INSEA
Institut National des Statistiques et
d'Economie Appliquée

Année Universitaire 2023-2024

Rapport de projet

Réaliser par :

Omar El Araby
Ilyas Chaatouf

Série chronologiques : méthode de Box et Jenkin

Encadrer par :
Mme FADOUA BADAOUI

Sommaire :

1-Introduction :

Ce rapport présente l'application de la méthode de Box et Jenkins à la série chronologique mensuelle du taux d'inflation aux UK sur la période de janvier 1993 à décembre 2017. L'objectif est de construire un modèle ARIMA adéquat pour prédire le taux d'inflation futur et d'évaluer ses performances.

2-Description et Préparation des données :

2-1 Description des donnees :

La série chronologique utilisée dans cette analyse représente le taux d'inflation mensuel aux United Kingdom , exprimé en pourcentage. Les données proviennent de la base de données FRED (Federal Reserve Economic Data) de la Réserve fédérale UK. La série couvre une période de 25 ans, de janvier 1993 à décembre 2017.

2-2 Préparation des donnees :

Ce script Python utilise diverses bibliothèques telles que pandas, scikit-learn, matplotlib et statsmodels pour effectuer une analyse de série chronologique du taux d'inflation

```
import pandas as pd
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.arima.model import ARIMA
from sklearn.metrics import mean_squared_error, mean_absolute_error
import numpy as np
from statsmodels.tsa.stattools import adfuller
from pandas.plotting import autocorrelation_plot
```

Le fichier des données doit être téléchargé dans un format compatible avec python, Le format .xlsx est généralement recommandé :

```
chemin_fichier = "C:/Users/dell/Documents/1A/S2/P2/sc/inflation_uk.xlsx"

donnees = pd.read_excel(chemin_fichier)
donnees.columns  = ['Mois','Inflation']
```

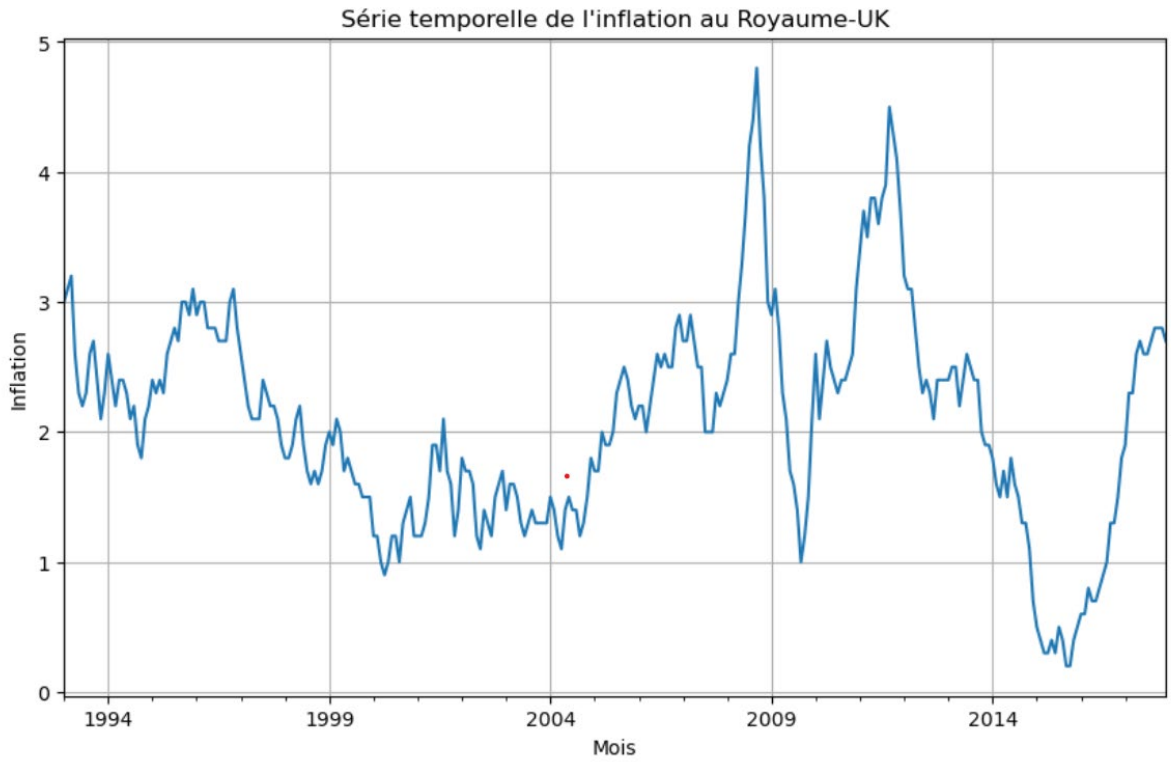
3-Visualisation de la série temporelle et ses composantes :

