Auteurs : MARINGER Léa, MOUTTAPA Arnaud

Modélisation et Prévisions des Revenus Trimestriels de Google

---

Rapport dans le cadre du cours de *Modèles de Prévisions*
(3ème année de cycle ingénieur - Data Science)

Octobre 2023

Table des matières

Introduction

Les technologies permettent de récolter et conserver de plus en plus de données. En particulier, de nombreuses données temporelles sont à notre disposition, que l'on caractérise de *séries temporelles*. Elles se définissent comme une suite d'observations portant sur un même objet et recueillies au fil du temps à des instants équidistants et sont largement utilisées dans tous les domaines : la finance, l'économie, la météorologie, l'ingénierie ou encore la santé publique. En effet, il peut s'agir des températures moyennes mensuelles à Paris, du nombre de nouveau cas de Covid-19 par jour au Mexique ou encore du prix de l'indice boursier S&P 500 par 5 minutes.

Mathématiquement, une série de données temporelles peut être définie comme une réalisation d'une famille de variables aléaloires ou comme un processus stochastique ($\{X_t\}$ avec $t \in \mathbb{Z}$). Il est alors possible d'ajuster aux données des modèles statistiques qui permettront de prévoir, ou au moins estimer, les valeurs futures afin d'anticiper certaines situations ou certaines actions à réaliser.

Dans ce rapport, nous nous intéresserons en particulier aux revenus trimestriels de l'entreprise Google et à des modèles statistiques comme ARMA et ARIMA. Dans une première partie, nous introduirons le modèle ARMA et présenterons des simulations effectuées en R. Dans un second temps, nous étudierons les revenus de Google, en passant par la stationnarisation de la série, l'ajustement de différents modèles et la prévision de valeurs futures.

1 Modèle ARMA et simulation

Dans cette partie, nous définirons rapidement le modèle ARMA, expliquerons comment effectuer des prévisions à partir de ce dernier et nous présenterons un exemple de simulation en R.

1.1 Définition du modèle ARMA

Un modèle ARMA (*autoregressive-moving-average*) est constitué d'une partie autorégressive (AR), qui dépend des observations passées, ainsi que d'une partie moyenne mobile (MA), qui dépend des erreurs passées [1]. Précisément, un processus $(X_t)_{t\in\mathbb{N}^*}$ peut être modélisé par un modèle ARMA d'ordres (p, q) s'il peut s'écrire sous la forme (1).

$$X_t = c + \sum_{i=1}^{p} \phi_i \cdot X_{t-i} + \sum_{j=1}^{q} \theta_j \cdot \epsilon_{t-j} + \epsilon_t \tag{1}$$

où $c \in \mathbb{R}$ est une constante, $\{\phi_i\}_{1\leq i\leq p}$ et $\{\theta_j\}_{1\leq j\leq q}$ sont respectivement les coefficients de la partie autorégréssive et de la partie moyenne mobile, et $(\epsilon_t)_{t\in\mathbb{N}}$ est un bruit blanc centré de variance σ_ϵ^2.

Le modèle ARMA permet de modéliser un plus large ensemble de processus que les modèles AR ou MA, qui peuvent manquer de précision, notamment lorsque l'on dispose de peu d'observations. Herman Wold a d'ailleurs montré que tout processus stationnaire pouvait être modélisé par un modèle ARMA (dont les ordres p et q sont à trouver) [4].

Pour rappel, on peut également écrire un processus ARMA(p, q) sous la forme (2).

$$\Phi(B) \cdot X_t = c + \Theta(B) \cdot \epsilon_t \tag{2}$$

où B est l'opérateur retard et les polynômes Φ et Θ sont définit par (3). Le processus est stationnaire si la partie autorégréssive l'est, c'est à dire si le module de toutes les racines du polinôme Φ est strictement supèrieur à 1.

$$\Phi(B) = I - \sum_{i=1}^{p} \phi_i \cdot B^i \ , \quad \Theta(B) = I + \sum_{j=1}^{q} \theta_j \cdot B^j \tag{3}$$

1.2 Prévisions avec un modèle ARMA

La modèlisation d'un processus permet d'effectuer des prévisions pour les valeurs futures d'un processus. Dans le cas d'un processus $(X_t)_{t\in\mathbb{N}^*}$ modélisé par un modèle ARMA d'ordres (p, q), on peut définir $\hat{X}_T(h)$ la prévision du processus $(X_t)_{1\leq t\leq T}$ à l'horizon $h > 0$ par (4).

$$\hat{X}_T(h) = \sum_{i=1}^{p} \phi_i \cdot \hat{X}_{T+h-i} + \sum_{j=1}^{q} \theta_j \cdot \hat{\epsilon}_{T+h-j} \tag{4}$$

où $\hat{\epsilon}_{T+h-j}$ vaut ϵ_{T+h-j} si $i \geq k$, 0 sinon, et $\hat{X}_{T+h-j}$ vaut X_{T+h-j} si $j \geq k$, $\hat{X}_T(h-j)$ sinon. Cela vient du fait que $\hat{X}_T(h)$ est définit par l'espérance la variable X_{T+h} connaissant les observations et les erreurs passées (5).

$$\hat{X}_T(h) = \mathbb{E}(X_{T+h}|X_1, X_2, ..., X_T, Z_{-1}) \tag{5}$$

avec $Z_{-1} = \{X_{-1}, X_{-2}, ..., X_{-p}, \epsilon_{-1}, \epsilon_{-2}, ..., \epsilon_{-q}\}$. Nous pouvons alors calculer récursivement $\hat{X}_T(1)$, $\hat{X}_T(2)$, etc. jusqu'à $\hat{X}_T(h)$.

Par exemple, considérons le processus stationnaire suivant :

$$Y_t = 0.2Y_{t-1} + 0.5Y_{t-2} + 0.3\epsilon_{t-1} + 0.5\epsilon_{t-2} + \epsilon_t \tag{6}$$

Il s'agit d'un processus ARMA(2, 2) dont les coefficients de la partie auto-régréssive et de la partie moyenne mobile sont respectivement $\{0.2, 0.5\}$ et $\{0.3, 0.5\}$. Il est alors possible de calculer les prévisions à l'horizon $h = 5$ grâce aux formules suivantes et aux observations et aux erreurs passées. En supposant que nous disposons des observations $\{Y_t\}_{1\leq t\leq T}$ et des résidus associés $\{\epsilon_t\}_{1\leq t\leq T}$ et en utilisant (4),

$$\hat{Y}_T(1) = 0.2Y_T + 0.5Y_{T-1} + 0.3\epsilon_T + 0.5\epsilon_{T-1}$$
$$\hat{Y}_T(2) = 0.2\hat{Y}_T(1) + 0.5Y_T + 0.5\epsilon_T$$
$$\hat{Y}_T(3) = 0.2\hat{Y}_T(2) + 0.5\hat{Y}_T(1)$$

$$\hat{Y}_T(4) = 0.2\hat{Y}_T(3) + 0.5\hat{Y}_T(2)$$
$$\hat{Y}_T(5) = 0.2\hat{Y}_T(4) + 0.5\hat{Y}_T(3)$$

Notons que lorsque h est supérieur à l'ordre de la partie moyenne mobile, la prévisions à l'horizon h ne contient plus de résidus.

1.3 Simulation et prévisions en R

1.3.1 Simulation d'un processus ARMA

Nous avons simulé en R le processus ARMA(2, 2) (6) définit dans la partie 1.2 avec des résidus ϵ_t qui suivent une loi normale centrée réduite. Notons que ce processus est stationnaire car le module des racines du polynôme associé à la partie autorégressive est strictement supérieur à 1.

$$\Phi(x) = 1 - 0.2 \cdot x + 0.5 \cdot x^2$$
$$x_1 \approx -1.63 \text{ et } x_2 \approx 1.22$$

