

Projet Séries Temporelles Analyses

Sheïma Mebarka, Ahmed Abbadi et Yuxuan Zeng

Année 2023-2024

Sommaire

Présentation du jeu de données	2
Sujet de l'étude	2
Description des variables	3
Références des différents jeux de données utilisés et bibliographie.....	3
Packages.....	3
I) Prétraitement des données	4
1) Préliminaires.....	4
2) Gestion de la non-stationnarité.....	4
II) Identification de plusieurs modèles et estimation des paramètres	8
1) Modèle MA.....	8
2) Modèle AR.....	10
3) Modèle ARMA.....	13
4) Choix du modèle.....	16
5) Résidus	17
III) Prédiction des valeurs futures	20
1) Intervalles de prédictions pour les 26 données les plus récentes.....	20
2) Matrice de corrélation et variables explicatives.....	21
3) Création d'un nouveau modèle.....	25
4) Prédictions	26
IV) Conclusion	27

Présentation du jeu de données

Sujet de l'étude

Notre projet consiste en l'analyse d'une série temporelle réalisé en Python. Celle-ci provient d'une base de données publique présentant le coût des actions mondiales de plusieurs entreprises influentes. Cette base de données offre un historique complet des performances boursières des marques les plus renommées au monde, avec des mises à jour quotidiennes. Les données couvrent la période du 1er janvier 2000 à aujourd'hui, fournissant ainsi une chronologie complète des informations boursières de diverses marques mondiales.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Date	Open	High	Low	Close	Volume	Dividends	Stock Splits	Brand_Na	Ticker	Industry_T	Country
2	2023-09-20	4.84000015	4.90999985	4.63000011	4.67000008	7441900	0	0	peloton	PTON	fitness	usa
3	2023-09-20	397.049988	397.98999	386.119995	386.299988	3866600	0	0	netflix	NFLX	entertainment	usa
4	2023-09-20	564.349976	569.219971	562.659973	563.830017	1311500	0	0	costco	COST	retail	usa
5	2023-09-20	138.550003	139.369995	135.199997	135.289993	46263700	0	0	amazon	AMZN	e-commerce	usa

Afin de mener à bien notre étude, nous avons modifié la base de données pour conserver comme variable d'intérêt le coût final de l'action, c'est-à-dire le prix de revient unitaire (PRU), qui permet de connaître le coût réel des actions. Chaque enregistrement comprend la date et l'heure exactes de l'observation sur une fréquence d'une semaine, éléments essentiels pour une analyse de série temporelle. Nous avons choisi de nous concentrer spécifiquement sur Apple, une marque emblématique dans le secteur de la technologie. Cette analyse peut fournir des connaissances précieuses sur les tendances du marché et le comportement des investisseurs vis-à-vis de l'une des entreprises les plus influentes au monde. Pour créer notre tableau, nous avons concaténé les deux colonnes correspondant respectivement à la date et au prix final de l'action. En raison de la grande quantité de données disponibles et de leur caractère non pertinent pour la prévision des coûts futurs, nous avons choisi de nous concentrer exclusivement sur les données de 2018 à aujourd'hui, totalisant ainsi 260 valeurs. Nous obtenons ainsi la série temporelle X_t suivante :

	Close
	Date
2018-09-16	53.321823
2018-09-23	52.300876
2018-09-30	53.350076
2018-10-07	54.633438
2018-10-14	52.845827

Dans le but de disposer de variables explicatives, susceptibles de nous éclairer sur les prédictions de nos valeurs futures, nous avons conservé les données concernant les prix de clôture de certaines enseignes particulièrement influentes sur le marché, telles que Microsoft, Google et Cisco. Nous avons aussi pris en compte d'autres variables explicatives que nous citerons plus bas.

Nous avons séparé cette base de données en 2 jeux de données. Une base de training de 260 observations et une base de test de 26 observations.

Références des différents jeux de données utilisés et bibliographie :

<https://www.kaggle.com/datasets/nelgiriyewithana/world-stock-prices-daily-updating/datan>

<https://moncoachdata.com/blog/modele-arima-avec-python/>

<https://www.aquiladata.fr/insights/mieux-comprendre-les-methodes-de-prevision-des-series-chronologiques/>

Gouriéroux, C., Monfort, A. (1995) *Séries temporelles et modèles dynamiques*. Economica.

Ladiray, D., Quenneville, B. (2000) *Désaisonnaliser avec la méthode X – 11*

Polycopié de cours de séries temporelles « Times Series Analysis » Olivier Wintenberger, Frédéric Guilloux

Description des variables

Une variable d'intérêt et un indicateur de temps

Date : Date correspondant à la journée au cours de laquelle le coût de l'action évolue
(Format : aaaa-mm-jj)

Close : Le prix réel/de clôture de l'action

6 variables explicatives au total

High : Le prix le plus élevé atteint au cours de la journée à cette date

Low : Le prix le plus bas atteint au cours de la journée à cette date

Volume : Le volume des transactions, c'est-à-dire le nombre d'actions négociées à cette date

Microsoft_close : Les prix de clôture du groupe Microsoft

Google_close : Les prix de clôture du groupe Google

Cisco_close : Les prix de clôture du groupe Cisco

Bibliothèques python utilisées :

```
Pandas
matplotlib
statsmodels
numpy
scipy
seaborn
```

I) Prétraitement des données

1) Préliminaires

Avant de débuter notre projet, des ajustements ont été apportés à notre jeu de données afin de le rendre propice à une analyse temporelle pertinente. Nous avons initialement retenu uniquement deux variables, à savoir la date (*Date*) et la clôture (*Close*), correspondant respectivement à t et X_t . Suite à cette sélection, nous avons converti la date au format Date de Python. Ensuite, la série a été fractionnée en deux segments distincts : l'ensemble d'entraînement et l'ensemble de test. Plutôt que d'opter pour les 10 dernières valeurs, nous avons préféré attribuer 10% de notre jeu de données à la partie test en raison du volume conséquent de données, soit 26 valeurs au total, visant ainsi une précision accrue dans nos prédictions.

Par la suite, une vérification des doublons a été réalisée pour ne conserver que la première occurrence, et les valeurs manquantes ont été traitées. Ces étapes préliminaires ont ouvert la voie au lancement effectif du projet.

2) Gestion de la non-stationnarité

La série temporelle initiale, non différenciée, est représentée sous la notation D_t et affichée graphiquement pour une observation visuelle.

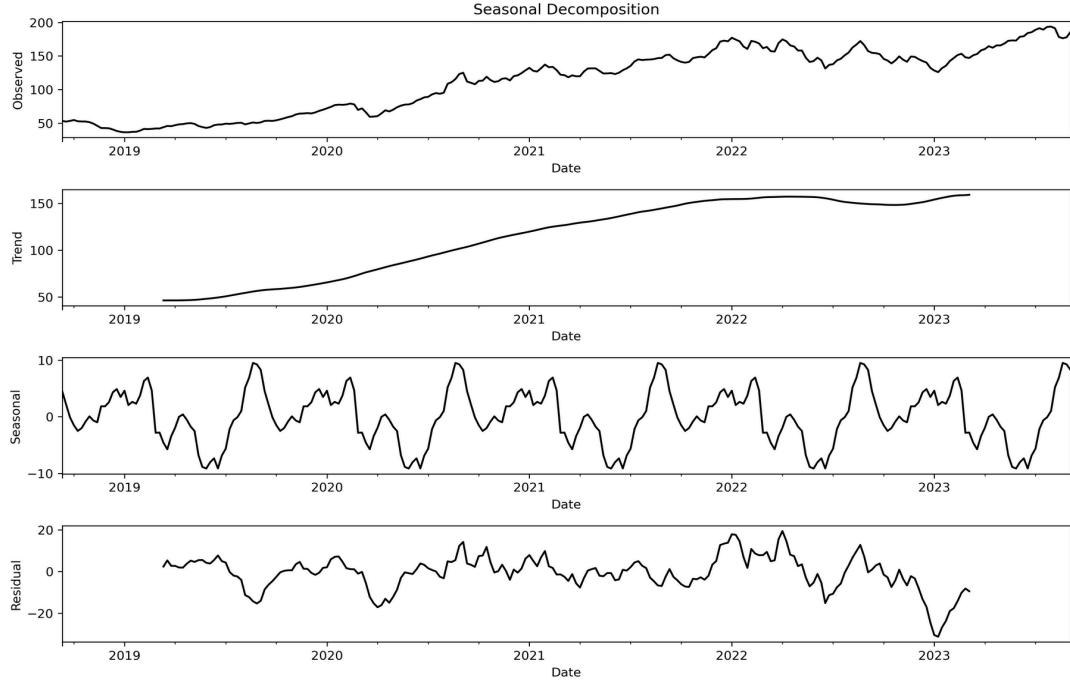


Notre objectif est d'obtenir une transformation, notée X_t , des données D_t de manière à rendre X_t stationnaire. La décomposition générale des séries chronologiques se présente sous la forme suivante :

$$D_t = f(t) + S_t + X_t, t \geq 1$$

Ici, $f(t)$ représente la composante tendance, une fonction déterministe du temps t , S_t est la composante saisonnière avec une période $S_t + T = S_t$ pour une certaine période T , et X_t constitue notre série temporelle, potentiellement stationnaire. Cette stationnarité sera évaluée ultérieurement.

Par la suite, une décomposition saisonnière est réalisée pour examiner l'existence d'une tendance et/ou d'une saisonnalité, ainsi que la possible présence de résidus.



Pour confirmer ces observations, nous appliquons le test de racine unitaire (ADF). Cette méthode est fréquemment utilisée pour déterminer si les données de séries chronologiques possèdent une racine unitaire, c'est-à-dire si elles sont stationnaires ou non. Les résultats obtenus sont les suivants :

```

Test Statistic           -0.626705
p_value                 0.864833
#Lags Used              1.000000
Number of Observations Used 258.000000
Critical Value (1%)      -3.455953
Critical Value (5%)       -2.872809
Critical Value (10%)      -2.572775
dtype: float64

```

====> The time series is not stationary.

L'observation de la p-value révèle qu'elle est supérieure à 0,05, indiquant ainsi que la série temporelle initiale n'est pas stationnaire. Afin de remédier à cela, une étape de différenciation est nécessaire pour rendre notre série stationnaire.

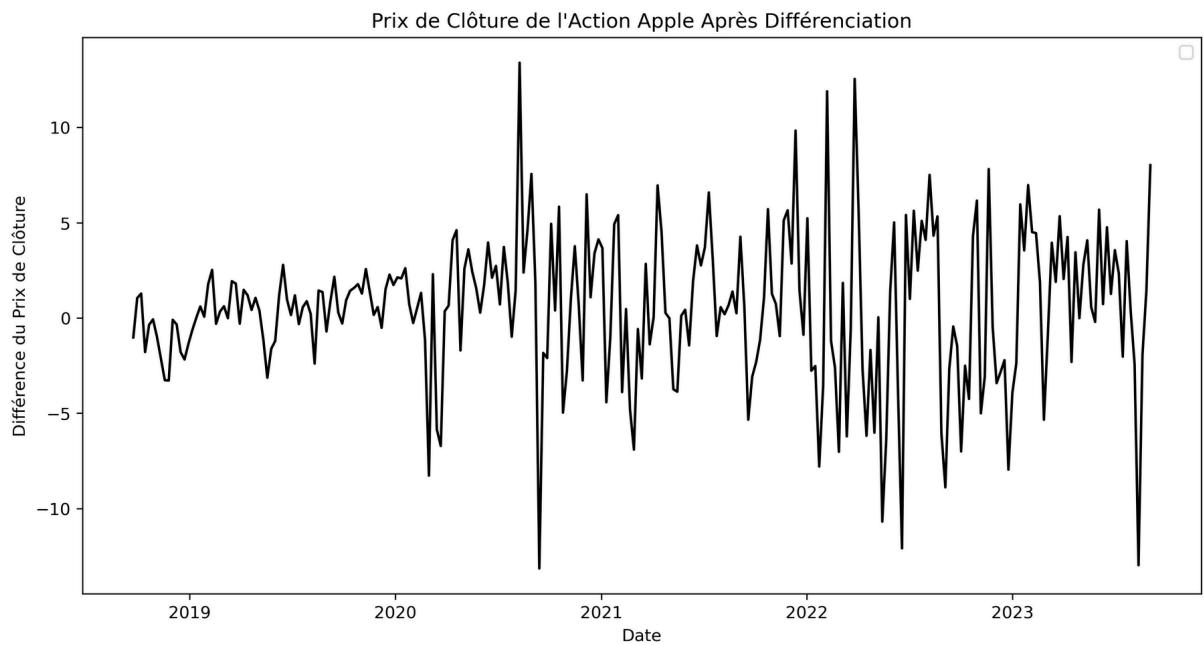
D'après le cours, on sait alors que la suite de terme $\Delta D_t = D_t - D_{t-1} = b + X_t - X_{t-1} = b + \Delta X_t$ est stationnaire si X_t l'est.

