Compte-rendu du TP 1

Séries temporelles - UKgas

Professeur : Grégory SOLER

EL KHAMLICHI Badreddine
EL KHALFIOUI Nadir

Polytech Lyon
MAM 4A
15/06/2023

Table des matières

1 Introduction

Nous étudions ici la série UKgas qui représente la consommation de gaz au Royaume-Uni de 1961 à 1986. Cette consommation est suivie trimestriellement, ce qui en fait une série de saisonnalité d'ordre 4. En analysant les données trimestrielles, nous pourrons identifier les motifs saisonniers récurrents et comprendre comment la consommation de gaz fluctue au fil des saisons.

Pour analyser et modéliser la série "UKgas", nous explorerons différentes techniques adaptées aux séries temporelles. Tout d'abord, nous utiliserons des méthodes graphiques pour visualiser les données et identifier les tendances, les schémas saisonniers et les éventuelles anomalies. Nous examinerons également des outils statistiques tels que la décomposition de la série en composantes (tendance, saisonnalité et résidus).

Ensuite, nous mettrons en œuvre des modèles de séries temporelles pour capturer les motifs et les tendances observés dans la série "UKgas". De plus, nous pourrions également explorer des modèles plus avancés, tels que les modèles SARIMA, pour mieux capturer les caractéristiques complexes et les non-linéarités potentielles de la série. Afin de mieux comprendre cette série, nous allons exploiter différents modèles et méthodes comme :

1. Lissage exponnentiel
2. Régression linéaire
3. SARIMA

L'objectif final de cette analyse de série temporelle sera de développer un modèle prédictif fiable pour estimer la consommation future de gaz au Royaume-Uni.

2 Description de la série

La série UKgas répertorie ainsi la consommation de gaz (en therms) au Royaume-Uni entre 1960 et 1986. Ces données sont récoltées trimestriellement, découpant ainsi chaque année en période de 3 mois chacune et donnant ainsi 4 périodes par an, pour un total de 108 récoltes de données.

Nous allons commencer par afficher le graphe de la série mais également celui de la série passée au log :

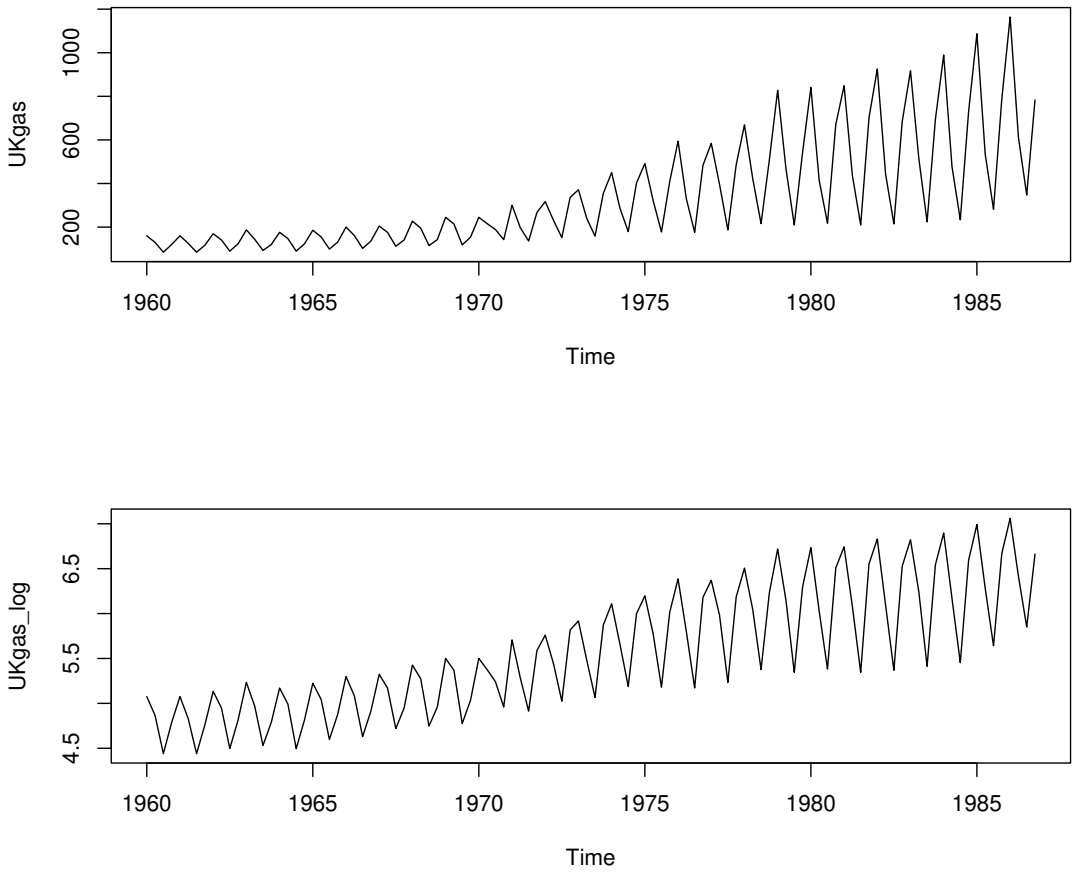

FIGURE 1 – Chronographe de UKgas et log(UKgas)

Tandis que le premier modèle est basé sur un modèle multiplicatif de la forme :

$$X_t = T_t * S_t * \varepsilon_t$$

nous passons la série au log afin d'obtenir un modèle additif de la forme suivante :

$$Y_t = log(T_t) + log(S_t) + log(\varepsilon_t)$$

Ce passage au logarithme permet de linéariser la série temporelle et de passer sur un modèle additif. On peut ainsi isoler une certaine tendance linéaire, la saisonnalité mais également des résidus.

2.1 Analyse de la décomposition UKgas

Nous utilisons ici la fonctione *decompose* de R qui permet de séparer la composante saisonnière, la tendance et les résidus de la série UKgas. Après application de cette fonction, nous obtenons les résultats suivants :

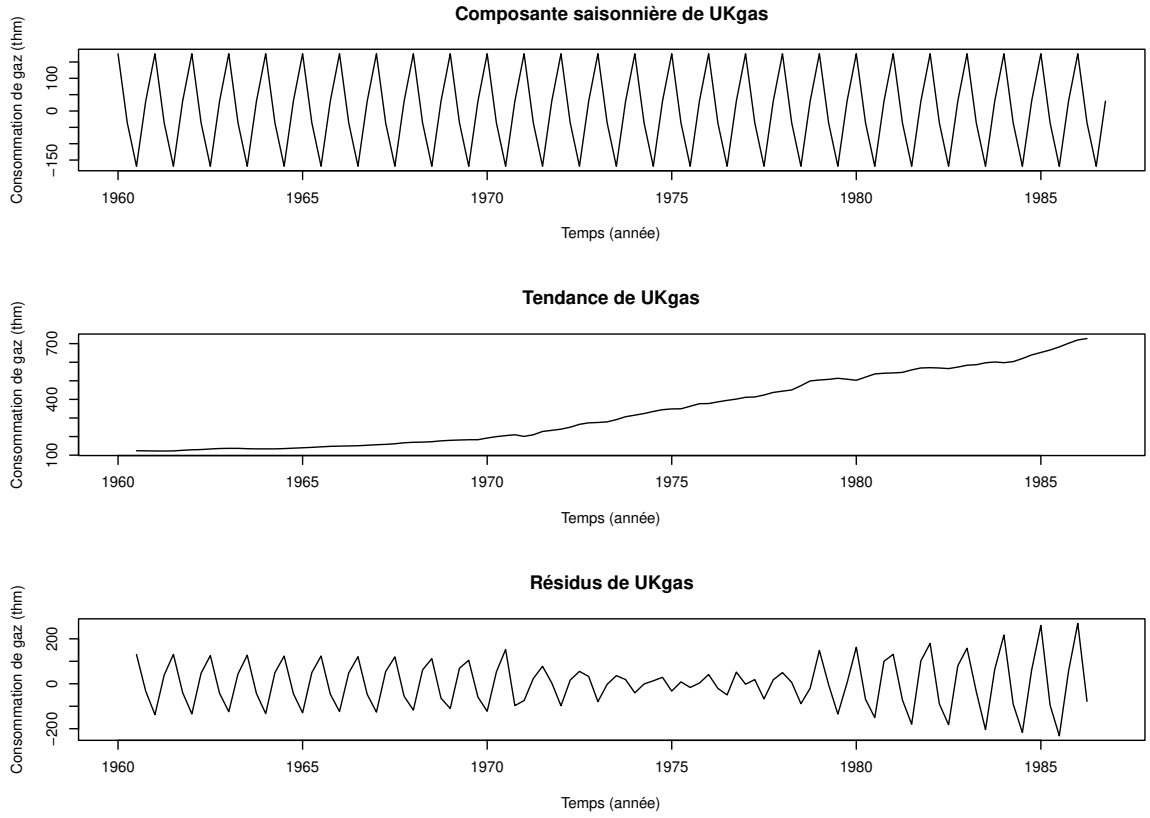

FIGURE 2 – Décomposition de la série UKgas en saisonnalité, tendance et résidus

On remarque visuellement que la tendance n'est pas réellement linéaire, d'où la nécessité de passer la série au log. La tendance est croissante tandis que la saisonnalité est annuelle. Cela peut être dûe au fait que les Anglais aient besoin de se chauffer en hiver car il fait froid, contrairement à l'été.

Afin de procéder à la prévision de la série, nous allons isoler d'un côté la série privée des données de la dernière année, que nous appellerons $UKgas\_train$, qui fera office de série d'entraînement. Grâce à cette dernière, nous établirons des modèles prévisionnels que nous utiliserons puis comparerons avec la série de test que nous appellerons $UKgas\_test$.