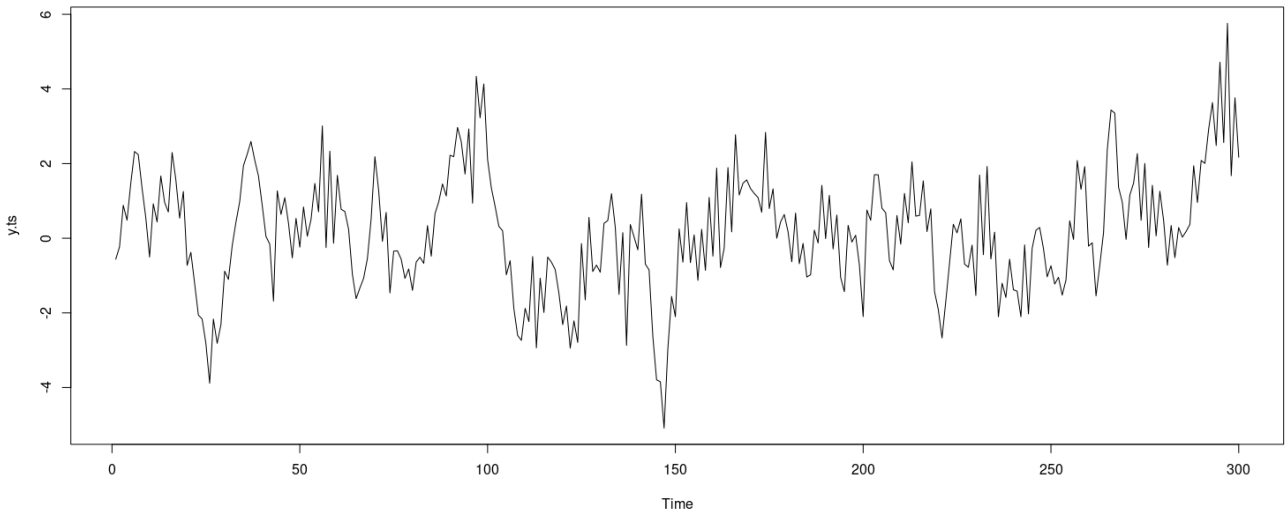


FIGURE 1 – Simulation de $n = 300$ observations d'un processus ARMA(2,2) de coefficients autorégressifs 0.2 et 0.5 et de coefficients de moyenne mobile 0.3 et 0.5.

En R, il existe une fonction *arima.sim()* du package *stat* qui permet de simuler un processus ARMA à partir d'un nombre d'observations, de coefficients autorégressifs, de coefficients

de moyenne mobile et variance des résidus au choix. Cependant, nous avons ici utilisé notre propre fonction qui construit le processus itérativement à partir des résidus (qui suivent une loi normale centrée).

Comme vu précedemment, le processus (6) est théoriquement stationnaire. Il est également possible de tester la stationnarité sur notre réalisation du processus avec le test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) (voir partie 2.3.1). Si l'on réalise plusieurs réalisations du processus et que nous leur appliquons le test KPSS, ce dernier affirme la stationnarité dans la plupart des cas. Il arrive toutefois que ce dernier indique que la série est non stationnaire, ce qui peut arriver en raison du caractère aléatoire de la procédure de simulation.