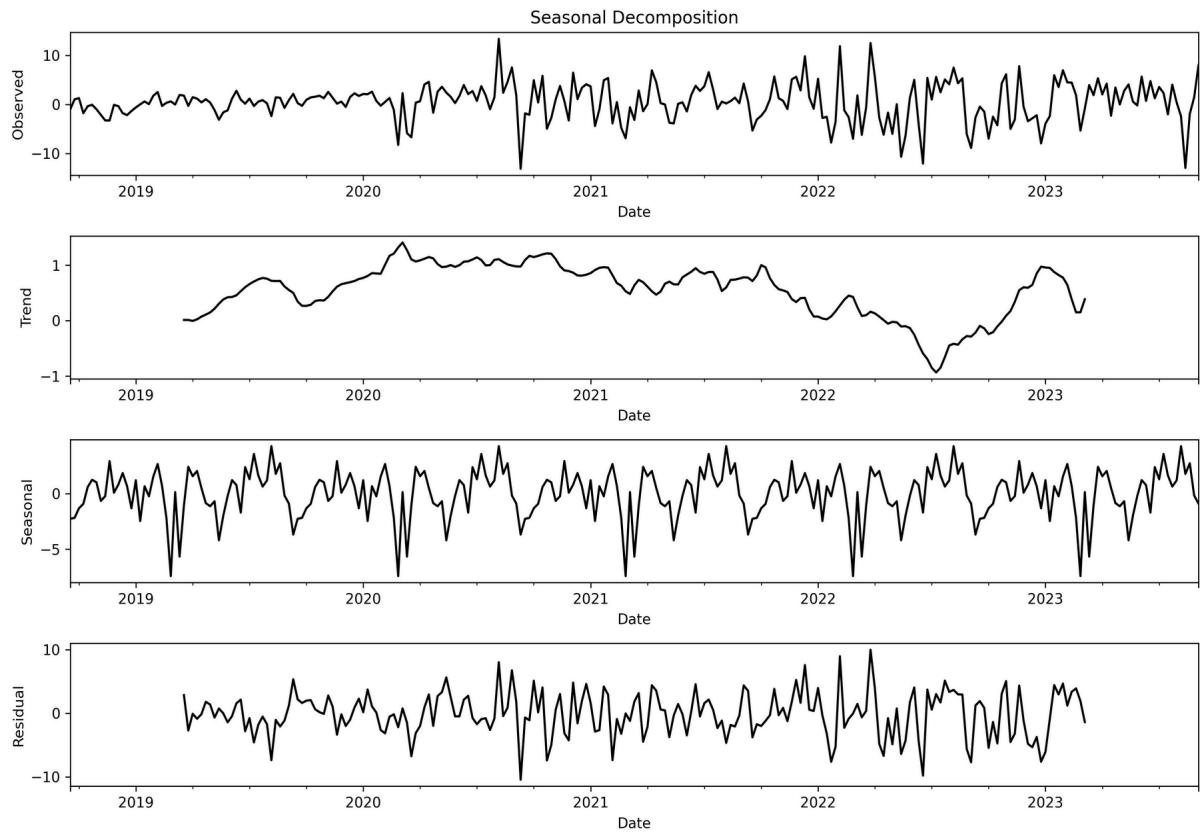
En considérant toutes ces remarques, on réalise une différenciation simple en prenant un lag de 1 et on obtient une série temporelle stationnaire d'expression :

$$X_t = \frac{D_t - D_{t-1}}{\sigma(D_t)}$$

avec $\sigma(D_t)$ l'écart-type de D_t utilisé ici pour améliorer la visualisation des résultats.

Ce qui suit présente la série temporelle différenciée ainsi que la décomposition saisonnière des données suite à cette différenciation :





Il est évident que la différenciation a nettement éliminé la tendance dans les données.
 Par la suite, un nouveau test ADF est effectué pour garantir la stabilité des données post-traitement différenciation.

```

Test Statistic      -1.214597e+01
p_value            1.610493e-22
#Lags Used        0.000000e+00
Number of Observations Used 2.580000e+02
Critical Value (1%) -3.455953e+00
Critical Value (5%) -2.872809e+00
Critical Value (10%) -2.572775e+00
dtype: float64

```

====> The time series is stationary.

La série étant stationnaire, nous pouvons alors commencer les ajustements des modèles.

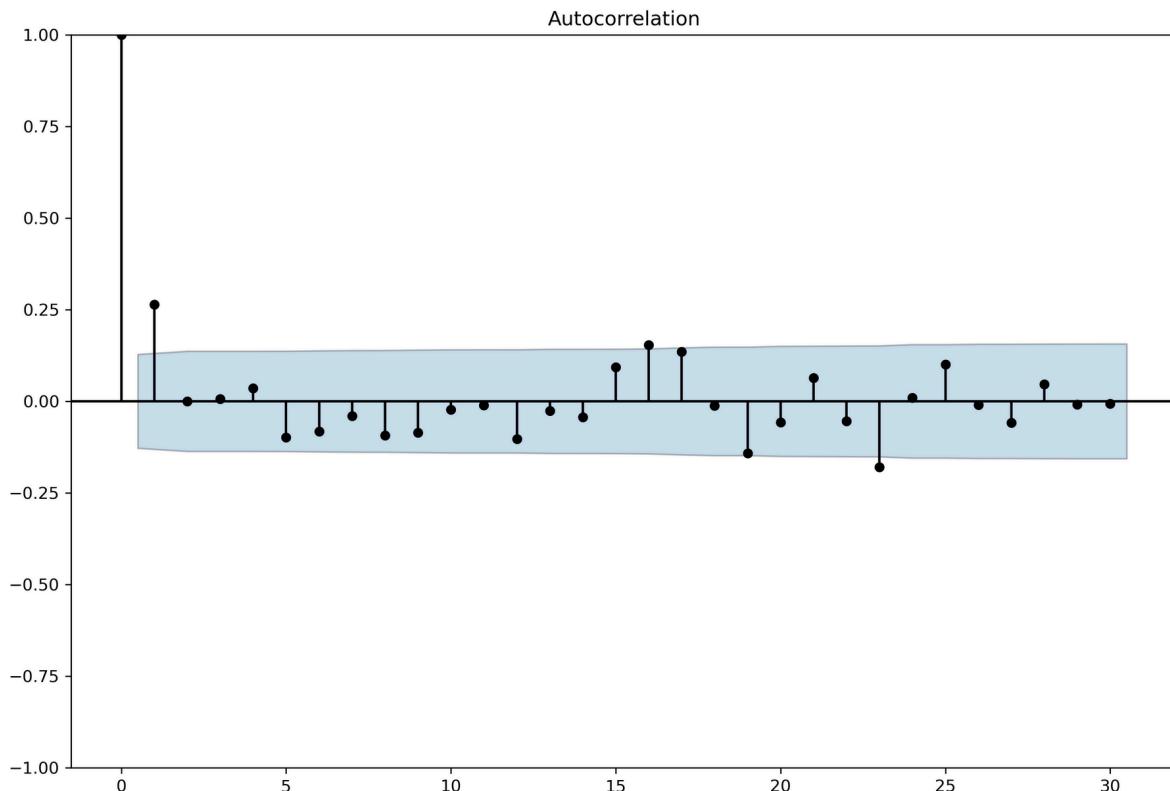
II) Identification de plusieurs modèles et estimation des paramètres

1) Modèle MA (Moving Average)

Un processus MA d'ordre q MA(q), où $q \in \mathbb{N} \cup \{\infty\}$, est une solution de l'équation :

$$X_t = Z_t + \gamma_1 Z_{t-1} + \dots + \gamma_q Z_{t-q}, \quad t \in \mathbb{Z}$$

On utilise une fonction d'autocorrélation (ACF) afin de déterminer l'ordre du modèle MA, dont le graphe est présenté ci-dessous :



On rappelle que l'autocorrélation est de la forme :

$$\rho_X(h) = \rho(X_t, X_{t+h}) = \frac{\gamma_X(h)}{\gamma_X(0)}, \text{ avec } \gamma_X(h) = \text{Cov}(X_t, X_{t+h})$$

Conformément au cours, si les estimations de l'autocorrélation $\hat{\rho}(h)$ semblent négligeables pour tout $h > q$, il est possible de modéliser les données avec un modèle MA d'ordre q, noté MA(q). L'ordre q étant généralement déterminé par la dernière composante significativement non nulle (en ne prenant pas en compte la première composante) on pourrait alors envisager l'ordre 1. Mais nous allons prendre en compte toutes les valeurs possibles par lecture graphique. Le graphique de la fonction d'autocorrélation (ACF) du modèle MA présente généralement une coupure au niveau du retard q. Cette coupure signifie que le graphique ACF tend vers zéro après un certain retard. Selon l'observation du graphique ACF, les valeurs possibles pour q seraient 1 ; 12 ; 17 et 23.

Pour déterminer l'ordre optimal du modèle MA, nous avons réalisé une série d'essais en expérimentant différentes valeurs d'ordre afin de trouver la configuration la plus appropriée. L'évaluation des performances d'ajustement, ainsi que l'utilisation de critères d'information tels que l'AIC (critère d'information d'Akaike), nous a permis d'explorer différentes options pour le modèle MA.

On obtient les valeurs d'AIC pour les différents ordres considérés comme pertinents précédemment :

Valeurs AIC pour chaque ordre MA :

