université angers

FACULTÉ
DES SCIENCES
Unité de formation et de recherche

Modélisation d'une série chronologique

- *Pr. Frédéric Proia*

- *Harold Ndangang*
- *Mohamed-Amine Goutali*

Plan du rapport :

I. Introduction :

II. Modélisation additive :

- Visualisation des données
- Décomposition de la série
- Analyse de la fluctuation
- Construction des modèles ARMA

III. Modélisation SARIMA :

- Stationnarisation de la série
- Construction des modèles SARIMA

IV. Prédiction :

V. Conclusion :

I. **Chapitre 0 : Introduction**

Une série chronologique, ou série temporelle, est une suite de valeurs numériques représentant l'évolution d'une quantité spécifique au cours du temps. La modélisation et les prévisions des séries chronologiques est un enjeu important dans des nombreux domaines : la démographie, le traitement de signal, l'économétrie financière …

Les objectifs des études des séries chronologiques peuvent être divisés en deux composantes principales :

- L'analyse descriptive des donnée et la modélisation.
- La prévision des valeurs futurs et l'estimation des risques.

Dans ce rapport, on va s'intéresser à une série chronologique non stationnaire.

Problématique :

On dispose d'une base de données réelles de la SNCF présentant le nombre de voyageurs mensuel pendant 216 mois consécutifs. Cela peut être considérée comme une série chronologique $(X_t)_{t=1,\dots,216}$.

On s'intéresse principalement à trouver des modèles :

- Additive
- SARIMA

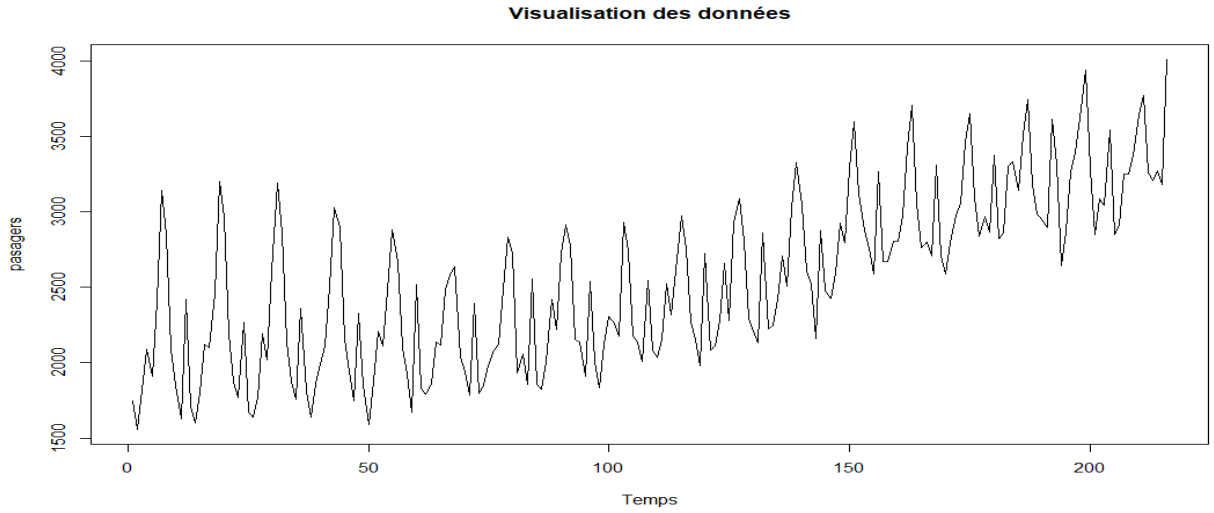
Pour modéliser notre série tout en respectant certains critères et pour faire de la prévision.

II. **Chapitre 1 : Modélisation additive**

Dans ce chapitre, nous allons chercher un modèle additif qui modélise au mieux notre série.

