ÉCONOMIE

Analyse de l'inflation

Une étude exhaustive de trois séries de l'IPC à l'aide du code R

Dans une réalisation exceptionnelle, Ruthvik, un élève de douzième année au DPSBN (Delhi Public School Bangalore North), a réalisé l'exploit remarquable de voir son document de recherche en économie publié dans le prestigieux Journal « Exploratio » aux États-Unis. Sous la direction experte d'un éminent professeur du MIT, Ruthvik a exploité la puissance des techniques de modélisation de séries chronologiques en utilisant le langage de programmation R pour approfondir les complexités des données de séries chronologiques économiques, en se concentrant particulièrement sur les indices des prix à la consommation (IPC), notamment CPIAUCSL, CPILFENS, et CPILFESL. Avec un dévouement sans faille, il a méticuleusement appliqué une approche analytique cohérente à toutes les séries CP, facilitant une comparaison robuste de la précision des prévisions des modèles simples et ARIMA. Les mesures d'évaluation, notamment l'erreur absolue moyenne (MAE) et l'erreur quadratique moyenne (RMSE), ont été habilement utilisées pour évaluer les performances du modèle. Les recherches de Ruthvik ne sont pas seulement un témoignage de ses prouesses scientifiques, mais aussi une source d'inspiration pour les étudiants et les universitaires. Son travail révolutionnaire fait progresser considérablement le domaine de l'analyse et des prévisions économiques. Ce qui justifie le relai ici de la publication

PAR RUTHVIK REDDY BOMMIREDDY, 15 OCTOBRE 2023

LEGENDE **Auteur :** Mentor de Ruthvik Reddy Bommireddy : Dr. Peter Kempthorne
Delhi Public School Bangalore North

Abstrait

Les séries chronologiques économiques sont considérées comme les baromètres économiques de la santé d'une économie. Ils fournissent des informations sur la performance et la stabilité des principaux indicateurs économiques, tels que l'inflation. Par conséquent, il est sage d'essayer de prévoir ces variables.

Dans cet article, nous mettons en œuvre des techniques de modélisation de séries temporelles à l'aide du langage de programmation R pour analyser les séries temporelles économiques liées à l'inflation et aux taux d'intérêt. Plus précisément, nous appliquons des modèles simples et ARIMA pour prévoir l'indice des prix à la consommation (IPC).

Pour cibler notre analyse, nous appliquerons notre méthode d'analyse à une série de l'IPC, puis utiliserons la même méthode pour les autres séries de l'IPC. Cela nous permettra de comparer l'exactitude de nos modèles dans différentes séries de l'IPC et d'identifier des modèles ou des tendances dans les données.

L'objectif de cette analyse est de comparer la précision des modèles simples et ARIMA pour la prévision de l'IPC. Nous utiliserons des mesures d'évaluation telles que l'erreur absolue moyenne (EMA) et l'erreur quadratique moyenne (RMSE) pour comparer les performances des deux modèles. Notre analyse permettra de déterminer quelle technique de modélisation est la mieux adaptée à la prévision de ces variables économiques.

1. Séries chronologiques historiques de l'inflation aux États-Unis

1.1 Vue d'ensemble des séries

Notre recherche analyse les séries chronologiques sur les prix à la consommation aux États-Unis maintenues par la Réserve fédérale de St. Louis. Plus précisément, nous analysons trois séries :

- Indice des prix à la consommation pour tous les consommateurs urbains : tous les articles de la moyenne urbaine des États-Unis (CPIAUCSL)
- Indice des prix à la consommation pour l'ensemble des consommateurs urbains : tous les articles moins les aliments et l'énergie dans la moyenne des villes américaines (CPILFESL)
- Indice des prix à la consommation pour l'ensemble des consommateurs urbains : tous les articles moins la nourriture et l'énergie dans la moyenne urbaine des États-Unis (CPILFENS)

Le CPILFENS et le CPIFESL sont tous deux des séries de l'IPC produites par le BLS, mais ils diffèrent en termes de couverture géographique, d'étendue de la couverture et de méthodologie. De plus, le CPILFESL est désaisonnalisé et le CPILFENS ne l'est pas. Il s'agit de séries chronologiques mensuelles allant de janvier 2011 à novembre 2022. La section suivante fournit une description des quatre séries.

1.2 Descriptions des séries

Référence : https://fred.stlouisfed.org/

1.2.1 Description de la CPIAUCSL

Source : U.S. Bureau of Labor Statistics
Unités :Indice 1982-1984=100, Fréquence corrigée des variations saisonnières : Mensuelle

CPIAUCSL est l'abréviation de Consumer Price Index for All Urban Consumers : All Items. Il s'agit d'une mesure de la variation moyenne dans le temps des prix payés par les consommateurs urbains pour un panier de biens et de services, y compris la nourriture, le logement, les vêtements, le transport et les soins médicaux, entre autres. Le CPIAUCSL est produit par le Bureau of Labor Statistics des États-Unis et est utilisé comme indicateur de l'inflation dans l'économie américaine.

1.32 Description du CPILFESL

Source : Communiqué du Bureau of Labor Statistics des États-Unis : Indice des prix à la consommation

Unités : Indice 1982-1984=100, Fréquence désaisonnalisée : Mensuelle

CPILFESL est l'abréviation de Consumer Price Index for All Urban Consumers : All Items Less Food and Energy (indice des prix à la consommation pour tous les consommateurs urbains : tous les articles moins de nourriture et d'énergie). Il s'agit d'une mesure de l'évolution moyenne dans le temps des prix payés par les consommateurs urbains pour un panier de biens et services, à l'exclusion des prix de l'alimentation et de l'énergie. Cette mesure est utilisée pour évaluer la tendance sous-jacente de l'inflation dans l'économie américaine, car les prix des denrées alimentaires et de l'énergie peuvent être très volatils et sujets à des changements saisonniers qui peuvent obscurcir la tendance globale de l'inflation. Le CPIFESL se concentre sur une gamme étroite de biens et de services, y compris l'alimentation, le logement, l'habillement et le transport.

1.35 Description des CPILFENS

Source : Communiqué du Bureau of Labor Statistics des États-Unis : Indice des prix à la consommation

Unités : Indice 1982-1984=100, non désaisonnalisé Fréquence : Mensuel

L'indice des prix à la consommation pour tous les consommateurs urbains : tous les articles moins l'alimentation et l'énergie est un agrégat des prix payés par les consommateurs urbains pour un panier type de biens, à l'exclusion de l'alimentation et de l'énergie. Cette mesure, connue sous le nom d'« IPC de base », est largement utilisée par les économistes parce que les prix de l'alimentation et de l'énergie sont très volatils. CPILFENS couvre un large éventail de biens et de services, y compris l'alimentation, le logement, l'habillement, le transport et les soins médicaux.

1.4 Extraire la liste des symboles

Pour mener notre analyse, nous avons utilisé un objet de série chronologique R nommé « economic_data2 ». Cette série chronologique trimestrielle s'étend du 2011-01-01 au 2022-11-30 et contient plusieurs colonnes représentant les trois séries de l'indice des prix à la consommation (IPC) : CPILFENS (indice des prix à la consommation pour tous les consommateurs urbains : tous les articles moins les aliments et l'énergie), CPIAUCSL (indice des prix à la consommation pour tous

les consommateurs urbains : tous les articles) et CPILFESL (indice des prix à la consommation pour tous les consommateurs urbains : tous les articles moins les aliments et l'énergie). Les objets R ont été créés en important des données directement à partir des données économiques de la Réserve fédérale (https : //fred.stlouisfed.org/).

Nous pouvons extraire les séries chronologiques individuelles de economic-data2 à l'aide du code suivant list_symbols <- unique (economic_data2$symbol)

2. Méthodologie

Dans ce document de recherche, nous avons étudié les propriétés des séries chronologiques de 3 séries d'IPC : CPILFENS, CPIAUCSL et CPILFESL.

Notre approche a été guidée par les principes énoncés dans le livre « Forecasting : Principles and Practice » de Rob J. Hyndman et George Athanasopoulos (Hyndman et Athanasopoulos, 2018). Pour effectuer l'analyse des données et la modélisation, nous avons utilisé la fonctionnalité fournie par le package « fpp2 » dans R. En appliquant les méthodologies décrites dans le livre de Hyndman et en mettant en œuvre le paquet « fpp2 », nous avons été en mesure de mener une analyse rigoureuse et robuste des séries chronologiques sur les données économiques.

Pour mener à bien notre étude, nous avons d'abord effectué une série d'analyses exploratoires de données en observant visuellement les diagrammes de séries chronologiques et leurs graphiques saisonniers, ainsi que la fonction d'autocorrélation (ACF). Par la suite, nous avons utilisé des modèles simples, à savoir naïfs et dérivants, pour prévoir les données. Après une enquête plus approfondie, nous avons constaté que nous pouvons rejeter l'hypothèse nulle selon laquelle les séries chronologiques sont stationnaires pour les trois séries, ce qui signifie que leurs propriétés statistiques changent au fil du temps.

Pour transformer les données en séries temporelles stationnaires, nous avons appliqué la première et la deuxième différenciation. Ensuite, nous avons appliqué la fonction auto.arima du package de prévisions dans R pour trouver le modèle le mieux adapté à chaque variable. Pour CPIAUCSL, le meilleur modèle est ARIMA(0,1,1) avec dérive, ce qui indique que la variable est influencée par sa propre valeur décalée et qu'elle a une tendance linéaire dans le temps. Pour CPILFENS, le meilleur modèle s'avère être ARIMA(2,1,0) avec dérive, ce qui indique que la variable est influencée par ses deux propres valeurs décalées et a

une tendance linéaire dans le temps. Pour CPILFESL, le meilleur modèle est ARIMA(1,2,1), ce qui indique que la variable est influencée par sa propre valeur décalée, qu'elle a une tendance quadratique au fil du temps et qu'elle est affectée par une moyenne mobile décalée d'ordre un.