MA(1): AIC = 1145.4446466592003

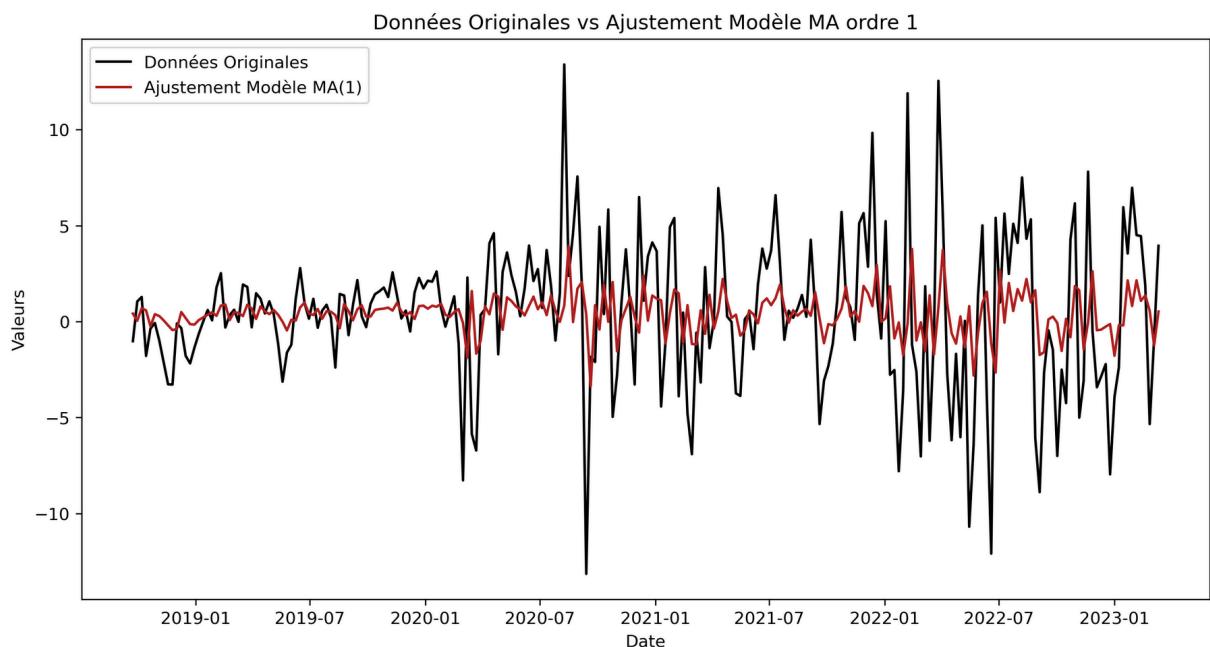
MA(12): AIC = 1158.753345773495

MA(17): AIC = 1160.5213907429643

MA(23): AIC = 1163.8690566999846

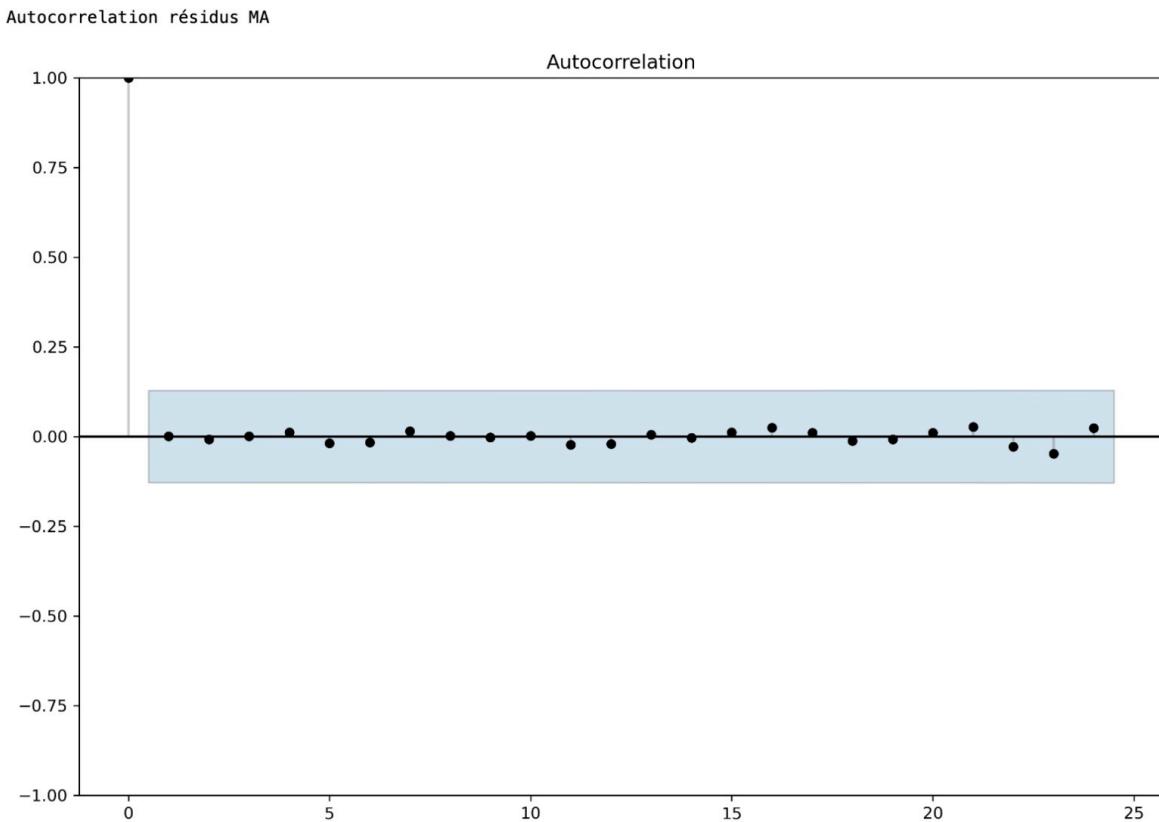
Après avoir exploré diverses valeurs d'ordre pour le modèle MA, nous avons identifié que le modèle MA d'ordre 1 démontrait les meilleures performances dans l'ajustement de notre série temporelle. Ce choix est motivé par plusieurs facteurs, incluant la qualité de l'ajustement, les performances des critères d'information comme l'AIC, et notre compréhension des caractéristiques des données.

Ensuite, nous analysons ce que l'on obtient avec ce modèle MA(1) en l'adaptant aux données différencierées. Pour cela, on représente ci-dessus la solution du modèle MA(1) (en rouge) ainsi que notre série temporelle (X_t) (en noir).



Enfin, nous appliquons une fois de plus la fonction d'autocorrélation (ACF) pour évaluer si les résidus présentent raisonnablement un comportement de Bruit Blanc.

On représente le graphe ci-dessous :



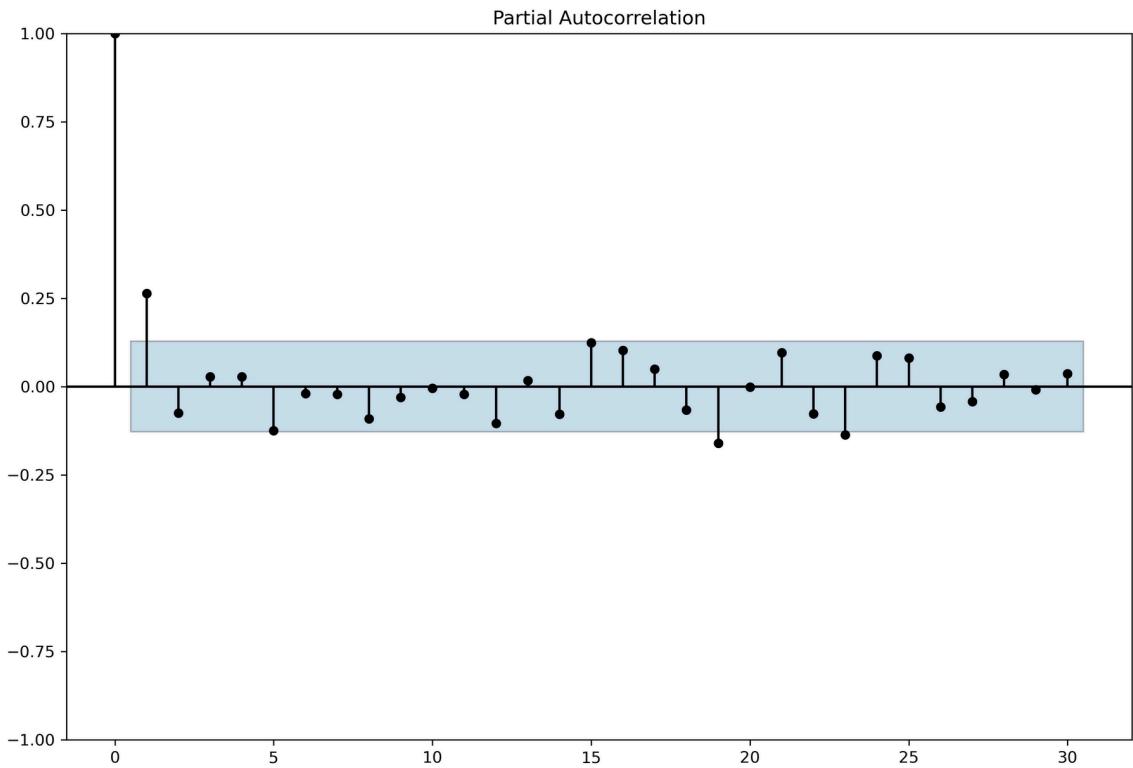
D'après le graphe ACF des résidus, les coefficients d'autocorrélation présentent des valeurs très proches de zéro et ne semblent pas présenter de tendance évidente. Dans ce contexte, ces coefficients peuvent être considérés comme négligeables et suffisamment proches de bruit blanc, ayant ainsi peu d'impact sur les résultats de notre modèle. Par conséquent, nous conservons l'hypothèse selon laquelle le modèle MA(1) est approprié.

2) Modèle AR (AutoRegressive)

La série temporelle (X_t) satisfait un modèle AR(p), où $p \in \mathbb{N} \cup \{\infty\}$, si et seulement si c'est une solution de l'équation

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t, t \in \mathbb{Z}$$

On utilise une fonction d'autocorrélation partielle (PACF) afin de déterminer l'ordre du modèle AR, dont le graphe est présenté ci-dessous :



L'autocorrélation partielle d'ordre h est définie comme suit :
(Sous la convention que $\Pi_0(X_0) = 0$)

$$\tilde{\rho}_X(h) = \rho_X(X_0 - \Pi_{h-1}(X_0), X_h - \Pi_{h-1}(X_h)), h \geq 1$$

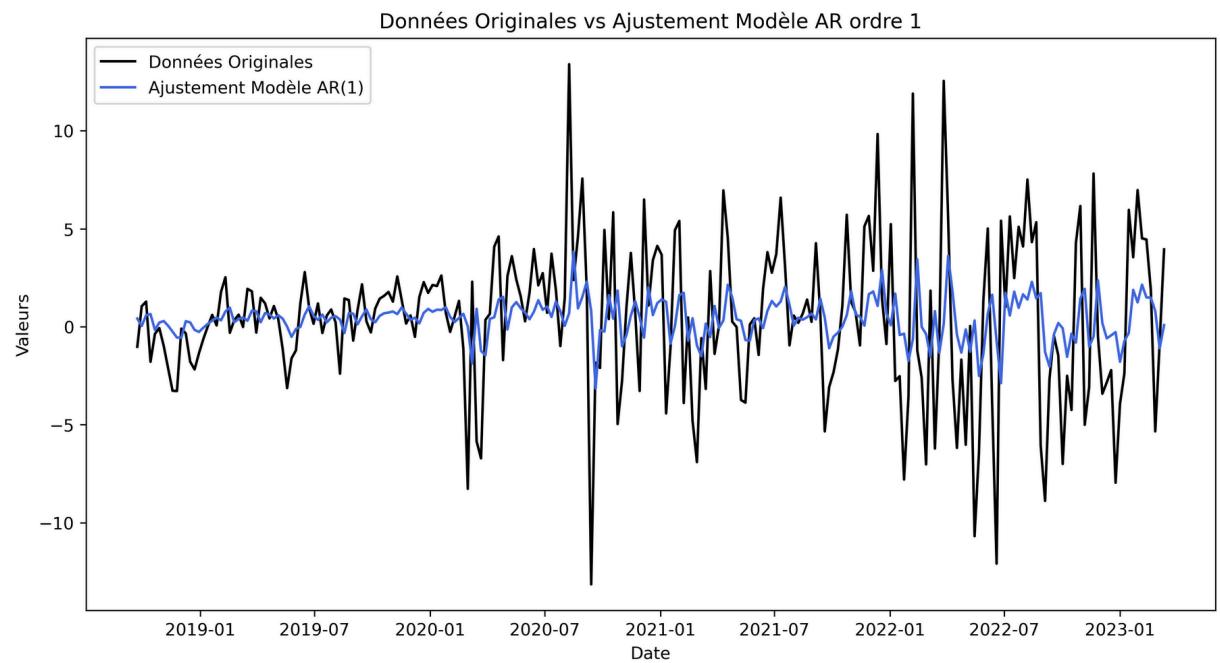
Où $\Pi_{h-1}(X_0)$ est la projection de X_0 sur l'étendue linéaire de (X_1, \dots, X_{h-1}) .

Pour identifier l'ordre optimal du modèle AR, nous avons entrepris une série d'essais en testant diverses valeurs d'ordre pour déterminer la configuration la plus adaptée. L'évaluation des performances d'ajustement, couplée à l'utilisation de critères d'information tels que l'AIC (critère d'information d'Akaike), nous a guidés dans l'exploration de différentes options pour le modèle AR.