Différentes méthodes verront le jour à travers ce projet :

1. Lissage exponentiel
2. Régression linéaire
3. SARIMA selon la méthodologie de Box-Jenkins vue en cours

3 Modélisation par lissage exponentiel

3.1 Méthode du lissage exponentiel : explications

Cette méthode suppose que notre série peut être approximée par une variable a_t. Nous allons chercher à résoudre l'équation suivante :

$$\min_c \sum_{i=1}^{T-1} a^i (X_{T-i} - c)^2 = 1$$

La série X_T peut alors s'écrire de la manière suivante :

$$X_T = c = (1-a) \sum_{t=0}^{T-1} a^t X_{T-t} \text{ avec } 0 < a < 1$$

3.2 Comparaison des trois méthodes de prévision

	LES	LED	HW	HWM	EtsAuto
RMSE	320.86	307.7	41.85	52.31	53.26
MAPE	0.366	0.383	0.065	0.091	0.093

Après avoir analysé les résultats des trois méthodes de prévision utilisées (LES, LED et HW), nous avons observé des différences significatives en termes de précision. Les valeurs RMSE (Root Mean Square Error) et MAPE (Mean Absolute Percentage Error) ont été calculées pour évaluer les performances de chaque méthode.

Rappelons que le but de la modélisaion est de trouver un modèle précitif minimisant les valeurs du RMSE et du MAPE.

Parmi les trois méthodes testées, la méthode HW (Holt-Winters) est celle qui a donné les résultats les plus satisfaisants. Elle a obtenu le RMSE le plus faible (41.85) mais également le MAPE le plus bas (0.065), indiquant ainsi une meilleure précision dans la prédiction des valeurs futures. Cette méthode a également été utilisée pour effectuer une prévision à l'horizon de 4 ans, comme illustré dans la Figure 1. On peut observer que les prévisions de la méthode HW suivent de près les valeurs réelles, démontrant ainsi sa capacité à capturer les tendances et les variations saisonnières de la série temporelle.

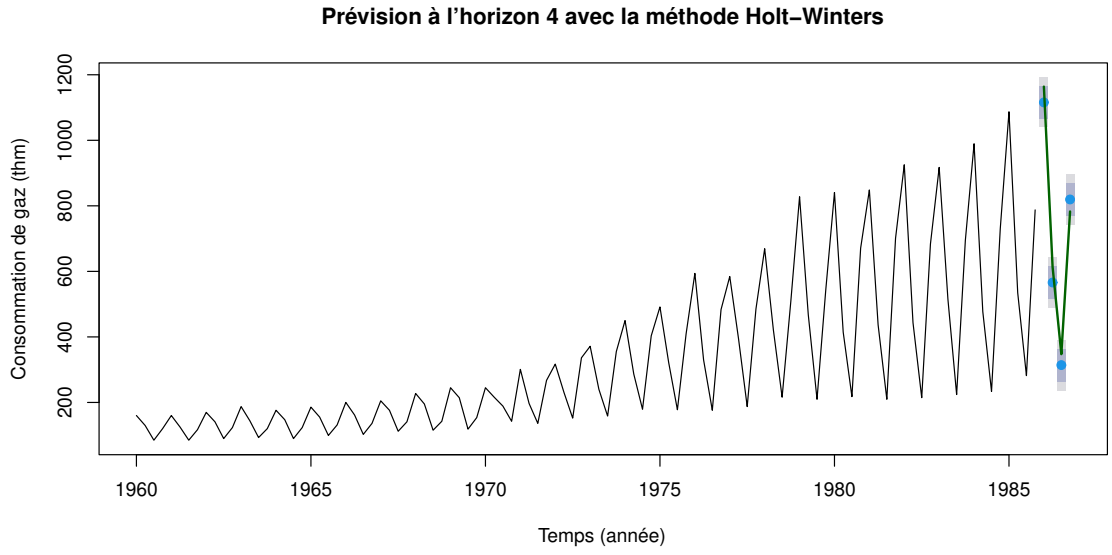

FIGURE 3 – Prévision à l'horizon 4 avec Holt-Winters

3.3 Prédiction

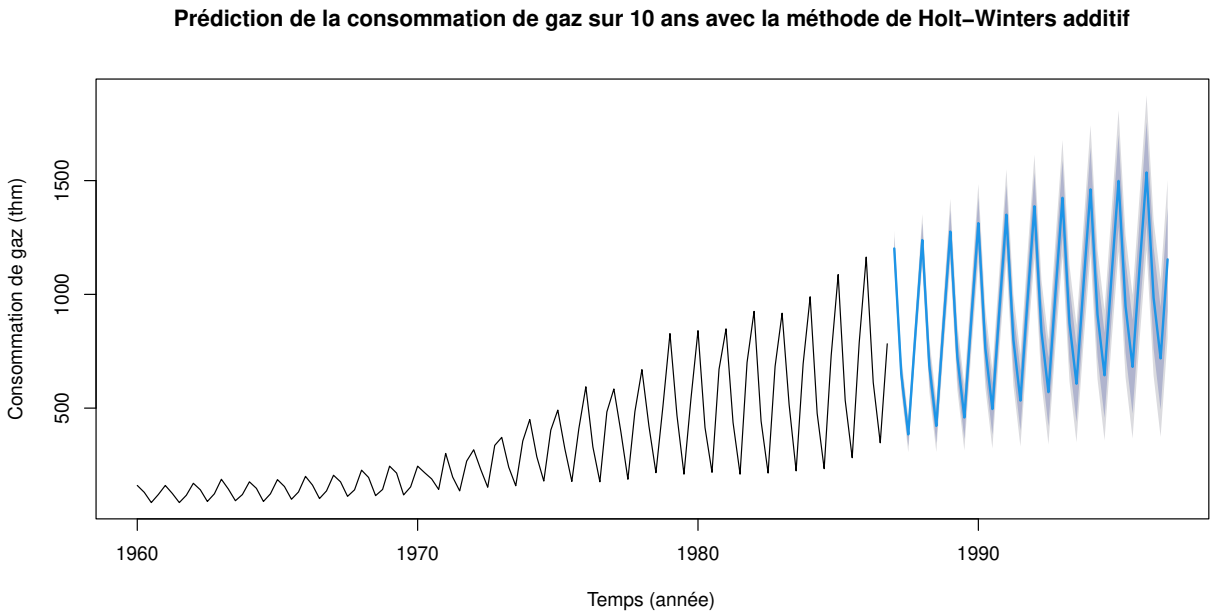

FIGURE 4 – Prédiction sur 10 ans avec Holt-Winters additif

En utilisant la méthode HW (Holt-Winters) avec une approche additive, nous avons également réalisé une prédiction sur une période de 10 ans, comme présenté dans la Figure 4. Cette prédiction offre une vision à long terme de l'évolution de la série temporelle. Cependant, il est important de noter que les prédictions à plus long terme sont soumises à une plus grande incertitude, car elles sont plus sensibles aux changements et aux influences externes sur une période prolongée.

En conclusion, la méthode de lissage exponentiel avec la méthode HW (Holt-Winters) s'est révélée être la plus performante pour la prédiction de la série temporelle étudiée. Cependant, il est essentiel de prendre en compte les limites et les incertitudes liées à toute méthode de prévision, en particulier lorsqu'on se projette sur des horizons temporels plus longs.

4 Modélisation par régression linéaire

L'objectif de cette partie est de modéliser et de prévoir la série temporelle "UKgas" en utilisant la régression linéaire. Pour capturer au mieux les variations de la série, nous appliquons le logarithme à celle-ci. De plus, nous divisons les données en ensembles d'entraînement et de test.

Nous commençons par analyser la tendance et la saisonnalité de la série en utilisant des fonctions trigonométriques. Nous estimons la tendance en créant une séquence linéaire de points représentée par la variable "Trend". Ensuite, nous déterminons la composante saisonnière en utilisant des fréquences cycliques de cosinus et de sinus calculées à partir de la tendance. Cela nous permet d'obtenir la matrice de régresseurs "Regresseur" en combinant la tendance et la composante saisonnière.

4.1 Modèle automatique avec lm

Nous commençons par ajuster un modèle de régression linéaire automatique en utilisant la fonction "lm" :

$$fit\_lm <- lm(UKgas\_trainLog\ ., data = Regresseur)$$

Les mesures d'information AIC et BIC associées à ce modèle sont respectivement de -44.12304 et -28.25669.

Afin d'évaluer la pertinence d'un modèle de régression linéaire ou d'un modèle autorégressif, il est essentiel d'analyser les résidus afin de s'assurer qu'ils présentent les caractéristiques d'un bruit blanc, c'est-à-dire qu'ils sont indépendants et sans autocorrélation. Pour ce faire, nous utilisons différentes analyses graphiques telles que le lagplot, l'ACF (fonction d'autocorrélation) et le PACF (fonction d'autocorrélation partielle).

