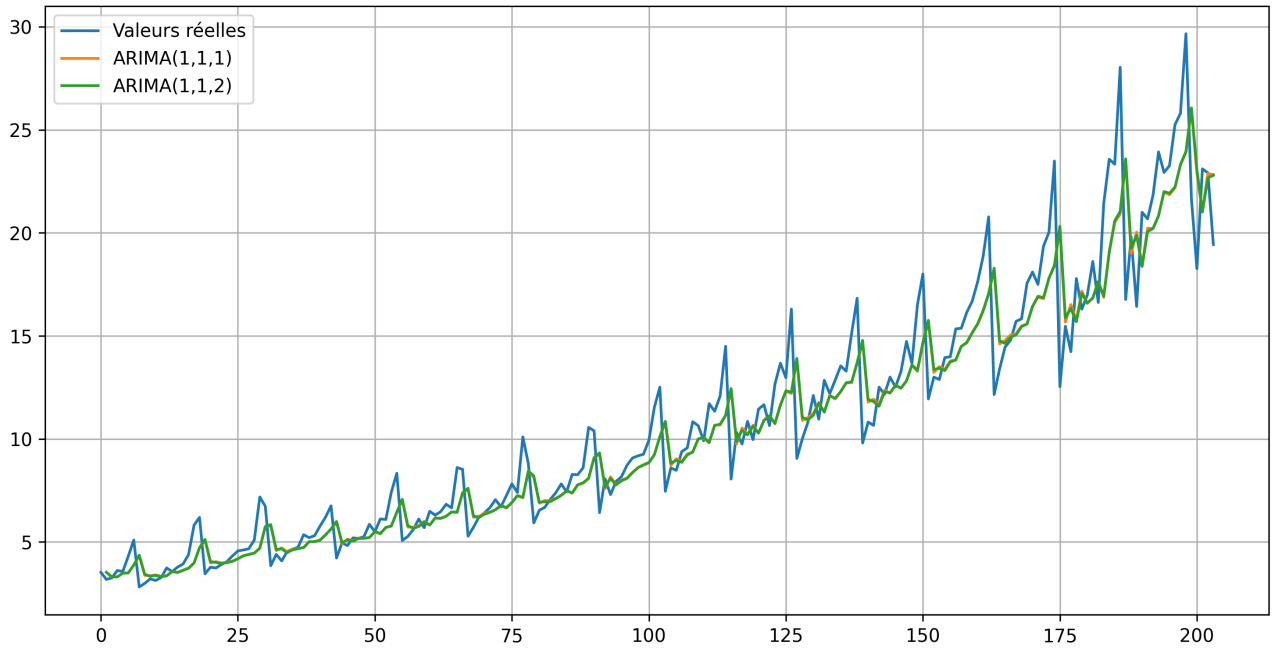

FIGURE 18 – Comparaison des prédictions ARIMA(1,1,1) vs ARIMA(1,1,2)

L'analyse de ce graphique permet de déterminer quel modèle offre le meilleur compromis entre précision et stabilité des prévisions. Un modèle performant est capable de reproduire la tendance générale de la série tout en limitant les écarts entre valeurs observées et valeurs prédites.

Si les deux modèles présentent des performances proches, le principe de parcimonie conduit à privilégier le modèle le plus simple, à savoir le modèle $ARIMA(1,1,1)$. En revanche, si l'ajout d'un terme de moyenne mobile supplémentaire améliore significativement la qualité des prédictions, le modèle $ARIMA(1,1,2)$ peut être retenu.

Ainsi, la construction et l'évaluation du modèle ARIMA permettent de valider les choix méthodologiques effectués tout au long de l'analyse et constituent une étape déterminante avant l'utilisation du modèle à des fins de prévision.

2.6 Conclusion de la modélisation ARIMA

Dans cette étude, nous avons suivi une démarche méthodique et stochastique pour construire un modèle ARIMA adapté à la série temporelle mensuelle fournie. Chaque étape, depuis la vérification de la stationnarité jusqu'à l'évaluation finale du modèle, a été guidée par une analyse statistique rigoureuse et une observation attentive des graphiques de la série, des fonctions ACF et PACF, ainsi que des résidus des modèles.

La série originale s'est révélée non stationnaire, nécessitant une différenciation d'ordre $d = 1$ pour stabiliser la moyenne et la variance. L'analyse de la PACF a indiqué que la dépendance temporelle est principalement concentrée sur le premier retard, justifiant le choix de $p = 1$ pour la composante autorégressive. De même, l'ACF a montré qu'un unique terme de moyenne mobile suffisait pour capturer l'influence des chocs passés, ce qui a conduit à retenir $q = 1$.

Deux modèles ont ensuite été ajustés, $ARIMA(1,1,1)$ et $ARIMA(1,1,2)$, afin d'évaluer l'impact de l'ajout d'un terme supplémentaire de moyenne mobile. L'analyse des résidus a confirmé que les deux

modèles produisent des erreurs proches d'un bruit blanc, sans autocorrélations résiduelles significatives. La comparaison graphique des valeurs réelles et prédites a montré que les deux modèles se superposent quasiment, reproduisant avec précision la dynamique observée de la série.

Ainsi, bien que l'ajout d'un terme MA supplémentaire dans le modèle $ARIMA(1,1,2)$ n'apporte qu'une amélioration marginale, les deux modèles restent performants. Conformément au principe de parcimonie, le modèle $ARIMA(1,1,1)$ peut être considéré comme suffisant pour représenter la série et effectuer des prévisions fiables, tout en garantissant une bonne stabilité et une interprétation simple des paramètres.

En conclusion, cette modélisation illustre l'importance de combiner analyses graphiques, tests statistiques et ajustements méthodiques pour construire un modèle ARIMA robuste, capable de capturer la structure temporelle des données tout en restant parcimonieux et interprétable.

Nos feuilles de calculs

Nous avons trouvé pertinent de vous mettre ci-dessous, un lien vous amenant à notre Google Sheets partagé. Vous y trouverez tous les calculs et graphiques de la Partie 1.

— **Google Sheets** : Cliquez sur moi !

Remerciements

Nous tenions sincèrement à vous remercier pour tous vos cours ainsi que pour ce projet, qui a été pour nous vraiment très intéressant. On y est resté certes très longtemps, parfois même plus que ce qu'on pensait au départ, mais au final c'était très enrichissant et on a beaucoup appris en le faisant. Merci pour votre suivi et pour tout ce que ce travail nous a apporté.

Petits "mots" de fin

Nous nous excusons par avance pour la longueur de ce rapport. Il semblerait que notre série ait pris un peu trop de liberté... et qu'elle ait décidé de s'étirer sur plusieurs pages !

Nous espérons néanmoins que ce nombre de pages restera une anomalie ponctuelle dans la saisonnalité des rapports que vous aurez à corriger, et surtout qu'il ne s'agira pas d'une tendance de long terme.