Valeurs AIC pour chaque ordre AR :
AR(1): AIC = 1146.1486679428604
AR(12): AIC = 1157.631067515938
AR(19): AIC = 1160.8015049084506

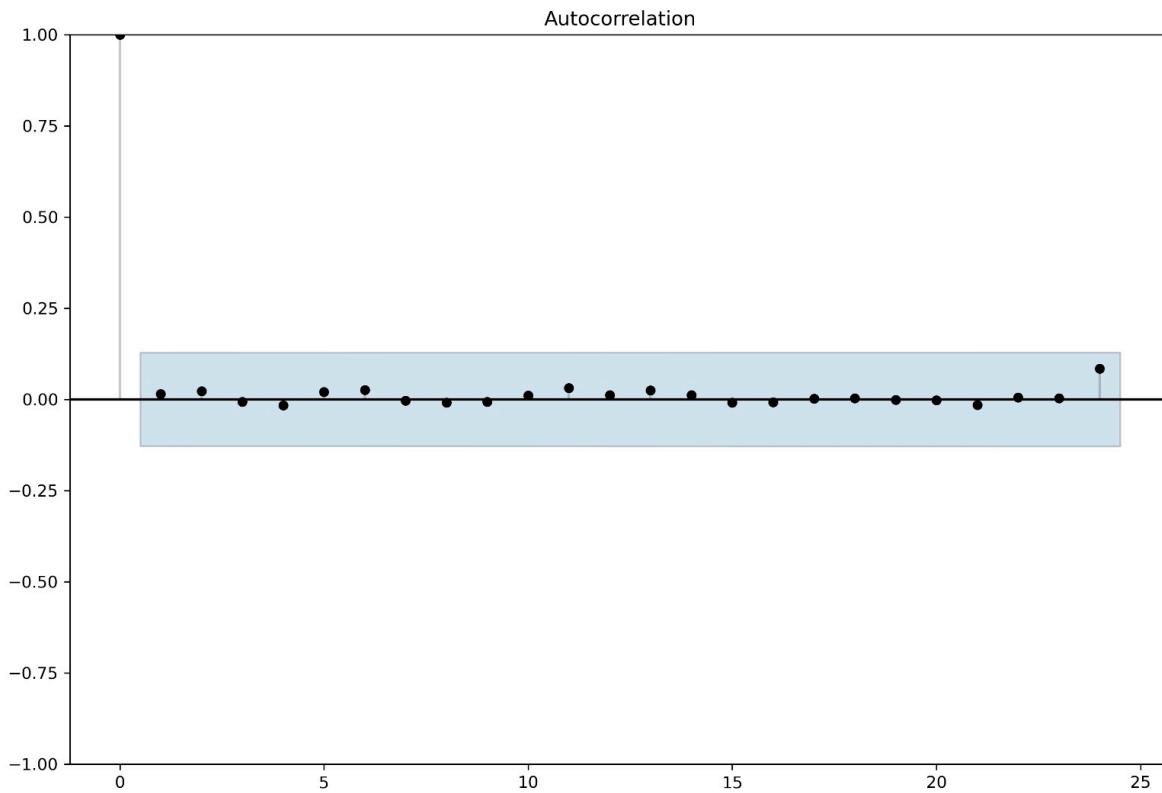
Après avoir exploré diverses valeurs d'ordre pour le modèle AR, nous avons observé que le modèle AR d'ordre 1 AR(1) présentait les performances optimales pour ajuster notre série temporelle. Cette conclusion découle de la plus faible valeur d'AIC associée au modèle de premier ordre.

De même que pour le modèle MA, nous adaptons notre modèle à nos données différencierées. Pour cela, on représente ci-dessus la solution du modèle AR(1) (en bleu) ainsi que notre série temporelle (X_t) (en noir).



Afin de vérifier si les résidus sont bien raisonnablement Bruits blancs, nous réalisons un test ACF, dont la représentation graphique est affichée ci-dessus :

Autocorrelation résidus AR



D'après le graphe PACF des résidus, les coefficients d'autocorrélation présentent des valeurs très proches de zéro et ne semblent pas présenter de tendance évidente. Dans ce contexte, ces coefficients peuvent être considérés comme négligeables et suffisamment proches de bruit blanc, ayant ainsi peu d'impact sur les résultats de notre modèle. Par conséquent, nous conservons l'hypothèse selon laquelle le modèle AR(1) est approprié.

3) Modèle ARMA (Moving average & AutoRegressive)

Pour trouver les ordres optimum du modèle ARMA(p, q), nous allons nous baser sur le Critère d'Information d'Akaike (AIC). L'AIC égal à

$$AIC = \frac{1}{n} L_n^0 (\hat{\theta}_n(p, q)) + \frac{2(p + q)}{n}$$

est basé sur la log-vraisemblance pénalisée. Ce critère nous permet de comparer les modèles entre eux, le plus approprié est en général celui avec l'AIC le plus faible. A l'aide d'une fonction nous générerons la matrice AIC à partir de nos données différencierées. Nous focalisons notre attention sur la partie supérieure de la matrice, où chaque ligne (i, j) correspond à la valeur de l'AIC pour le modèle ARMA(i, j). Cela s'explique par le fait que ces valeurs sont considérablement plus élevées que celles situées en dessous, rendant ainsi inutile l'examen de l'intégralité de la matrice. Pour faciliter la lecture, nous organisons les valeurs d'AIC en ordre croissant, associées à leurs ordres respectifs, disposées en colonnes.

AIC Matrix:		
	Order (p,q)	AIC
1	(0, 1)	1286.266943
20	(3, 2)	1286.313292
6	(1, 0)	1287.590836
2	(0, 2)	1288.235890
12	(2, 0)	1288.238728
7	(1, 1)	1288.241105
8	(1, 2)	1289.615466
32	(5, 2)	1289.806638
3	(0, 3)	1289.871457
18	(3, 0)	1290.075590
13	(2, 1)	1290.166735
30	(5, 0)	1290.212930
4	(0, 4)	1290.277755
14	(2, 2)	1290.451203
9	(1, 3)	1290.913450
11	(1, 5)	1290.915496
26	(4, 2)	1291.055738
22	(3, 4)	1291.117977
31	(5, 1)	1291.171048
27	(4, 3)	1291.386579
5	(0, 5)	1291.636945
24	(4, 0)	1291.867867
15	(2, 3)	1291.950941
28	(4, 4)	1291.951970
19	(3, 1)	1292.055556
10	(1, 4)	1292.067605
29	(4, 5)	1292.235306
17	(2, 5)	1292.749132
25	(4, 1)	1292.818165
21	(3, 3)	1293.537214
33	(5, 3)	1293.670696
16	(2, 4)	1293.783998
23	(3, 5)	1294.329101
34	(5, 4)	1294.665834
35	(5, 5)	1295.132258
0	(0, 0)	1302.433591

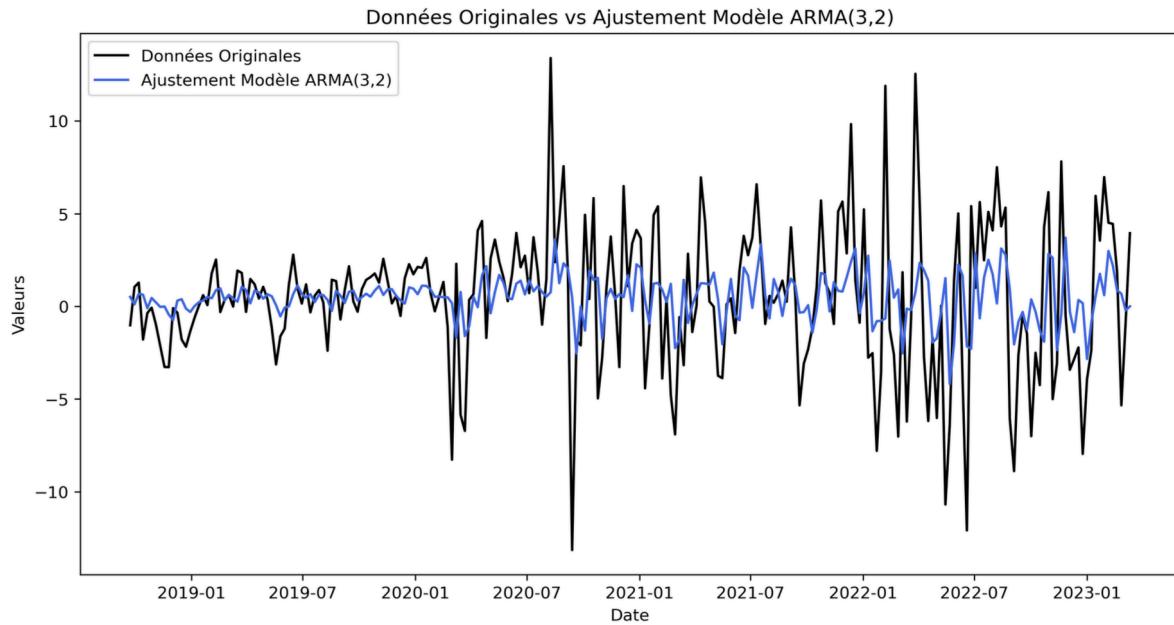
Voici les trois indices d'AIC les plus bas identifiés :

Top Three Models:		
	Order (p,q)	AIC
1	(0, 1)	1286.266943
20	(3, 2)	1286.313292
6	(1, 0)	1287.590836

Le coefficient (0,1) de la matrice affiche le plus bas AIC, ce qui suggère initialement que cet ordre serait optimal pour le modèle ARMA. Cependant, il est important de noter que l'AIC n'est qu'un critère parmi d'autres pour évaluer la pertinence d'un ordre. D'autres facteurs doivent être pris en considération. Par exemple, les valeurs d'AIC pour les coefficients (0,1) et (3,2) sont très proches.

De plus, il est intéressant de noter que le modèle ARMA(0,1) équivaut à un modèle MA(1), expliquant ainsi l'égalité entre l'AIC de MA(1) et celui du modèle ARMA(0,1). Par conséquent, la décision finale ne peut reposer uniquement sur l'AIC.

Afin de faire un choix éclairé, comme pour nos autres modèles, nous examinons également l'aspect graphique de l'ajustement des modèles à nos données différencierées. Dans ce contexte, nous présentons ci-dessus la solution du modèle ARMA(0,1) (en bleu), la comparant visuellement à notre série temporelle (X_t) (en noir).



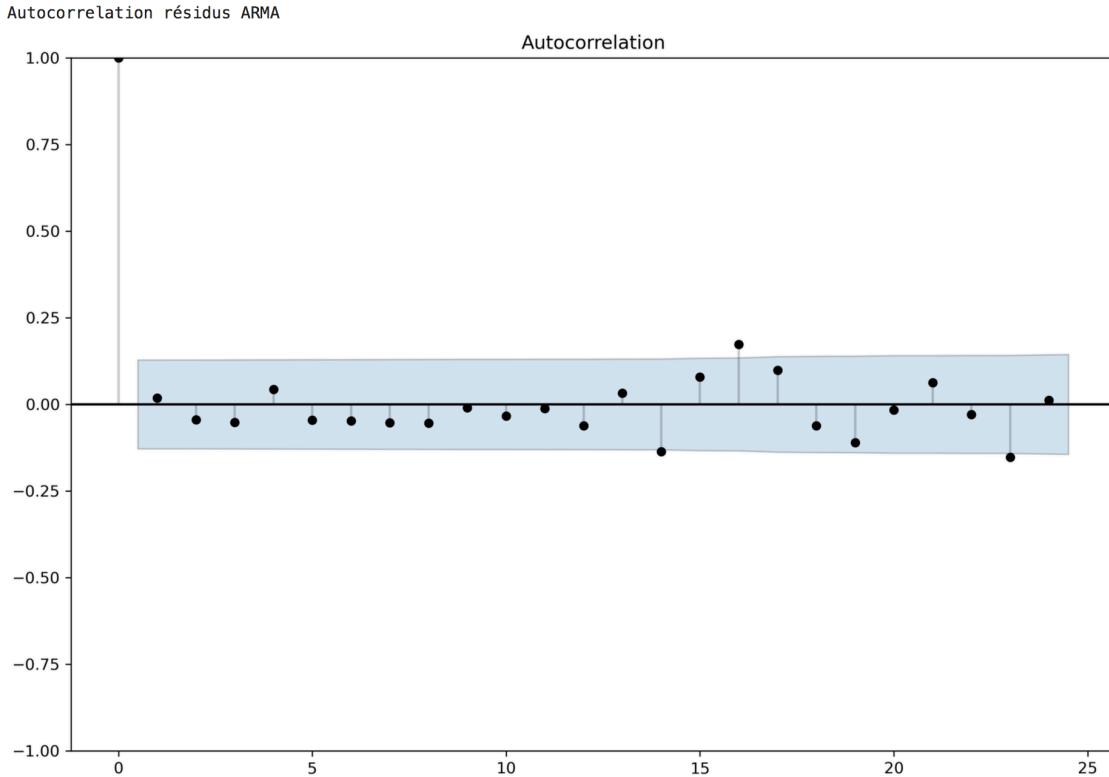
L'examen visuel indique que le modèle ARMA(3,2) s'ajuste de manière plus appropriée à nos données différencierées par rapport au modèle ARMA(0,1) (consultez le graphe d'ajustement du modèle MA(1) à la page 9). En tenant compte des facteurs mentionnés précédemment, ainsi que de celui-ci, notre choix se porte donc sur celui-ci.

	coef	std err	z	P> z	[0.025	0.975]
const	0.5144	0.349	1.473	0.141	-0.170	1.199
ar.L1	-0.0920	0.067	-1.363	0.173	-0.224	0.040
ar.L2	-0.8801	0.031	-28.846	0.000	-0.940	-0.820
ar.L3	0.2962	0.065	4.552	0.000	0.169	0.424
ma.L1	0.3701	0.037	9.931	0.000	0.297	0.443
ma.L2	0.9910	0.055	17.959	0.000	0.883	1.099
sigma2	13.2916	1.056	12.582	0.000	11.221	15.362

Nous obtenons ainsi le modèle suivant :

$$X_t = -0,0920 \cdot X_{t-1} - 0,8801 \cdot X_{t-2} + 0,2962 \cdot X_{t-3} + Z_t + 0,3701 \cdot Z_{t-1} + 0,9910 \cdot Z_{t-2}$$

Nous allons maintenant vérifier si ce modèle est approprié.