Nous avons ensuite appliqué l'erreur quadratique moyenne (RMSE), l'erreur absolue moyenne (MAE) et l'erreur absolue moyenne en pourcentage (MAPE). Les résultats nous ont montré que les modèles ARIMA fournissent des prévisions pour les trois variables avec des mesures d'erreur relativement faibles.

Dans la section suivante, nous nous pencherons sur les analyses détaillées des trois séries chronologiques économiques, à savoir CPIAUCSL, CPILFENS et CPILFESL. Cette section fournit un aperçu complet des méthodologies et du code R utilisés pour générer les résultats. En incluant le code R, nous visons à permettre à nos lecteurs de reproduire les analyses présentées dans cet article ou de les appliquer à d'autres données de séries chronologiques. Nous fournissons une ressource pratique pour tous ceux qui souhaitent explorer davantage nos méthodologies ou étendre les analyses à différents ensembles de données de séries chronologiques.

Grâce à cet examen détaillé des analyses et à l'inclusion du code R, nous espérons que les connaissances acquises dans le cadre de nos recherches pourront être facilement appliquées et exploitées par l'ensemble de la communauté.

3. Analyse de l'indice des prix à la consommation pour l'ensemble des consommateurs urbains : tous les articles moins l'alimentation et l'énergie dans la moyenne des villes américaines (CPILFESL)

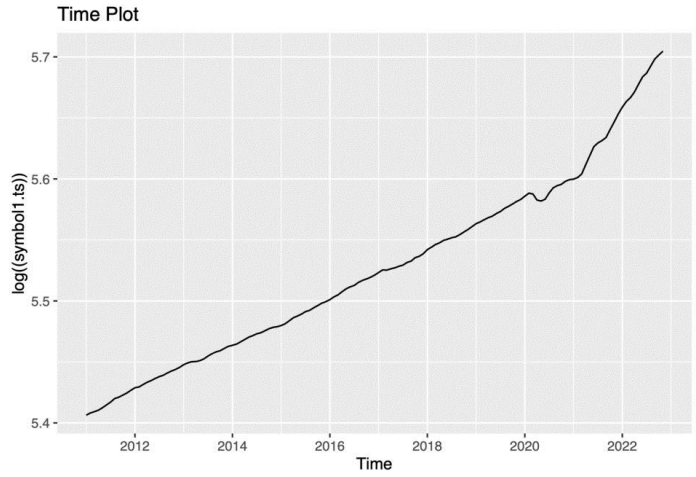
3.1 Extraire la série chronologique et tracer le graphique chronologique, le graphique saisonnier et la fonction d'auto-corrélation

Tout d'abord, extrayez la série et créez un objet ts :

```
symbol1<- list_symbols[2]

data.symbol1<- filter(economic_data2, symbol==symbol1)

symbol1.ts<- ts(data.symbol1$price, frequency=12, start=c(2011,1))
```

Tracé temporel

```
autoplot(log((symbol1.ts)))+
  ggtitle("Time Plot")
```



En examinant les données, nous avons fait plusieurs observations sur les tendances du CPILFESL :

Tout d'abord, nous avons noté que le CPILFESL a chuté en 2020. Cela indique une baisse de l'inflation, car une baisse de l'IPC suggère que les prix des biens et des services diminuent.

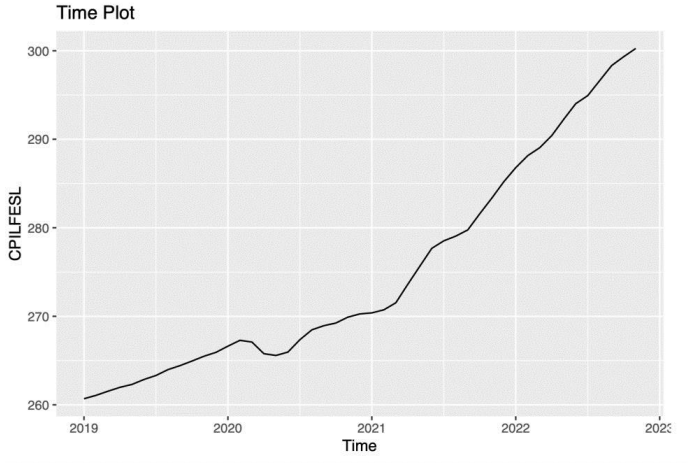
Deuxièmement, nous avons observé une augmentation considérable du CPILFESL au cours des années 2021 et 2022. Cela indique une augmentation de l'inflation, car l'IPC augmente.

Troisièmement, nous avons remarqué qu'avant l'année 2020, l'indice augmentait à un rythme linéaire constant. Cela suggère un taux d'inflation relativement stable au cours de cette période, les prix des biens et des services augmentant à un rythme soutenu.

Enfin, nous avons observé qu'après 2020, l'indice a augmenté à un rythme beaucoup plus rapide. Cela indique une augmentation soudaine de l'inflation au cours de cette période, qui pourrait être due à divers facteurs tels que des changements dans l'offre et la demande, des changements dans les politiques gouvernementales ou d'autres facteurs externes.

L’objectif principal de la prédiction de l’indice est d’évaluer la sensibilité de nos modèles de prévision à l’utilisation de données antérieures à l’année 2020. Fondamentalement, nous voulons comprendre comment l’exactitude et la fiabilité de nos prédictions sont affectées par l’inclusion de données à partir de 2020 et avant.

```
autoplot(window(symbol1.ts,start=c(2019,1)))+

 ggtitle("Time Plot") +

 ylab("CPILFESL")+

 xlab("Time")
```



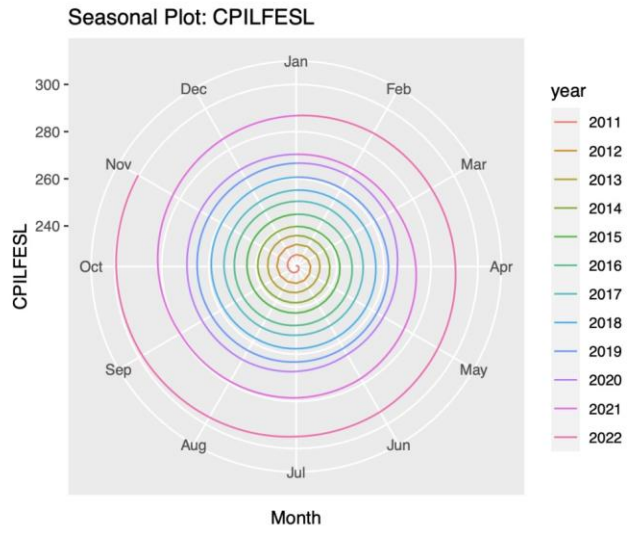
Graphique de l’heure saisonnière :

Un graphique saisonnier est une représentation graphique d’une série chronologique qui fournit des informations sur ses modèles saisonniers. Il permet d’identifier les modèles ou cycles répétitifs dans les données qui se produisent sur des intervalles de temps fixes, tels que des modèles quotidiens, mensuels ou annuels.

Nous utilisons le code R suivant pour tracer le graphique saisonnier

```
ggseasonplot(symbol1.ts, polar= TRUE)+

 ylab("CPILFESL")+
```

ggtitle("Seasonal Plot: CPILFESL")



La parcelle saisonnière est mise en œuvre par le package « fpp2 ». Il s'agit d'une technique unique en son genre de visualisation de données qui représente les valeurs de séries chronologiques en coordonnées polaires. En utilisant des coordonnées polaires, le graphique saisonnier fournit une représentation compacte et simple des modèles saisonniers dans les données. Dans ce graphique, l'angle correspond au mois de l'année, tandis que le rayon représente le temps écoulé depuis le début de la série.

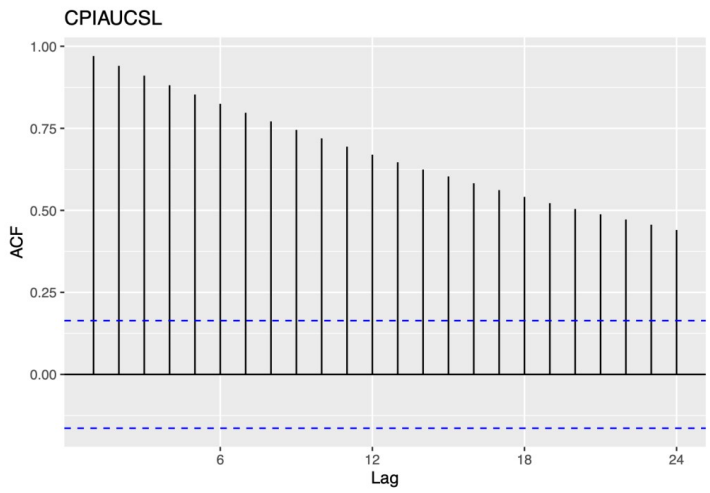
Nous avons fait les observations suivantes :

Tout d'abord, nous avons constaté qu'il y a un saut important de la variable CPILFESL chaque année. Cela indique qu'il y a un changement important dans le taux d'inflation d'une année à l'autre. Cela peut être dû à divers facteurs.