Nous commençons par examiner le lagplot du modèle :

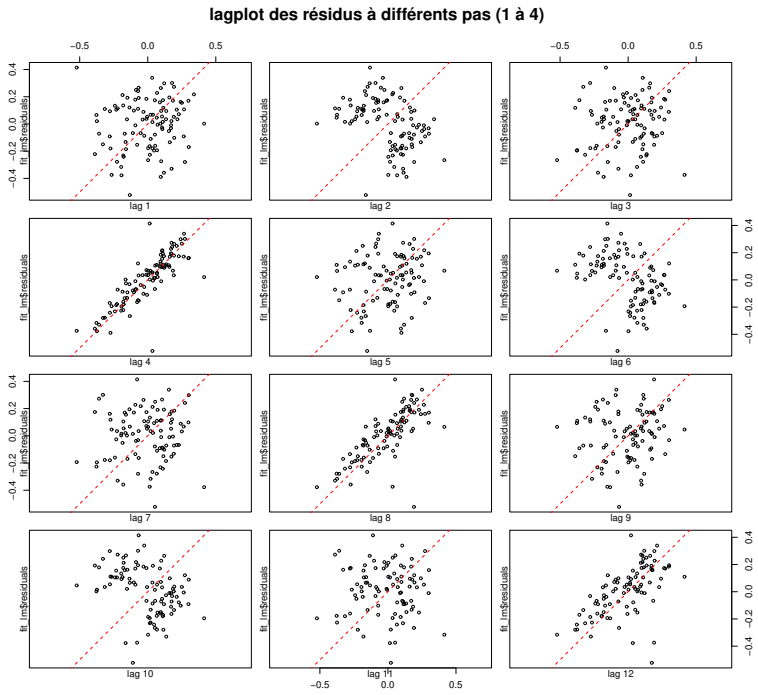

FIGURE 5 – Lagplot pour le modèle de régression linéaire auto

Le lagplot permet de visualiser la corrélation entre les résidus successifs du modèle à différents décalages (lags). Si les résidus sont des bruits blancs, nous nous attendons à ce qu'il n'y ait aucune corrélation significative entre les résidus à des décalages différents. Ainsi, si le lagplot révèle une corrélation significative à certains lags, cela suggère que le modèle présente une autocorrélation dans les résidus, ce qui peut indiquer que le modèle ne capture pas complètement les motifs temporels dans les données. Ici nous observons que les résidus sont corrélés pour tous les lags modulo 4. Nous n'avons donc pas une décorrélation des résidus à tous les décalage. Cela va nous suggéré également d'émettre l'hypothèse d'utilisation d'un modèle autorégressif d'ordre 4 (AR(4)).

Nous affichons également les ACF et PACF de notre modélisation :

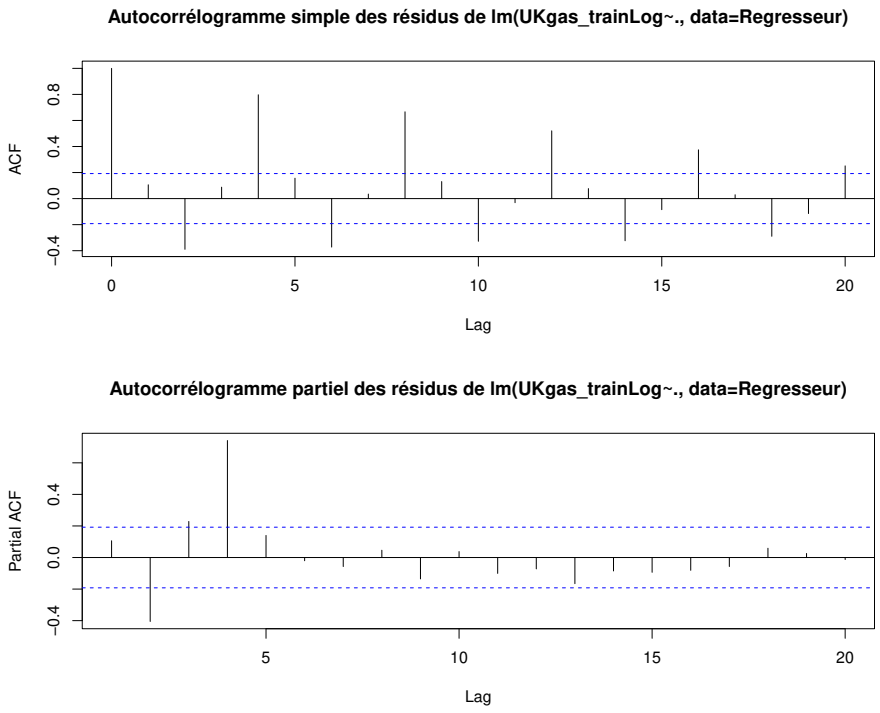

FIGURE 6 – ACF et PACF des résidus du modèle de régression linéaire automatique

Ensuite, nous utilisons l'ACF (fonction d'autocorrélation) pour mesurer la corrélation linéaire entre les résidus à différents lags. L'ACF nous permet d'évaluer l'autocorrélation des résidus, c'est-à-dire la corrélation entre un résidu et les résidus décalés dans le temps. Si les résidus sont véritablement des bruits blancs, nous nous attendons à ce que l'ACF soit proche de zéro pour tous les lags, ce qui indiquerait l'absence d'autocorrélation. Cependant, si l'ACF présente des valeurs significativement différentes de zéro pour certains lags, cela suggère une autocorrélation dans les résidus, ce qui indique la présence de motifs temporels non capturés par le modèle.

De même, nous utilisons le PACF (fonction d'autocorrélation partielle) pour mesurer la corrélation linéaire entre les résidus à différents lags, tout en éliminant l'influence des lags intermédiaires. Le PACF nous permet de détecter les retards spécifiques qui sont significativement corrélés avec les résidus, ce qui est essentiel pour identifier les motifs temporels importants. Si le PACF montre une corrélation significative uniquement pour un retard spécifique, cela suggère que ce retard spécifique est important pour expliquer les motifs temporels dans les données.

Dans notre cas, en examinant le PACF de notre modèle, nous remarquons que le PACF s'annule à r = 4, ce qui indique une corrélation significative uniquement pour un retard de 4. Cette observation confirme l'idée d'utiliser un modèle autorégressif d'ordre 4 (AR(4)), où les quatre valeurs précédentes de la série temporelle sont utilisées comme variables explicatives. Cette approche permet de capturer les motifs temporels importants et d'améliorer la performance de notre modèle de régression linéaire.

On a également ici fais des test statistique (test du port manteau) et qqplot des résidus, et ils satisfont pas tous l'hyothèse de non autocorrélation, de tout manière on a vu avec les prédénts tests graphuique qu'on peut songer à un meilleur modèle, celui avec un AR(4).

Nous présentons également la prévision réalisée avec le modèle de régression linéaire :

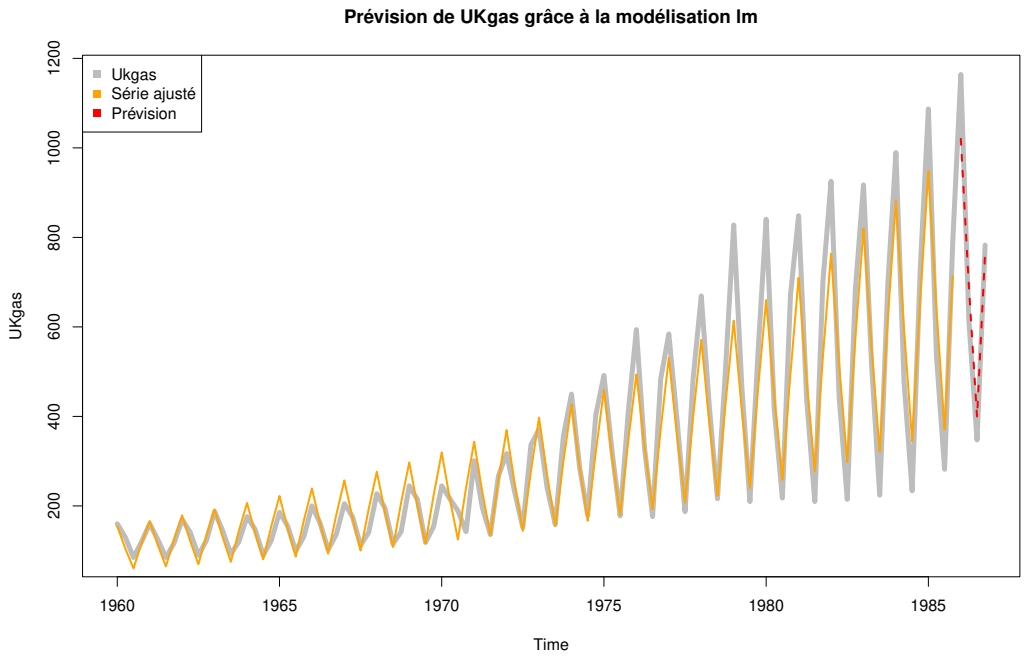

FIGURE 7 – Graphique d'ajustement et prévision de la série UKgas avec modèle de RL auto

Nous constatons que la prévision ne correspond pas parfaitement aux données observées. On peut donc à présent passée à une nouvelle modélisation qui nous a été suggéré avec l'étude de notre modélisation actuelle.

4.2 Choix d'un modèle de régression avec un AR(4)

Sur la base des résultats précédents, nous décidons de modéliser la série en utilisant un modèle autorégressif d'ordre 4 (AR(4)) avec les mêmes variables explicatives que dans le modèle de régression linéaire précédent. Pour ajuster ce modèle, nous utilisons la fonction "Arima()" avec la commande suivante :

$$fit\_ar4 <- Arima(UKgas\_trainLog, order = c(4,0,0), xreg = as.matrix(Regressors))$$

Les mesures d'information AIC et BIC associées à ce modèle sont respectivement de -158.8564 et -132.4125. On voit qu'on a bien minimiser ces indicateurs par rapport à nos AIC et BIC du précédent modèle lm.