D'après le graphe ACF des résidus, nous observons une certaine volatilité dans les résultats de l'ACF, mais ces derniers restent globalement dans l'intervalle de confiance. Cette observation pourrait résulter d'un échantillon de données relativement important, ce qui pourrait conduire les tests à détecter des corrélations faibles mais non nulles. De plus, les coefficients d'autocorrélation présentent des valeurs très proches de zéro et ne semblent pas présenter de tendance évidente. Dans ce contexte, ces coefficients peuvent être considérés comme négligeables et suffisamment proches de bruit blanc, ayant ainsi peu d'impact sur les résultats de notre modèle. Par conséquent, nous conservons l'hypothèse selon laquelle le modèle ARMA(3,2) est approprié.

4) Choix du modèle

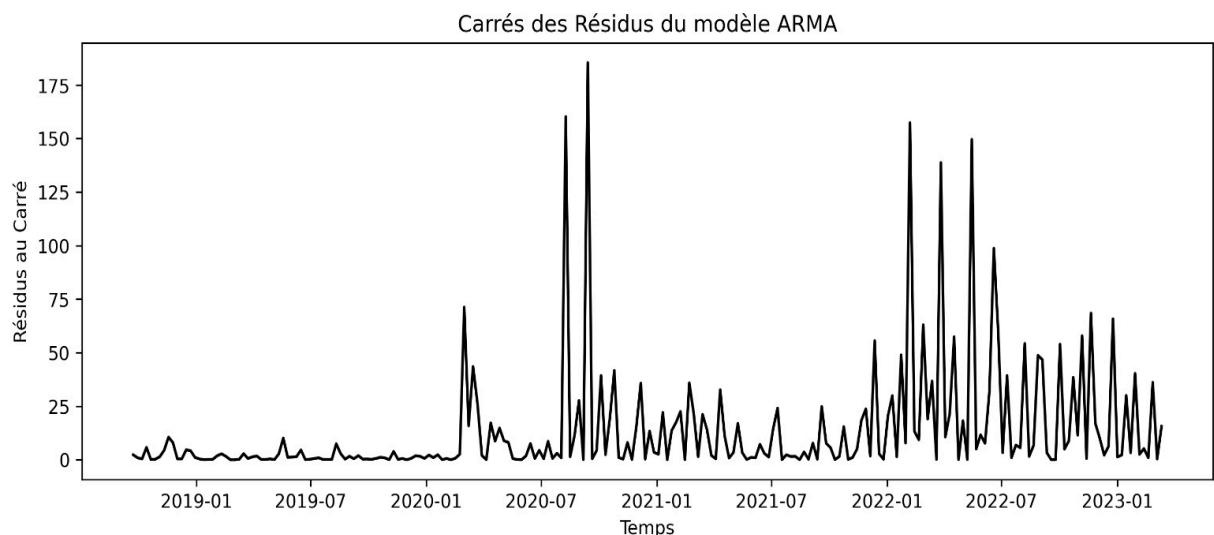
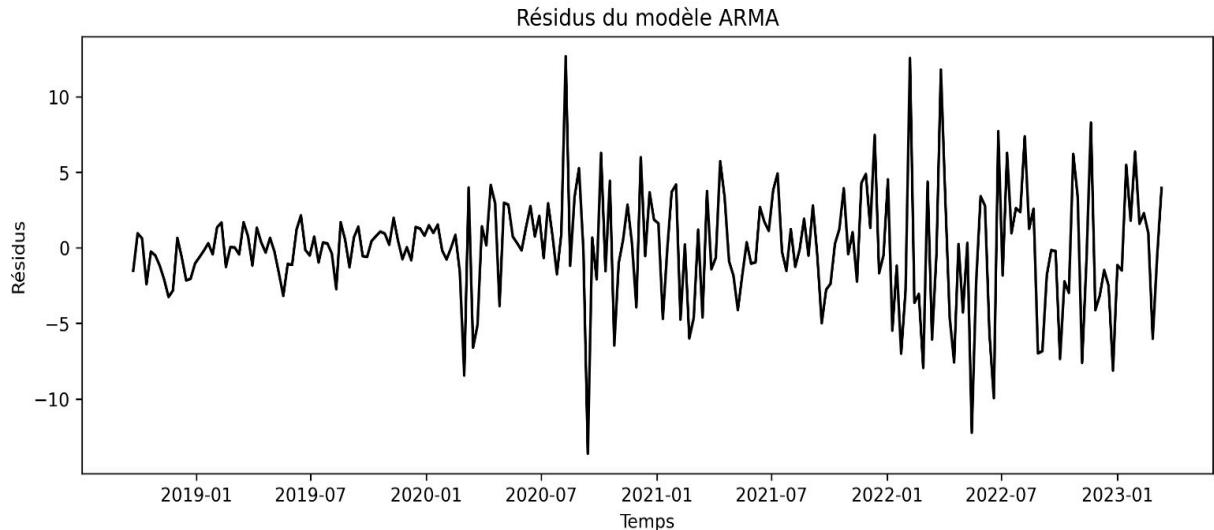
Nous allons comparer les 3 modèles que nous avons trouvés pour ne retenir que le plus approprié. À cette fin, rappelons les valeurs d'AIC pour les trois modèles considérés : MA(1), AR(1), et ARMA(3,2).

MA(1): AIC =	1286.2669434574527
AR(1): AIC =	1287.5908359277587
ARMA(3,2): AIC =	1286.3077649957831

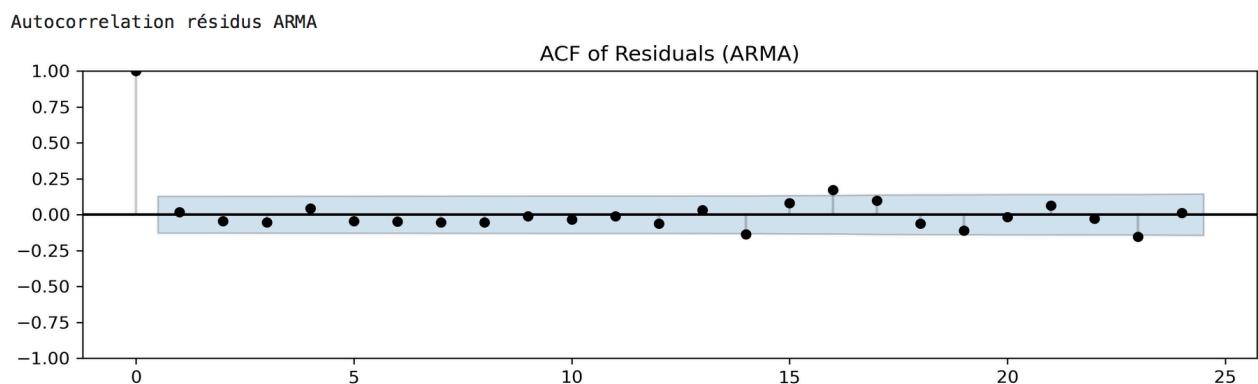
Compte tenu des explications antérieures et des valeurs d'AIC très similaires entre le modèle MA(1) et le modèle ARMA(3,2), notre choix se porte sur ce dernier. Il affiche une performance plus avantageuse en termes d'ajustement aux données différencierées.

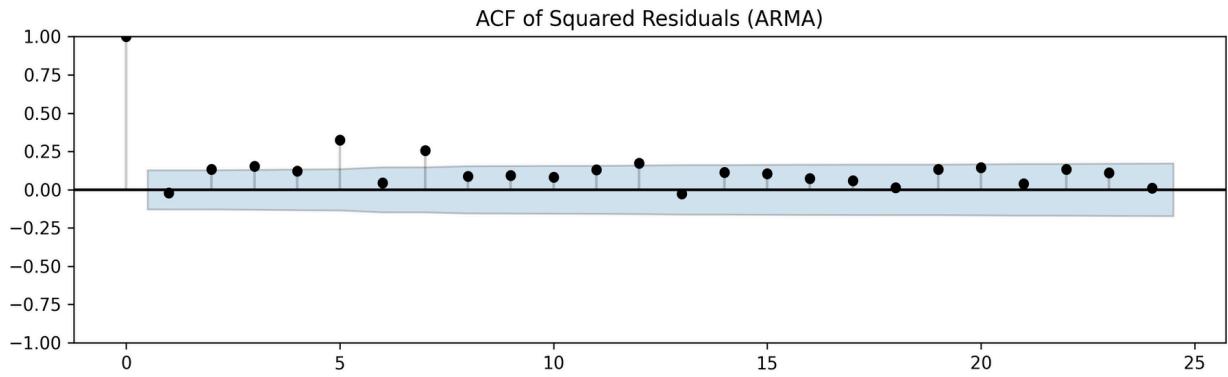
5) Résidus

Nous devons maintenant vérifier si les résidus de notre modèle ARMA(3,2) sont gaussiens. On affiche en premier leur distribution.



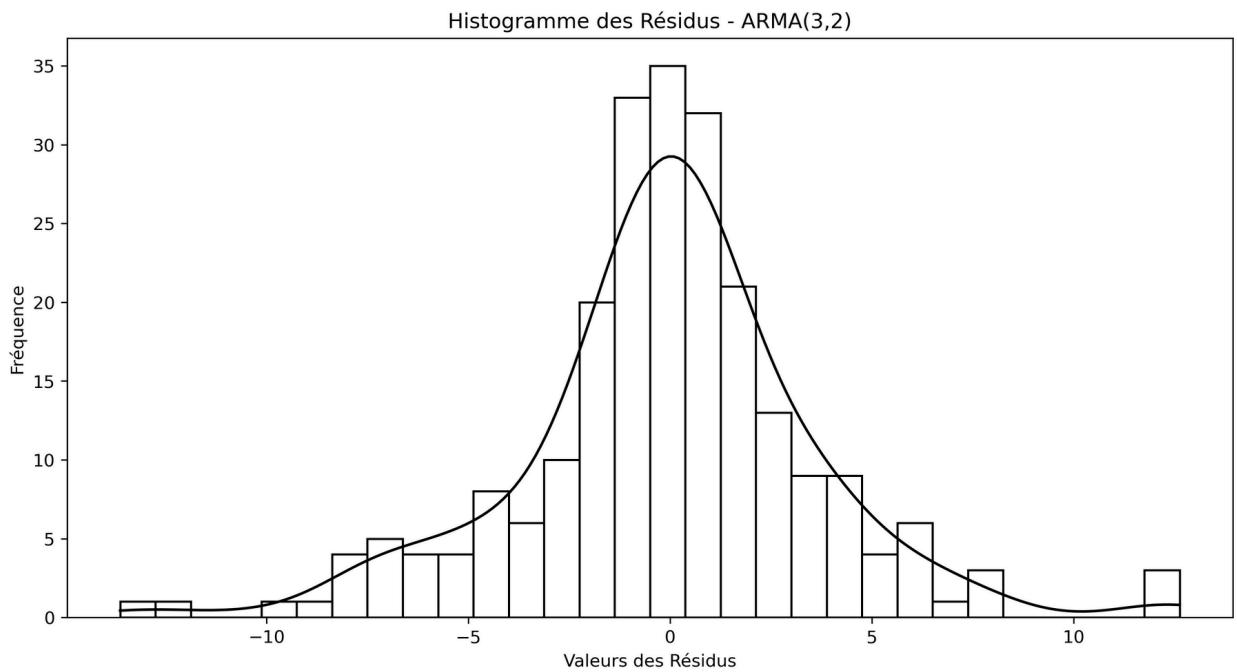
Puis on affiche leur graphes ACF.

