Rapport de projet

Séries Chronologiques : Prétraitement, analyse descriptive, modèle prédictif et simulation d'un jeu de données

Abdoulaye TRAORE

21 February - 07 March 2025

Professeur principal :
M. Yannig GOUDE
Associate Professor, Laboratoire de
Mathematiques, Universite Paris-Sud

Table des matières

Introduction

Ce projet s'inscrit dans le cadre des cours de séries chronologiques APM_4STA2_TA dispensé à ENSTA Paris.

L'objectif de cette étude est d'analyser les tendances du prix de l'or entre 2010 et 2024 et de développer un modèle prédictif basé sur des méthodes de séries temporelles, notamment les techniques de lissages exponentiels.

Pour cela, nous utiliserons un jeu de données provenant de Kaggle contenant des informations financières sur plusieurs actifs, dont l'or.
L'analyse se concentrera exclusivement sur cette matière première.
Pour plus d'informations, consultez le jeu de données sur Kaggle : Financial Data on Kaggle.

La méthodologie suivra plusieurs étapes :

- Description de la source suivi d'un prétraitement des données pour corriger les valeurs manquantes et harmoniser leur fréquence.
- Analyse descriptive des données pour identifier les tendances et caractéristiques principales du prix de l'or.
- La modélisation s'appuiera sur des méthodes de séries temporelles, notamment le lissage exponentiel, dans le but de fournir un modèle prédictif : étude de la décomposition en tendance, cycles, partie aléatoire, étude de la stationnarité.
- Enfin, des prévisions seront réalisées sur un échantillon test afin de mesurer la pertinence du modèle proposé.

Ce rapport couvre toutes les étapes, de la préparation des données à l'évaluation des résultats, en mettant l'accent sur la rigueur méthodologique.

1 Prétraitement des Données et Description de la source

1.1 Source et Description des Données

Les données utilisées dans cette étude proviennent d'un jeu de données financier en accès libre sur la plateforme Kaggle [1].
Ce dataset, publié en Octobre 2024 et actualisé en Décembre 2024, contient des informations sur divers indices boursiers, matières premières, indicateurs économiques et taux de change.

Nous nous intéressons exclusivement au prix de l'or, historiquement perçu comme une réserve de valeur, protégeant la richesse sur de longues périodes, notamment en période d'inflation, de dévaluation monétaire ou d'instabilité économique.

Le dataset couvre une période allant de 2010 à 2024 et contient des données de marché représentant les variations du prix de l'or sous forme de séries temporelles. Il comprend plusieurs indicateurs financiers relatifs à l'or, tels que :

— **gold_open** : Prix d'ouverture de l'or pour une journée donnée.
— **gold_high** : Prix le plus élevé atteint dans la journée.
— **gold_low** : Prix le plus bas enregistré dans la journée.
— **gold_close** : Prix de clôture de l'or en fin de journée.
— **gold_volume** : Volume d'échanges de l'or sur la période donnée.

date	sp500 open	sp500 high	sp500 low	sp500 close	sp500 volume	sp500 high-low
2010-01-14	114.49	115.14	114.42	114.93	115646960.0	0.7199999999999989
2010-01-15	114.73	114.84	113.2	113.64	212252769.0	1.6400000000000006
2010-01-18						
2010-01-19	113.62	115.13	113.59	115.06	138671890.0	1.539999999999992
2010-01-20	114.28	114.45	112.98	113.89	216330645.0	1.4699999999999989

FIGURE 1.1 – Présentation d'un échantillon du dataset pour sp500

Il est important de noter que les données financières sont enregistrées à différentes granularités temporelles (les jours d'enregistrement sont quelques peu irréguliers).
Cette disparité peut engendrer des incohérences et des valeurs manquantes qu'il est nécessaire de traiter en amont pour garantir une analyse fiable.

1.2 Prétraitement des Données

Afin d'assurer la qualité des données avant toute analyse, plusieurs étapes de prétraitement ont été mises en œuvre dont :

1. https://www.kaggle.com/datasets/franciscogcc/financial-data

— **Suppression des valeurs manquantes** : Certaines entrées du dataset contiennent des valeurs absentes (`NaN`). Les observations correspondantes ont été identifiées et supprimées : 3904 observations à 3719 après traitement.

— **Conversion de la colonne date** : Les dates ont été converties en format `Date` afin de permettre une manipulation efficace et une bonne structuration des séries temporelles.