Afin de valider la pertinence de ce modèle, nous effectuons les mêmes analyses que précédemment. Tout d'abord, nous examinons le lagplot des résidus du modèle AR(4). Le lagplot nous permet d'évaluer l'autocorrélation des résidus à différents retards :

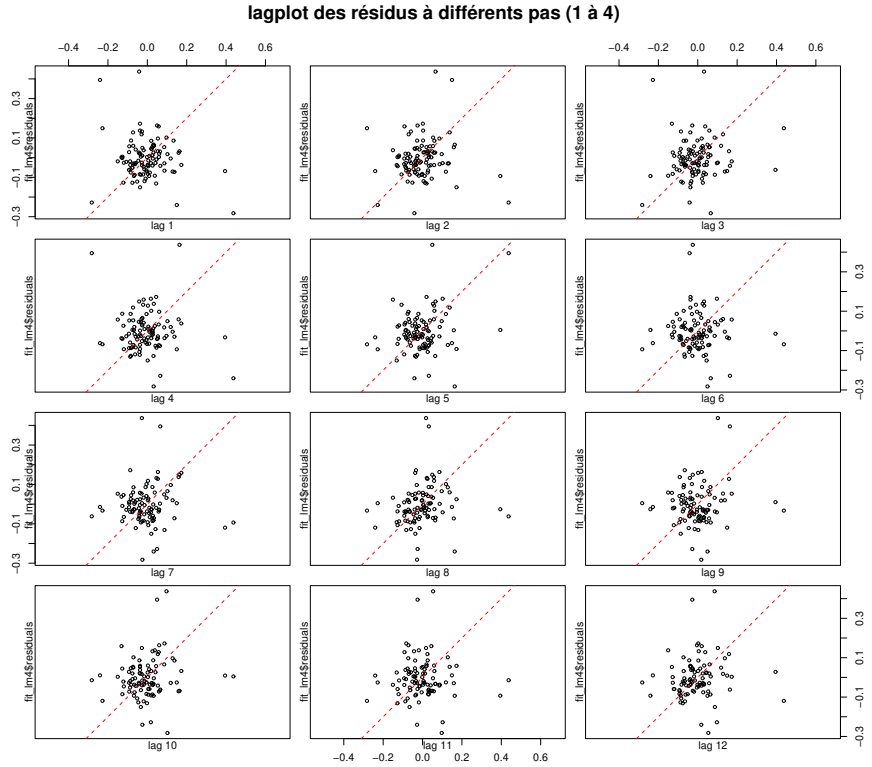

FIGURE 8 – Lagplot pour le modèle de régression linéaire avec un AR(4)

En observant le lagplot, nous constatons que les résidus semblent être des bruits blancs, sans autocorrélation significative.

Ensuite, nous examinons l'ACF (fonction d'autocorrélation) et le PACF (fonction d'autocorrélation partielle) des résidus du modèle AR(4).

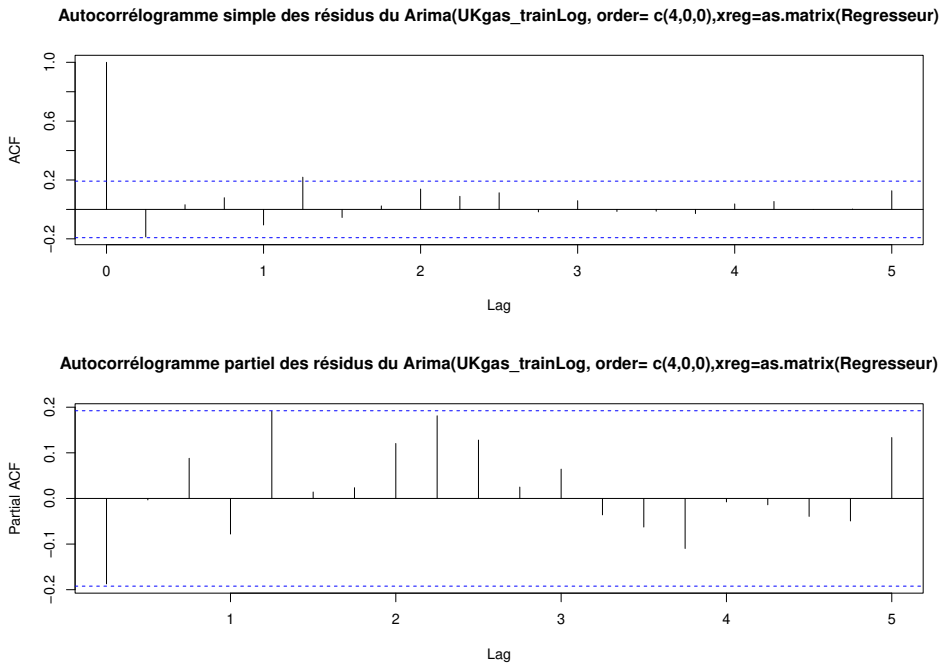

FIGURE 9 – ACF et PACF des résidus du modèle de régression linéaire avec un AR(4)

Ces analyses nous permettent de vérifier s'il reste une autocorrélation résiduelle dans les résidus du modèle. En analysant les graphiques de l'ACF et du PACF des résidus, nous remarquons que les motifs sont similaires à ceux d'un bruit blanc, ce qui confirme notre hypothèse selon laquelle les résidus sont bien des bruits blancs.

Pour compléter l'analyse des résidus, nous traçons le QQ plot. Le QQ plot nous permet de vérifier si les résidus suivent approximativement une distribution normale.

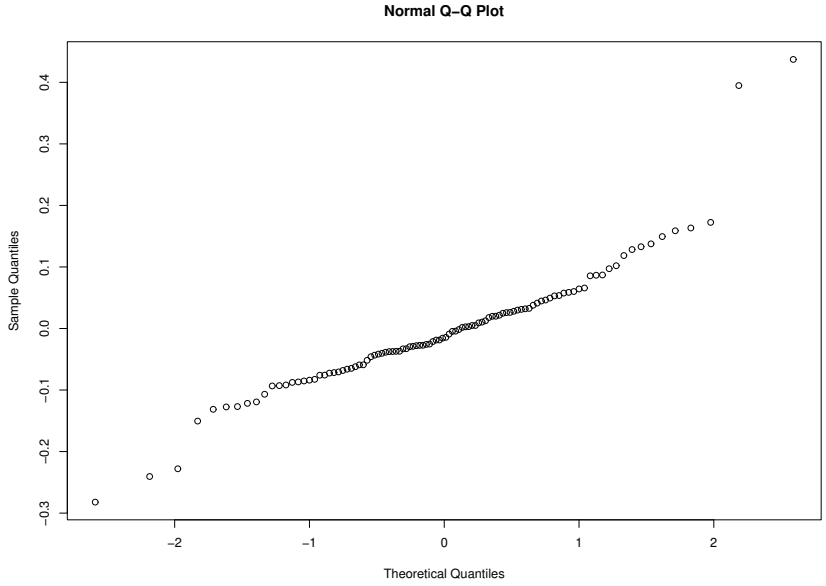

FIGURE 10 – Graphique d'ajustement et prévision de la série UKgas avec modèle de RL Ar(4)

En observant le QQ plot, nous remarquons une tendance linéaire, ce qui suggère que les résidus du modèle AR(4) suivent une distribution normale, caractéristique d'un bruit blanc. Nous observons une tendance linéaire dans le QQ plot, ce qui suggère que les résidus du modèle AR(4) suivent une distribution normale, caractéristique d'un bruit blanc.

Enfin, nous effectuons le test de portmanteau (Box-Ljung test) pour évaluer l'autocorrélation résiduelle dans les résidus du modèle AR(4). Les tests de portmanteau nous donnent des statistiques de test et des valeurs de p pour différents retards. En analysant les résultats des tests de portmanteau pour différents lags, nous observons les statistiques de test et les valeurs de p associées. Par exemple, pour un retard de 1, la statistique de test est X-squared = 3.753 avec un degré de liberté de 1 et une valeur de p de 0.05271. Ces résultats nous indiquent que l'autocorrélation résiduelle n'est pas significative jusqu'à ce retard. Des conclusions similaires sont tirées pour les retards 2, 3 et 4, avec des valeurs de p supérieures à 0.05. Cela suggère que l'autocorrélation résiduelle n'est pas statistiquement significative dans les résidus du modèle AR(4).

En conclusion, sur la base de l'ensemble des analyses effectuées, nous validons le choix du modèle AR(4) pour modéliser la série. Les résidus du modèle AR(4) semblent être des bruits blancs, sans autocorrélation significative, et suivent approximativement une distribution normale. De plus, les tests de portmanteau ne révèlent pas d'autocorrélation résiduelle significative. Ces résultats renforcent notre confiance dans la capacité du modèle AR(4) à capturer les motifs temporels dans les données de la série temporelle.

Enfin, nous présentons la prévision réalisée avec le modèle AR(4) :

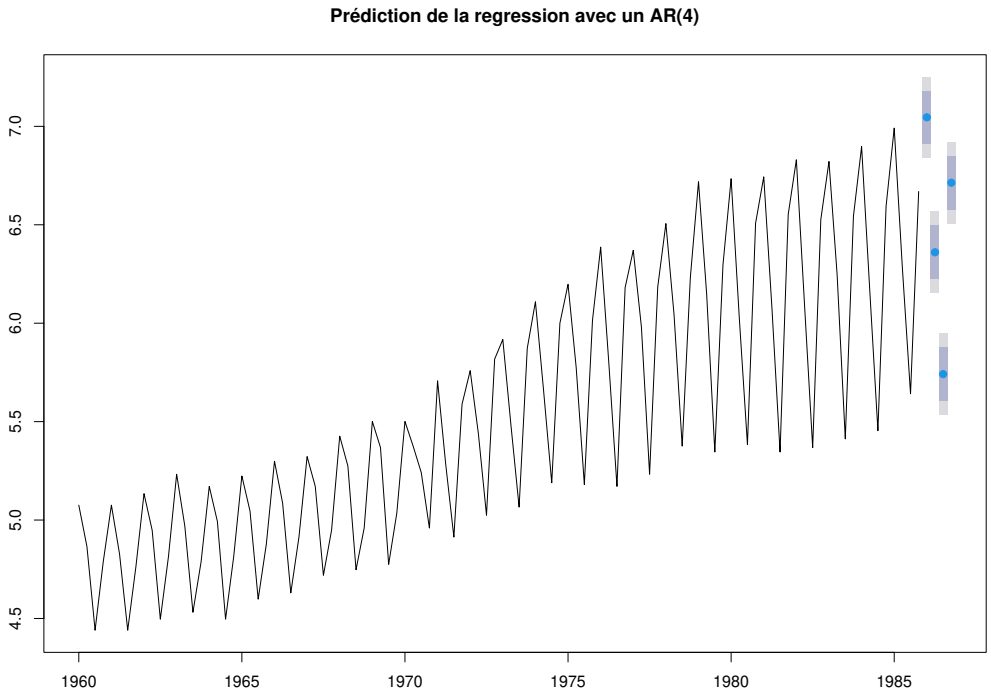

FIGURE 11 – Prédiction avec régression linéaire avec AR(4)

Nous constatons que la prévision du modèle AR(4) correspond beaucoup mieux aux données observées que celle du modèle de régression linéaire automatique.