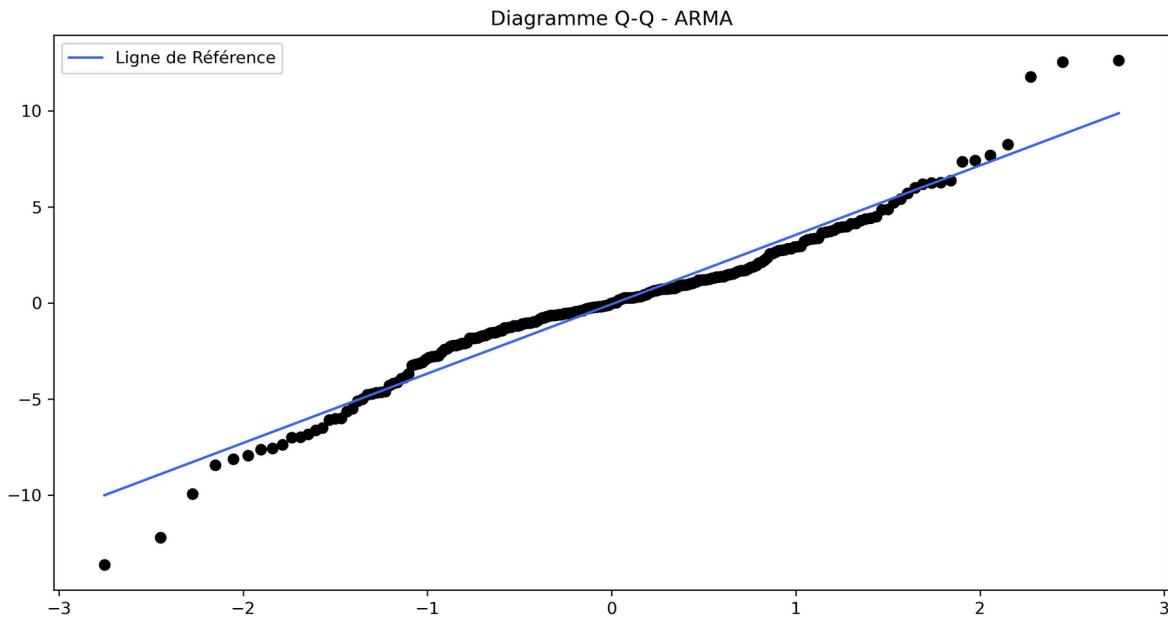




Il ne semble pas rester d'autocorrélation dans les résidus car nous n'observons aucunes composantes significatives non nulles mises à part deux dans le graphe des résidus au carré.

Nous allons maintenant effectuer un test Shapiro-Wilk ainsi qu'afficher l'histogramme des résidus et le diagramme Q-Q plot afin d'étudier la normalité des résidus.





L'histogramme peut nous amener à supposer une normalité des résidus.

Ceci peut être conforté par le diagramme QQ-plot qui montre un alignement assez satisfaisant des points sur la bissectrice, à l'exception de certaines valeurs qui s'en éloignent légèrement. On peut alors émettre l'hypothèse que les résidus suivent une distribution normale.

Le test Shapiro-Wilk nous donne les résultats suivants :

```
Résultats du Test de Shapiro-Wilk - ARMA : ShapiroResult(statistic=0.9610891342163086, pvalue=5.496301
582752494e-06)
```

On remarque alors que la p-value est largement inférieur à 0,005, ce qui amène à rejeter l'hypothèse selon laquelle les résidus suivent une distribution normale.

Afin d'avoir une évaluation la plus robuste possible de la normalité des résidus ARMA, nous effectuons combiner nos résultats précédents avec le test de Kolmogorov-Smirnov. Ce test compare la distribution empirique des données à une distribution théorique et donc nous permettre de conclure quant à la distribution normale de nos résidus. Il est certes moins puissant que le test de Shapiro-Wilk mais il est moins sensible à la taille de l'échantillon, ce qui dans notre cas peut s'avérer révélateur. Il affiche les résultats suivants :

```
Résultats du Test de Kolmogorov-Smirnov - AR : KstestResult(statistic=0.21642138886384632, pvalue=4.219481917029026
e-10, statistic_location=-1.6182034447750164, statistic_sign=1)
```

Ce test nous montre une fois de plus que la p-value est largement inférieur à 0,05 et nous rejetons alors l'hypothèse selon laquelle les résidus sont gaussiens. Malgré cela, compte tenu du fait que l'analyse graphique offre une perception intuitive des données et que les tests statistiques peuvent parfois être sensibles à la taille de l'échantillon et/ou aux valeurs aberrantes, nous privilégions cette l'approche graphique. Nous allons tout de même effectuer des prédictions avec ce modèle.

III) Prédiction des valeurs futures

1) Intervalles de prédictions pour les 26 données les plus récentes

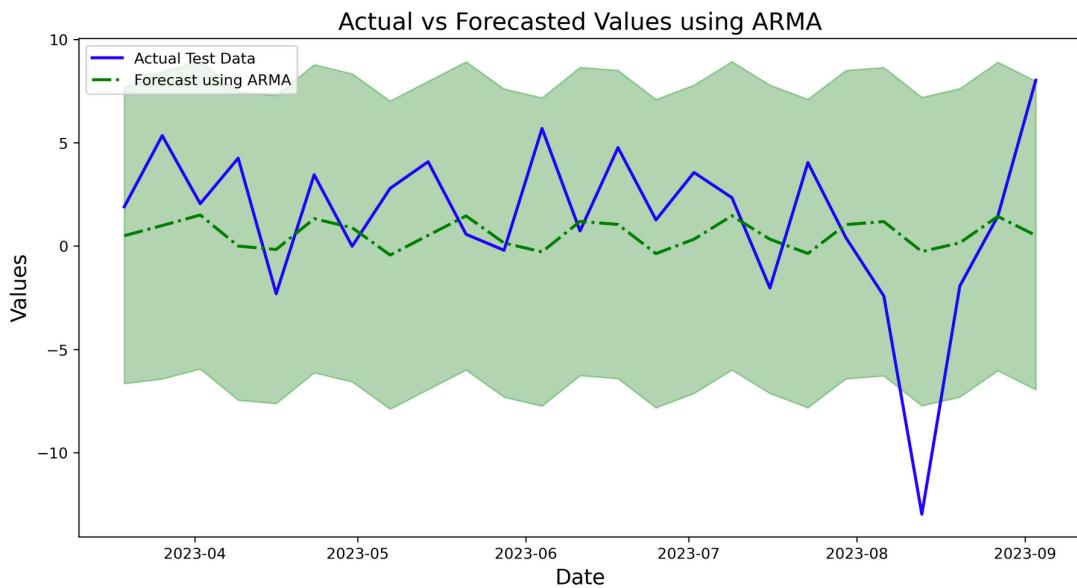
Dans cette partie, nous allons prédire les valeurs des 26 prochaines observations. Nous avons choisi 26 observations et non 10 car cela semblait plus adapté à la taille de notre échantillon de données. A l'aide de notre modèle, nous allons prédire 26 valeurs soit 10% de nos données et les comparer à nos valeurs réelles d'avril 2023 à septembre 2023. Les prédictions sont bien entendues à faire sur notre série originelle et non différenciée. Pour repasser sur nos données réelles, nous devons appliquer l'opération inverse du prétraitement, c'est à dire :

$$D_t = X_t + D_{t-d}$$

Avec $d=1$ car nous avons différencié une fois notre série. Nous avons alors :

$$D_t = X_t + D_{t-1}$$

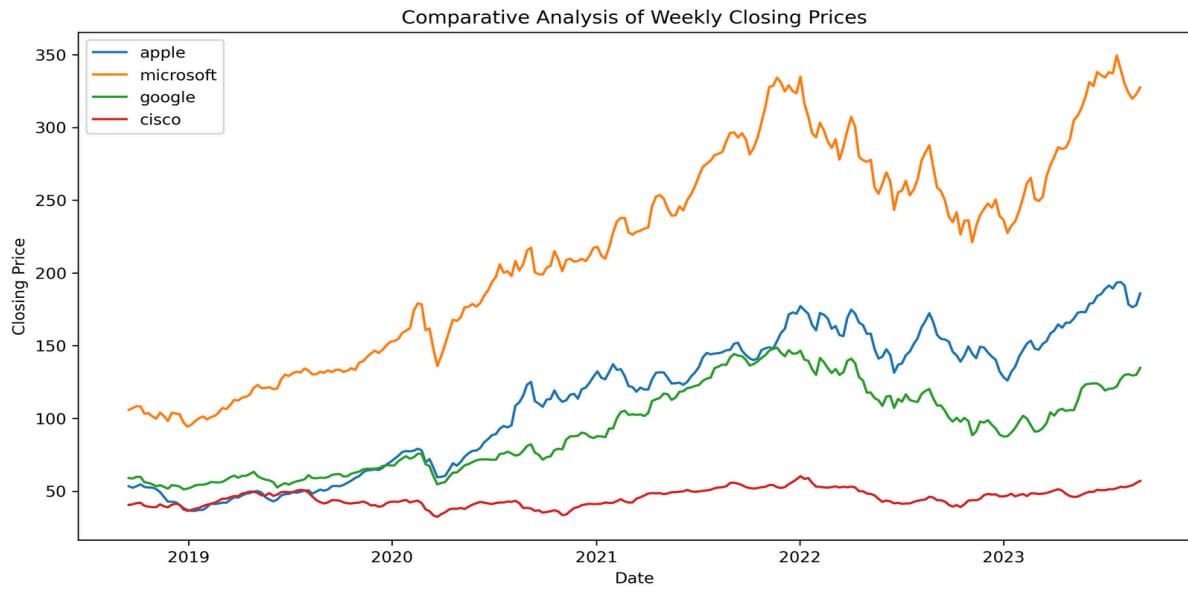
Nous avons choisi un intervalle de confiance à 95%. On présente nos prédictions sur le graphique suivant avec nos valeurs réelles de test (en bleu) et les prédictions de notre modèle ARMA(3,2) (en vert).



Au vu de la non normalité des résidus, on pouvait s'attendre à ce que nos prédictions assez loin des 26 dernières valeurs de notre série. Nous avons fait le choix de conserver ce modèle tout en rejetant l'hypothèse de la distributivité normale des résidus tout en étant conscient du risque que cela pouvait avoir sur nos prédictions. En revanche, bien que les prédictions soient loin d'être parfaites, elles ne sont pas catastrophiques. On peut remarquer que nos prédictions approchent doucement nos valeurs comme aux périodes fin avril 2023 - début avril 2023 ; fin avril 2023 ; mi-juin 2023 etc. De plus pour ce type de données, il est assez probable de ne pas avoir de prédictions très précises. Nous allons alors tenter d'améliorer notre modèle et de prédire nos valeurs futures par un autre moyen qui nous l'espérons sera plus performant.