Deuxièmement, nous avons observé qu'il n'y a pas de chevauchement dans les données des saisons. Cela signifie que les points de données de chaque saison sont distincts et ne se chevauchent pas. Cette tendance pourrait avoir des implications importantes pour notre analyse, car elle suggère qu'il peut y avoir des tendances saisonnières ou des modèles qui sont uniques à chaque saison.

Autocorrélation et bruit blanc :

```
ggAcf(symbol1.ts)+

  ggtitle(symbol0)
```



Sur la base de nos observations des données, nous avons identifié plusieurs tendances importantes qui donnent un aperçu de la nature de la série chronologique :

Tout d'abord, nous avons noté qu'il y avait une lente diminution de la fonction d'autocorrélation (ACF) à mesure que les décalages augmentaient.

Deuxièmement, nous avons comparé les pics de l'ACF aux valeurs attendues pour une série de bruit blanc. Plus précisément, nous nous attendions à ce que 95 % des pics de l'ACF se situent dans la fourchette de ±2/T, où T est la longueur

de la série chronologique. Dans ce cas, T est égal à 11*12, soit 132. Cela signifie que nous nous attendons à ce que les pics de l'ACF se situent dans la fourchette de ±0,17

Cependant, nous avons constaté que les pics de l'ACF se situent en fait au-dessus de +0,17 et franchissent les limites bleues. Cela indique que la série n'est pas du bruit blanc, car il existe une corrélation significative entre les points de données qui ne peut pas être expliquée par le seul hasard.

Le modèle ACF observé, où l'autocorrélation diminue progressivement à mesure que le décalage augmente, est typique des séries chronologiques non stationnaires. Ce modèle est souvent observé dans les modèles de séries chronologiques tels que les marches aléatoires. Dans la section 3.2, nous étudions les caractéristiques des séries temporelles non stationnaires, en particulier les marches aléatoires, et analysons leur comportement à l'aide des modèles naïfs et dérivants. De plus, la section 3.2 donne un aperçu de la

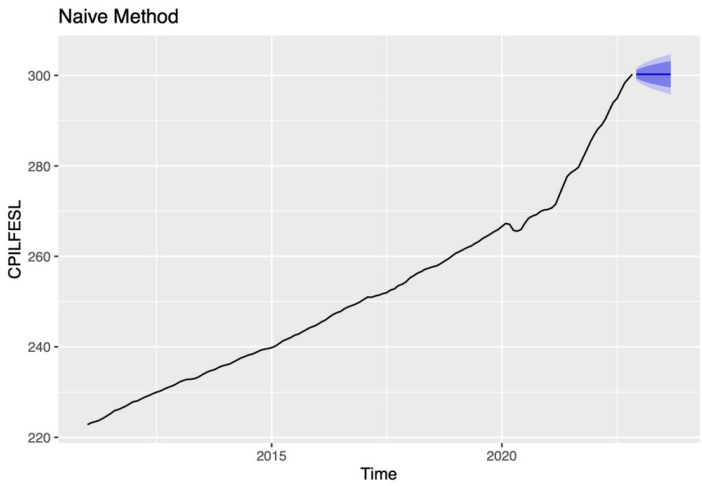
pertinence de ces modèles pour saisir la dynamique de séries temporelles non stationnaires, telles que les marches aléatoires, et met en lumière leurs capacités à prévoir dans de telles situations.

3.2 Évaluer des modèles de prévision simples

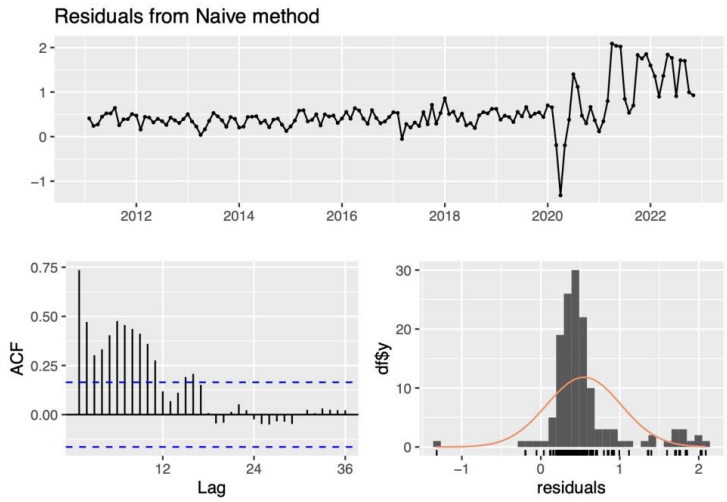
3.2.1 Marche aléatoire avec naïf

En utilisant le package R « fpp2 », appliquons la méthode naïve

```
symbol1.ts.naive=naive(symbol1.ts)

autoplot(symbol1.ts.naive)+

 ylab(symbol1)+

 ggtitle("Naive Method")
```




```
checkresiduals(symbol1.ts.naive)
```



##

Ljung-Box test

##

data: Residuals from Naive method

Q* = 340.41, df = 24, p-value < 2.2e-16

##

Model df: 0.   Total lags used: 24

La méthode naïve de prévision suppose que toutes les valeurs de prévision futures seront les mêmes que celles de la précédente ou de la dernière observation. Cependant, cette méthode n'est pas utile

En ce qui concerne la méthode naïve saisonnière, cette approche n'est utile que lorsque les données sont de nature saisonnière, ce qui n'est pas le cas pour ce jeu de données particulier. Au lieu de cela, les données semblent présenter davantage une tendance au fil du temps. Par conséquent, l'utilisation de la méthode naïve saisonnière ne fournirait pas d'informations ou de prévisions significatives.

Pour nous assurer de l'adéquation du modèle naïf dans la capture des modèles sous-jacents des séries chronologiques non stationnaires, nous avons effectué une analyse résiduelle. Les valeurs résiduelles sont les différences entre les

valeurs observées et les prédictions du modèle correspondant. L'analyse des résidus nous permet d'évaluer la capacité du modèle à capturer la variation restante dans les données.

Dans cette analyse, nous avons utilisé la fonction checkresiduals() qui fournit divers graphiques et tests pour évaluer les valeurs résiduelles du modèle. Ces tracés et tests aident à déterminer si le modèle capture de manière adéquate les modèles inhérents et le caractère aléatoire des données de séries chronologiques.

En examinant les graphiques, tels que la fonction d'autocorrélation (ACF) et l'histogramme des résidus, nous pouvons évaluer si les résidus présentent des modèles restants ou un comportement systématique. Idéalement, les résidus devraient être non corrélés, normalement distribués et avoir une variance constante.

En plus de l'inspection visuelle, la fonction checkresiduals() effectue également des tests statistiques, y compris le test de Ljung-Box, pour vérifier l'autocorrélation significative dans les résidus. Une autocorrélation significative indique que le modèle ne parvient pas à capturer une structure sous-jacente dans les données.

En analysant les données, nous avons observé que la variabilité des résidus est relativement faible dans la première moitié de la série, mais qu'elle augmente dans la seconde moitié. Ce modèle indique que la prise des logarithmes de la série est appropriée pour appliquer des modèles de séries chronologiques stationnaires. Cette approche peut aider à stabiliser la variance des données et faciliter l'application de diverses techniques statistiques.

Après avoir examiné les données, nous avons constaté que la prise de logarithmes ne semble pas améliorer la stationnarité de la série. Cependant, nous avons constaté que le fait de prendre les deuxièmes différences de la série, par opposition aux premières, conduit à une plus grande stationnarité. Cela suggère que la tendance dans les données n'est pas linéaire et peut nécessiter des techniques de modélisation plus complexes.

Maintenant, nous allons vérifier les erreurs de la méthode naïve

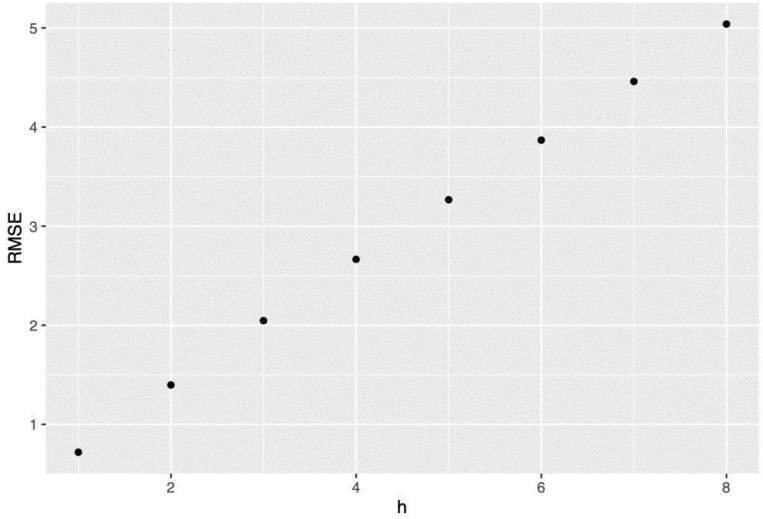
```
e <- tsCV(symbol1.ts, forecastfunction=naive, h=8)
```

 # Compute the RMSE values and remove missing values

rmse <- sqrt(colMeans(eˆ2, na.rm = T))

 # Plot the RMSE values against the forecast horizon data.frame(h = 1:8, RMSE = rmse) %>%

ggplot(aes(x = h, y = RMSE)) + geom_point()



Tout au long de notre analyse, nous fournirons des valeurs RMSE qui sont directement comparables à la série chronologique originale. Cela nous permet d'évaluer le niveau de précision atteint par les modèles de prévision et de faire des évaluations éclairées quant à leur adéquation aux applications pratiques et aux processus de prise de décision.