Ces étapes garantissent la cohérence et la fiabilité des données, permettant ainsi une modélisation plus robuste.

1.3 Mise en forme des Données

Après le prétraitement, les données ont été structurées sous la forme d'une table comportant les colonnes suivantes :

— `date` : Date de l'enregistrement.

— `gold_open` : Prix d'ouverture de l'or.

— `gold_high` : Prix maximal atteint dans la journée.

— `gold_low` : Prix minimal enregistré dans la journée.

— `gold_close` : Prix de clôture de l'or en fin de journée.

— `gold_volume` : Volume des transactions sur l'or.

Les données ont été organisées sous forme de série temporelle avec une fréquence adaptée aux jours de trading (252 jours par an en moyenne selon les recherches).
Ce format facilite les analyses de tendance et la modélisation prédictive.

2 Analyse descriptive des données

2.1 Statistiques descriptives globales

L'ensemble de données étudié couvre la période allant de **janvier 2010 à octobre 2024**. Les statistiques descriptives globales sont les suivantes :

— **Prix minimum** : 100.5 USD

— **Prix maximum** : 253.9 USD

— **Prix moyen** : 145.5 USD

— **Prix médian** : 137.7 USD

— **Volume moyen échangé** : 9 658 138

Ces valeurs indiquent une **forte croissance** du prix de l'or sur la période analysée, avec des fluctuations visibles.

```
> head(financial_data)  # Voir les premières lignes du dataset
        date gold_open gold_high gold_low gold_close gold_volume
1 2010-01-14    111.51    112.37   110.79     112.03    18305238
2 2010-01-15    111.35    112.01   110.38     110.86    18000724
4 2010-01-19    110.95    111.75   110.83     111.52    10467927
5 2010-01-20    109.97    110.05   108.46     108.94    17534231
6 2010-01-21    108.48    108.78   106.61     107.37    25747831
7 2010-01-22    106.93    107.68   106.01     107.17    24209966
```

Structure et les premières valeurs du jeu de données

```
> str(financial_data)  # Structure du dataset après suppression
'data.frame':   3719 obs. of  6 variables:
 $ date       : Date, format: "2010-01-14" "2010-01-15" "2010-01-19" "2010-01-20" ..
 $ gold_open  : num  112 111 111 110 108 ...
 $ gold_high  : num  112 112 112 110 109 ...
 $ gold_low   : num  111 110 111 108 107 ...
 $ gold_close : num  112 111 112 109 107 ...
 $ gold_volume: num  18305238 18000724 10467927 17534231 25747831 ...
 - attr(*, "na.action")= 'omit' Named int [1:185] 3 23 57 78 99 124 144 170 228 249
  ..- attr(*, "names")= chr [1:185] "3" "23" "57" "78" ...
```

Structure interne du dataframe

```
> summary(financial_data)  # Statistiques descriptives générales
      date              gold_open       gold_high       gold_low       gold_close     gold_volume
 Min.   :2010-01-14   Min.   :100.9   Min.   :101.0   Min.   :100.2   Min.   :100.5   Min.   : 1436508
 1st Qu.:2013-09-24   1st Qu.:120.6   1st Qu.:121.0   1st Qu.:120.2   1st Qu.:120.6   1st Qu.: 5795310
 Median :2017-06-05   Median :137.6   Median :138.1   Median :137.0   Median :137.7   Median : 8087993
 Mean   :2017-06-03   Mean   :145.5   Mean   :146.1   Mean   :144.8   Mean   :145.5   Mean   : 9658138
 3rd Qu.:2021-02-11   3rd Qu.:167.8   3rd Qu.:168.4   3rd Qu.:167.1   3rd Qu.:167.8   3rd Qu.:11567291
 Max.   :2024-10-23   Max.   :253.1   Max.   :253.9   Max.   :252.5   Max.   :253.9   Max.   :93698108
>
```