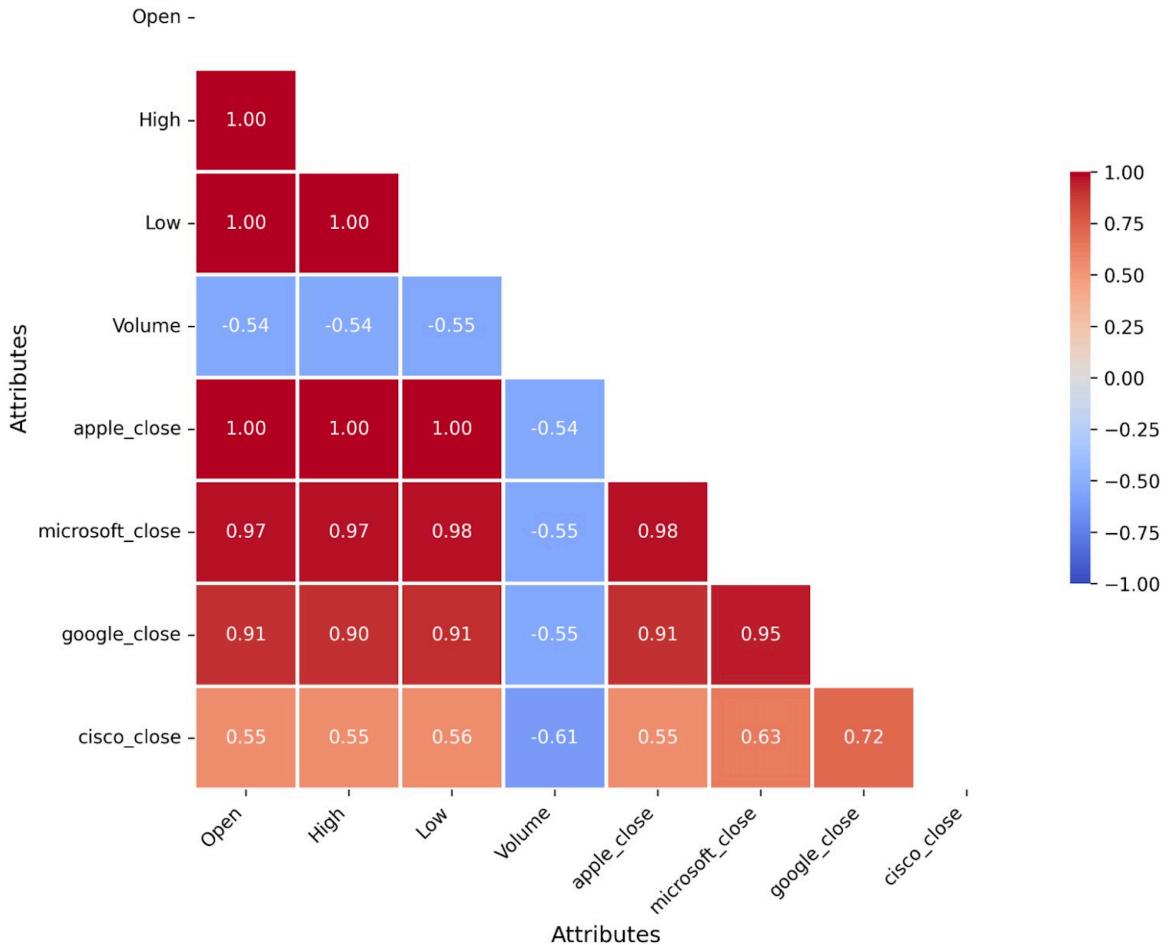
2) Matrice de corrélation et variables explicatives

Avant de sélectionner nos variables explicatives nous observons les représentations temporelles des prix de clôture de Microsoft, Google, et Cisco, les comparant entre elles et avec celle d'Apple, notre variable d'intérêt.



On remarque que les prix de clôture suivent une tendance similaire à ceux d'Apple, à l'exception de Cisco qui semble avoir une trajectoire légèrement différente. Cette observation suggère un possible impact sur notre variable d'intérêt. Nous confirmerons cela via une matrice de corrélation dans le but de sélectionner les variables explicatives ayant un impact significatif sur *Apple_Close*, réduisant ainsi le nombre de variables pour optimiser notre modèle de prévision.

Lower Triangular Correlation Matrix



La matrice de corrélation révèle que *Microsoft_Close* présente le coefficient de corrélation linéaire le plus élevé avec notre variable d'intérêt, *Apple_Close*, excluant *High* et *Low*. Suivant de près, *Google_Close* affiche également une corrélation significative. Étant donné que le coefficient de corrélation pour *High* et *Low* est de 1.00, suggérant des valeurs fréquemment égales, nous les excluons pour éviter toute distorsion potentielle des prédictions. Des variables excessivement similaires compliqueraient l'interprétation des rôles dans la prédiction, rendraient les résultats peu fiables et obscurciraient les conclusions sur l'impact de chaque variable. Ainsi, nous les retirons.

Cisco_Close est fortement corrélée à *Microsoft_Close* et *Google_Close* mais compte tenu des explications données sur son allure graphique et que l'on dispose déjà de variables explicatives fortement corrélées avec notre variable d'intérêt, nous ne le conservons pas. Bien que *Volume* présente une corrélation de 0,54 avec *Apple_Close*, elle reste également corrélée aux trois autres variables, suggérant une influence potentielle sur notre variable d'intérêt. Ainsi, nous conservons ces trois dernières variables explicatives : *Microsoft_Close*, *Google_Close* et *Volume*.

De la même manière que pour notre variable d'intérêt, nous allons procéder à la vérification de la stationnarité de nos variables explicatives.

Après application du test ADF on obtient les résultats suivants :

Microsoft :

```
microsoft ADF-Test
Test Statistic           -0.833539
p_value                  0.809071
#Lags Used               1.000000
Number of Observations Used 258.000000
Critical Value (1%)      -3.455953
Critical Value (5%)       -2.872809
Critical Value (10%)      -2.572775
dtype: float64

====> The time series is not stationary.
```

Google :

```
google ADF-Test
Test Statistic           -0.792211
p_value                  0.821399
#Lags Used               1.000000
Number of Observations Used 258.000000
Critical Value (1%)      -3.455953
Critical Value (5%)       -2.872809
Critical Value (10%)      -2.572775
dtype: float64

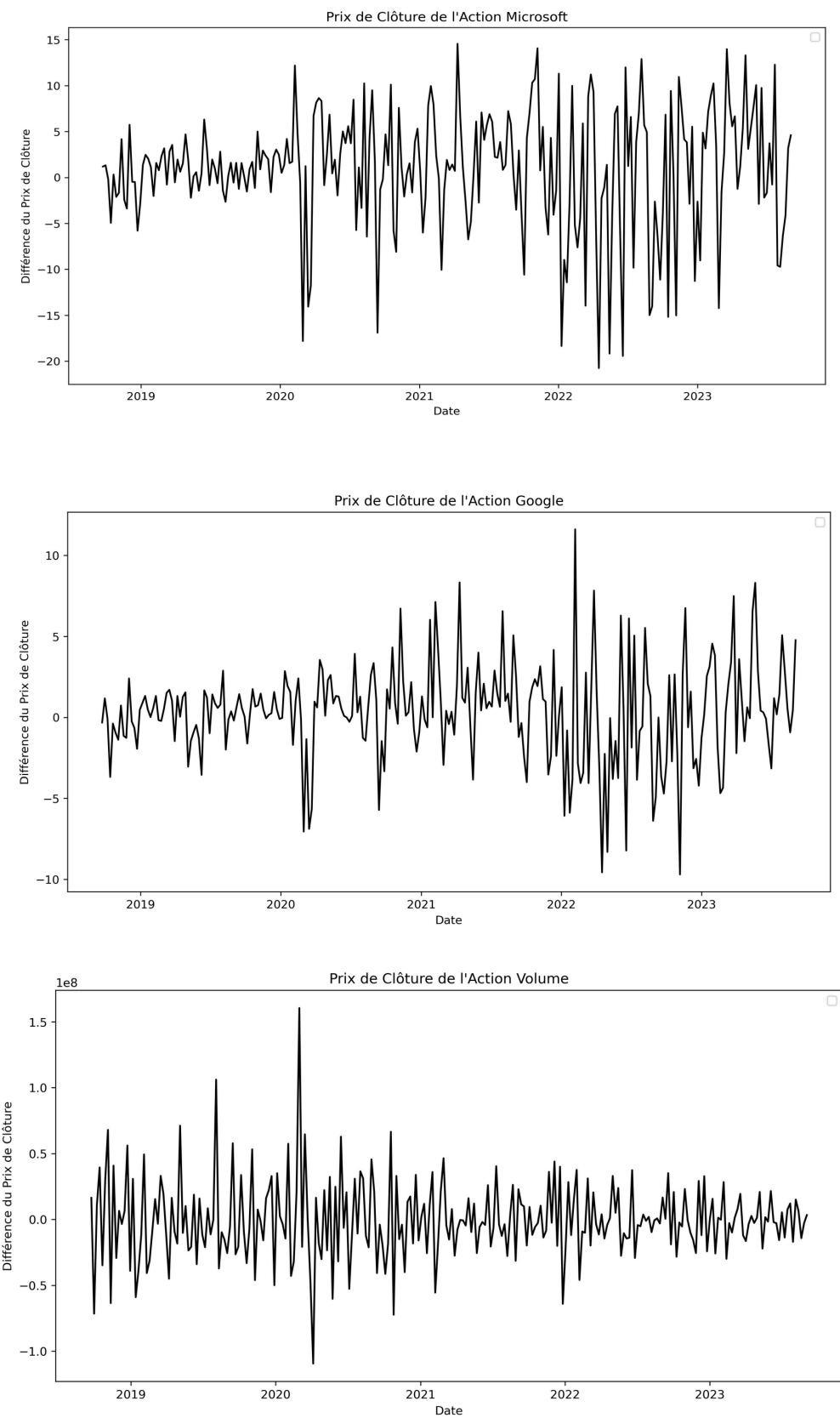
====> The time series is not stationary.
```

Volume :

```
Apple trading Volume ADF-Test
Test Statistic           -2.387410
p_value                  0.145321
#Lags Used               13.000000
Number of Observations Used 246.000000
Critical Value (1%)      -3.457215
Critical Value (5%)       -2.873362
Critical Value (10%)      -2.573070
dtype: float64

====> The time series is not stationary.
```

Les trois n'étant pas stationnaires, nous allons procéder à la différenciation de celle-ci, exactement de la même manière qu'en première partie. On affiche ci-dessous les nouvelles représentations graphiques après différenciation ainsi que les nouveaux résultats des test ADF.



```

google ADF-Test après diff
Test Statistic           -1.349225e+01
p_value                  3.089910e-25
#Lags Used               0.000000e+00
Number of Observations Used 2.580000e+02
Critical Value (1%)      -3.455953e+00
Critical Value (5%)       -2.872809e+00
Critical Value (10%)      -2.572775e+00
dtype: float64

microsoft ADF-Test après diff
Test Statistic           -1.359755e+01
p_value                  1.981652e-25
#Lags Used               0.000000e+00
Number of Observations Used 2.580000e+02
Critical Value (1%)      -3.455953e+00
Critical Value (5%)       -2.872809e+00
Critical Value (10%)      -2.572775e+00
dtype: float64

```

====> The time series is stationary.

====> The time series is stationary.

```

volume ADF-Test après diff
Test Statistic           -6.055915e+00
p_value                  1.245326e-07
#Lags Used               1.600000e+01
Number of Observations Used 2.420000e+02
Critical Value (1%)      -3.457664e+00
Critical Value (5%)       -2.873559e+00
Critical Value (10%)      -2.573175e+00
dtype: float64

```

====> The time series is stationary.