Le résultat du test KPSS pour le processus simulé dans la partie 1.3.1 est donné Figure 2. La statistique vaut 0.2468, ce qui est inférieur à la valeur critique à 1% 0.739 et confirme donc bien la stationnarité de la série simulée.

```
########################
# KPSS Unit Root Test #
########################

Test is of type: mu with 5 lags.

Value of test-statistic is: 0.2468

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

FIGURE 2 – Statistique de KPSS calculée pour le processus simulé dans la partie 1.3.1.

1.3.2 Prévisons

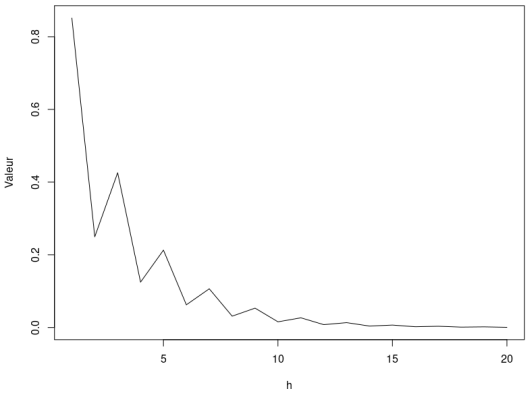
La série a ensuite été séparée en séries d'entrainement ($\frac{14}{15}$ de la série) et de test ($\frac{1}{15}$ de la série) afin d'effectuer des prévisions et de les comparer aux valeurs réelles. En effet, un modèle ARMA peut être utile pour réaliser des prévisions uniquement à court terme : les prévisions tendent très rapidement vers 0 .

Des prévisions pour les 20 valeurs de la série de test sont alors effectuées grâce à la formule théorique de la partie 1.2 d'une part (Figure 4a). D'autre part, un modèle ARMA d'ordres $p = 2$ et $q = 2$ est ajusté au processus simulé. Les coefficients donnés par ce modèle sont donnés par Figure 3. Remarquons qu'ils sont proches mais ne sont pas égaux aux coefficients théoriques. Cette légère différence se justifie par, d'une part, le fait que le processus simulé est une simulation avec une part d'aléatoire, d'autre part, le fait que le modèle n'est ajusté que sur une partie du processus simulé (série d'entrainement).

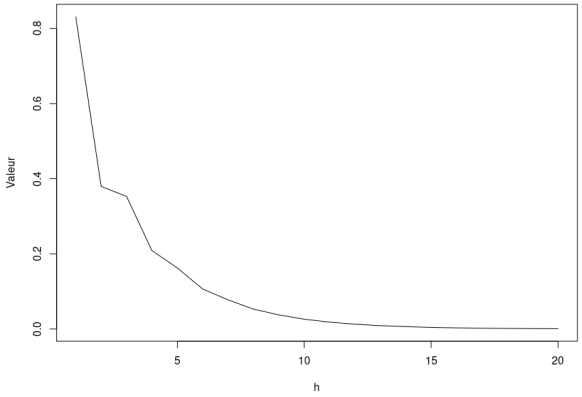
```
      ar1       ar2       ma1       ma2
0.2645939 0.3039849 0.2214321 0.5333157
```

FIGURE 3 – Coefficients trouvés par le modèle ARMA(2, 2) ajusté au processus simulé.

Des prévisions sont alors effectuées également avec ce modèle ajusté, grâce à la fonction *predict()* en R, pour la série de test (Figure 4b). Les différentes prévisions sont présentées Figure 4. Nous constatons que l'allure des prévisions est semblable et les valeurs prédites oscillent légérement dans le cas des prévisions avec la formule théorique. Cela peut s'expliquer par la différence des coefficients ou la méthode de prévisions utilisée par le modèle ARMA ajusté en R. De plus, à l'échelle des valeurs réelles de la série, les prédictions données par les 2 méthodes sont quasiment identiques (Figure 5).



(a) Avec la formule théorique de la partie 1.2.



(b) Avec le modèle ARMA ajusté au processus en R.

.

FIGURE 4 – Prévisions de $h = 20$ valeurs futures pour le processus simulé dans la partie 1.3.1.

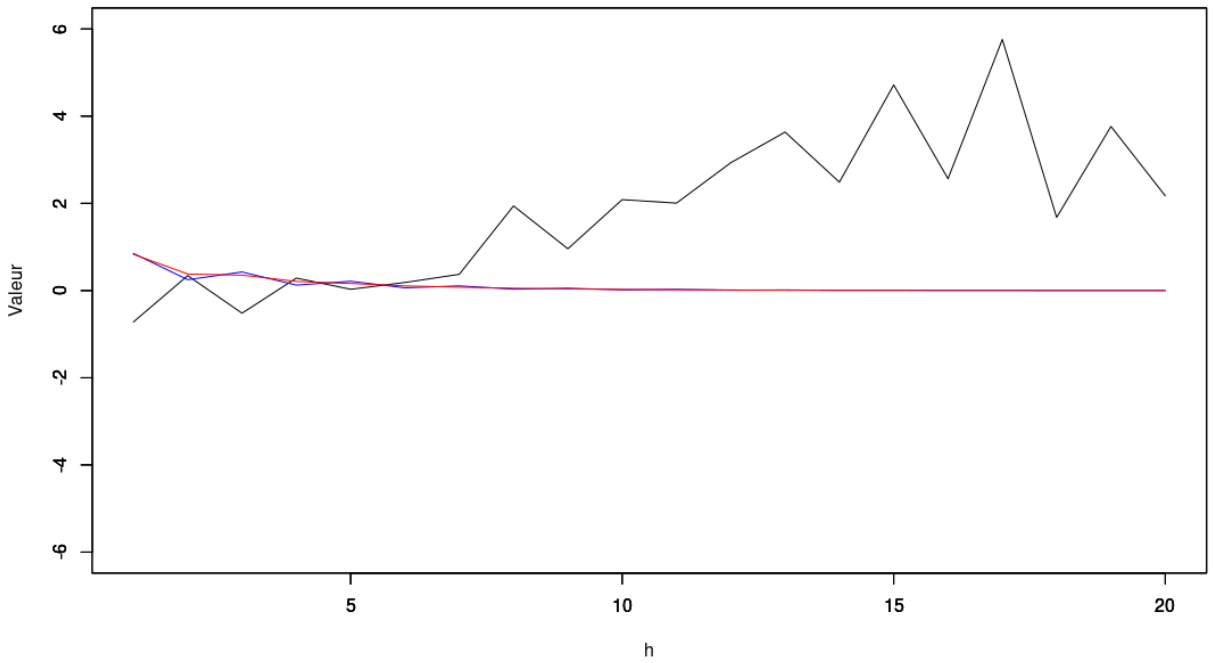


FIGURE 5 – Prévions grâce à la formule théorique de la partie 1.2 (en bleu) et grâce au modèle ARMA ajusté en R (en rouge). La série réelle (de test) est affichée en noir.

2 Etude des revenus trimestriels de Google

Les résultats d'une entreprise sont très importants à analyser pour cette dernière. En effet, ils permettent de comprendre si les perfomances de l'entreprise sont bonnes et/ou si la stratégie doit être adaptée. Ces données permettent aussi aux investisseurs d'évaluer le potentiel d'une entreprise. Nous avons alors décidé de nous intéresser aux revenus trimestriels de l'entreprise Google. Google est une entreprise mondialement connue pour ses services technologiques, fondée en 1998 et désormais détenue par le groupe *Alphabet* depuis août 2015.

2.1 Présentation des données

Les données étudiées sont issus du site *Statista* (accessible ici) et couvrent les revenus de Google par trimestre, du premier trimestre de 2008 au second trimestre de 2023. Nous avons donc une série de $T = 62$ observations. Les données sont des valeurs entières exprimant les revenus en millions de dollars américains. Les données ont été téléchargées sous format *.xls* puis convertit en ficher *.csv* afin d'en faciliter le traitement.

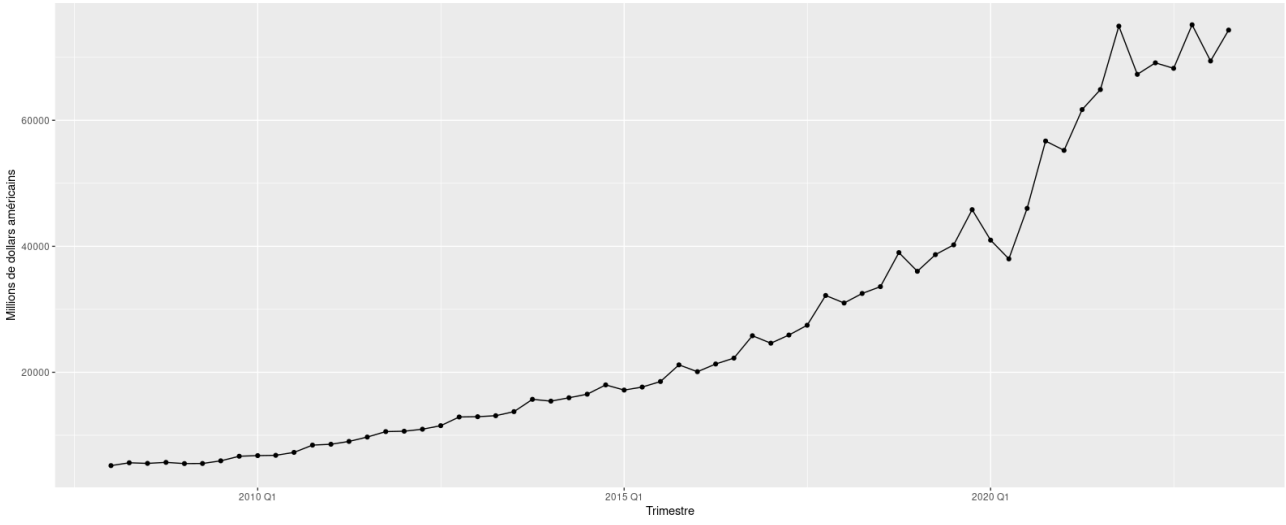


FIGURE 6 – Graphique des revenus trimestriels de Google (données initiales).

Les données semblent contenir une tendance croissance exponentielle ou polynômiale, ainsi

qu'une saisonnalité de période 4 (période d'un an). Nous observons un pique aux derniers trimestres de chaque année, probablement dû aux fêtes de Noël. En effet, plus de 80% des revenus de Google sont générées par la publicité et cette dernière joue un rôle considérable lors des achats effectués pour les fêtes de fin d'année [3]. De plus, nous constatons une croissance inhabituelle sur une durée de 2 ans à partir de 2020 (Figure 6). Cela est peut être une conséquence du Covid-19 et des confinements que le virus a engendré. En effet, les ménages sont limités dans leurs sorties et leurs voyages et passent beaucoup plus de temps à la maison, d'où une croissance des dépenses dans les produits technologiques et un temps plus important exposées aux publicités numériques. Enfin, une stagnation des revenus de Google est observée sur l'année 2022, possiblement en lien avec la guerre en Ukraine et la forte inflation.

2.2 Modèles de références

Avant tout, il est toujours intéressant de commencer par appliquer des modèles de référence simples pour effectuer des prédictions afin de voir s'ils permettent de modèliser la série correctement avant de se tourner vers des modèles plus complexes si necessaire. Les méthodes les plus simples, et parfois sous estimées notamment pour les prévisions à court terme, sont par exemple la méthode de la moyenne, les méthodes naïves (saisonnière ou non) ou encore la méthode de la dérive.