4.3 Conclusion, choix du modèle et prédiction sur 5 ans

En résumé, nous avons ajusté deux modèles de régression linéaire pour modéliser la série UKgas. Le premier modèle, "fit_lm", utilise une méthode automatique de sélection des variables explicatives, tandis que le deuxième modèle, "fit_ar4", est un modèle AR(4) avec les variables explicatives correspondant aux quatre valeurs précédentes de la série.

Les résultats indiquent que le modèle AR(4) offre un meilleur ajustement aux données que le modèle "fit_lm", avec des mesures d'information AIC et BIC plus faibles. De plus, l'analyse des résidus confirme que le modèle AR(4) produit des résidus semblables à un bruit blanc, tandis que le modèle de régression linéaire présente une autocorrélation significative dans les résidus.

En termes de prévision, le modèle AR(4) présente également de meilleures performances, avec une prévision plus précise qui correspond davantage aux données observées.

Par conséquent, nous retenons le modèle AR(4) comme le modèle final pour la modélisation et la prévision de la série UKgas.

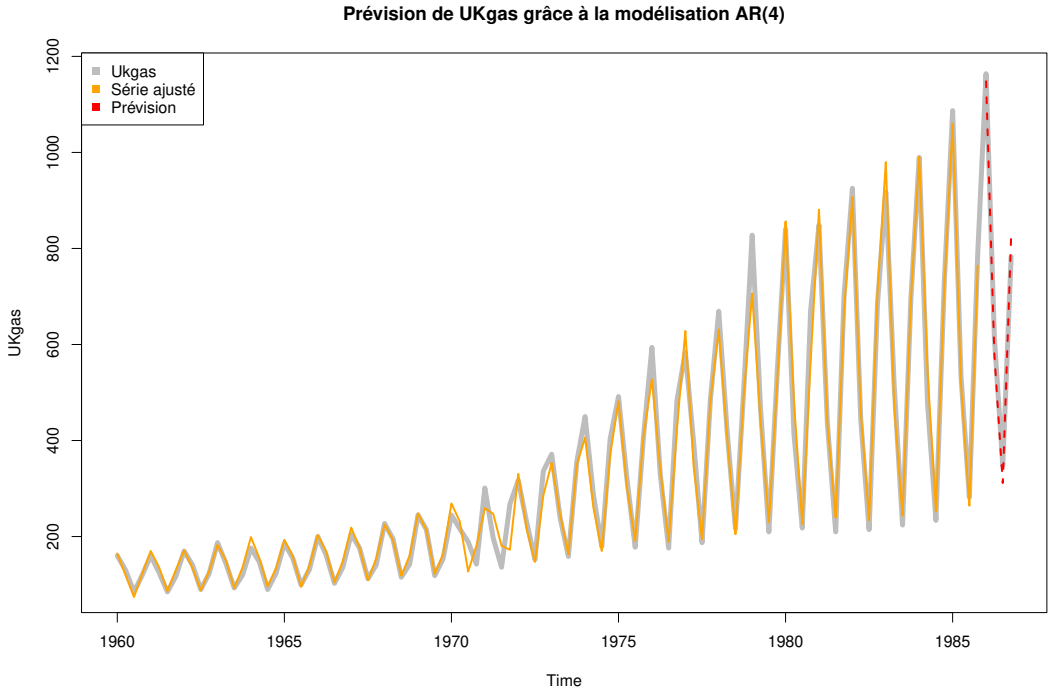

FIGURE 12 – Prévision de la série UKgas avec modèle régression linéaire Ar(4)

En conclusion, le modèle AR(4) est capable de capturer les tendances et les variations saisonnières de la série UKgas, et il permet de faire des prévisions précises.

Sur la base des différentes caractéristiques et des résultats obtenus, nous avons sélectionné le modèle AR(4) pour effectuer nos prévisions. Ce modèle intègre une tendance linéaire et une composante saisonnière estimée à partir des données. En combinant ces informations, nous ajustons le modèle AR(4) pour obtenir des prévisions sur une période de 5 ans.

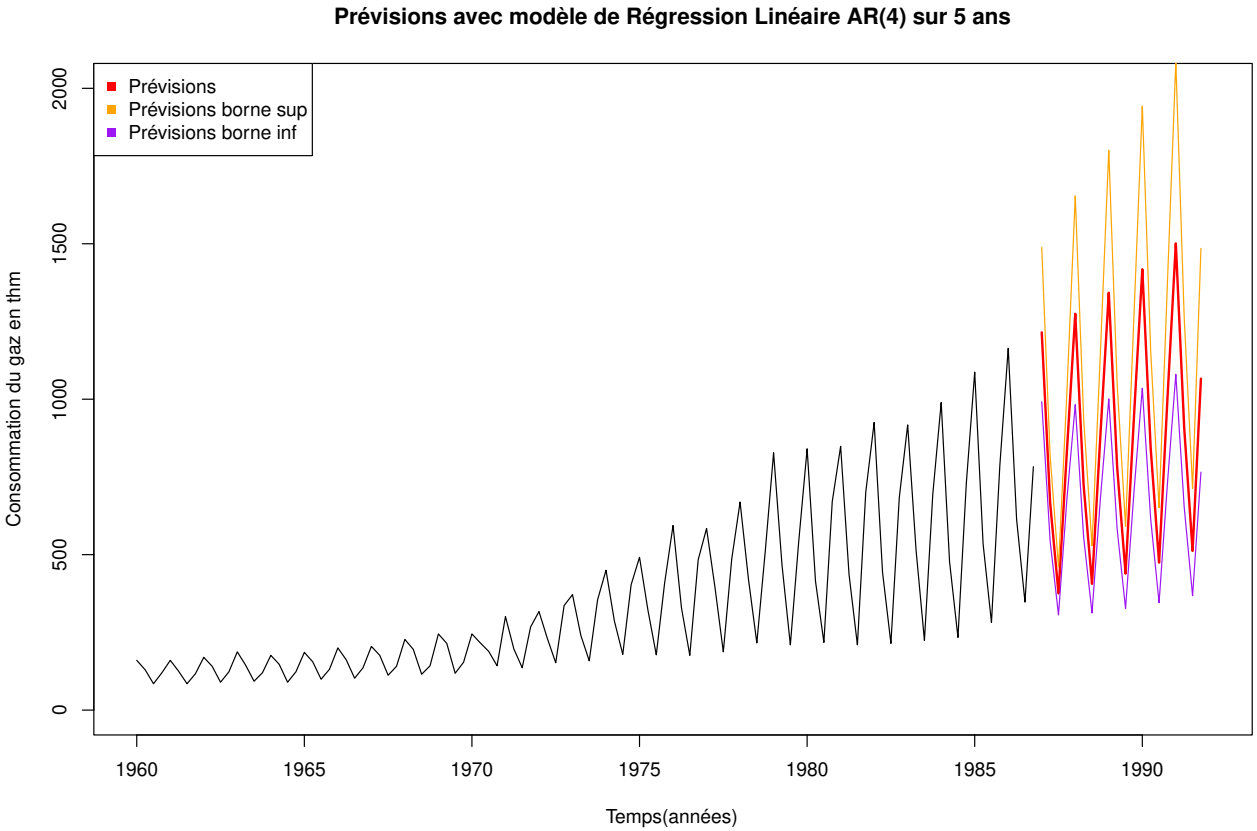

FIGURE 13 – Prévision de UKgas sur 5 ans

Le graphique affiche les données historiques ainsi que les prévisions pour les 5 prochaines années. La ligne rouge représente les prévisions, tandis que les lignes orange et violette représentent les bornes supérieure et inférieure des intervalles de confiance. Ces intervalles de confiance nous permettent d'estimer la variabilité des prévisions.

5 Modélisation SARIMA avec la méthodologie de Box-Jenkins

On a utilisé la fonction diff() avec l'argument differences = 1 pour effectuer cette différenciation. Ensuite, on a examiné les graphiques de la série différenciée et de ses fonctions ACF et PACF. On a constaté une corrélation significative au lag 4 dans l'ACF, ce qui suggère une saisonnalité d'ordre 4.

Dans cette partie, nous allons essayer de trouver le meilleur modèle pour prédire la série UKgas.

5.1 Découpage de la série et recherche des paramètres

Nous commençons par diviser la série "UKgas" en deux parties : la première partie ("UKgas.M") comprenant les données de 1960 à 1985 et la deuxième partie ("UKgas.T") comprenant les données à partir de 1986.

Cette division nous permet de réserver une période pour l'ajustement du modèle et une autre période pour l'évaluation des prédictions. Cela nous permet d'évaluer la capacité du modèle à généraliser sur des données futures non utilisées lors de l'ajustement et savoir si les modèles prédictifs sont cohérents avec la série de données.

Nous avons également appliqué le logarithme aux données de la première partie ("UKgas.M") pour atténuer les variations importantes et ainsi linéariser la tendance.