3-1 Division de données :

```
Taille de l'échantillon d'apprentissage : 240
Taille de l'échantillon de validation : 60
```

Nous avons utilisé les premiers 80 % des observations pour entraîner et tester divers modèles tout en conservant les observations restantes pour valider le modèle final .

3-2 Visualisation de la série :



- L'inflation montre une tendance générale **décroissante** de 1993 à 2017.
- On observe des périodes de légère augmentation de l'inflation, notamment de 2005 à 2008.
- La tendance décroissante est plus prononcée de 1996 à 2002 et de 2011 à 2015.

4-Analyse de la série :

TESTE DE DICKY-FULLER :

```
Augmented Dicky-Fuller Test
ADF Test Statistics : -2.3334896125608386
p-value : 0.1613942353277148
# of lags : 12
Num of Observations : 287
Weak evidence against null hypothesis
Fail to reject null hypothesis
Data has a unit root and is non-stationary
```

D'après le test de Dickey-Fuller Augmenté (ADF) que vous avez effectué, les données présentent **vraisemblablement une racine unitaire et ne sont donc pas stationnaires**. Voici une interprétation détaillée des résultats pour votre jeu de données :

- **Statistique du test :** -2.3334896125608386

  Cette valeur négative est un bon signe pour la stationnarité. Cependant, sa valeur absolue (-2.33) n'est pas assez faible pour rejeter l'hypothèse nulle d'une racine unitaire à un niveau de signification courant (généralement 5 %).

- **Valeur p :** 0.1613942353277148

  La valeur p est supérieure au niveau de signification habituel (0,05). Cela signifie que nous **ne rejetons pas l'hypothèse nulle**. En termes plus simples, les preuves contre la stationnarité des données ne sont pas statistiquement significatives.