La méthode de la moyenne donne pour toute prévision future la moyenne des observations passées (7). Notons que la prévision à l'horizon h prend la même valeur pour toute valeur de h.

$$\forall h \in \mathbb{N}^* \quad \hat{X}_{T+h} = \frac{1}{T}\sum_{i=1}^{T} X_i \tag{7}$$

La méthode naïve consiste quant à elle à prendre pour valeur prédite la dernière valeur observée. En effet, en supposant que la dernière observation est faite à l'instant T, la prédiction à l'instant $T + h$ pour tout horizon h est $\hat{X}_{T+h} = X_T$. Il existe également une méthode naïve pour les séries saisonnières. La période d'une série saisonnière sera notée m. Dans ce cas, chaque valeur future estimée prend la valeur de la dernière valeur observée à la saison

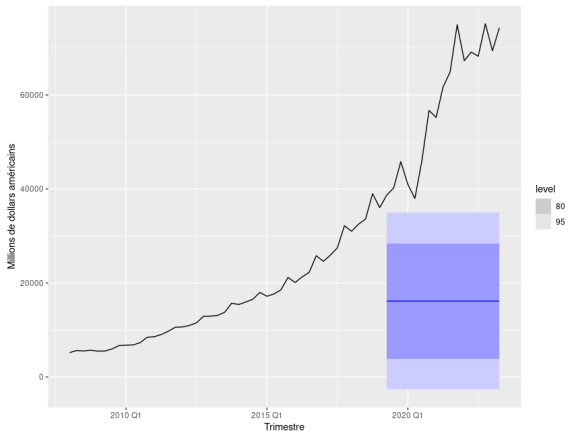
concernée (8). Par exemple, cette méthode prédira pour le dernier trimestre des années 2023, 2024, 2025, etc. le même revenu qu'au trimestre de l'année 2022.

$$\hat{X}_{T+h} = X_{T+h-m(\lfloor \frac{h-1}{m} \rfloor+1)} \tag{8}$$

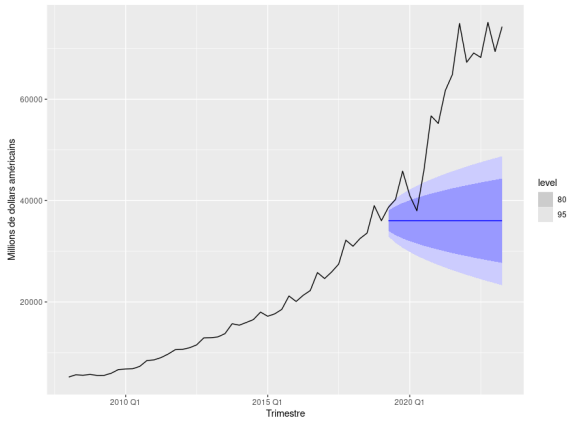
Nous avons alors ajusté ces 4 modèles à notre série et effectué des prévisions. Afin d'évaluer les prévisions, nous avons commencé par séparer le jeu de données en données d'entrainement ($\frac{3}{4}$ des données soit 45 observations, allant du premier trimestre de 2008 au premier trimestre de 2019) et en données de test (($\frac{1}{4}$ des données soit 17 observations, allant du second trimestre de 2019 au second trimestre de 2023).

Les prévisions ne coincident dans aucun des 4 cas avec les valeurs réelles de la série (Figure 8) et les valeurs réelles ne se trouvent même pas dans les intervalles de confiance des prévisions des différentes méthodes (Figure 7). En effet, la méthode de la moyenne et les méthodes naïves (saisonnière et non saisonnière) ne prennent pas en compte la tendance de la série (Figure 7a, 7c et 7b). La méthode de la dérive capture quant à elle la croissance de la série mais pas son caractère polynômial/exponentiel (Figure 7d). Aussi, la méthode naïve saisonnière est la seule à capturer la saisonnalité de la série (Figure 7c).

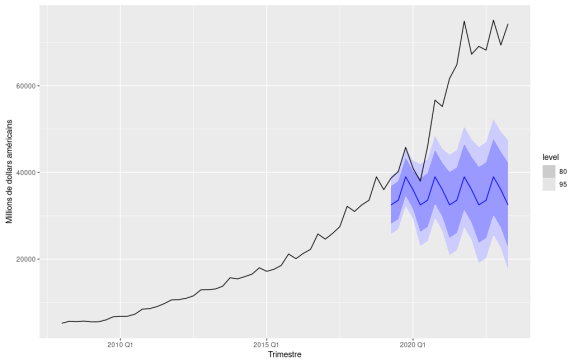
De manière général, ces modèles de référence ne sont pas performants sur le long terme, surtout lorsque la série est composée d'une tendance dans le cas de la moyenne et des modèles naïfs. La méthode de la dérive peut permettre de capturer une tendance mais linéaire uniquement et la méthode naïve saisonnière de capturer une saisonnalité lorsque que la série n'a pas de tendance. En somme, ce n'est pas le cas pour la série des revenus de Google. Il pourrait alors être intéressant d'appliquer une transformation logarithmique afin de rendre la tendance linéaire et/ou d'extraire les composantes tendanciels et saisonnières afin d'appliquer par exemple des méthodes de lissage, mais ce n'est pas le sujet de cette étude. Puisque ces modèles ne semblent pas performants dans notre étude, nous allons nous intéresser à des modèles plus complexes (SARIMA).



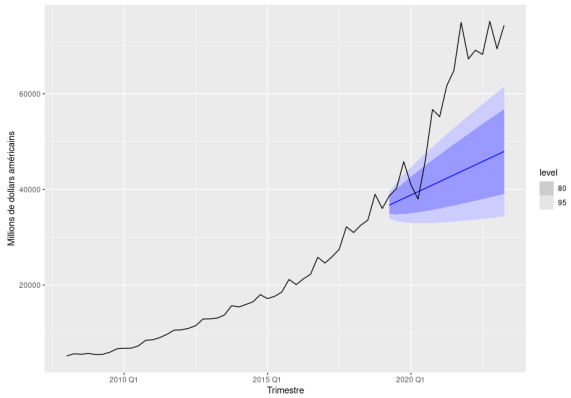
(a) Prévisions avec la méthode de la moyenne.



(b) Prévisions avec la méthode naïve.



(c) Prévisions avec la méthode naïve saisonnière.



(d) Prévisions avec la méthode de la dérive.

FIGURE 7 – Prévisions réalisées avec les 4 modèles de référence (moyenne, naïf, naïf saisonnier et dérive) et leurs intervalles de confiance à 80% (en bleu foncé) et 95% (en bleu clair).

2.3 Stationnarisation

2.3.1 Test de KPSS

Le Test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) est un test statistique de racine unitaire qui permet de tester si une série temporelle est stationnaire (hypothèse nulle). Le test tend à exprimer la série comme la somme d'une tendance déterministe, d'une marche aléatoire et d'une erreur stationnaire [2].

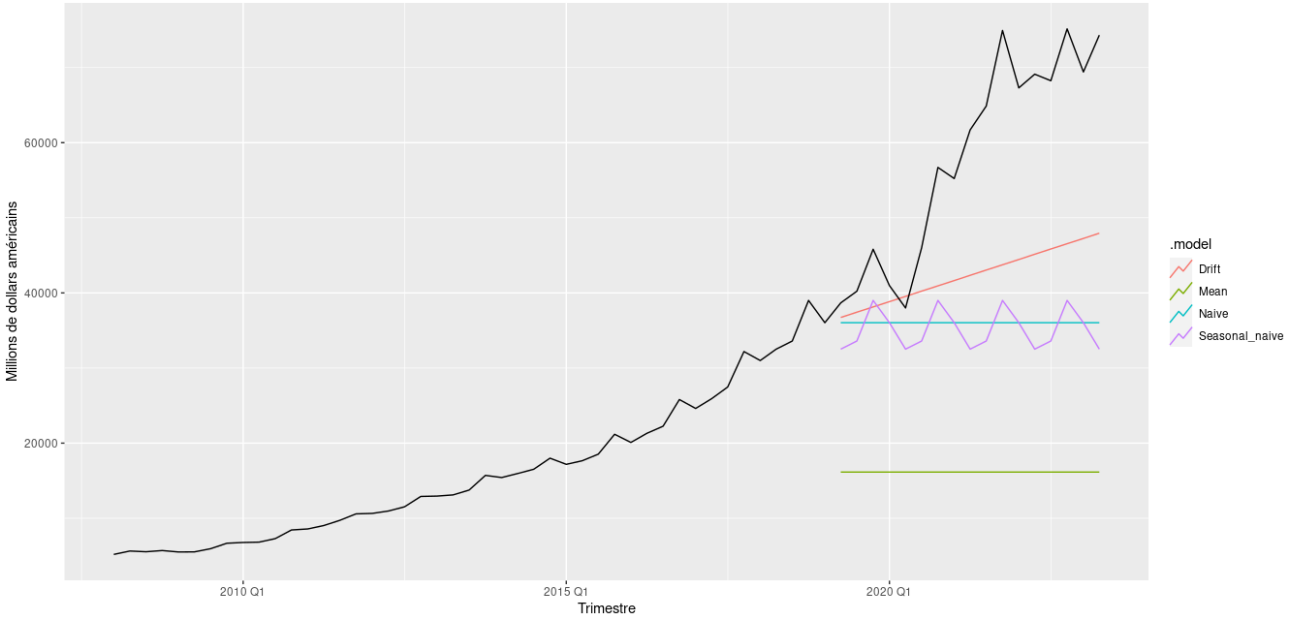


FIGURE 8 – Graphique des prévisions trimestriels de Google avec les 4 modèles de référence (moyenne en vert, naïf en bleu, naïf saisonnier en violet et dérive en rouge).

En R, la fonction *ur.kpss()* du package *urca* permet de réaliser le KPSS test et c'est cette dernière que nous utiliserons dans la suite.

2.3.2 Transformation logarithmique

Suite aux constats fait dans la partie 2.1, nous avons effectué une transformation logarithmique sur la série. C'est à cette nouvelle série que nous ferons référence dans la suite. En effet, cette transformation permet d'une part de rendre la tendance plus linéaire, d'autre part de stabiliser la variance, notamment au niveau des années 2020 et 2021 où une croissance particulière forte était observée. Notons qu'une chute inhabituelle des revenus de Google reste observable au second trimestre de l'année 2020.

La saisonnalité de période 4 reste, comme attendue, observable. De plus, nous pouvons observer la tendance à travers la fonction d'autocorrélation qui décroit lentement (Figure 10a). La série n'est donc pas encore stationnaire. Le test de KPSS confirme cela avec une statistique valant 1.6339, valeur supérieure à la valeur critique à 1% 0.739 (rejet de l'hypothèse nulle de stationnarité) (Figure 11).

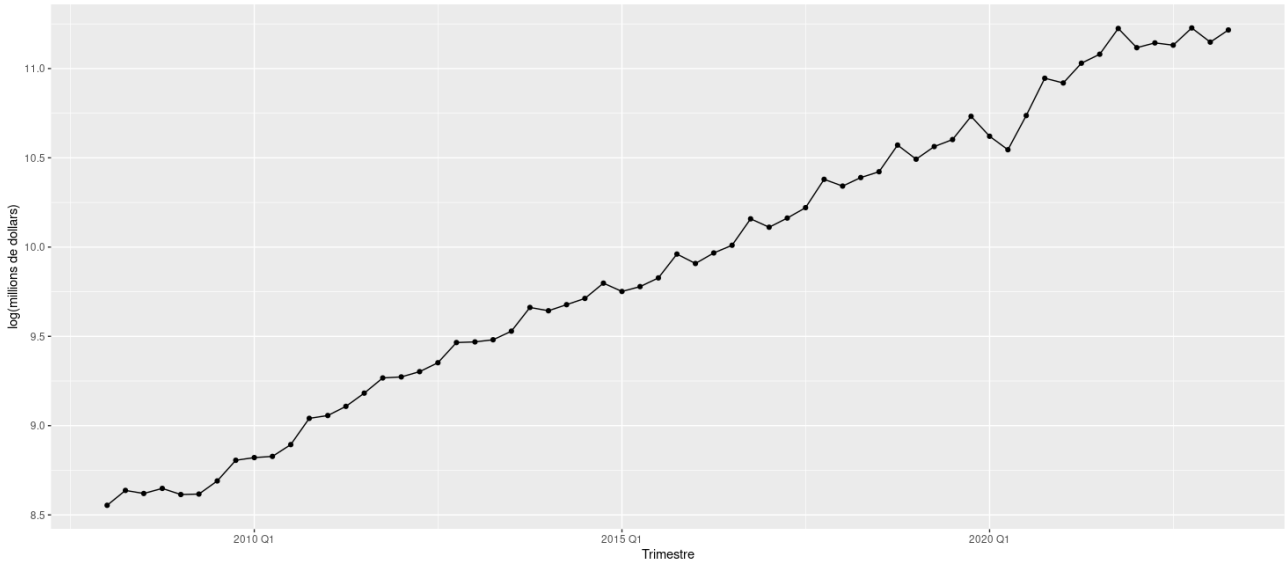
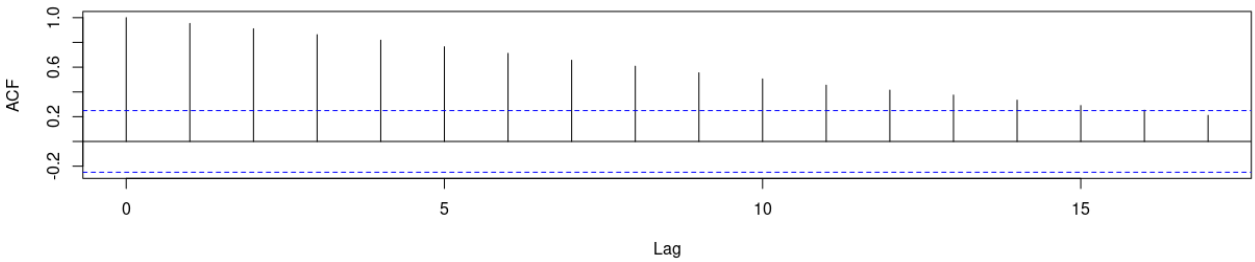
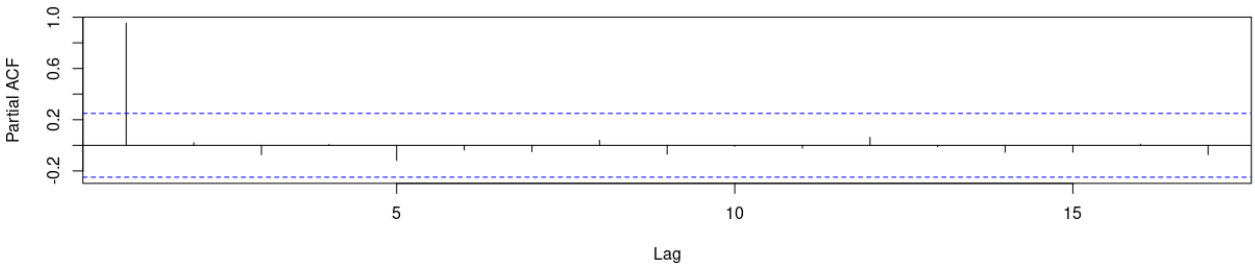


FIGURE 9 – Graphique de la série des logarithmiques des revenus de Google.



(a) Autocorrélations.



(b) Autocorrélations partielles.

FIGURE 10 – Autocorrélations (en haut) et autocorrélations partielles (en bas) de la série loarithmique.