Nous obtenons finalement la base suivante :

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits	Brand_Name	Ticker	Industry_Tag	Country
20/09/2023 00:00	179.25999450683594	179.6999969482422	175.39999389648438	175.49000549316406	58436200.0	0.0	0.0	apple	AAPL	technology	usa
20/09/2023 00:00	138.0800018310547	138.0800018310547	133.6199951171875	133.74000549316406	29927500.0	0.0	0.0	google	GOOGL	technology	usa
20/09/2023 00:00	329.510009765625	329.5899963378906	320.510009765625	320.7699990136719	21436500.0	0.0	0.0	microsoft	MSFT	technology	usa
20/09/2023 00:00	55.95000076293945	56.04999923706055	55.439998626708984	55.5	12219900.0	0.0	0.0	cisco	CSCO	technology	usa
19/09/2023 00:00	177.52000427246094	179.6300048828125	177.1300048828125	179.07000732421875	51826900.0	0.0	0.0	apple	AAPL	technology	usa
19/09/2023 00:00	137.4199981689453	138.41000366210938	136.6199951171875	138.0399932861328	20353700.0	0.0	0.0	google	GOOGL	technology	usa
19/09/2023 00:00	326.1700134277344	329.3900146484378	324.510009765625	328.6499938964844	16505900.0	0.0	0.0	microsoft	MSFT	technology	usa
19/09/2023 00:00	56.18000030517578	56.209999084472656	55.459999084472656	55.84000015258789	10748500.0	0.0	0.0	cisco	CSCO	technology	usa
18/09/2023 00:00	176.47999572753906	179.3800048828125	176.1699981689453	177.97000122070312	67257600.0	0.0	0.0	apple	AAPL	technology	usa
18/09/2023 00:00	136.6100061035156	139.16000366210938	136.6100061035156	138.2100067138672	21861300.0	0.0	0.0	google	GOOGL	technology	usa
18/09/2023 00:00	327.79998779296875	330.3999938964844	326.3599853515625	329.0599975585937	16834200.0	0.0	0.0	microsoft	MSFT	technology	usa

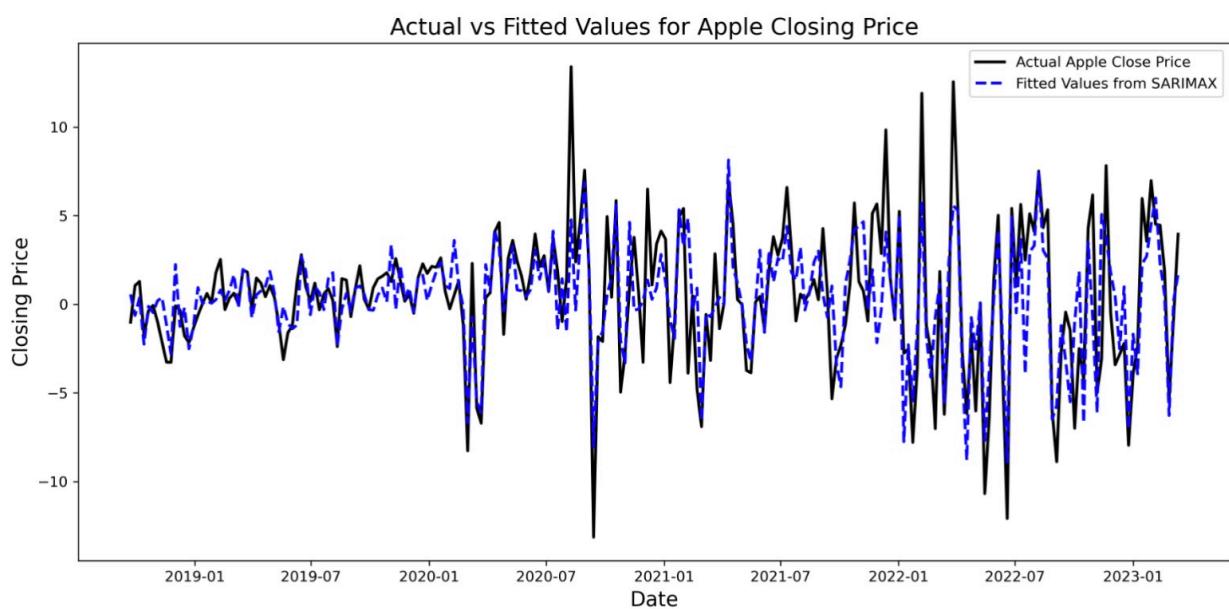
A présent, toutes nos variables explicatives ont été prétraitées, nous allons pouvoir passer à la création de notre modèle de prédiction.

3) Création du modèle

En deuxième partie, nous avons fixé le modèle ARMA(3,2) qui semblait offrir des prédictions acceptables mais perfectibles. Nous le peaufinons en intégrant les données antérieures de nos variables explicatives et de notre variable d'intérêt. Pour élaborer ce nouveau modèle, nous avons utilisé la méthode SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables). Ce modèle étend la classe des modèles ARMAX en incorporant des composantes saisonnières, tout en intégrant des variables explicatives exogènes pour renforcer la précision de la prédiction.

Nous avons débuté avec un modèle SARIMAX de base, en se basant sur notre modèle ARMA(3,2) et en impliquant des composantes saisonnières. Afin d'améliorer la capacité prédictive, nous avons intégré des variables explicatives pertinentes, à savoir le volume des transactions ainsi que les coûts des actions de Microsoft et Google, en raison de leur corrélation significative avec les coûts des actions d'Apple. La fusion des éléments ARMAX avec les variables explicatives a donné naissance à un modèle SARIMAX étendu, enrichi par des composantes saisonnières.

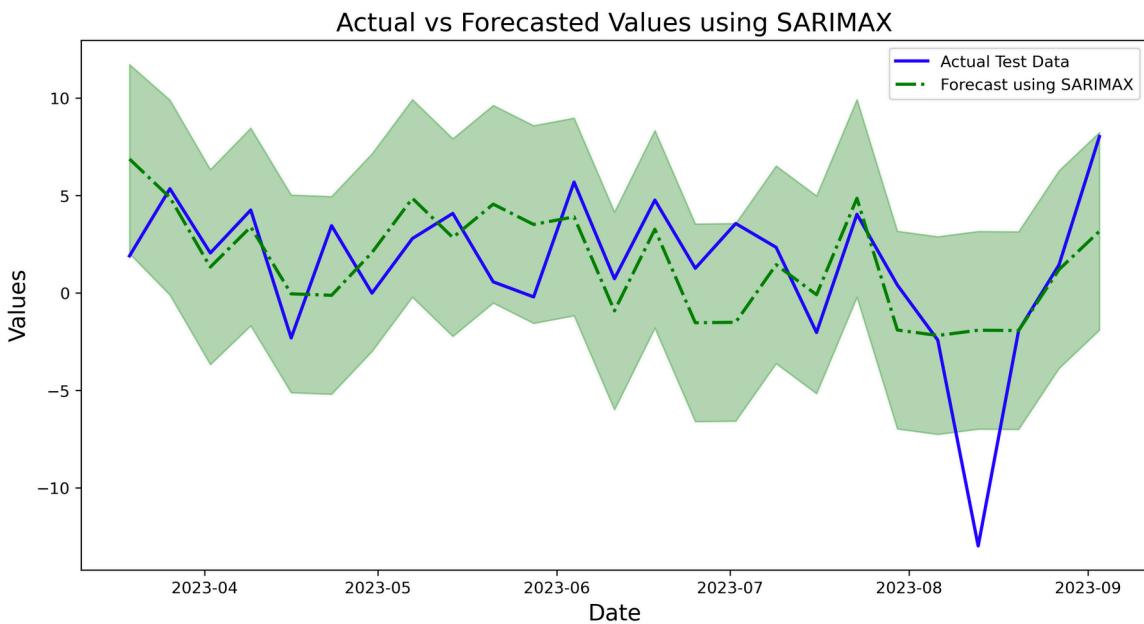
Afin de garantir la généralisation du modèle, nous l'avons ajusté sur des données indépendantes, évaluant sa capacité à prédire avec précision des observations non utilisées pendant l'entraînement. Cette adaptation est représentée graphiquement sur notre ensemble d'entraînement comme suit :



<Figure size 3600x2400 with 0 Axes>

4) Prédictions

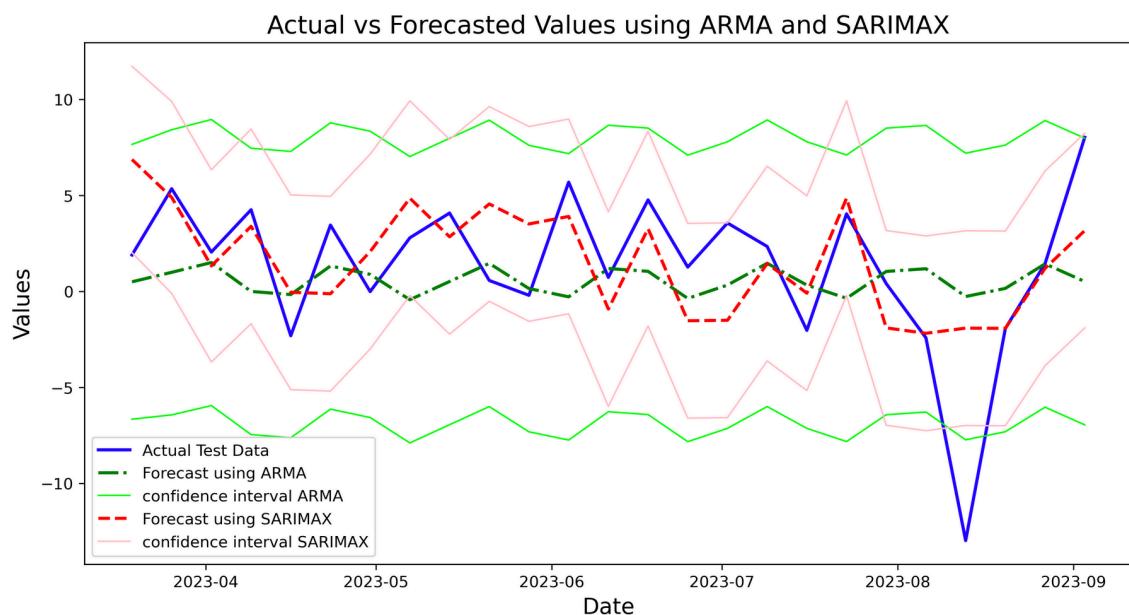
Nous allons maintenant pouvoir travailler sur le jeu test afin de voir si ce nouveau modèle permet de prédire de manière assez fiable l'évolution future de notre série temporelle. De la même manière que précédemment on utilise un intervalle de confiance à 95% et on représente graphiquement nos prédictions comme suit :



Les prédictions semblent bonnes. Ce nouveau modèle intégrant les variables explicatives, démontrent une amélioration significative. Elles se distinguent par leur précision marquée et un rapprochement progressif avec notre série d'observations au fil du temps. Ce modèle semble s'ajuster efficacement à l'augmentation des données.

IV) CONCLUSION

Pour conclure, nous allons comparer nos 2 modèles de prédiction à savoir le modèle ARMA(3,2) fixé en deuxième partie et celui obtenu à partir de ce dernier et des variables explicatives. On pourra également étudier la précision de chacun d'entre eux vis-à-vis des données réelles observées.



Au cours de notre analyse des modèles de prédiction pour les coûts des actions d'Apple, notre étude a mis en lumière deux approches distinctes : le modèle ARMA(3,2) et le modèle SARIMAX. L'objectif était d'évaluer leur capacité à anticiper les variations des coûts des actions d'Apple, en incorporant des variables explicatives pertinentes.

Le modèle ARMA(3,2) a été notre point de départ, fournissant des prédictions acceptables. Cependant, nous avons optimisé son potentiel en adoptant le modèle SARIMAX, incluant des variables explicatives externes. Les résultats ont été concluants, montrant que le modèle SARIMAX, avec l'ajout judicieux de variables explicatives, surpassait significativement le modèle ARMA(3,2). Les prédictions du modèle SARIMAX se sont distinguées par leur précision accrue, s'ajustant de manière adaptative aux données et présentant un rapprochement notable avec la série d'observations.

Cette amélioration souligne l'efficacité du modèle SARIMAX dans la prédiction des coûts des actions d'Apple. Toutefois, l'ouverture à d'autres améliorations et explorations reste présente. Des considérations telles que l'ajout de nouvelles variables explicatives, l'optimisation des paramètres du modèle, ou l'exploration de méthodes plus avancées peuvent être envisagées pour affiner davantage nos prédictions.

En résumé, cette étude met en lumière les avantages du modèle SARIMAX et ouvre la voie à des recherches futures visant à perfectionner nos capacités de prévision dans le domaine des séries temporelles financières.