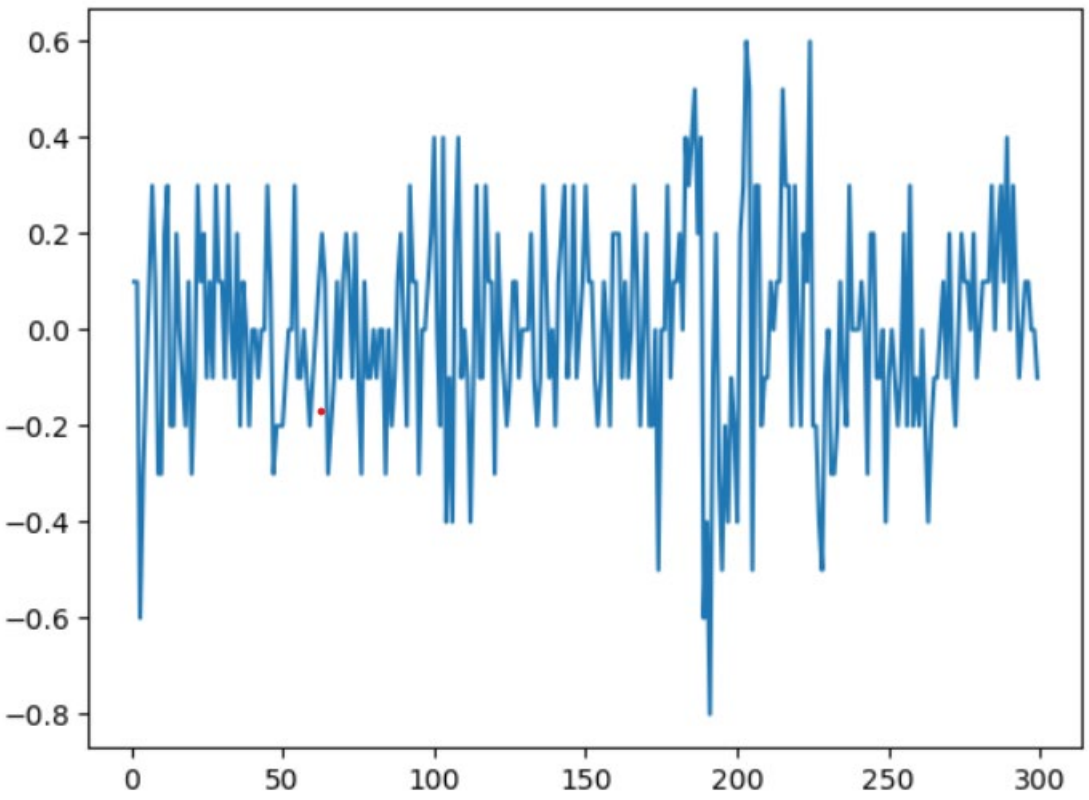
- **Nombre de retards :** 12

  Cela indique que le test a utilisé 12 retards (valeurs passées) pour tenir compte d'une éventuelle autocorrélation dans vos données. Le nombre approprié de retards peut être déterminé par différentes méthodes, et 12 pourrait être le choix optimal dans ce cas.

- **Nombre d'observations :** 287

  Cela signifie que le test a été effectué sur un ensemble de données comportant 287 observations.

Les données analysées initialement n'étaient pas exploitables directement car elles variaient dans le temps (non stationnaires). On a donc utilisé la différenciation pour les stabiliser :



Le graphique montre l'évolution des données après différenciation. La ligne horizontale indique que les données sont désormais stationnaires (propriétés stables). L'allure de la courbe permet d'analyser les variations de la série originale.

TESTE DE DICKY-FULLER :

```
Augmented Dicky-Fuller Test
ADF Test Statistics : -6.779314819548393
p-value : 2.523131312366415e-09
# of lags : 11
Num of Observations : 287
Strong evidence against null hypothesis
reject null hypothesis
Data has no unit root and is stationary
```

'après le test de Dickey-Fuller Augmenté (ADF) que vous avez effectué, vos données d'inflation sont stationnaires. Voici une interprétation détaillée des résultats :

•Statistique du test ADF : -6.779314819548393 : Cette valeur négative est beaucoup plus importante que la valeur critique habituelle pour un niveau de signification de 5 %. Cela indique une forte preuve contre l'hypothèse nulle d'une racine unitaire.

•Valeur p : 2.523131312366415e-09

•Cette valeur p est extrêmement faible, bien inférieure au niveau de signification de 5 %. Cela signifie que nous pouvons rejeter l'hypothèse nulle avec une grande confiance. En