```
########################
# KPSS Unit Root Test #
########################

Test is of type: mu with 3 lags.

Value of test-statistic is: 1.6339

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

FIGURE 11 – Statistique de KPSS calculée sur la série logarithmique.

2.3.3 Différenciation première

Afin d'éliminer la tendance, une différenciation première est d'abord effectuée. Autrement dit, pour chaque observation X_t de la série d'origine, la valeur précédente de la série X_{t-1} lui est soustraite (9).

$$\forall t \in [\![2\ ;\ T]\!]\quad Y_t = \nabla X_t = (I - B)X_t = X_t - X_{t-1} \tag{9}$$

où $(Y_t)_{t\in[\![2\ ;\ T]\!]}$ est la série différenciée. Remarquons que la série différenciée n'est plus constituée que de $T-1$ observations (une valeur de moins que la série initiale). En effet, la différenciation pour la première observation n'existe pas. Le graphique de la série différenciée est donné par Figure 12.

La série différenciée semble être de moyenne constante et de variance plus ou moins constante, à l'exception d'une variation plus élevée en 2021. Cependant, le graphique de la série semble toujours montrer la saisonnalité, avec des pics à intervalles réguliers. Cela est également observable sur les autocorrélations : seules celles d'ordre multiple de 4, la période, sont significativement non nulles (Figure 13). Sur le graphique des autocorrélations partielles de la série différenciée, nous constatons également une autocorrélation partielle significative d'ordre 4 (Figure 13b).

Puisque la saisonnalité n'a pas encore été éliminée de la série, la série différenciée n'est

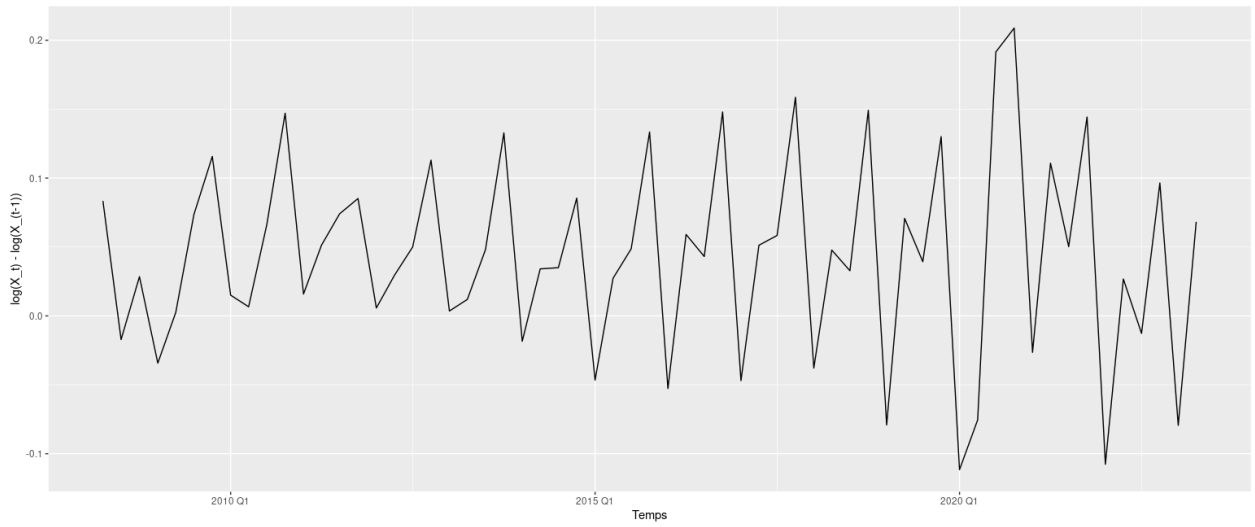
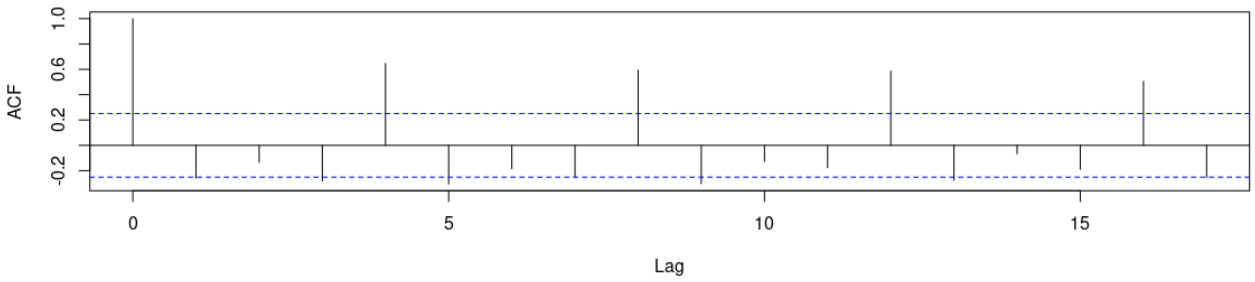
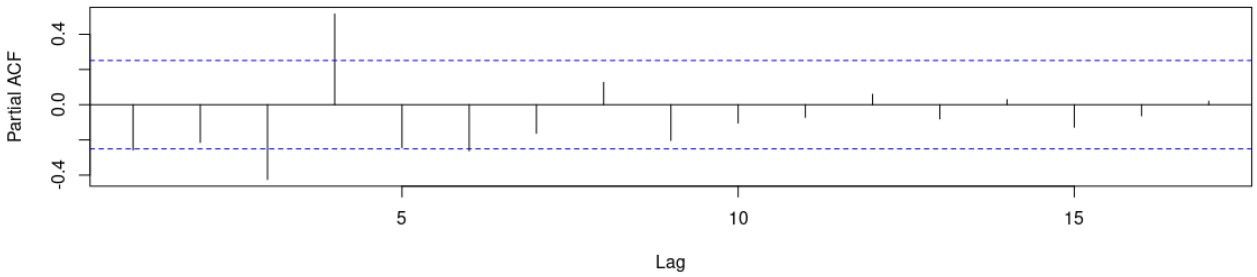


FIGURE 12 – Graphique de la différence première appliquée à la série logarithmique des revenus de Google.