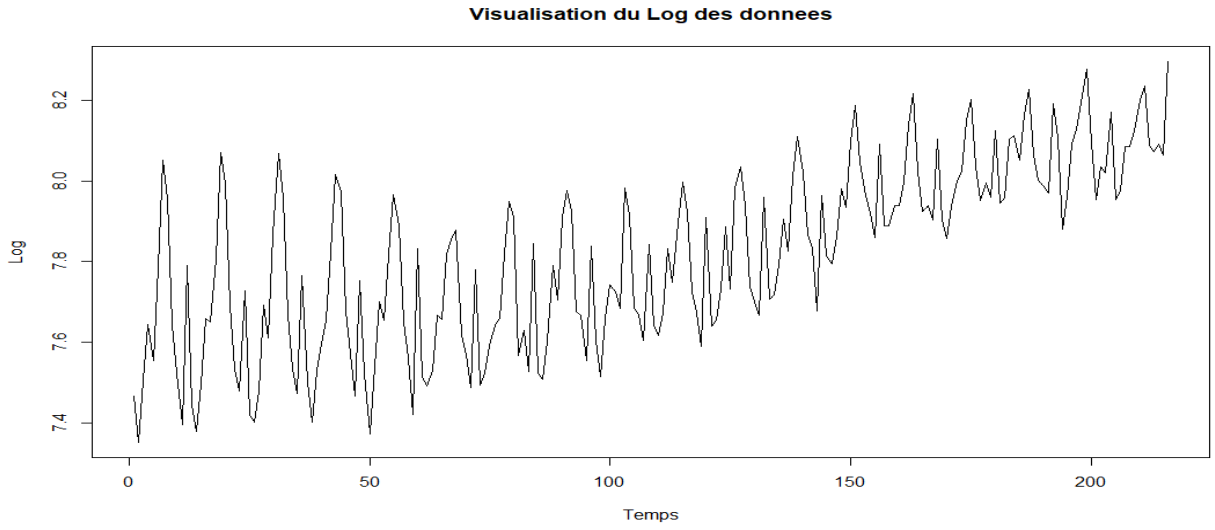
1. *Visualisation des données* :

On commence par la visualisation de la série $(X_t)_{t=1,\dots,216}$:



D'après le graphe ci-dessus, on constate que la série n'est pas stationnaire.

D'autre part, sa variance est très élevée. Donc on décide d'appliquer le Log à la série. On note $(Y_t) = (\log(X_t))$.

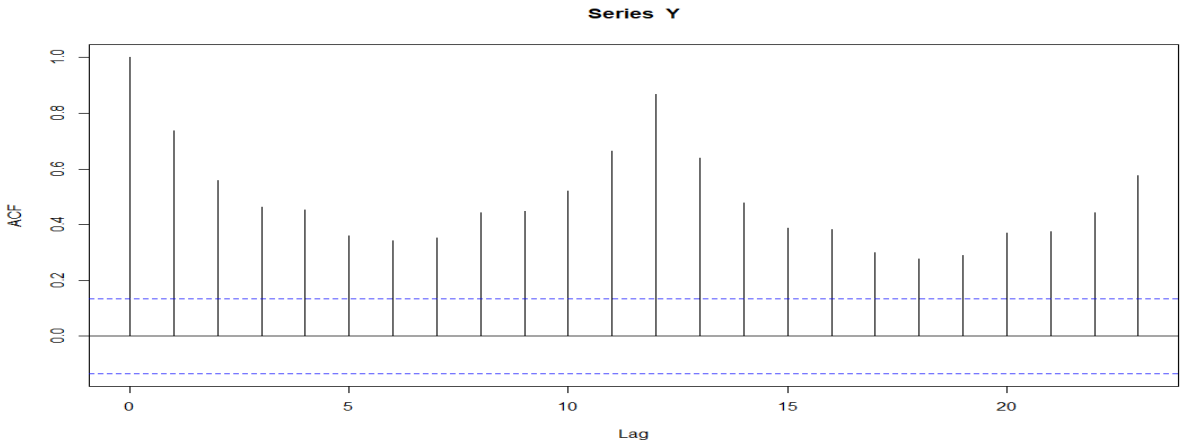


On passe d'une variance de 318521.1 pour la série de base vers une variance de 0.05, ce qui est raisonnable.

2. *Modélisation additive de la série Log* :

Vu que la série (Y_t) n'est pas stationnaire, on va se ramener à une décomposition tendance, saisonnalité et fluctuation.

Afin de décomposer cette série, nous avons besoin de la fréquence de la saisonnalité. Pour ce fait, nous regardons les autocorrélations empiriques représentées dans le graphe ci-dessous.