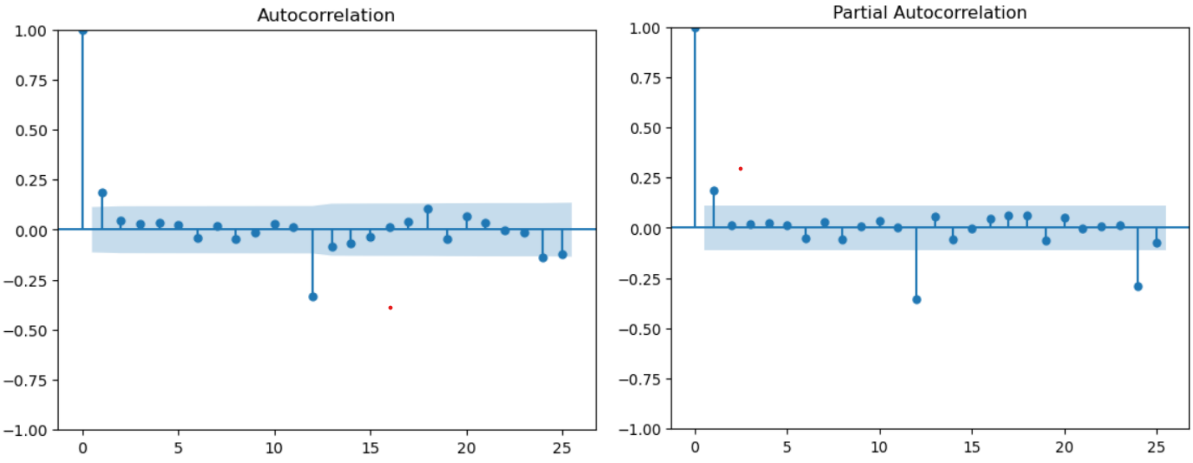
d'autres termes, il existe des preuves statistiquement significatives que les données d'inflation sont stationnaires.

•Nombre de retards : 11 :Cela indique que le test a utilisé 11 retards (valeurs passées) pour tenir compte d'une éventuelle autocorrélation dans vos données. Le nombre approprié de retards peut être déterminé par différentes méthodes, et 11 pourrait être le choix optimal dans ce cas.

5-Application de la méthode Box et Jenkin :

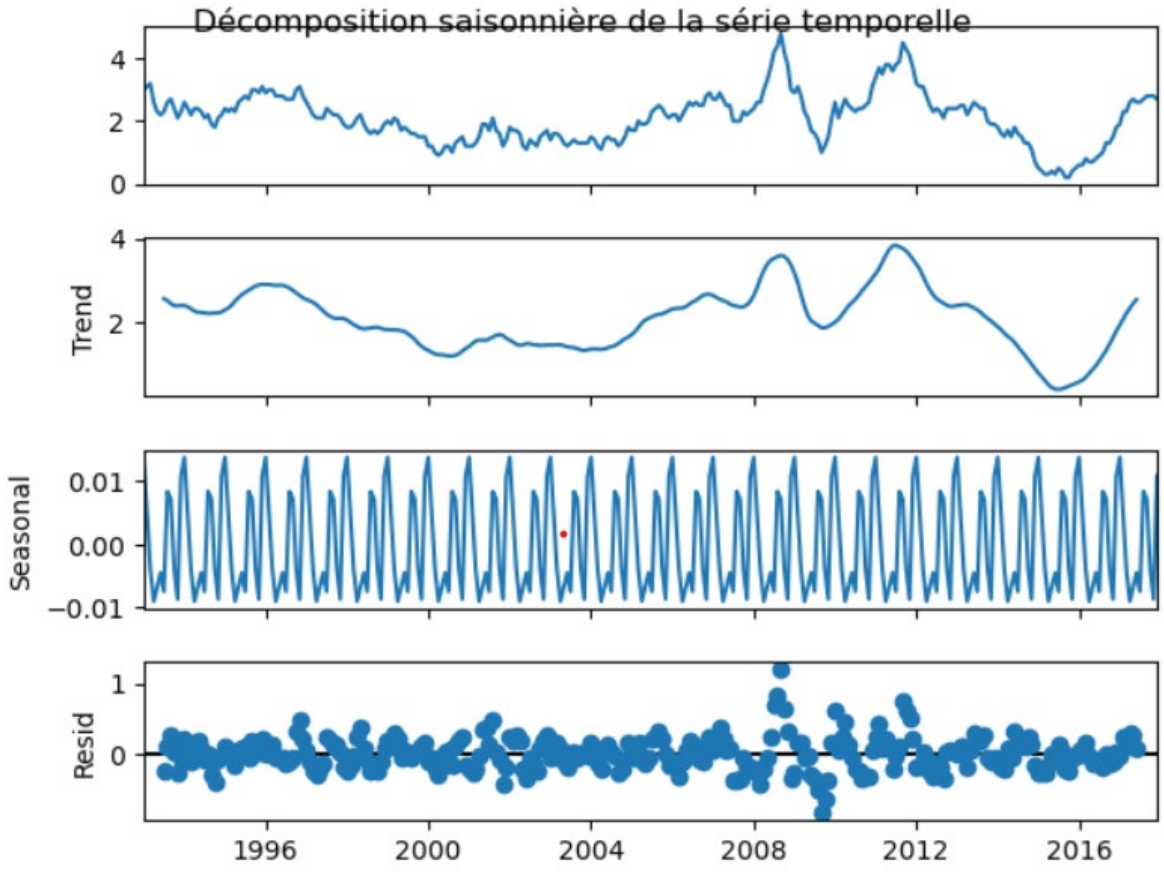
5-1 Autoccorelation :