(a) Autocorrélations.



(b) Autocorrélations partielles.

FIGURE 13 – Autocorrélations (en haut) et autocorrélations partielles (en bas) de la série différenciée.

pas stationnaire. Il est intéressant d'observer en revanche que le KPSS donne une valeur de statistique en dessous des valeurs critiques (à 1% mais aussi à 10%), ce qui indique que la

série est stationnaire (Figure 14). Plusieurs raisons peuvent être à l'origine de ce résultat, notamment le fait que la série est relativement courte (61 observations) ou encore le fait que le test KPSS s'intéresse surtout à la stationnarité tendancielle et peut alors parfois passer à côté de la non-stationnarité liée à la saisonnalité. Une alternative au test de KPSS peut alors être le test de Dickey-Fuller augmenté (ADF). Il s'éffectue en R grâce à la fonction *adf.test()* du package *tseries*. Nous avons alors appliqué ce test à la série différenciée et obtenons une p-valeur de 0.2074 supèrieur à la valeur 0.05. L'hypothèse nulle (stationnarité de la série) est donc rejetée et la série n'est pas stationnaire.

```
#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 3 lags.

Value of test-statistic is: 0.0658

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

FIGURE 14 – Statistique de KPSS calculée sur la série différenciée.

```
        Augmented Dickey-Fuller Test

data:  gr_diff$Revenue
Dickey-Fuller = -2.9084, Lag order = 3, p-value = 0.2074
alternative hypothesis: stationary
```

FIGURE 15 – Résultat du test de Dickey-Fuller Augmenté sur la série différenciée.

2.3.4 Différentiation saisonnière

Afin d'éliminer maintenant la saisonnalité de notre série, une différenciation saisonnière est effectuée. Il s'agit de retirer à chaque observation la valeur précédente de la même saison (10). Ici, on applique la différenciation saisonnière à la série différenciée (d'ordre 1) calculée

précedemment. La Figure 16 montre le graphique de la nouvelle série.

$$\forall t \in [\![m+1\ ;\ T]\!]\quad \nabla_m Y_t = (I - B^m)Y_t = Y_t - Y_{t-m} \tag{10}$$

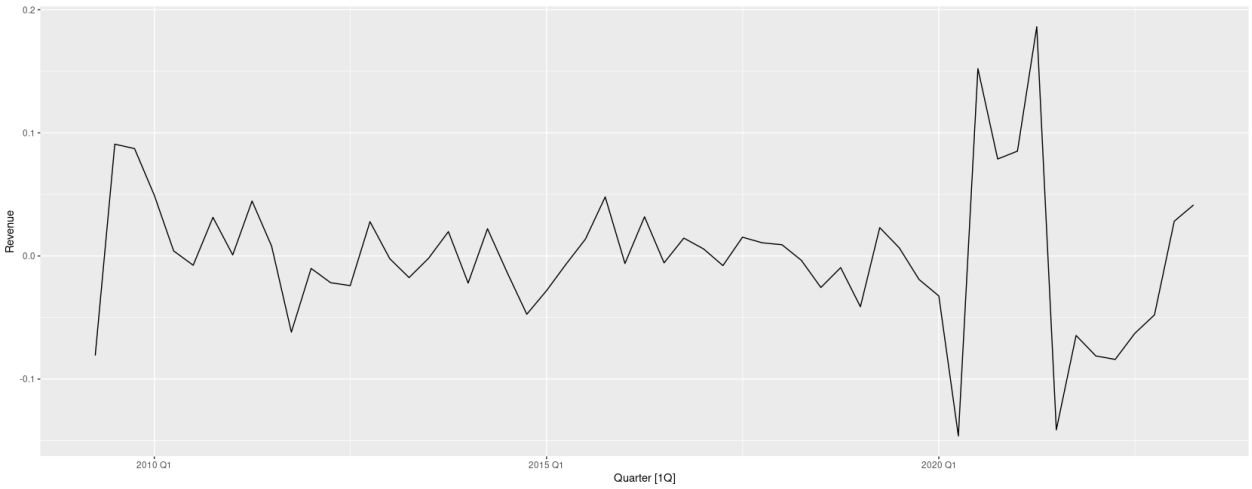


FIGURE 16 – Graphique de la différenciation saisonnière appliquée à la série différenciée des revenus de Google.

Graphiquement, nous voyons que la saisonnalité a disparu. La variance est relativement constante, encore une fois à l'exception de l'année 2021 où une variation plus élevée est observée. Le test de KPSS donne une statistique de 0.0846 qui est supérieur à la valeur critique à 1% 0.739 (Figure 17). Cela indique que la série est désormais stationnaire.

Dans la suite, on considèrera alors une differenciation première et une différenciation saisonnière (de période 4) pour l'estimation des modèles statistiques.

2.4 Choix des modèles

Dans cette partie, la démarche de recherche du meilleur modèle est présentée. Nous travaillerons sur la série logarithmique et, suite a l'analyse de la stationnarité qui précède, choisissons des modèles avec différence d'ordre 1 et différenciation saisonnière d'ordre 1 et de pèriode 4. Les graphiques des autocorrélations (ACF) et des autocorrélations partielles (PACF) obtenues (Figure 18) permettent de définir un premier modèle. La significativé des

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 3 lags.

Value of test-statistic is: 0.0846

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

FIGURE 17 – Statistique de KPSS calculée sur la série avec différeciation première et différenciation saisonnière.

coefficients du modèle, obtenue grâce à un test de Student (fonction *t_test()* du package *caschrono* en R), permet d'affiner petit à petit les paramètres de notre modèle afin d'obtenir finalement un modèle optimal. Enfin, nous comparons les performances de notre modèle avec celles du modèle donné en R par la fonction *auto.arima()* du package *forecast*.

2.4.1 Tests et critères d'évaluation des modèles

Avant tout, introduisons ici les tests des Ljung-Box et de Shapiro-Wilk, ainsi que le critère d'information d'Akaike (AIC) que nous utiliserons dans la suite afin d'analyser les résidus et évaluer les différents modèles.

Le test Ljung-Box est un test de type portmanteau permettant de tester si des variables aléatoires suivent une loi de bruit blanc. L'hypothèse nulle de ce test est que la statistique de test Q suit une loi χ^2_h. Le test de Ljung-Box est utilisé car il est plus efficace pour des échantillons de petite taille. En effet, la statistique utilisée a une distribution de probabilité mieux approchée par un χ^2 que la statistique de Box-Pierce. La statistique du test de Ljung-Box est définie par (11).

$$Q^*(h) = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \tag{11}$$

où h est le retard, n le nombre d'observations de la série et $\hat{\rho}_k$ l'autocorrélation empirique

d'ordre k. Ici, le test de Ljung-Box sera appliqué aux résidus afin de déterminer si ces derniers suivent une loi de bruit blanc. Sous l'hypothèse H_0, les résidus suivent une loi χ^2_{h-q} où q est le nombre de paramètres estimés par le modèle.

Le test de Shapiro-Wilk permet quant à lui de déterminer si des variables aléatoires suivent une loi normale. Nous utilisons ce test ici sur les résidus, après avoir constaté qu'ils n'étaient pas corrélés, dans le but de déterminer s'ils suivent une loi de bruit blanc gaussien. En R, le test de Shapiro-Wilk peut être réalisé à l'aide de la fonction *shapiro.test()* du package *stats*.

Enfin, l'AIC (Akaïve Information criterion) sera utilisé pour mesurer la qualité des modèles statistiques et comparer les modèles. L'AIC est défini par (12).

$$AIC = -2 \cdot log(L) + 2 \cdot (p + q + k + 1) \tag{12}$$

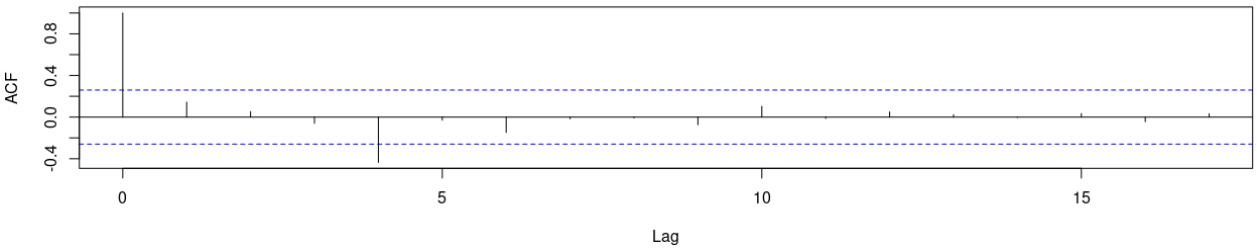
où L est la vraissemblance des données, p et q sont les ordres des parties autorégressive et moyenne mobile et k vaut 1 si une constante est introduite dans le modèle, 0 sinon.

2.4.2 Modèle 1 : ARIMA(0,1,0)(1,1,1)$_4$

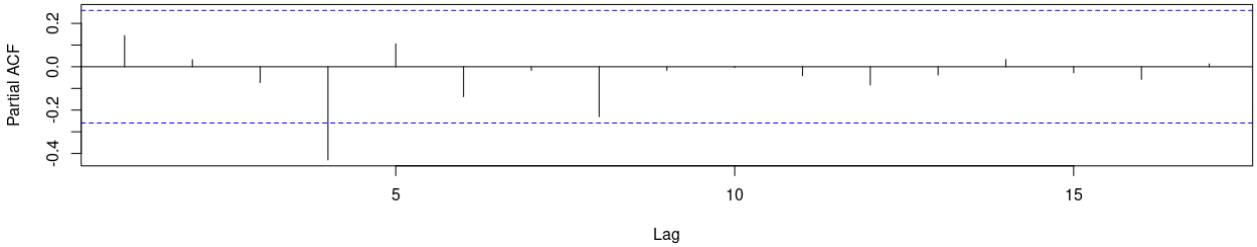
Pour le choix des paramètres du premier modèle, nous observons les graphiques des ACF et PACF (Figure 18). L'ordre des parties autorégressives sont à chercher sur le graphique des autocorrélations partielles, tandis que l'ordre des parties moyenne mobile s'observent sur le graphique des autocorrélations. Ici, nous constatons que seul l'autocorrélation partielle d'ordre 4 est significativement non nulle, d'où un ordre autorégressif nul et un ordre autorégressif saisonnier de 1. De même, seul l'autocorrélation d'ordre 4 semble significative, d'où le choix d'un ordre 0 pour la partie moyenne mobile et un ordre 1 pour la partie moyenne mobile saisonnière. Finalement, un modèle $ARIMA(0,1,0)(1,1,1)_4$ est ajusté.

Le modèle retenu est ajusté à série logarithmique et les coefficients obtenus sont donnés par la Figure 19. Le test de Student est appliqué à ces derniers afin de tester leur significativité est les résultats du test sont présentés par la Figure 20.

L'AIC vaut -155.91 (Figure 19). La p-valeur du coefficient autorégressif saisonnier d'ordre



(a) Autocorrélations.



(b) Autocorrélations partielles.

FIGURE 18 – Autocorrélations (en haut) et autocorrélations partielles (en bas) de la série avec différenciation première et différenciation saisonnière.