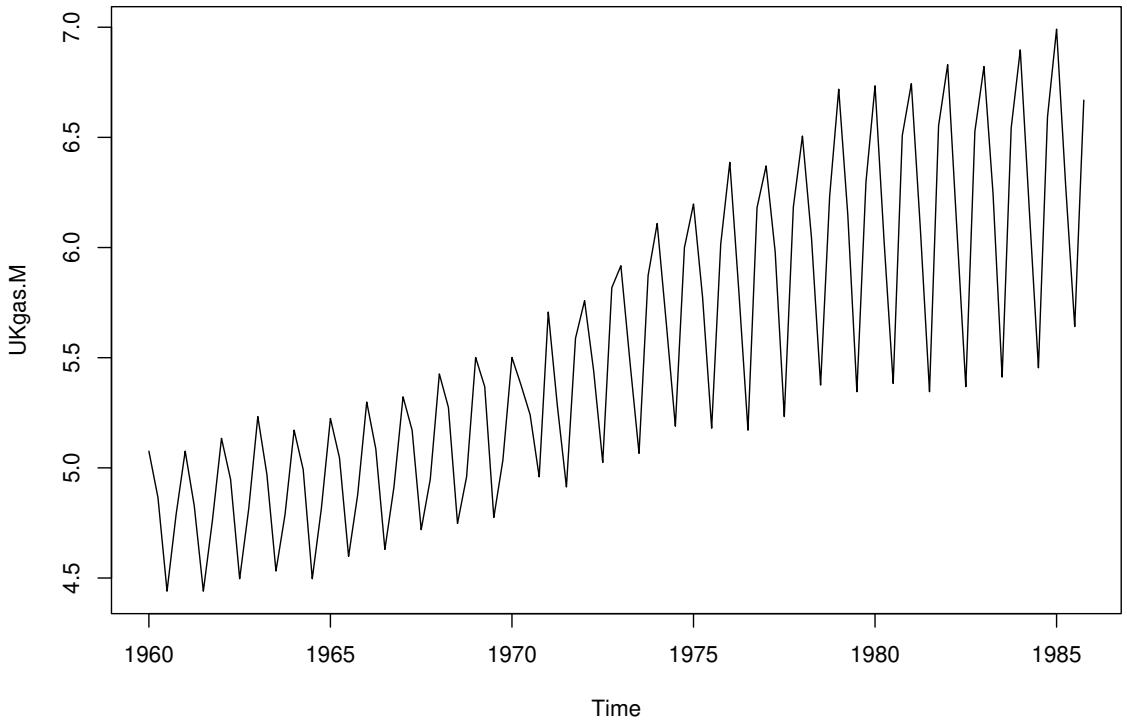

FIGURE 14 – Série passée au log avec tendance linéaire

Nous effectuons ensuite des différenciations sur la série "UKgas.M" pour stationnariser la série. Une première différenciation est faite afin de nous permettre d'éliminer la tendance

présente dans les données. Une deuxième différenciation, avec un décalage (lag) de 4, est utilisée pour capturer la saisonnalité d'ordre 4.

Désormais, nous modélisons l'ACF (AutoCorrelation Function) et le PACF (Partial AutoCorrelation Function) afin de trouver respectivement le "q" (paramètre de Moyenne Mobile (MA)) et "p" (paramètre de la partie autorégressive (AR)).

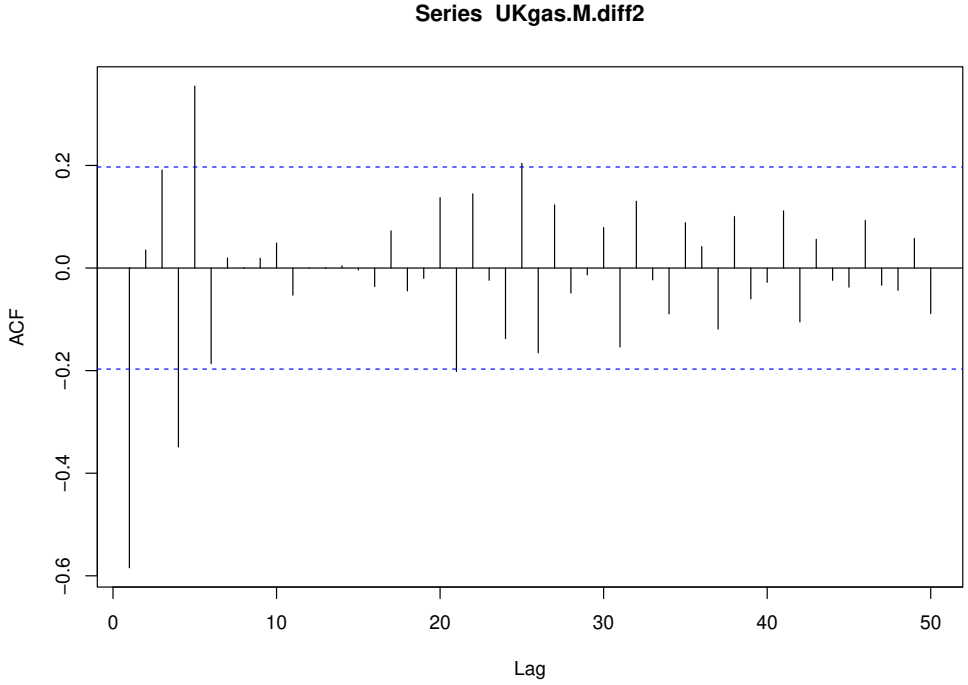

FIGURE 15 – ACF de la série UKgas différenciée 2 fois

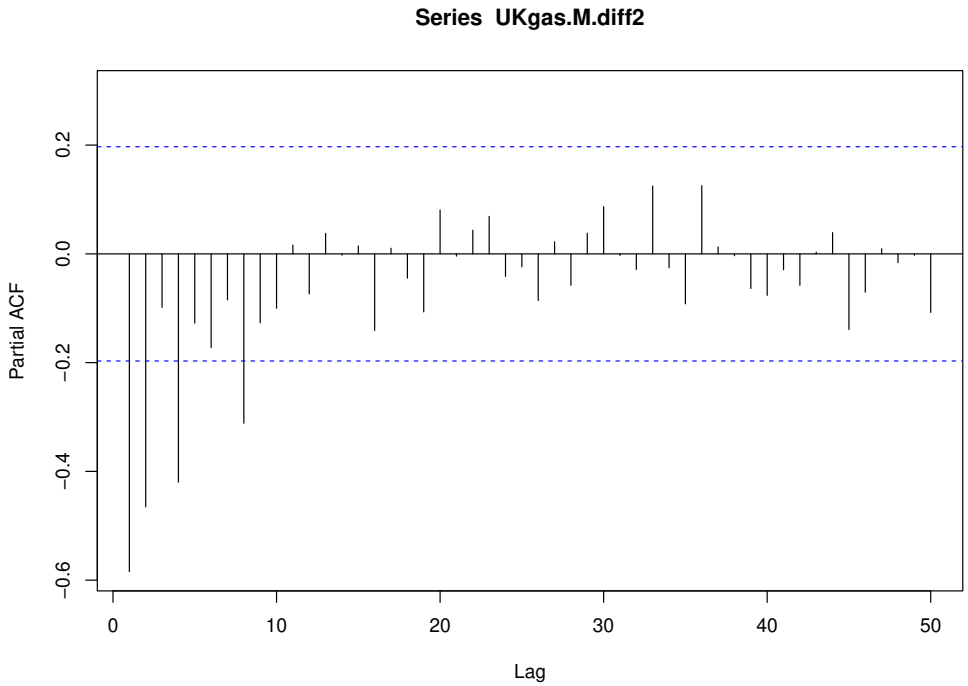

FIGURE 16 – PACF de la série UKgas différenciée 2 fois

A partir des graphiques, nous choisissons donc un premier modèle basé sur un ARMA(8,5). Regardons désormais les différents modèles SARIMA que nous pouvons modéliser.

5.2 Recherche de la modélisation optimale

Afin de déterminer le modèle adéquat et le plus optimal, nous utiliserons le Critère d'Information d'Akaike (AIC). Le but va être de le minimiser dans chaque modèle et celui qui obtiendra le plus petit AIC sera candidat au modèle final.

5.2.1 AR(8) : modèle le plus simple

Nous allons explorer dans un premier temps un modèle parcimonieux de type AR(8). Ce dernier a l'avantage d'utiliser une modélisation très simple qu'est l'AR. La modélisation nous donne un ACF et PACF acceptables dont les corrélation avec les X_{t-h} sont nulles :

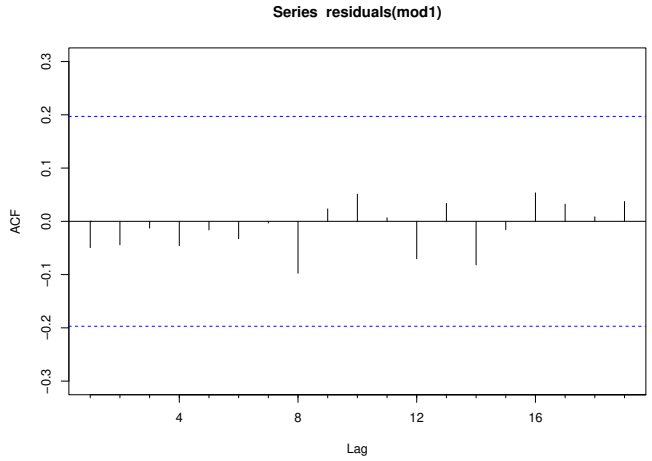

FIGURE 17 – ACF de la série AR(8)

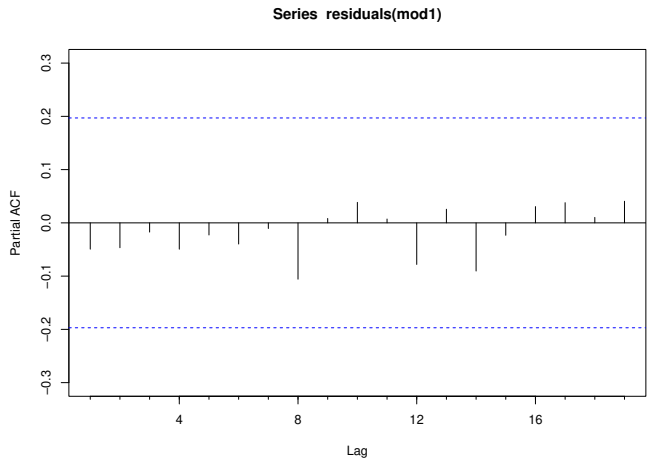

FIGURE 18 – PACF de la série AR(8)

Nous remarquons que dans les 2 graphes, il n'y a aucune corrélation significative pour chacun des h entre X_t et X_{t-h}. Pour ce modèle, nous obtenons un $AIC = -156.5176$.