Résumé statistique des variables numériques

FIGURE 2.1 – Statistiques descriptives globales

2.2 Visualisation des données

2.2.1 Évolution du prix de clôture de l'or

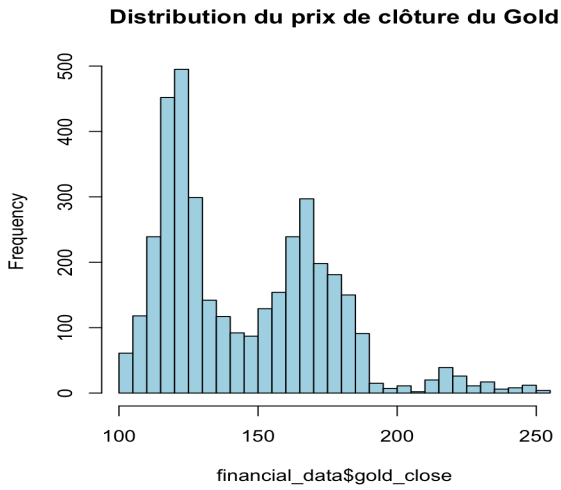
La figure ci-dessous figure 2.2 montre l'évolution du prix de clôture de l'or entre **2010 et 2024**. On observe une saisonnalité des données ainsi qu'une tendance générale haussière, bien que certaines baisses soient visibles, probablement en lien avec des événements économiques majeurs.

FIGURE 2.2 – Évolution du prix de clôture du Gold (2010-2024)

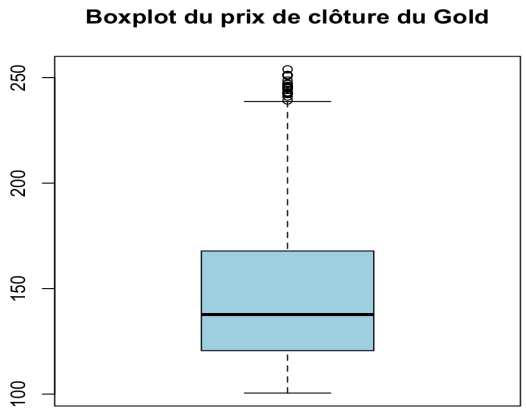
2.2.2 Distribution des prix de clôture

L'histogramme du prix de clôture (figure 2.3.1) représente la distribution des prix de clôture. La majorité des prix se situent entre **120 et 170 USD**, avec quelques occurrences de prix extrêmes.

Le boxplot du prix de clôture (figure 2.3.2) quant à lui permet de détecter des valeurs aberrantes. Certains points hors des moustaches suggèrent des variations inhabituelles du prix, probablement causées par des périodes de crises économiques ou de forte spéculation.



Distribution des prix de clôture du Gold

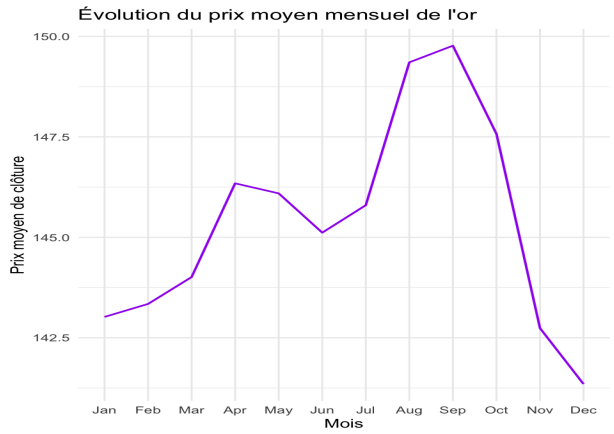


Boxplot du prix de clôture du Gold

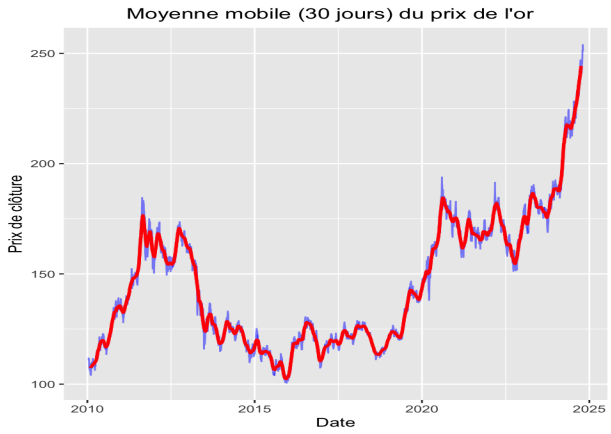
FIGURE 2.3 – Distribution des prix de clôture

2.2.3 Tendance mensuelle des prix de l'or et saisonnalités

L'analyse de la saisonnalité du prix de l'or révèle une hausse progressive durant le premier quadrimestre (143.125 → 146.25 $), probablement due aux achats post-fêtes et à la reprise économique. Une baisse entre avril et juin s'explique par une stabilisation des marchés et des hausses de taux d'intérêt.
Ensuite, une forte augmentation jusqu'en septembre (150 $) est stimulée par les incertitudes économiques estivales.
Enfin, une chute progressive mène à un minimum en décembre (140.625 $), influencée par les prises de bénéfices, la conversion des actifs en liquidités et la demande accrue en dollars en fin d'année.