```
Call:
arima(x = google_train$Revenue, order = c(0, 1, 0), seasonal = list(order = c(1,
    1, 1), period = 4))

Coefficients:
         sar1     sma1
      0.1536  -0.6346
s.e.  0.3667   0.2605

sigma^2 estimated as 0.0009889:  log likelihood = 80.95,  aic = -155.91
```

FIGURE 19 – Coefficients obtenus pour l'ajustement du modèle ARIMA$(0,1,0)(1,1,1)_4$

```
> t_stat(model)
             sar1      sma1
t.stat 0.418894 -2.436166
p.val  0.675294  0.014844
```

FIGURE 20 – Résultats du test de Student appliqué aux coefficients du modèle 1.

1 (*sar1*), vaut 0.68, ce qui est supérieur au risque de 5%. L'hypothèse nulle ne peut donc pas être rejetée et le coefficient *sar1* n'est pas significatif. À l'inverse, le coefficient de moyenne

mobile saisonier d'ordre 1 (*sma1*) possède une p-valeur de 0.01 donc l'hypothèse nulle est rejetée pour ce coefficient, ce qui signifie que ce dernier est significatif. Nous considérons donc maintenant le modèle $ARIMA(0,1,0)(0,1,1)_4$.

2.4.3 Modèle 2 : ARIMA(0,1,0)(0,1,1)$_4$

Puisque la partie autorégressive saisonnière n'est pas significative dans le modèle précedent, cette dernière est retirée ici. Le nouveau modèle est ajusté et est donc un $ARIMA(0,1,0)(0,1,1)_4$. Il ne reste donc qu'un coefficient moyenne mobile saisonnier d'ordre 1 et ce dernier est donné Figure 21. La valeur de l'AIC est de -157.72, ce qui est indique une amélioration du modèle par rapport au modèle précédent (modèle 1).

```
Call:
arima(x = google_train$Revenue, order = c(0, 1, 0), seasonal = list(order = c(0,
    1, 1), period = 4))

Coefficients:
         sma1
      -0.5490
s.e.   0.1832

sigma^2 estimated as 0.000991:  log likelihood = 80.86,  aic = -157.72
```

FIGURE 21 – Coefficients obtenus pour l'ajustement du modèle $ARIMA(0,1,0)(0,1,1)_4$.

```
> t_stat(model2) # Coefficient significatif
             sma1
t.stat -2.997558
p.val   0.002722
```

FIGURE 22 – Résultats du test de Student appliqué au coefficient du modèle 2.

Le résultat du test de Student appliqué au coefficient est dans la Figure 22. La p-valeur pour le coefficient moyenne mobile saisonnier d'ordre 1 (*sma1*) est de 0.00. L'hypothèse nulle de non significativité du coefficient est ainsi rejetée, ce qui signifie que le coefficient est significatif.

Maintenant que le coefficient du modèle est significatif, la non-corrélation des résidus doit être vérifiée. Pour cela, le test de portmanteau avec la statistique de Ljung-Box est appliqué

```
> ljung_box(model2$residuals)
  lb_stat lb_pvalue
0.7258882 0.3942192
```

FIGURE 23 – Résultat du test de Ljung-Box sur les résidus du modèle 2.

```
        Shapiro-Wilk normality test

data:  model2$residuals
W = 0.96932, p-value = 0.2736
```

FIGURE 24 – Résultat du test de Shapiro-Wilk sur les résidus du modèle 2.

aux résidus du modèle 2. Ses résultats sont donnés par la Figure 23. La statistique de test est de 0.73 et la p-valeur associée est de 0.39, ce qui est supérieur à la valeur 0.05. Ainsi, l'hypothèse nulle est acceptée avec un risque de 5% de se tromper. Cela signifie que les résidus sont non corrélés.

Nous appliquons ensuite le test de Shapiro-Wilk afin de vérifier que les résidus sont gaussiens. Les résultats sont donnés par la Figure 24. La p-valeur associée à la statistique de test est de 0.27, ce qui est supérieur à la valeur 0.05. L'hypothèse nulle (les résidus suivent une loi de bruit blanc gaussien) est donc acceptée avec un risque d'erreur de 5%.

Finalement, le coefficient du modèle obtenu est significatif et les résidus sont non corrélés et sont des bruits blancs gaussiens. Ce modèle étant satisfaisant, une comparaison avec le modèle proposé par la fonction *auto.arima()* du package *forecast* est effectuée dans la partie qui suit.

2.4.4 Modèle 3 : ARIMA(2,0,1)(0,1,1)$_4$ (avec *auto.arima()*)

La fonction *auto.arima* détermine automatiquement les meilleurs paramètres pour ajuster un modèle ARIMA à une série donnée. Toutefois, ce n'est pas toujours le meilleur modèle parmi tous les modèles et il n'est donc pas impossible d'obtenir un meilleur modèle avec d'autres paramètres. La fonction est appliquée sur la série logarithmique des revenus de Google (sans différences) et les résultats obtenus sont présentés sur la Figure 25. La fonction propose un modèle $ARIMA(2,0,1)(0,1,1)_4$. Nous notons que l'AIC est de -162.89, ce qui

```
Series: google_train
ARIMA(2,0,0)(0,1,1)[4] with drift

Coefficients:
         ar1      ar2     sma1   drift
      1.1482  -0.4003  -0.4367  0.0456
s.e.  0.1588   0.1884   0.1826  0.0027

sigma^2 = 0.0009108:  log likelihood = 86.45
AIC=-162.89   AICc=-161.18   BIC=-154.32
```

FIGURE 25 – Résultat de la fonction auto.arima appliquée sur la série logarithmique des revenus de Google.

```
> t_stat(model3)
             ar1       ar2       sma1    drift
t.stat 7.232567 -2.124477 -2.391248 17.05561
p.val  0.000000  0.033630  0.016791  0.00000
```

FIGURE 26 – Résultat des tests de Student appliqués aux coefficients du modèle 3.

est meilleur que l'AIC de notre modèle 2 (-157.72). Aussi, la Figure 26 montre que tous les coefficients sont significatifs (test de Student).

Le test de Ljung-Box donne une p-valeur de 0.54, ce qui est supèrieur à la valeur 5% donc l'hypothèse nulle est acceptée et les résidus sont non-corrélés (Figure 27). Aussi, Le test de Shapiro-Wilk donne une p-valeur de 0.15 supèrieur à 5%. Ainsi, nous pouvons affirmer que les résidus suivent une loi de bruit blanc gaussien (Figure 28).