En examinant les résultats de la prévision, il devient clair que l'erreur de prévision a tendance à augmenter à mesure que l'horizon de prévision augmente. Cela signifie que la précision de nos prédictions a tendance à diminuer à mesure que nous essayons de faire des prévisions plus loin dans l'avenir.

C'est assez courant parce qu'à mesure que nous faisons des prévisions dans l'avenir, les prévisions deviennent de plus en plus difficiles en raison de l'incertitude et de la variabilité des résultats.

e <- tsCV(symbol1.ts, forecastfunction = naive, h=1)

```
sqrt(mean(eˆ2, na.rm=TRUE))
```

```
## [1] 0.7189534
```

```
sqrt(mean(residuals(naive(symbol0.ts))ˆ2, na.rm=TRUE))
```

```
## [1] 0.9239518
```

La fonction « tsCV » est un outil utile pour évaluer la précision d'un modèle de prévision de séries chronologiques. Il s'agit de l'abréviation de « validation croisée de séries chronologiques » et est utilisée pour estimer l'erreur de prévision d'un modèle. La fonction « tsCV » calcule les erreurs de prévision en divisant les données de la série chronologique en ensembles d'apprentissage et de test. Il ajuste le modèle à l'ensemble d'apprentissage et génère des prévisions pour l'ensemble de test correspondant. Les erreurs de prévision sont ensuite calculées en comparant les valeurs prédites aux valeurs réelles de l'ensemble de test.

La comparaison de l'erreur de validation croisée au RMSE des résidus permet d'évaluer les performances du modèle. Les résidus sont les différences entre les valeurs observées et les prédictions en une étape correspondantes générées par le modèle. Ces résidus représentent les erreurs de prédiction dans l'échantillon.

Idéalement, l'erreur de validation croisée devrait être inférieure au RMSE des résidus. Cela suggère que le modèle fonctionne bien non seulement sur les données sur lesquelles il a été formé, mais également sur les données invisibles. Si l'erreur de validation croisée est supérieure au RMSE des résidus, cela indique que le modèle peut surajuster les données d'entraînement.

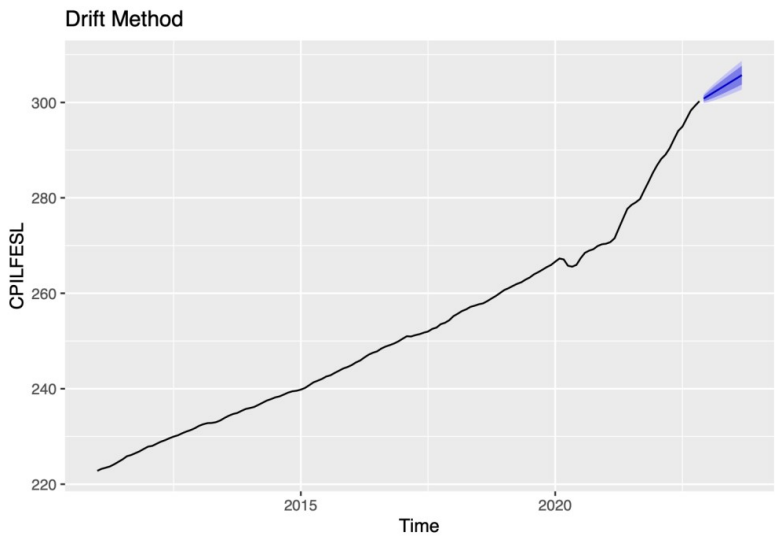
Dans ce cas, le RMSE de l'erreur de validation croisée est inférieur au RMSE des résidus, ce qui suggère que le modèle fonctionne plutôt bien.

3.2.2 Marche aléatoire avec dérive

Nous appliquons maintenant la méthode de la dérive

```
autoplot(rwf(symbol1.ts, drift=TRUE),
  series="Drift")+
```

```
ylab(symbol1)+
  ggtitle("Méthode Drift")
```



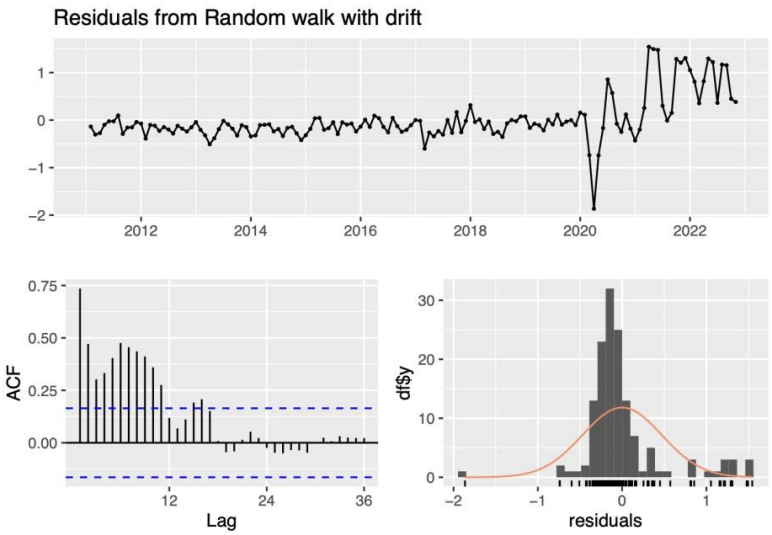
Nous faisons ici référence à une méthode de prévision connue sous le nom de méthode de dérive. La méthode de dérive consiste à prendre la dernière observation des données et à y ajouter la quantité de changement au fil du temps (c'est-à-dire la dérive) des données. La valeur résultante est ensuite utilisée comme prévision pour les périodes futures.

La composante dérive des données est une mesure de la tendance à augmenter ou à diminuer les données au fil du temps. En utilisant cette dérive dans les prévisions, nous sommes en mesure de prendre en compte la tendance globale des données et de faire des prédictions plus précises sur les valeurs futures.

```
fit0 = rwf(symbol1.ts, drift=TRUE)
summary(fit0)
```

```
##
## Méthode de prévision : marche aléatoire avec dérive
##
## Informations sur le modèle :
## Appel : rwf(y = symbol1.ts, drift = TRUE)
##
## Dérive : 0,5455 (se 0,0394)
## sd résiduel : 0,47
##
## Mesures d'erreur :
```

ME RMSE MAE
Ensemble d'entraînement -6.605036e-15 0,4683447 0,2966886 -0,007742377 0,1129022
MASE ACF1
Entraînement set 0,04767245 0,7353021
##
Prévisions :
##
décembre 2022
janvier 2023
février 2023
mars 2023
avril 2023
mai 2023
juin 2023
juillet 2023
août 2023
septembre 2023
Point Forecast Lo 80 Hi 80 Lo 95 Hi 95
300.8065 300.2020 301.4109 299.8821 301.7309
301.3520 300.4942 302.2098 300.0401 302.6639
301.8974 300.8432 302.9517 300.2851 303.5097
302.4429 301.2214 303.6644 300.5748 304.3111
302.9884 301.6180 304.3588 300.8926 305.0842
303.5339 302.0276 305.0401 301.2303 305.8375
304.0794 302.4469 305.7118 301.5828 306.5759
304.6248 302.8738 306.3758 301.9469 307.3027
305.1703 303.3069 307.0337 302.3205 308.0201 305.7158
303.7451 307.6865 302.7019 308.7297
Vérifier les résidus (fit0)



##
Test de Ljung-Box
##
données : Résidus de marche aléatoire avec dérive
Q* = 340,41, df = 23, valeur p < 2,2e-16
##
Modèle df : 1. Décalages totaux utilisé : 24 précision (fit0)
ME RMSE MAE MPE MAPE
Ensemble d'entraînement -6.605036e-15 0.4683447 0.2966886 -0.007742377 0.1129022
MASE ACF1
Ensemble d'entraînement 0.04767245 0.7353021

La sortie du modèle « fit0 » nous fournit des prévisions ponctuelles et des intervalles de prévision à l'aide de la fonction « rwf ». La fonction « rwf » calcule des prévisions ponctuelles sur la base du modèle de marche aléatoire avec dérive et fournit des intervalles de prévision qui capturent l'incertitude des prévisions. Nous avons utilisé la fonction checkresiduals() qui fournit divers tracés et tests statistiques pour évaluer les résidus du nouveau modèle. Ces graphiques et tests nous aident à déterminer si le modèle s'adapte correctement aux données.

Nous avons fait les observations importantes suivantes :

Premièrement, nous pouvons constater qu'à mesure que l'horizon de prévision (h) augmente, la largeur des intervalles de prévision augmente également. Cet élargissement des intervalles de prévision est attendu car la prévision devient plus difficile à mesure que nous projetons plus loin dans le futur. L'incertitude

accrue sur des horizons temporels plus longs se traduit par des intervalles plus larges pour tenir compte de la variabilité potentielle des valeurs prévues.

Deuxièmement, le coefficient de dérive estimé est de 0,5455 avec une erreur standard de 0,0394. Cela indique que la série chronologique a une tendance linéaire dans le temps. De plus, l'écart type résiduel est de 0,47, ce qui représente l'ampleur moyenne des différences entre les valeurs observées et les valeurs prévues.