Cela pourrait correspondre à un excellent modèle parcimonieux, utilisant que très peu de paramètres et obtenant un excellent AIC.

5.2.2 ARMA(8,5) : modèle empirique

Dans cette section, nous allons modéliser le SARIMA issu des observations faites sur la série $UKgas$ différenciée 2 fois. Cette série peut être modélisée par un ARMA(8,5) simple, sans composante saisonnière.

Cette dernière présente un ACF et un PACF montrant une non corrélation entre la série au temps t et ses évènement passés :

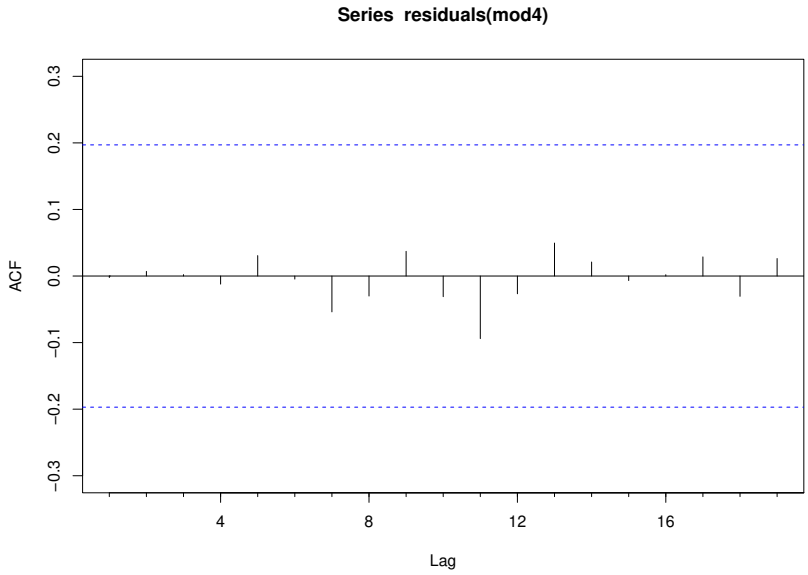

FIGURE 19 – ACF de la série ARMA(8,5)

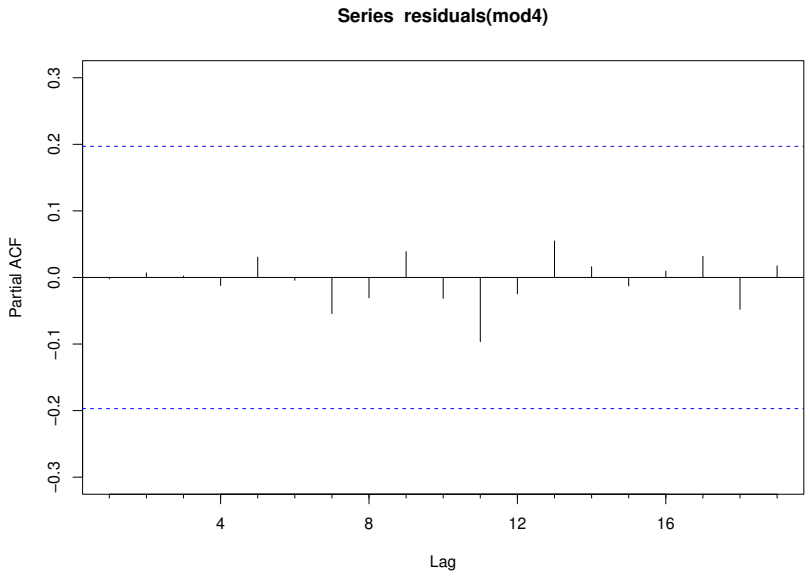

FIGURE 20 – PACF de la série ARMA(8,5)

Ce modèle nous donne un AIC plus faible, ce dernier étant égal à -150.85. Nous rappelons que le but de ces modélisations est d'obtenir un AIC le plus petit possible. De ce fait, nous avons d'ores et déjà une petite idée du modèle à privilégier pour le modèle prédictif final, à savoir le modèle précédent.

5.2.3 Quid des modèles SARIMA ?

Jusqu'à présent, nous avons exploré des modèles simples, à savoir l'AR et l'ARMA. Cependant, nous avons également travaillé, dans le cadre scolaire, sur des modèles de type SARIMA (Seasonal AutoRegressive Integrating Moving Average).

Un SARIMA se modélise lorsqu'une série ARIMA possède une composante saisonnière en plus. Dans notre cas, nous différencions une fois en tendance pour qu'elle soit linéaire et une autre fois en saisonnalité pour stationnariser la série.

Aucun des modèles précedents nous dirige vers une SARIMA, les modèles étant déjà parcimonieux et avec des ACF et PACF ne suggérant aucune amélioration. Nous choisissons ainsi le modèle d'un AR(8) pour la série 2 fois différenciée.

5.3 Choix du modèle final

Après étude des différents modèles ARMA, nous choisissons l'AR(8) pour la série 2 fois différenciée, obtenant le meilleur AIC. Nous modélisons ainsi la série UKgas à l'aide d'un :

$$SARIMA(8,1,0)(0,1,0)_4$$

Afin de savoir si le modèle possède les meilleurs paramètres, nous étudions les p-values associées à chacune des valeurs des paramètres à l'aide de la commande suivante :

$$round(pnorm(-abs(mod\_final\$coef), sd = sqrt(diag(mod\_final\$var.coef))), 4)$$

De cette manière, à l'aide de cette commande, nous allons pouvoir déterminer l'importance mais surtout la validité des paramètres de notre modèle final.

Après avoir exécuté la commande précédente, nous obtenons le résultat suivant sur les p-values des paramètres :

ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8
0.0000	0.0000	0.0000	0.0000	0.0002	0.0003	0.0012	0.0004

Chacune des p-values est inférieure 0.05 suggérant une bonne modélisation mais surtout une importance de chacun des paramètres de la modélisation. Ce test de la p-value nous permet de valider le modèle final basé sur un $SARIMA(8,1,0)(0,1,0)_4$.

5.4 Validation du modèle final $SARIMA(8,1,0)(0,1,0)_4$

Nous avons au préalable modéliser différents modèles prédictifs puis avons choisi finalement un modèle $SARIMA(8,1,0)(0,1,0)_4$. De ce modèle, nous avons déterminer l'AIC comme étant le meilleur des modèles testés suite à la différenciation de la série et de l'analyse ACF et $PACF$.

Nous allons extraire les résidus et tester leur normalité et vérifier si ce sont des bruits blancs. Afin de le faire, nous modélisons les résidus et vérifions qu'ils soient bien décorrélés à chaque saisonnaalité à l'aide du lagplot et vérifions la normalité des résidus grâce au qqnorm.

Regardons dans un premier temps le lagplot du $SARIMA(8,1,0)(0,1,0)_4$:

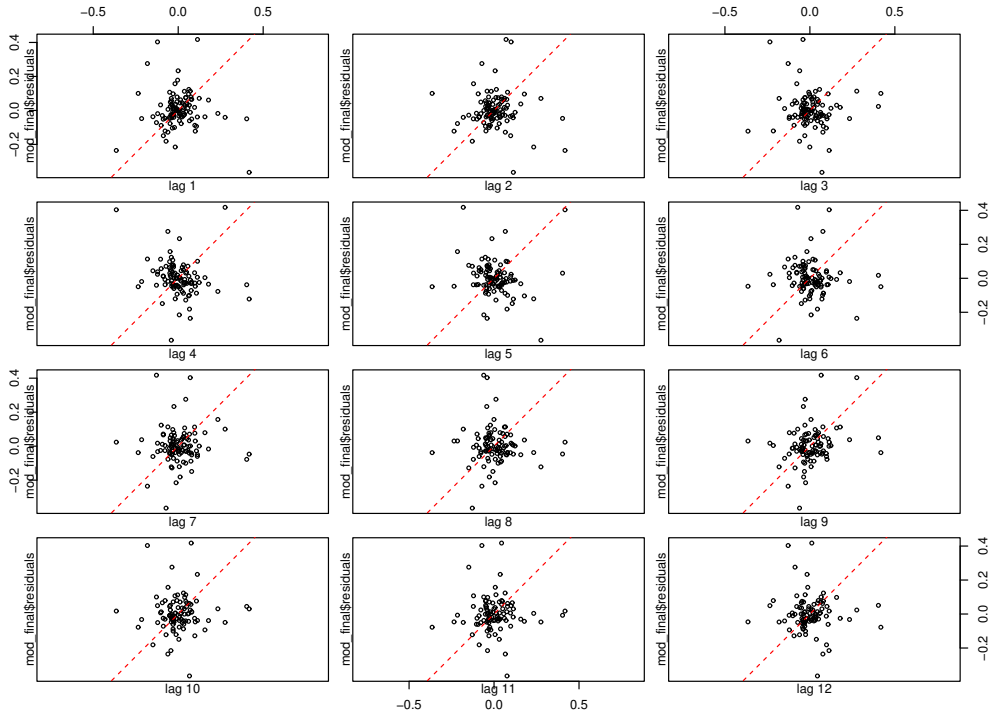

FIGURE 21 – Lagplot des résidus du $SARIMA(8,1,0)(0,1,0)_4$

Puis la normalité des résidus :