```
> ljung_box(model3$residuals)
  lb_stat lb_pvalue
0.3710189 0.5424494
```

FIGURE 27 – Résultat du test de Ljung-Box sur les résidus du modèle 3.

Finalement, le modèle proposé par la fonction *auto.arima()* semble meilleur que le modèle 2 que nous avons obtenu. En effet, son AIC est meilleur. De plus, tous les coefficients du modèle sont significatifs et les résidus sont non-corrélés et suivent une loi de bruit blanc gaussien. Notons que ce modèle ne réalise pas de différenciation première contrairement à ce

```
        Shapiro-Wilk normality test

data:  model3$residuals
W = 0.96232, p-value = 0.1497
```

FIGURE 28 – Résultat du test de Shapiro-Wilk sur les résidus du modèle 3.

que nous avons fait.

2.5 Prévisions

Le modèle 2 et le modèle 3 sont alors utilisés pour réaliser des prévisions sur la période du deuxième trimestre de 2019 au deuxième trimestre de 2023 (série de test). Rappelons que la série a été tranformée en appliquant le logarithme. Les prévisions obtenues sont affichées sur la Figure 29. Les prévisions données par les deux modèles ont la même allure et sont quasiment identiques. Nous remarquons que la saisonnalité est bien capturée par les deux modèles. Cependant, la forte tendance inhabituelle sur les années 2020 et 2021 est entraîne un écart entre les prévisions et la série réelle (logarithmique).

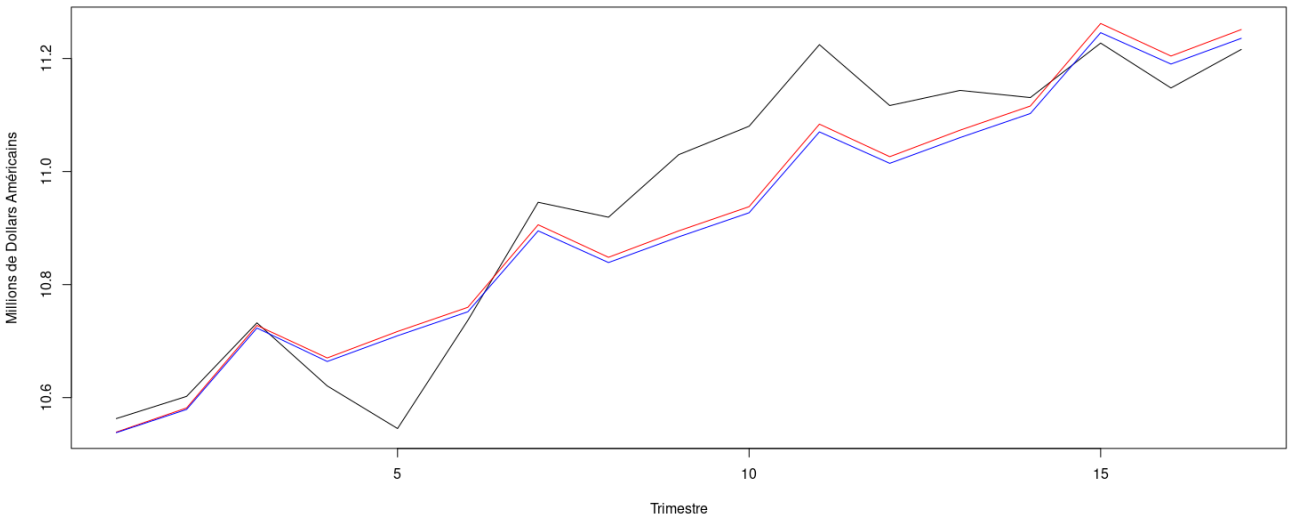


FIGURE 29 – Prévisions obtenues sur la période T2 2019 - T2 2023 avec le modèle 2 (en rouge) et le modèle 3 (en bleu) (série logarithmique).

Enfin, nous avons transformé les prévisions des deux modèles en appliquant la fonction

exponentielle à chacune des valeurs. En effet, nous avions effectué une transformation logarithmique sur les données afin de rendre la tendance linéaire et stabiliser la variance et devons donc repasser aux valeurs réelles. La Figure 30 donne les prévisions finales sur la série réelle initiale en millions de dollars américains.

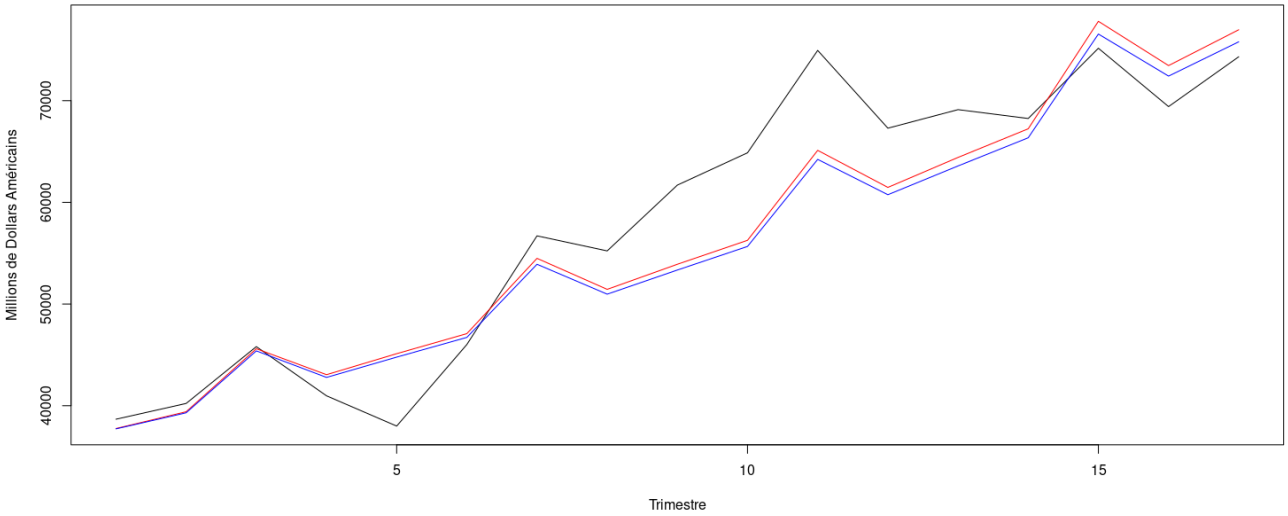


FIGURE 30 – Prévisions obtenues sur la période T2 2019 - T2 2023 avec le modèle 2 (en rouge) et le modèle 3 (en bleu) des revenus de Google en millions de dollars américains.

L'erreur moyenne absolue (MAE) a été calculé pour les deux modèles sur la période de test et ces dernières sont les mêmes à 10^{-2} prêt : 194.1485 pour le modèle 2 et 194.1604 pour le modèle 3. Finalement, les performances des deux modèles sont très similaires et le modèle 3 obtenus par la fonciton *auto.arima()* n'est pas significativement meilleur que le modèle 2 que nous avons trouvé.

Conclusion

En conclusion, nous avons commencé ce rapport par une rapide partie introductive des modèles ARMA avec une simulation en R. Surtout, nous nous sommes intéréssé aux revenus de Google, entreprise qui joue un rôle clé dans le quotidien du monde entier et sur le marché des services et des technologies. C'est pour cette raison qu'il est enrichissant de se pencher sur ses revenus, mais surtout de pouvoir modéliser la série afin de réaliser des estimations pour le futur de l'entreprise. En effet, les modèles que nous avons obtenus semblent bien modéliser la série et pourraient être utilisés dans ce but. Soulignons tout de même que les prévisions de nos modèles sont sensibles aux perturbations exceptionnelles comme nous pouvons l'observer avec la forte croissance des revenus de l'entreprise en 2020 et 2021.

Comme complément à cette étude, il pourrait être intéressant de tenter de modéliser la série par d'autres méthodes, notamment des méthodes de lissage exponentielle.

Références

[1] JJ Daudin, C Duby, S Robin, and P Trécourt. Analyse de séries chronologiques. *INA-PG, Mathématiques*, 1996.

[2] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root : How sure are we that economic time series have a unit root ? *Journal of econometrics*, 54(1-3) :159–178, 1992.

[3] Statista (site web). Distribution of google segment revenues from 2017 to 2022, Février 2023. Consulté le : 01/11/2023.

[4] Herman Wold. A study in the analysis of stationary time series, 1938.