Différenciation appliquée: pour stabiliser les données temporelles initialement instables.



Tracé ACF — meilleur décalage pour le paramètre MA, q= 2

Tracé PACF — meilleur décalage pour le paramètre AR, p = 2



Le panneau supérieur montre le taux inflation d'origine . Il fluctue beaucoup au fil du temps, avec plusieurs récessions et reprises évidentes.

Le panneau du milieu montre la composante de tendance .La composante tendancielle est une évolution douce et à long terme du taux de inflation. Il montre que le taux de inflation a globalement diminué au fil du temps, même s'il y a eu des hauts et des bas à court terme.

Le panneau inférieur montre la composante saisonnière .La composante saisonnière est la partie du taux de inflation qui est due à des facteurs saisonniers, Comme vous pouvez le constater, la composante saisonnière est assez régulière, le taux de inflation ayant tendance à être plus élevé en été.et plus faible en hiver .La somme des composantes tendancielles, saisonnières et irrégulières est égale au taux inflation initial.

La variance du résidu peut également être considérée comme constante.

5-2 Prévisions à l'aide du modèle ARIMA

Pour la modélisation à l'aide d'ARIMA, nous utilisons la fonction de différence et de journalisation pour prévoir le modèle saisonnier d'ARIMA. Le modèle utilise p, d, q (2, 1, 2) .

```
==============================================================================
Dep. Variable:                      y   No. Observations:                  300
Model:                 ARIMA(2, 1, 2)   Log Likelihood                42.669
Date:                Sun, 12 May 2024   AIC                            -75.338
Time:                        21:02:26   BIC                            -56.835
Sample:                    01-01-1993   HQIC                           -67.932
                         - 12-01-2017
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.5035      0.188      7.984      0.000       1.134       1.873
ar.L2         -0.5304      0.185     -2.873      0.004      -0.892      -0.169
ma.L1         -1.3568      0.401     -3.387      0.001      -2.142      -0.572
ma.L2          0.3571      0.233      1.534      0.125      -0.099       0.813
sigma2         0.0437      0.016      2.743      0.006       0.012       0.075
===================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):                 0.69
Prob(Q):                              0.86   Prob(JB):                         0.71
Heteroskedasticity (H):               1.42   Skew:                            -0.05
Prob(H) (two-sided):                  0.08   Kurtosis:                         3.21
===================================================================================
```

Ce résultat montre les résultats de l'ajustement d'un modèle SARIMAX pour prédire les changements du taux d'Inflation aux UK à l'aide de données de 1993 à 2017. Voici un aperçu des points clés :