Évolution moyenne des prix de l'or par mois



Moyenne mobile (30 jours) du prix de l'or

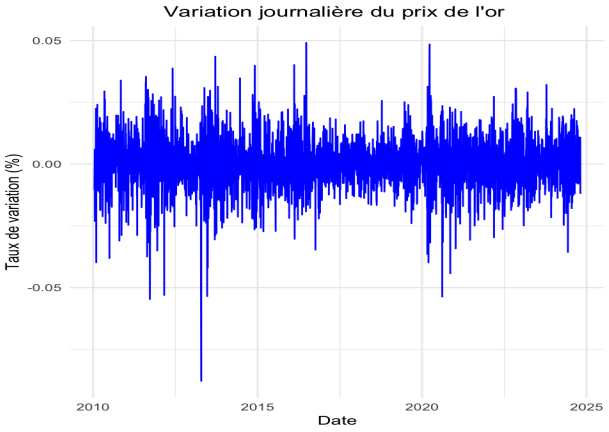
FIGURE 2.4 – Tendance mensuelle des prix de l'or

L'application d'une moyenne mobile sur 30 jours permet d'éliminer les fluctuations court-terme et de dégager une tendance sous-jacente. Comme souligné précédemment, les tendances haussières et baissières montrent que le prix de l'or réagit aux événements économiques et financiers globaux.

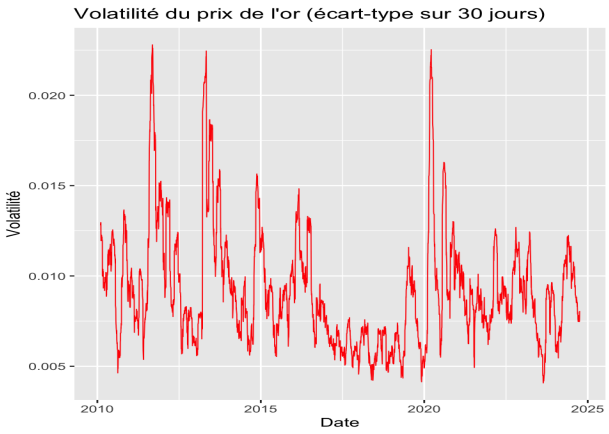
2.3 Analyse de la volatilité

Le graphique ci-dessous (figure 2.5.1) illustre les variations journalières du prix de l'or en pourcentage entre 2010 et 2025.
On observe des fluctuations constantes, traduisant la nature volatile du marché de l'or. Des pics extrêmes apparaissent, souvent en réaction à des événements économiques majeurs (crises financières, décisions des banques centrales, tensions géopolitiques). Après 2015, l'intensité des variations semble plus stable, bien que certaines périodes, comme 2020, révèlent une volatilité accrue, probablement liée à la crise du COVID-19 et aux incertitudes des marchés.

Le graphique suivant (figure 2.5.2) représente l'évolution de la volatilité du prix de l'or, mesurée par l'écart-type des variations sur une période de 30 observations.
On note plusieurs pics marqués de volatilité, notamment entre 2011-2013 et 2020. Après 2015, la volatilité a tendance à diminuer, indiquant une stabilisation relative du marché sauf en 2020-2021 en réponse probablement à la pandémie de Covid-19.



Variation journalière du prix de l'or



Volatilité du prix de l'or (écart-type sur 30 jours)

FIGURE 2.5 – Analyse de la volatilité

3 Modélisation et Prévision

3.1 Analyse de la Stationnarité : Test de Dickey-Fuller

Le test de Dickey-Fuller a révélé que la série temporelle n'était pas stationnaire avec une p-value de 0.9896. Une différenciation d'ordre 1 a été appliquée pour rendre la série stationnaire avec une p-value de 0.01.

```
        Augmented Dickey-Fuller Test

data:  financial_data$gold_close
Dickey-Fuller = -0.32848, Lag order = 15, p-value = 0.9896
alternative hypothesis: stationary
```

Test de stationnarité avant différenciation

```
        Augmented Dickey-Fuller Test

data:  na.omit(financial_data$gold_close_diff)
Dickey-Fuller = -16.068, Lag order = 15, p-value = 0.01
alternative hypothesis: stationary
```

Test de stationnarité après différenciation

FIGURE 3.1 – Analyse de la stationnarité

3.2 Décomposition de la Série Temporelle

La figure ci-dessous présente la décomposition de la série étudiée permettant d'analyser ses différentes composantes : **tendance** (évolution à long terme), **saisonnalité** (fluctuations périodiques) et **composante aléatoire** (variabilité imprévisible).