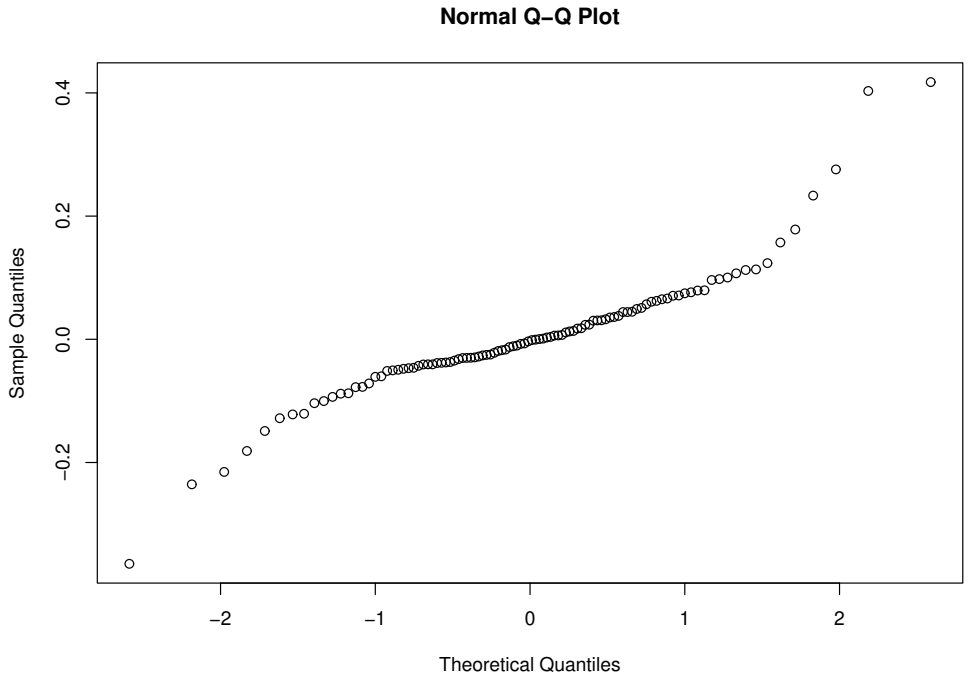

FIGURE 22 – Qqnorm des résidus du $SARIMA(8,1,0)(0,1,0)_4$

Nous remarquons que le qqnorm est linéaire et nous pouvons donc admettre que les résidus suivent une loi normale, faisant d'eux des bruits blancs.

Maintenant que le modèle est entièrement validé, nous pouvons modéliser la série prédite sur 5 ans par exemple :

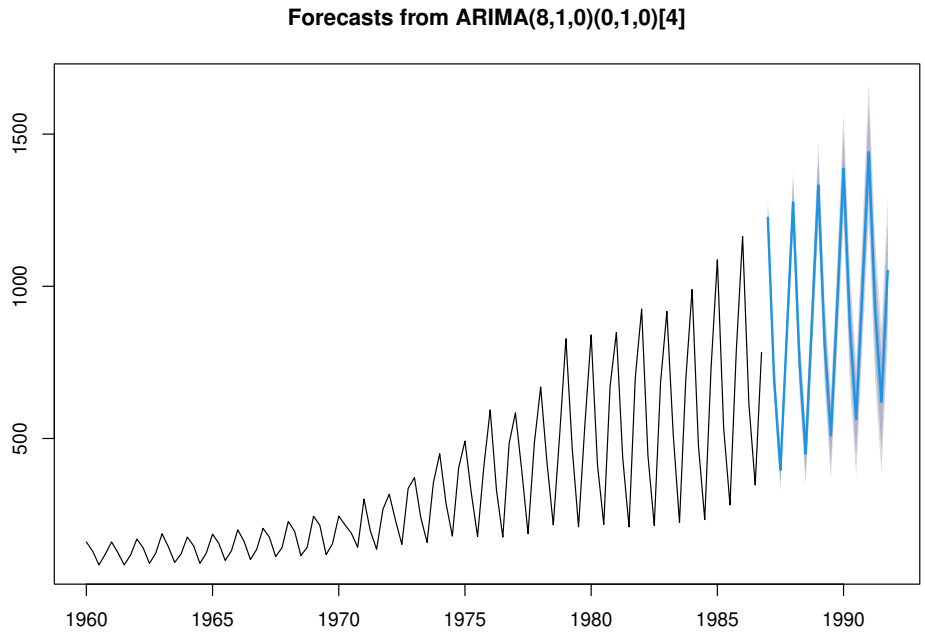

FIGURE 23 – Prédiction du modèle $SARIMA(8,1,0)(0,1,0)_4$

Cette figure de prédiction est obtenue à l'aide de la fonction *forecast* contenue dans le package sur R éponyme. Prenons donc la moyenne de la prédiction, sa borne supérieure et sa borne inférieure afin de modéliser au mieux la prédiction ainsi que son intervalle de confiance :

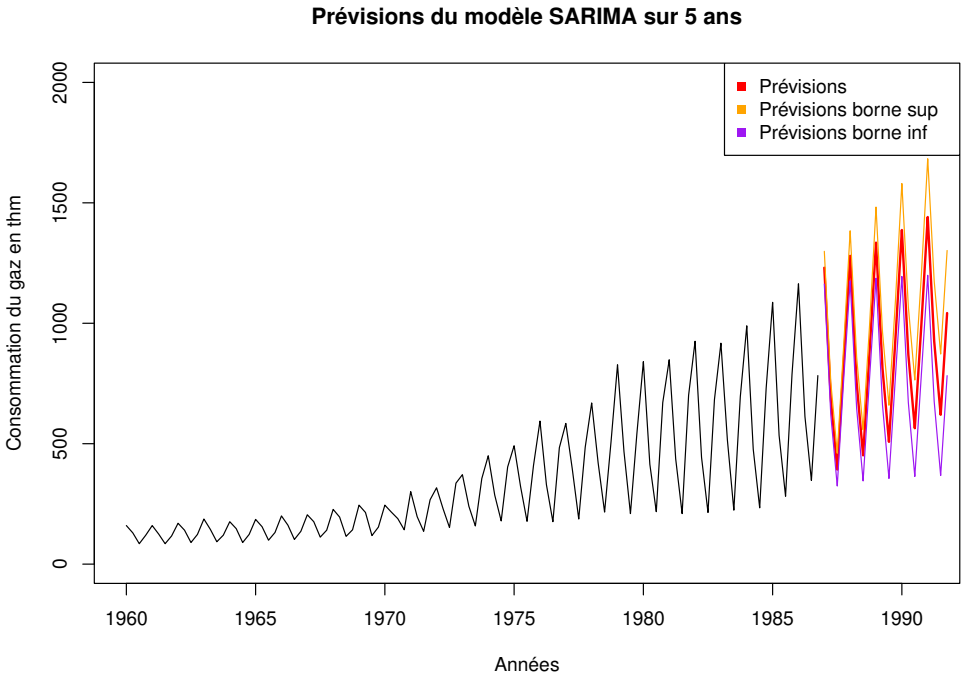

FIGURE 24 – Prédiction du modèle $SARIMA(8,1,0)(0,1,0)_4$

6 Choix du modèle final

Après avoir effectué une analyse approfondie de la série temporelle UKgas en utilisant différentes méthodes telles que le lissage exponentiel, le modèle de Holt-Winters, la régression linéaire avec un AR(4) et le modèle SARIMA, nous avons comparé les performances de ces modèles en utilisant les métriques RMSE (Root Mean Square Error) et MAPE (Mean Absolute Percentage Error).

	HW	RL_Ar4	SARIMA
RMSE	41.85	33.23928	61.79265
MAPE	0.065	0.05649634	0.09705765

Le tableau fourni présente les résultats de cette comparaison, montrant les valeurs de RMSE et MAPE obtenues pour chaque méthode.

Le RMSE mesure l'écart moyen entre les valeurs réelles de la série et les valeurs prédites par le modèle. Plus le RMSE est faible, plus les prédictions du modèle sont proches des données réelles. Dans notre cas, nous avons obtenu les valeurs suivantes :

— Pour le lissage exponentiel (HW) : un RMSE de 41.85.
— Pour la régression linéaire avec un AR(4) (RL_Ar4) : un RMSE de 33.23928.
— Pour le modèle SARIMA : un RMSE de 61.79265.

Le MAPE quantifie l'erreur moyenne en pourcentage entre les valeurs réelles et les valeurs prédites. Un MAPE plus faible indique une meilleure précision du modèle. Les résultats obtenus sont les suivants :

— Pour le lissage exponentiel (HW) : un MAPE de 0.065.
— Pour la régression linéaire avec un AR(4) (RL_Ar4) : un MAPE de 0.05649634.
— Pour le modèle SARIMA : un MAPE de 0.09705765.

En analysant ces résultats, nous pouvons conclure que la méthode de régression linéaire avec un AR(4) a produit les résultats les plus favorables pour notre série UKgas. En effet, elle présente à la fois le RMSE et le MAPE les plus bas parmi les méthodes comparées.

Cela suggère que la régression linéaire avec un AR(4) est capable de capturer les motifs et les tendances de la série temporelle UKgas de manière plus précise que les autres méthodes évaluées.