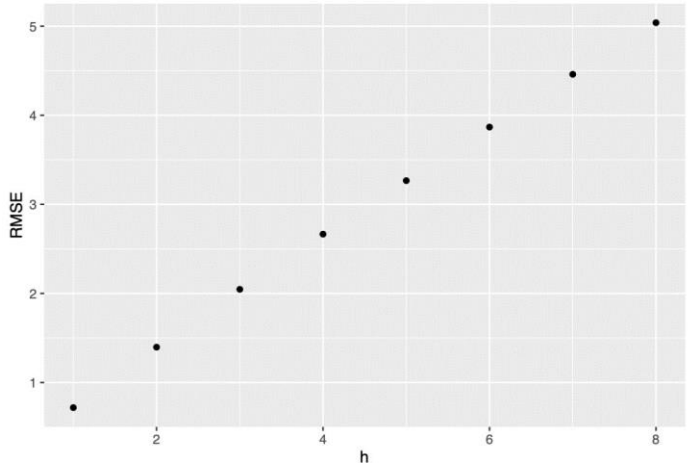
Troisièmement, le RMSE (erreur quadratique moyenne) est de 0,4683447, ce qui indique l'ampleur moyenne des erreurs de prévision. Le MAE (erreur absolue moyenne) est de 0,2966886, ce qui représente l'ampleur absolue moyenne des erreurs. Le MPE (pourcentage d'erreur moyen) est de -0,007742377. Le MAPE (pourcentage d'erreur absolu moyen) est de 0,1129022, reflétant l'écart en pourcentage moyen des prévisions. Le MASE (erreur d'échelle absolue moyenne) est de 0,0476724. L'ACF1 (coefficient d'autocorrélation de premier ordre) est de 0,7353021, ce qui suggère une autocorrélation relativement élevée.

Enfin, la statistique du test (Q*) est de 340,41, ce qui indique une preuve significative d'autocorrélation.

Dans l'ensemble, les résultats suggèrent que même si le modèle de marche aléatoire avec dérive capture la tendance linéaire dans la série chronologique, il existe toujours des autocorrélations dont le modèle ne prend pas en compte. Le développement de modèles ARIMA abordera ce problème ci-dessous.

Vérifions maintenant les erreurs dans la méthode de dérive

```
e <- tsCV(symbol1.ts, forecastfunction=naive, h=8)
# Compute the RMSE values and remove missing values
rmse <- sqrt(colMeans(e^2, na.rm = T))
# Plot the RMSE values against the forecast horizon
data.frame(h = 1:8, RMSE = rmse) %>%
ggplot(aes(x = h, y = RMSE)) + geom_point()
```




```
e <- tsCV(symbol1.ts, rwf, drift=TRUE, h=1)
sqrt(mean(eˆ2, na.rm=TRUE))
## [1] 0,4723834
sqrt(mean(residuals(rwf(symbol1.ts, dérive = VRAI))ˆ2, na.rm = VRAI))
## [1] 0,4683447
```

En comparant les valeurs RMSE de dérive et naïf, nous pouvons voir que le RMSE du modèle de dérive (0,4683447) est inférieur au RMSE du modèle naïf (0,9239518). Cela indique que le modèle de dérive est plus performant en termes de précision des prévisions, car il présente une ampleur moyenne d'erreurs inférieure à celle du modèle naïf.

Le RMSE plus faible dans le modèle de dérive suggère que l'intégration de la tendance linéaire (dérive) dans le modèle de prévision améliore sa capacité de prévision. Le modèle de dérive prend en compte l'évolution systématique des séries chronologiques au fil du temps, ce qui permet d'obtenir des prévisions plus précises que le modèle naïf, qui ne suppose aucune dérive.

Par conséquent, sur la base des valeurs RMSE, le modèle de dérive est préféré au modèle naïf pour prévoir cette série chronologique.

3.3 Ajustement des modèles ARIMA

L'hypothèse nulle de notre analyse est que les données des séries chronologiques sont stationnaires. Cela signifie que les propriétés statistiques des données, telles que la moyenne et la variance, restent constantes dans le temps. Pour analyser ces données de séries chronologiques, nous utiliserons un modèle de moyenne mobile intégrée auto-régressive (ARIMA) qui sera utilisé pour identifier les tendances et les modèles dans les données et faire des prédictions sur les valeurs futures.

Dans le langage de programmation R, nous pouvons utiliser la fonction auto.arima() pour identifier et spécifier un modèle ARIMA adapté à nos données. Cette fonction automatise le processus de sélection du modèle approprié en recherchant le modèle le mieux adapté en fonction d'une variété de critères statistiques.

```
bibliothèque(urca)
goog %>% ur.kpss() %>% summary()
##
## ########################
## # KPSS Test de racine unitaire #
## ########################
##
## Le test est de type : mu avec 7 décalages.
##
## La valeur de la statistique de test est : 10,7223
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
```

Sur la base de notre analyse statistique, nous avons constaté que nous pouvons rejeter l'hypothèse nulle selon laquelle la série chronologique ou les données sont stationnaires. Cette conclusion est basée sur le fait que la valeur de notre statistique de test s'est avérée supérieure aux valeurs critiques.

Une série chronologique ou des données stationnaires est une série dans laquelle les propriétés statistiques telles que la moyenne et la variance ne changent pas au fil du temps. En revanche, une série chronologique ou des données non stationnaires ont des propriétés statistiques qui changent avec le temps. Le fait que nous ayons rejeté l'hypothèse nulle indique que la série chronologique ou les données considérées ne sont pas stationnaires, ce qui signifie que leurs propriétés statistiques changent avec le temps.

Nous allons maintenant appliquer la technique de différenciation. Ce faisant, nous pouvons améliorer la précision et la fiabilité de notre analyse et prendre des décisions plus éclairées basées sur la modélisation d'une transformation de la série chronologique qui satisfait à un test de cohérence avec la stationnaire.

Le test KPSS (Kwiatkowski-Phillips-Schmidt-Shin) est un test statistique utilisé pour évaluer la stationnarité d'une série temporelle. La stationnarité fait référence à la propriété d'une série chronologique où ses propriétés statistiques (telles que la moyenne et la variance) restent constantes dans le temps.

Dans le cadre du test KPSS, l'hypothèse nulle suppose que la série temporelle est stationnaire. Si la statistique du test dépasse la valeur critique, cela indique une preuve contre l'hypothèse nulle. Cela implique que la série est probablement non stationnaire.

Examinons maintenant le résultat du test KPSS

```
bibliothèque (urca)
symbol1.ts %>% ur.kpss() %>% résumé()
##
## ######################
## # Test de racine unitaire KPSS #
## ######################
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 2,8269
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
```

```
ndiffs(symbol1.ts)
## [1] 2 # pour des séries entières, ndiffs() suggère une différenciation du second ordre.
```

La valeur statistique du test est de 2,8269 et elle est comparée aux valeurs critiques à différents niveaux de signification. Les valeurs critiques aux niveaux de signification de 10 %, 5 %, 2,5 % et 1 % sont respectivement 0,347, 0,463, 0,574 et 0,739.

Dans ce cas, la valeur statistique du test de 2,8269 dépasse toutes les valeurs critiques. Par conséquent, nous rejetons l'hypothèse nulle de stationnarité, suggérant que la série chronologique pourrait ne pas être stationnaire.

La fonction ndiffs() est également appliquée pour déterminer le nombre de différences requis pour obtenir la stationnarité de la série entière. Dans ce cas, ndiffs(symbol1.ts) suggère une différenciation du second ordre pour obtenir la stationnarité.

```
bibliothèque(urca)
symbol1.ts[1:108] %>% ur.kpss() %>% summary()
##
## ###################### ###
## # Test de racine unitaire KPSS #
## ########################
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 2,2539
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
ndiffs(symbol1.ts[1:108])
## [1] 2
# pour une série entière, ndiffs() suggère une différenciation du second ordre.
```

Ci-dessus, le sous-ensemble symbol1.ts[1:108] est utilisé, ce qui signifie que seules les 108 premières observations sont prises en compte. Auparavant, l'intégralité de la série chronologique symbol1.ts était analysée.

La période de 1:108 a été spécifiquement sélectionnée pour s'adapter aux données de la série chronologique, qui incluent les taux d'inflation jusqu'en 2020. Nous avons identifié une distorsion dans l'indice des prix à la consommation (IPC) à un certain moment au cours de cette période, et nous utilisé ce point comme seuil pour notre analyse. Ce faisant, nous avons pu concentrer notre analyse sur un sous-ensemble de données pertinent et informatif, ce qui nous a permis de faire des observations plus précises et significatives sur les tendances de l'inflation au fil du temps.

```
bibliothèque (urca)
symbol1.ts[1:108] %>% ur.kpss() %>% résumé()
```

Le résultat montre que dans les deux cas, la statistique du test dépasse les valeurs critiques, ce qui indique une preuve contre l'hypothèse nulle de

stationnarité. La fonction ndiffs() suggère que 2 différences peuvent être nécessaires pour atteindre la stationnarité dans les deux cas.

Dans l'ensemble, la différence dans les résultats est principalement due au sous-ensemble de la série chronologique considéré, tandis que l'interprétation des résultats des tests reste la même.