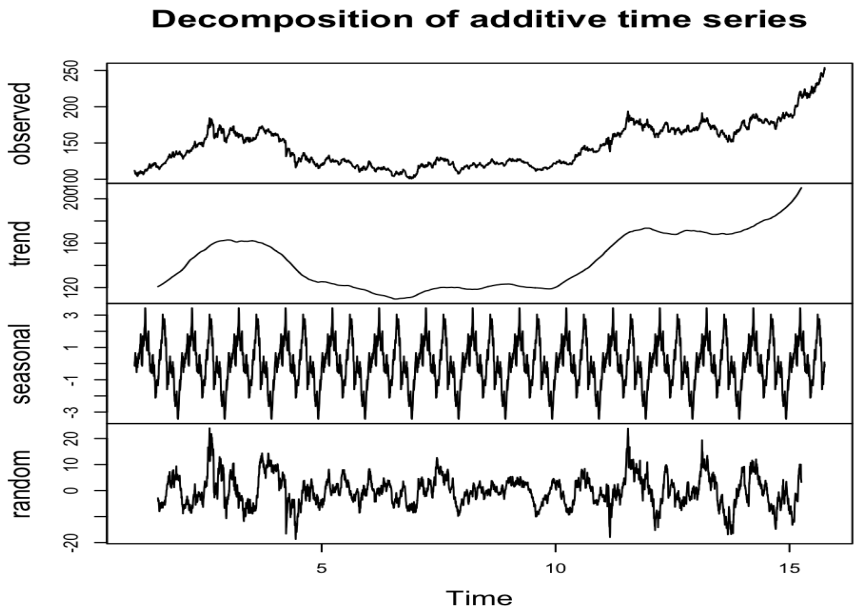


FIGURE 3.2 – Décomposition de la série temporelle

Série observée

La première courbe représente la **série temporelle originale** déjà présentée dans les sections précédentes.

Tendance (trend)

La seconde courbe met en évidence la tendance sous-jacente, révélant trois phases distinctes : Une hausse initiale suivie d'un plateau, une période de stagnation, déjà identifiée dans l'évolution mensuelle des prix (2.2.3) ainsi qu'une forte reprise en fin de période, en cohérence avec les pics identifiés dans l'analyse saisonnière.

Ces observations confirment l'idée que le prix de l'or suit des cycles économiques marqués avec une **tendance générale à la hausse**.

Saisonnalité (seasonal)

La composante saisonnière montre une régularité dans les variations, indiquant des effets récurrents : une saisonnalité **additive**. Comme observé précédemment dans l'analyse de la saisonnalité (2.2.3), les hausses marquées en fin d'été et les baisses en fin d'année sont bien capturées ici.
Cela souligne l'importance d'intégrer la saisonnalité dans un modèle prédictif.

Composante aléatoire (random)

Enfin, la dernière courbe représente la composante aléatoire, qui regroupe les variations imprévisibles. On remarque une **volatilité accrue** à certaines périodes, notamment en début et fin de série, ce qui correspond aux moments d'incertitude observés en 2.3.

3.3 Modèles de Lissages Exponentiels

Le modèle de lissage exponentiel retenu est **la méthode de Holt-Winters** car :

— L'amplitude des variations saisonnières ne dépend pas du niveau de la série comme vu dans la décomposition de la série temporelle (additive) ,

— La courbe de tendance dans la décomposition montre une évolution non constante du prix de l'or, avec des phases de croissance et de stagnation.

— Il existe des variations irrégulières dans la composante résiduelle.

Les méthodes de lissage exponentiel simple et double ne conviendraient donc pas.

On effectue **la modélisation et les prévisions que sur la période de 2024** comportant 205 observations avec une fréquence mensuelle fixé à 17 en ignorant les jours non-ouvrables, compte tenu de la granularité des observations. Cette période conserve la saisonnalité et la tendance globale observées en amont.
Les données ont été divisées avec pour la modélisation 95% et 5% pour la prévision.