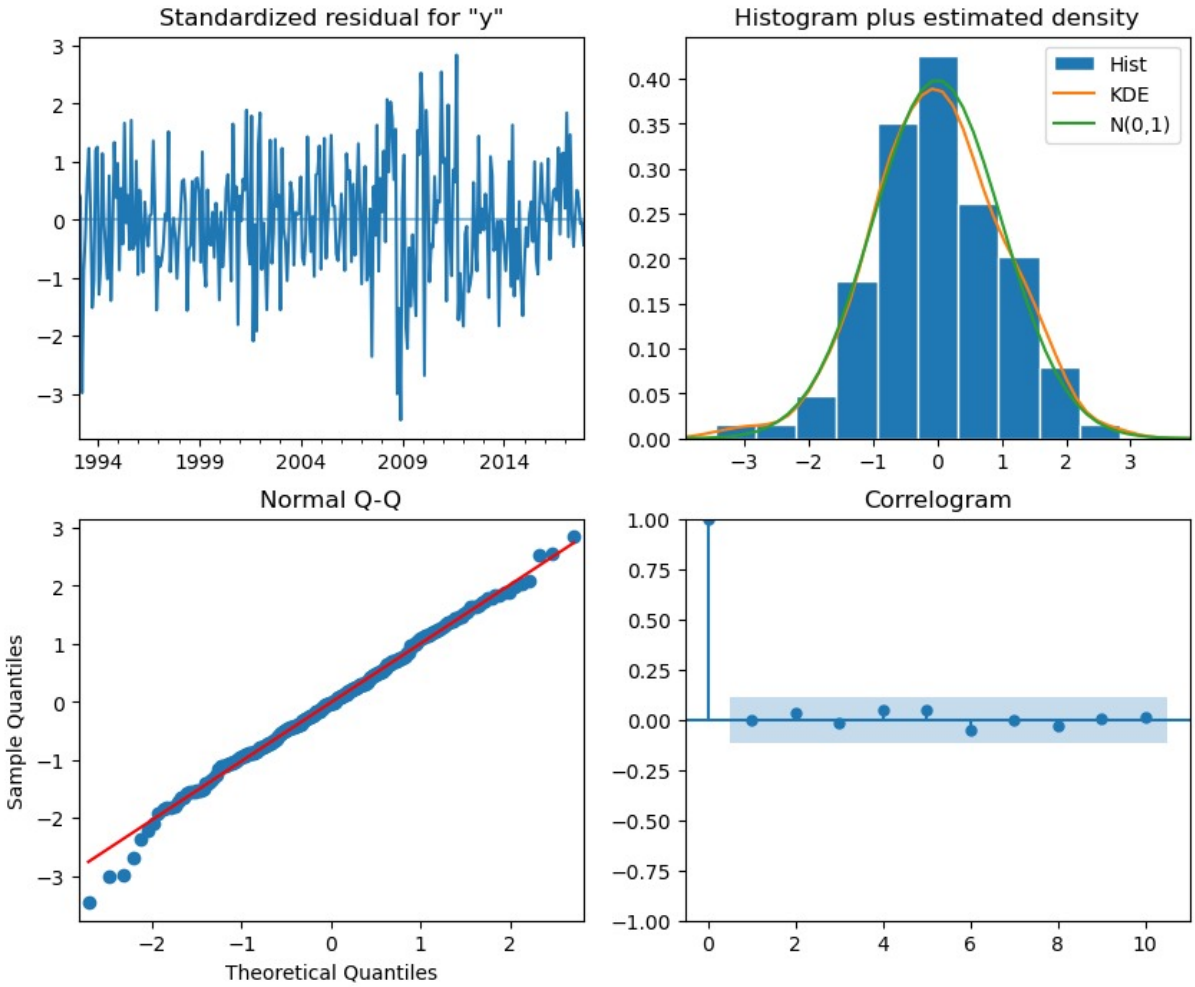
- **Modèle:** ARIMA(2, 1, 2) - Cela indique que le modèle utilise les deux dernières valeurs retardées de la variable dépendante (AR - AutoRegressive) pour la prévision, tient compte de la différenciation d'ordre 1 (I - Integrated) et intègre les deux derniers termes d'erreur (MA - Moving Average) dans le modèle.
- **Log vraisemblance:** 42.669 - Une valeur plus élevée indique un meilleur ajustement du modèle aux données.
- **AIC, BIC, HQIC:** -75.338, -56.835, -67.932 - Ces critères d'information permettent de comparer des modèles ARIMA différents. Généralement, des valeurs plus négatives indiquent un meilleur ajustement.
- **Période d'échantillonnage:** 01-01-1993 - 12-01-2017 - Ceci indique la période utilisée pour ajuster le modèle.
- **Type de matrice de covariance:** opg - Ceci concerne la méthode utilisée pour estimer la variance des erreurs du modèle.