Appliquons maintenant le test kpss pour la série originale, pour la première série différentielle et la deuxième série différentielle pour les comparer côte à côte.

```
##
## ######################
## # Test de racine unitaire KPSS #
## ############# ##########
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 2,8269
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
##
## ######################
## # Test de racine unitaire KPSS #
## ############# ##########
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 1,2285
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739

# Test KPSS pour la série de premières différences
symbol1_diff1 < - diff(symbol1.ts)
kpss_diff1 <- ur.kpss(symbol1_diff1)
résumé(kpss_diff1)
##
```

```
## ######################
## # Test de racine unitaire KPSS #
## ############# ##########
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 0,0269
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
# Test KPSS pour la série de secondes différences
symbol1_diff2 <- diff(symbol1.ts, différences = 2)
kpss_diff2 <- ur.kpss(symbol1_diff2)
summary(kpss_diff2)

##
## ########### ############
## # Test de racine unitaire KPSS #
## ######################
##
## Test est de type : mu à 4 décalages.
##
## La valeur de la statistique de test est : 0,0269
##
## Valeur critique pour un niveau de signification de
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
```

Dans le premier cas, nous pouvons observer que la statistique de test pour la série originale (2,8269) est supérieure aux valeurs critiques à tous les niveaux de signification. Cela suggère que la série originale n'est pas stationnaire.

Cependant, lorsque l'on prend la première différence de la série, la statistique de test diminue à 1,2285. Même si elle reste supérieure aux valeurs critiques, elle indique un degré de non-stationnarité inférieur à celui de la série originale.

Enfin, la série de deuxième différence présente une statistique de test nettement inférieure de 0,0269, ce qui est bien inférieur aux valeurs critiques. Cela indique que la série de secondes différences peut être considérée comme stationnaire.

Dans l'ensemble, les résultats du test KPSS suggèrent que la série de deuxième différence présente l'approximation la plus proche de la stationnarité parmi les trois séries testées.

```
library(urca)
 # Test KPSS pour la série chronologique originale
kpss_original <- ur.kpss(symbol1.ts[1:108])
summary(kpss_original)
##
## ############### #########
## # Test de racine unitaire KPSS #
## ######################
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 2,2539
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739

# Test KPSS pour la série chronologique de première différence
symbol1_diff1 <- diff(symbol1.ts[1:108])
kpss_diff1 <- ur.kpss(symbol1_diff1)
résumé(kpss_diff1)

##
## ##################### ###
## # Test de racine unitaire KPSS #
## ######################
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 0,54
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
```

```
# Test KPSS pour la série chronologique de seconde différence
symbol1_diff2 <- diff(symbol1.ts[1:108], différences = 2)
kpss_diff2 <- ur.kpss(symbol1_diff2)
summary(kpss_diff2)

##
## ################ #######
## # Test de racine unitaire KPSS #
## ######################
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 0,0241
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
```

Dans le premier cas, la statistique du test (2,2539) est supérieure aux valeurs critiques à tous les niveaux de signification. Cela indique que nous pouvons rejeter l'hypothèse nulle, suggérant que la série originale n'est pas stationnaire. Dans le deuxième cas, la statistique du test (0,54) est inférieure aux valeurs critiques à tous les niveaux de signification. Cela suggère que nous ne pouvons pas rejeter l'hypothèse nulle de stationnarité pour la série en différences premières. Ainsi,

la série en différences premières est considérée comme stationnaire.
Dans le troisième cas, la statistique du test (0,0241) est nettement inférieure aux valeurs critiques.

niveaux de signification. Cela suggère que nous ne pouvons pas rejeter l'hypothèse nulle de stationnarité pour les séries en différences secondes.

La première série différenciée est cohérente avec la stationnarité lorsque seules les 108 premières valeurs sont incluses dans la série. La preuve apparente de secondes différences avec l'ensemble de la série est peut-être due à la forte volatilité de la série après la période 108.

Maintenant, appliquons la fonction auto.arima() et la fonction checkresiduals()

```
symbol1.ts.arima<- auto.arima(symbol1.ts)
```

résumé(symbol1.ts.arima)
Série : symbol1.ts
ARIMA(2,2,2)(0,0,2)[12]
##
Coefficients :
ar1 ar2 ma1 ma2 sma1 sma2
1,3288 -0,7466 -1,7074 0,8814 -0,2547 -0,1773
se 0,0785 0,0702 0,0625 0,0780 0,0988 0,0937
##
si gma^2 = 0,07624 : log de vraisemblance = -17,12
AIC=48,24 AICc=49,08 BIC=68,88
##
Mesures d'erreur de l'ensemble d'entraînement :
ME RMSE MAE MPE MAPE MASE
Ensemble d'entraînement 0,01957902 0,2682821 0,1744252 0,006985389 0,06737252 0,02802695
ACF 1
Ensemble d'entraînement 0,0259118
checkresiduals (symbole1. ts.arima)

##
Test de Ljung-Box
##
données : résidus d'ARIMA(2,2,2)(0,0,2)[12]
Q* = 22,222, df = 18, valeur p = 0,2223
##
Modèle df : 6. Total des décalages utilisés : 24
Les modèles ARIMA saisonniers sont désignés par la notation ARIMA(p, d, q)(P, D, Q)[h], où les différences saisonnières d'ordre D sont incorporées à l'aide des termes ARMA(P, Q).

Dans cette notation, les lettres minuscules (p, d, q) représentent les composantes non saisonnières du modèle. Le paramètre « p » désigne l'ordre de la composante autorégressive (AR), qui capture la dépendance de l'observation actuelle par rapport à ses valeurs précédentes. Le paramètre « d » représente l'ordre de différenciation requis pour atteindre la stationnarité dans la partie non saisonnière de la série. Le paramètre « q » indique l'ordre de la composante de moyenne mobile (MA), qui modélise la dépendance de l'observation actuelle sur les erreurs résiduelles des observations précédentes.

En revanche, les lettres majuscules (P, D, Q) représentent les composantes saisonnières du modèle. Le paramètre « P » indique l'ordre de la composante autorégressive saisonnière (SAR), qui capture la dépendance de l'observation actuelle par rapport à ses valeurs saisonnières précédentes. Le paramètre « D » représente l'ordre de différenciation requis pour atteindre la stationnarité dans la partie saisonnière de la série. Le paramètre « Q » indique l'ordre de la composante de moyenne mobile saisonnière (SMA), qui modélise la dépendance de l'observation actuelle sur les erreurs résiduelles des observations saisonnières précédentes.

Le paramètre 'h' représente l'horizon de prévision, indiquant le nombre de pas de temps futurs pour lesquels nous souhaitons générer des prévisions.

La sortie de la fonction auto.arima() suggère un modèle ARIMA(2,2,2)(0,0,2)(12) pour les données de séries chronologiques. Dans cette notation, le premier ensemble de nombres (2,2,2) représente les composantes non saisonnières du modèle ARIMA. Plus précisément, cela indique que le modèle comprend un terme autorégressif (AR) d'ordre 2, un terme de différenciation (I) d'ordre 2 et un terme de moyenne mobile (MA) d'ordre 2. Le deuxième ensemble de nombres (0,0, 2) représente les composantes saisonnières du modèle ARIMA. Dans ce cas, le modèle inclut un terme d'ordre 2 de moyenne mobile saisonnière (SMA). Les termes saisonniers tiennent compte des tendances et des fluctuations qui se produisent à intervalles réguliers (dans ce cas, avec une période de 12).

De plus, la valeur p du test de Ljung Box est supérieure à 0,05. Nous ne parvenons donc pas à rejeter l'hypothèse nulle d'absence d'autocorrélation. Cela suggère que les résidus du modèle ARIMA ne présentent pas d'autocorrélation significative, ce qui conforte l'adéquation du modèle à capturer les modèles restants dans les données. De plus, ACF ne franchit pas les lignes bleues, ce qui suggère qu'il n'y a pas d'autocorrélation significative dans les résidus. Cela indique que le modèle a correctement capturé les modèles sous-jacents dans les données.

Maintenant, appliquons la fonction auto.arima() et la fonction checkresiduals() pour le sous-ensemble de la série originale

```
symbol1.ts.arima<- auto.arima(symbol1.ts[1:108])
résumé(symbol1.ts.arima)
## Série : symbol1.ts[1:108]
```

ARIMA(1,2,1)
##
Coefficients :
ar1 ma1
0,1956 -0,9525
se 0,1014 0,0344
##
sigma^2 = 0,02133 : log de vraisemblance = 55,94
AIC=-105,88 AICc=-105,64 BIC=-97,89
##
Mesures d'erreur de l'ensemble d'entraînement :
ME RMSE MAE MPE MAPE MASE
Ensemble d'entraînement 0,008633132 0,1433159 0,1138458 0,003132918 0
.04679805 0.28179
##ACF1
Ensemble de formation -0,009711438
vérifier les résidus (symbole1.ts.arima)

##
Test de Ljung-Box
##
données : résidus d'ARIMA (1,2,1)
Q* = 3,3287, df = 8, valeur p = 0,9121
##
Modèle df : 2. Total décalages utilisés : 10
L'analyse ci-dessus suggère que les données suivent un modèle ARIMA (1,2,1). La notation ARIMA se compose de trois nombres représentant respectivement le nombre de termes autorégressifs, intégrés et moyens mobiles. Pour ces données, le modèle est ARIMA(1,2,1), ce qui signifie qu'il existe un terme autorégressif, deux termes intégrés et un terme de moyenne mobile. Les composantes saisonnières (P, D, Q) ne sont pas présentes, ce qui suggère que le modèle ne capture aucun modèle saisonnier.

Ce modèle ARIMA(1,2,1) est approprié pour les données en fonction des modèles que nous avons observés dans les données. Ces modèles suggèrent qu'il existe une corrélation significative entre les points de données et que cette corrélation diminue avec le temps. En utilisant ce modèle, nous pouvons capturer efficacement ces modèles et faire des prédictions précises sur les tendances futures des données.