4 Simulation de prévision sur un échantillon test et Interprétations

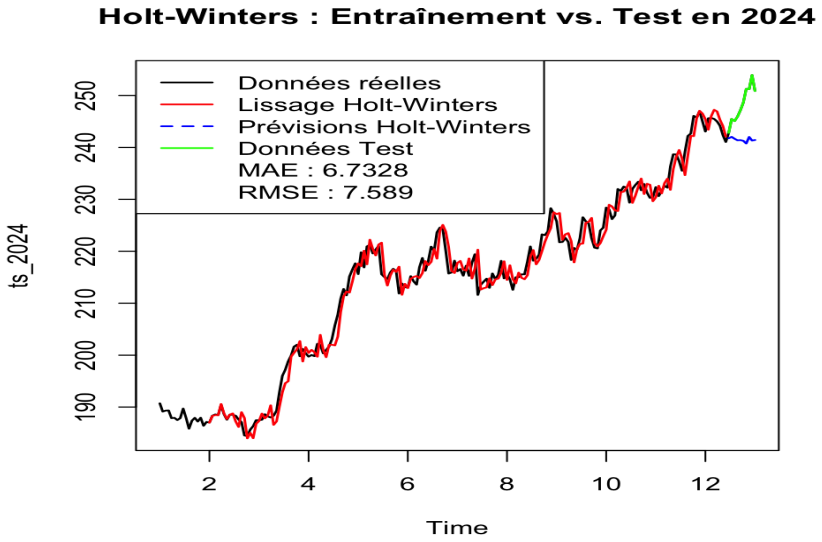


FIGURE 4.1 – Modélisation et prévision

Le lissage Holt-Winters ajuste bien les tendances et variations saisonnières des données réelles de 2024 avec toutefois une impression d'overfitting.

Les prévisions (courbe bleue) semblent légèrement sous-estimer la tendance haussière observée dans les données de test (courbe verte), ce qui indique que le modèle capte bien la dynamique globale mais pourrait manquer d'adaptabilité face à des fluctuations récentes.

Ce résultat n'est toutefois pas surprenant car la prévision des matières premières volatiles comme l'or requiert des outils mathématiques avancés (Calcul stochastique, méthode de Monte-Carlo) et est l'objet même des mathématiques financières.

Conclusion

Cette étude a permis d'analyser l'évolution du prix de l'or à travers une approche rigoureuse des séries temporelles. En passant par une phase de prétraitement minutieuse, nous avons nettoyé et structuré les données afin d'assurer une modélisation fiable. L'analyse descriptive a révélé des tendances marquées ainsi qu'une saisonnalité influençant les variations du prix, confirmant l'intérêt d'un modèle intégrant ces composantes.

Les résultats ont montré que le lissage Holt-Winters offre une bonne capacité d'adaptation aux fluctuations saisonnières, bien que les prévisions restent légèrement sous-estimées face aux tendances récentes. Ce décalage pourrait être amélioré par une optimisation plus fine des paramètres de lissage et l'intégration de méthodes plus avancées adaptées aux séries financières volatiles.

Les principaux enseignements de cette étude sont les suivants :

— L'importance du **prétraitement des données** et de la structuration temporelle pour garantir des résultats fiables.

— La **nécessité de prendre en compte la saisonnalité** dans la modélisation des séries temporelles financières.

— La **robustesse du modèle Holt-Winters** dans la capture des dynamiques de tendance et de saisonnalité.

— Les **limitations des modèles classiques** face aux variations brutales du prix de l'or, nécessitant des approches plus sophistiquées.

Cependant, cette étude présente certaines limitations, notamment dans la prise en compte des évènements macroéconomiques influençant le marché de l'or. Une perspective future consisterait à explorer des modèles hybrides intégrant des techniques plus avancées, comme les modèles ARIMA ou les modèles basés sur l'apprentissage automatique et les réseaux de neurones.

En définitive, cette analyse a permis d'apporter un regard approfondi sur les dynamiques du marché de l'or et a démontré l'intérêt des modèles de lissage exponentiel pour la prévision des tendances. Les pistes d'amélioration identifiées ouvrent la voie à des approches plus performantes, essentielles pour affiner les stratégies d'anticipation des fluctuations du prix de l'or.