Coefficients du modèle

- Le tableau présente les coefficients estimés pour les termes AR (ar.L1 et ar.L2), MA (ma.L1 et ma.L2) et la variance de l'erreur (sigma2).
- **coef** : valeur estimée du coefficient
- **std err** : erreur standard du coefficient - mesure la précision de l'estimation du coefficient.
- **z** : statistique de test - utilisée pour évaluer la signification statistique du coefficient (un coefficient proche de 0 aura un z faible).
- **P>|z|** : valeur de p associée au test z - indique la probabilité d'obtenir un z aussi extrême par hasard si le coefficient était vraiment nul. Une valeur de p inférieure à 0.05 indique que le coefficient est statistiquement significatif.
- **[0.025 0.975]** : intervalle de confiance à 95% du coefficient - il permet d'estimer la plage de valeurs dans laquelle se situe le véritable coefficient avec une probabilité de 95%.

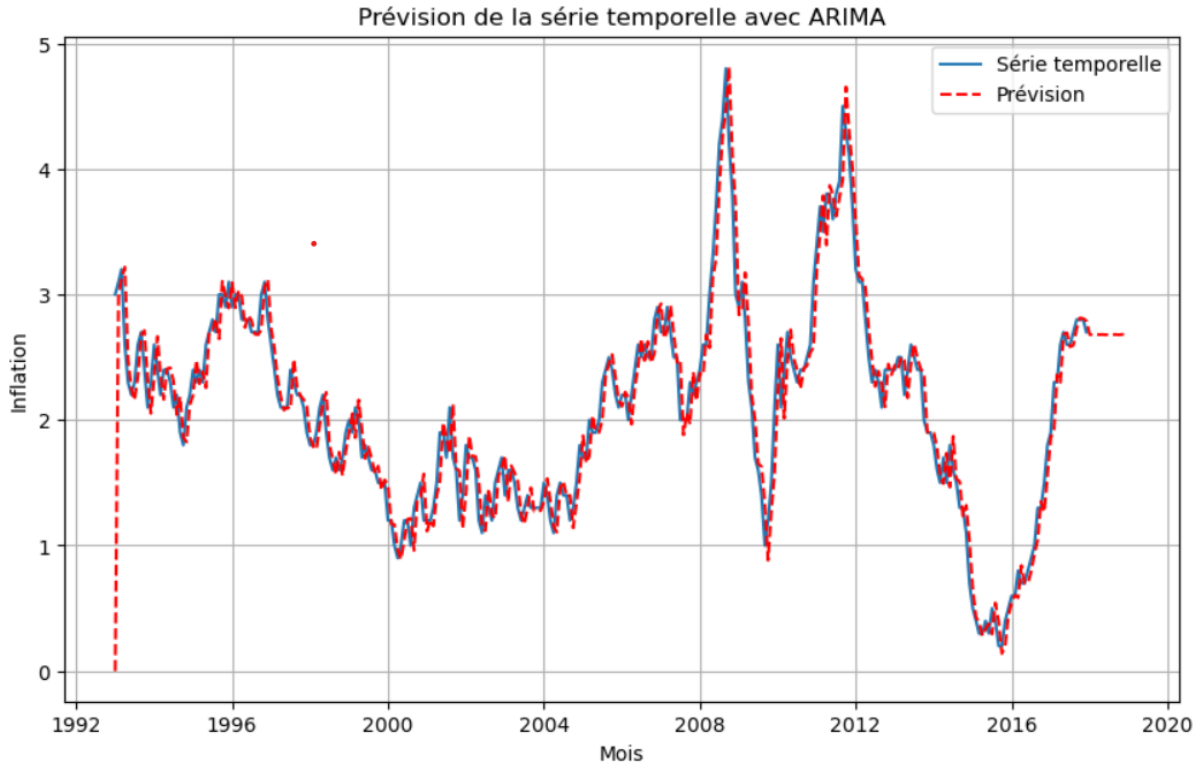
Tests de diagnostic

- **Ljung-Box (L1) et Jarque-Bera (JB)** : ces tests évaluent si les résidus du modèle (l'erreur entre les valeurs réelles et prédites) présentent une autocorrélation et une distribution normale, respectivement.
- **Prob(Q) et Prob(JB)** : valeurs de p associées aux tests. Une valeur de p supérieure à 0.05 indique que l'hypothèse nulle (pas d'autocorrélation ou distribution normale) ne peut pas être rejetée.
- **Heteroskedasticité (H)** : ce test vérifie si la variance des erreurs du modèle est constante au fil du temps.
- **Prob(H)** : valeur de p associée au test d'hétéroscédasticité. Une valeur de p supérieure à 0.05 indique que l'hétéroscédasticité n'est pas un problème.
- **Skew et Kurtosis** : ces mesures décrivent l'asymétrie et l'aplatissement de la distribution des résidus du modèle.



-Le résultat est une distribution de probabilité très similaire à une distribution gaussienne théorique. Les valeurs d'aplatissement et d'asymétrie se situent dans les paramètres optimaux. **Dans le cas où ces valeurs seraient en dehors des intervalles, nous pourrions effectuer une transformation box-cox sur la variable étudiée pour améliorer à la fois la forme des queues et la symétrie de la distribution.**

-Pour l'homocédasticité, l'un des moyens d'étude les plus utiles consiste à tracer un nuage de points sur notre domaine associé et à observer le comportement de la variance par rapport à la moyenne sur toute la série chronologique.




```