Comparaison de la sortie du modèle symbol1.ts avec le modèle symbol1.ts [1:108] :

En termes de RMSE, le deuxième modèle (symbole1.ts [1:108]) a une valeur inférieure (0,1433159) par rapport au premier modèle (symbole1.ts) avec un RMSE plus élevé (0,2682821). Cela suggère que le deuxième modèle fournit un meilleur ajustement aux données du sous-ensemble symbol1.ts [1:108] que le premier modèle ne le fait à l'ensemble de la série symbol1.ts.

De plus, nous pouvons considérer les résultats du test Ljung-Box. Le deuxième modèle a une statistique de test de Ljung-Box inférieure (3,3287) par rapport au premier modèle (22,222), ce qui suggère un meilleur ajustement en termes d'autocorrélation résiduelle. Dans les deux modèles, la valeur p est assez élevée, ce qui indique qu'il n'y a aucune preuve d'autocorrélation significative.

Compte tenu du RMSE inférieur, de la statistique de test Ljung-Box inférieure et du fait que le deuxième modèle capture les principales caractéristiques des données avec moins de paramètres, nous pouvons conclure que le modèle ARIMA (1,2,1) est adapté au symbole 1. ts [1:108] est meilleur que le modèle ARIMA(2,2,2)(0,0,2)[12] adapté à toute la série symbol1.ts. Ces résultats illustrent également comment l'inclusion des données volatiles à la fin de la série chronologique peut entraîner une sensibilité du modèle identifié.

4. Discussion des résultats

Dans cette section, je résumerai les résultats de modélisation pour trois des séries CPI : CPILFENS, CPIAUCSL, CPILFESL.

Pour CPILFENS, le modèle optimal est ARIMA(2,1,0) avec dérive. Cela signifie qu'il existe deux termes autorégressifs, une différence et aucun terme de moyenne mobile. La présence de dérive signifie que la série a une moyenne non nulle. Le modèle a une vraisemblance logarithmique de -15,3, un AIC de 38,6, un AICc de 38,99 et un BIC de 49,29. Les résidus ne montrent pas d'autocorrélation significative sur la base du test de Ljung-Box (valeur p = 0,4659).

Pour CPIAUCSL, le modèle optimal est ARIMA(0,1,1) avec dérive. Cela signifie qu'il n'y a pas de termes autorégressifs, une différence et un terme de moyenne mobile. La présence de dérive signifie que la série a une moyenne non nulle. Le modèle a un log de vraisemblance de -60,6, un AIC de 127,2, un AICc de 127,43

et un BIC de 135,22. Les résidus ne montrent pas d'autocorrélation significative sur la base du test de Ljung-Box (valeur p = 0,5011).

Pour CPILFESL, le modèle optimal est ARIMA(1,2,1). Cela signifie qu'il existe un terme autorégressif, deux différences et un terme de moyenne mobile. Le modèle a une log-vraisemblance de 55,94, un AIC de -105,88, un AICc de -105,64 et un BIC de -97,89. Les résidus ne montrent pas d'autocorrélation significative sur la base du test de Ljung-Box (valeur p = 0,9121).

En se référant aux tableaux de coefficients en annexe, les lecteurs peuvent avoir une idée plus claire de la façon dont ces estimations de paramètres contribuent au pouvoir prédictif des modèles choisis et cela conforte la logique derrière la sélection de modèles ARIMA spécifiques pour chaque série.

En comparant les trois modèles, nous pouvons voir qu'ils ont des ordres et des coefficients différents, reflétant les différentes caractéristiques des données. Les modèles ont également différentes valeurs AIC, AICc et BIC, qui peuvent être utilisées pour comparer la précision relative du modèle. En général, une valeur inférieure pour ces critères indique un meilleur ajustement, de sorte que le modèle ARIMA(1,2,1) pour CPILFESL présente le meilleur ajustement parmi les trois modèles. De plus, les résidus de tous les modèles ne présentent aucune autocorrélation significative basée sur le test de Ljung-Box, ce qui suggère que tous les modèles s'ajustent raisonnablement bien aux données.

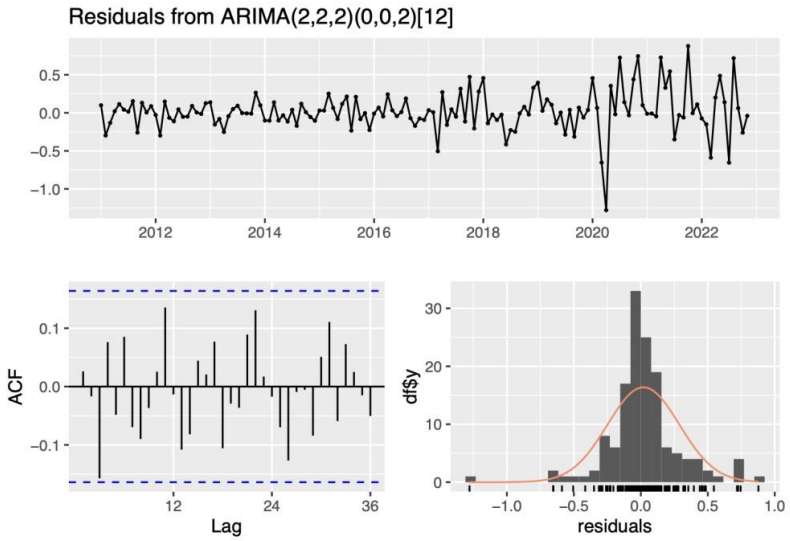
5. Conclusion

Dans ce document de recherche, nous avions pour objectif de modéliser les données de séries chronologiques d'inflation à l'aide de modèles de séries chronologiques simples et de modèles ARIMA, et d'appliquer ces modèles pour prévoir les valeurs futures des données de séries chronologiques. Nous avons collecté des données sur trois indicateurs économiques différents : l'indice des prix à la consommation pour tous les consommateurs urbains : tous les articles, sauf la nourriture et l'énergie (CPILFENS), l'indice des prix à la consommation pour tous les consommateurs urbains : tous les articles (CPIAUCSL) et l'indice des prix à la consommation pour tous les consommateurs urbains : All Items Less Shelter (CPILFESL), et les a modélisés à l'aide de modèles de séries chronologiques simples et de modèles ARIMA.

Nos résultats ont montré que les modèles ARIMA ont généralement de meilleurs résultats que les simples modèles de séries chronologiques en termes de

capacité à capturer les modèles et les tendances des données et à faire des prévisions précises. Plus précisément, le modèle ARIMA(2,1,0) avec dérive s'est avéré être le meilleur modèle pour le CPILFENS (Consumer Price Index for All Urban Consumers: All Items Less Food and Energy), le modèle ARIMA(0,1,1) avec dérive s'est avéré être le meilleur modèle pour CPIAUCSL (Indice des prix à la consommation pour tous les consommateurs urbains : tous les articles), et le modèle ARIMA (1,2,1) s'est avéré être le meilleur modèle pour CPILFESL (Indice des prix à la consommation pour tous). Consommateurs urbains : tous les articles moins le logement). Ces modèles ont pu capturer les tendances des données et faire des prédictions précises sur les valeurs futures de la série chronologique.

En termes d'orientations possibles pour les recherches futures, une orientation pourrait consister à explorer l'utilisation d'autres modèles de séries chronologiques au-delà d'ARIMA. Une autre solution pourrait consister à intégrer des facteurs externes susceptibles d'influencer les données des séries chronologiques, tels que les changements dans les politiques gouvernementales et les événements mondiaux. Dans l'ensemble, les résultats de cette recherche suggèrent que les modèles ARIMA peuvent être efficaces dans la modélisation et la prévision des données de séries chronologiques économiques, et que des recherches supplémentaires pourraient s'appuyer sur cette base pour améliorer l'exactitude et la précision de ces modèles.



Annexe

Résultats des deux autres modèles CPI utilisant une approche similaire :

1.1 Extraire les séries chronologiques de CPIAUCSL et CPILFENS

Extrayez la série et créez un objet ts : CPIAUCSL :

CPILFENS :

```
symbol4<- list_symbols[5]
data.symbol4<- filter(nomic_data2, symbol==symbol4)
symbol4.ts<- ts(data.symbol4$price, fréquence=12, start=c(2011,1))
```

1.2 Modèles ARIMA

1.2.1 Modèle ARIMA du CPIAUCSL :

Hypothèse nulle : la série chronologique/les données sont stationnaires.

Nous pouvons utiliser la fonction R auto.arima() pour identifier et spécifier un modèle de moyenne mobile intégrée auto-régressive (ARIMA) ajusté.

Dans ce cas, la valeur statistique du test est supérieure à la valeur critique à tous les niveaux de signification, ce qui suggère que la série temporelle symbol0.ts n'est pas stationnaire.

```
bibliothèque (urca)
symbol0.ts %>% ur.kpss() %>% résumé()
##
## ########################
## # Test de racine unitaire KPSS #
## ########################
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 2,6058
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
ndiffs(symbol0.ts)
## [1] 2
# pour des séries entières, ndiffs() suggère une différenciation du second ordre.
bibliothèque(urca)
symbol0.ts[1:108] %>% ur.kpss() %>% summary()
##
## ###################### ###
## # Test de racine unitaire KPSS #
## ########################
```

```
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 2,1774
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
ndiffs(symbol0.ts[1:108])
## [1] 1
# pour une série entière, ndiffs() suggère une différenciation du second ordre.
symbol0.ts.arima<- auto.arima(symbol0.ts)
summary(symbol0.ts.arima)
## Série : symbol0.ts
## ARIMA(0,2,2)
##
## Coefficients :
bibliothèque(urca)
##
##
## se 0,0794 0,0795
##
## sigma^2 = 0,3385 : log de vraisemblance = -123,02
## AIC=252,05 AICc=252,22 BIC=260,89
##
## Mesures d'erreur de l'ensemble d'entraînement :
## ME RMSE MAE MPE MAPE MASE
## Ensemble de formation 0,02149972 0,5736192 0,4093012 0,007226791
0,1626624 0,06733796
## ACF1
## Ensemble de formation -0,007051954
autoplot(symbol0.ts.arima)

vérifier les résidus (symbole0.ts.arima)


##
## Test Ljung-Box
##
## données : résidus d'ARIMA (0,2,2)
```

Q* = 16,696, df = 22, valeur p = 0,7799
##
Modèle df : 2. Total décalages utilisés : 24
symbol0.ts.arima<- auto.arima(symbol0.ts[1:108])
summary(symbol0.ts.arima)
Série : symbol0.ts[1:108]
ARIMA(0, 1,1) avec dérive
##
Coefficients :
ma1 dérive
0,3809 0,3518
se 0,0817 0,0567
##
sigma^2 = 0,1854 : log de vraisemblance = -60,6
AIC=127,2 AICc=127,43 BIC= 135.22
##
Mesures d'erreur de l'ensemble d'entraînement :
ME RMSE MAE MPE MAPE MASE
Ensemble d'entraînement 0,00159638 0,424554 0,3228736 0,0003542257 0,1355093 0,6720907
ACF1
Ensemble d'entraînement 0,03332243
autoplot(symbole 0.ts.arima)

vérifier les résidus (symbole0.ts.arima)

##
Test Ljung-Box
##
données : résidus d'ARIMA(0,1,1) avec dérive
Q* = 8,3315, df = 9, valeur p = 0,5011
##
Modèle df : 1 . Total des décalages utilisés : 10
Sur la base des modèles observés dans les données, nous avons déterminé qu'elles suivent un modèle ARIMA(0,1,1). Deux tendances spécifiques dans les données étayent cette conclusion.

Premièrement, la fonction d'autocorrélation partielle (PACF) est soit en décroissance exponentielle, soit sinusoïdale. Cela indique que les données dépendent de leurs propres valeurs passées et que la valeur actuelle est liée à la

valeur précédente avec un certain degré de corrélation. Deuxièmement, il y a un pic significatif au décalage 1 dans la fonction d'autocorrélation (ACF), mais aucun au-delà du décalage 1. Cela implique que les données ont une forte corrélation avec leur valeur passée immédiate, mais cette corrélation diminue rapidement à mesure que l'on s'éloigne de la valeur actuelle.

Dans l'ensemble, ces observations suggèrent que les données présentent un certain degré d'autocorrélation et de saisonnalité, et qu'un modèle de série chronologique comme ARIMA(0,1,1) serait approprié pour capturer avec précision ces tendances et faire des prédictions significatives.

1.2.2 Modèle ARIMA du CPILFENS :

Hypothèse nulle : la série chronologique/les données sont stationnaires.

Nous pouvons utiliser la fonction R auto.arima() pour identifier et spécifier un modèle de moyenne mobile intégrée auto-régressive (ARIMA) ajusté.

Dans ce cas, la valeur statistique du test est supérieure à la valeur critique à tous les niveaux de signification, ce qui suggère que la série temporelle symbol4.ts n'est pas stationnaire.

```
library(urca)
symbol4.ts %>% ur.kpss() %>% summary()
## # Test de racine unitaire KPSS #
## ################### ####
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 2,8246
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
ndiffs(symbol4.ts)
## [1] 2 # pour des séries entières, ndiffs() suggère une différenciation du second ordre.
bibliothèque (urca)
symbol4.ts[1:108] %>% ur.kpss() %>% résumé()
##
## ######################
```

```
## # Test de racine unitaire KPSS #
## ############# ##########
##
## Le test est de type : mu avec 4 décalages.
##
## La valeur de la statistique de test est : 2,2526
##
## Valeur critique pour un niveau de signification de :
## 10 % 5 % 2,5 % 1 %
## valeurs critiques 0,347 0,463 0,574 0,739
ndiffs(symbol4.ts[1:108])
## [1] 1
# pour une série entière, ndiffs() suggère une différenciation du second ordre.
symbol4.ts.arima<- auto.arima(symbol4.ts)
résumé(symbol4.ts.arima)
## Série : symbol4.ts
## ARIMA(2,2,2)(2,0,0)[12]
##
## Coefficients :
## ar1 ar2 ma1 ma2 sar1 sar2
## 1.2534 -0.6942 -1.5742 0.7216 0.3195 0.3412
## se 0.1310 0.0873 0.1515 0.1519 0.0848 0.0907
##
## sigma ^2 = 0,1333 : log de vraisemblance = -58,34
## AIC=130,68 AICc=131,52 BIC=151,32
##
## Mesures d'erreur de l'ensemble d'entraînement :
## ME RMSE MAE MPE MAPE MASE
## Ensemble d'entraînement 0,002739018 0,3546955 0,2447811 0,001036818
0,09470889 0,03932631
## ACF1
## Ensemble de formation 0.02117099
autoplot(symbol4.ts. arima)

vérifier les résidus (symbole4.ts.arima)

## Test Ljung-Box
##
## données : résidus d'ARIMA(2,2,2)(2,0,0)[12]
## Q* = 25,447, df = 18, valeur p = 0,1131
```

##
Modèle df : 6. Décalages totaux utilisés : 24
symbol4.ts.arima<- auto.arima(symbol4.ts[1:108])
summary(symbol4.ts.arima)
Série : symbol4.ts[1 : 108]
ARIMA(2,1,0) avec dérive
##
Coefficients :
ar1 ar2 dérive
0.5217 -0.4906 0.3999
se 0.0853 0.0850 0.0279
##
sigma^2 = 0.08015 : log de vraisemblance = -15,3
AIC=38,6 AICc=38,99 BIC=49,29 ##
Mesures d'erreur de l'ensemble d'entraînement :
##
ME RMSE MAE MPE MAPE MASE
Ensemble d'entraînement 0,002997304 0,277811 0,2278571 0,001007415 0,09316369 0,5187606
ACF1
Ensemble de formation -0.02481104
tracé automatique (symbole4.ts.arima)

vérifier les résidus (symbole4.ts.arima)

##
Test de Ljung-Box
##
données : résidus d'ARIMA(2,1,0) avec dérive
Q* = 7,6742, df = 8, valeur p = 0,4659
##
Modèle df : 2 . Total des décalages utilisés : 10

Ainsi, ARIMA(2,1,0) avec dérive signifie que le modèle a une composante autorégressive d'ordre 2, un degré de différenciation de 1, aucune composante de moyenne mobile et un terme constant. Les composants AR et MA capturent respectivement les modèles d'autocorrélation et de moyenne mobile dans les données, tandis que le terme de dérive capture la tendance ou le niveau global des données. Le terme de dérive dans le modèle ARIMA suggère qu'il peut y avoir

une tendance ou un changement systématique dans la moyenne des données au fil du temps qui n'est pas expliqué par les composantes AR ou MA.

Le tracé ACF montre une ligne croisant les lignes critiques bleues entre les décalages de 10 et 15, ce qui suggère qu'il existe encore une autocorrélation significative dans les résidus du modèle. Cela signifie que le modèle peut ne pas capturer entièrement toutes les informations pertinentes contenues dans les données. Il peut s'avérer nécessaire soit de modifier le modèle existant, soit d'essayer d'autres approches pour améliorer l'ajustement.

1.3 Discussion des modèles ARIMA basés sur les coefficients obtenus

1. Tableau des coefficients pour le modèle CPILFENS – ARIMA(2,1,0) avec dérive :
erreur standard d'estimation du coefficient AR1 0,5217 0,0853 AR2 -0,4906 0,0850 Dérive 0,3999 0,0279

2. Tableau des coefficients du modèle CPIAUCSL – ARIMA(0,1,1) avec dérive :

Coefficient Estimation Erreur type MA1 0,3809 0,0817 Dérive 0,3518 0,0567

3. Tableau des coefficients du modèle CPILFESL – ARIMA (1,2,1) :

Coefficient Estimation Erreur type AR1 0,1956 0,1014 MA1 -0,9525 0,0344

L'interprétation de ces estimations de paramètres peut offrir des informations précieuses sur la dynamique sous-jacente des séries chronologiques. Par exemple, les paramètres autorégressifs (AR) positifs correspondent souvent à l'élan ou à la persistance dans la série. Cela implique que la série a tendance à continuer d'évoluer dans la même direction que ses valeurs passées. En revanche, des paramètres AR négatifs suggèrent un retour à la moyenne, indiquant que la série a tendance à revenir vers sa moyenne au fil du temps.

De plus, des paramètres de moyenne mobile (MA) positifs indiquent que les erreurs de prévision récentes ont un effet de type dynamique sur la série. Cela signifie que les écarts par rapport aux prévisions ont tendance à être corrigés au cours des périodes ultérieures. De telles informations obtenues à partir des estimations des paramètres permettent une compréhension plus approfondie des données de séries chronologiques.

A propos de l'auteur

Ruthvik Reddy Bommireddy
Ruthvik est une élève de 12e année à l'école publique de Delhi, Bangalore Nord, en Inde. Son orientation académique se situe à l'intersection des sciences et de l'économie, motivée par le désir de créer un changement positif dans le monde à travers de futures entreprises. Ruthvik aime le basket-ball, le tennis et le patinage à roulettes, ainsi que le dessin et les esquisses. Il est également président et fondateur du club d'affaires de son école.