Erreur quadratique moyenne (RMSE) : 0.270360219892794
Erreur absolue moyenne (MAE) : 0.1733041206405144
```

Dans le processus d'évaluation des performances du modèle, nous avons calculé l'erreur quadratique moyenne (RMSE) et erreur absolue moyenne(MAE) comme mesure permettant de quantifier l'exactitude des prédictions du modèle. L'erreur RMSE calculée pour les valeurs prédites s'est avérée être de 0,2703. Et l'erreur MAE calculée pour les valeurs prédites est 0,1733 .

Le RMSE est une mesure statistique qui évalue l'ampleur moyenne des différences entre les valeurs prédites et observées. Dans le contexte de la prévision de séries chronologiques, une valeur RMSE inférieure signifie un meilleur ajustement du modèle aux données réelles.

6-conclusion :

Ce travail a permis d'appliquer la méthode de Box et Jenkins pour modéliser une série chronologique choisie. À travers les différentes étapes, nous avons pu mettre en œuvre les principaux éléments de cette méthodologie reconnue pour son efficacité dans la modélisation des séries temporelles.

Dans un premier temps, nous avons récupéré les données nécessaires à partir des langages Python. Ensuite, nous avons divisé ces données en un échantillon d'apprentissage et un échantillon de validation, afin de construire notre modèle sur l'échantillon d'apprentissage et de le valider sur l'échantillon de validation.

En créant un objet de type série temporelle et en représentant graphiquement la série, nous avons pu effectuer une analyse qualitative pour repérer d'éventuelles tendances et/ou saisonnalités. Cette étape préliminaire était cruciale pour comprendre la nature de la série et orienter nos choix de modélisation.

À travers l'analyse des corrélogrammes simple et partiel, nous avons examiné les autocorrélations et les autocorrélations partielles de la série pour différents décalages. Cette analyse nous a permis de sélectionner les ordres des modèles ARIMA à appliquer.

Ensuite, nous avons appliqué la méthode de Box et Jenkins pour estimer les paramètres des modèles ARIMA. Cette méthode itérative nous a conduits à sélectionner le modèle ARIMA le mieux adapté à notre série.

La représentation graphique de la série ainsi que de la prévision nous a permis de visualiser la performance de notre modèle. Enfin, nous avons évalué les prévisions obtenues à l'aide de différentes mesures d'erreur et interprété les résultats pour conclure sur la pertinence de notre modèle.