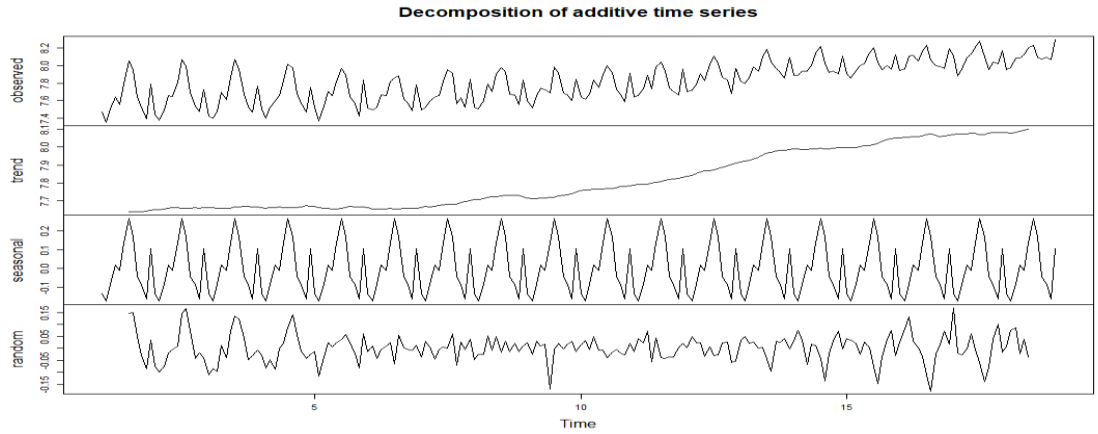
On constate que pour $\text{Lag} = 12$ la corrélation entre les variables est maximale. Donc, on va choisir une fréquence de 12 pour la saisonnalité (d'ailleurs, ce choix est adéquat aux données).

Ceci nous permettra de décomposer la série (Y_t) comme suit :

$$Y_t = m_t + s_t + \epsilon_t$$

Avec m_t : la tendance, s_t : la saisonnalité et ϵ_t la fluctuation.

Cette décomposition est représentée dans le graphe ci-dessous.



Dans la suite on s'intéresse à analyser et approximer la fluctuation (ϵ_t) .

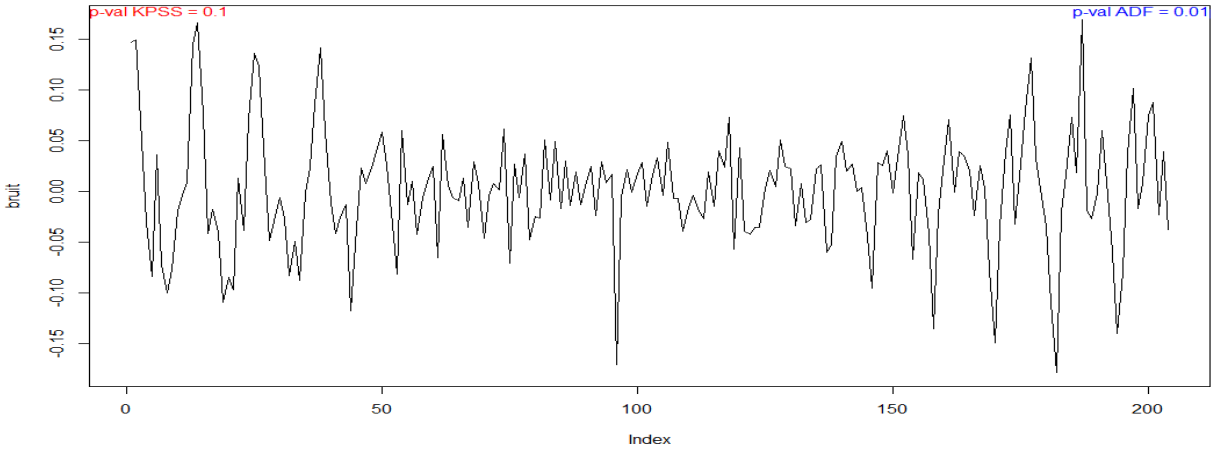
3. *<u>Analyse de la fluctuation</u>* :

En explorant les valeurs numériques de la fluctuation, on constate qu'il y a des valeurs marquantes (six premières et six dernières). Donc on les supprime.

On s'intéresse à étudier sa stationnarité. Pour ce fait, on va utiliser les deux tests de stationnarité :

i. *Test KPSS* :
H_0 : ' la trajectoire est issue d'un processus stationnaire' contre $H_1 = \overline{H_0}$

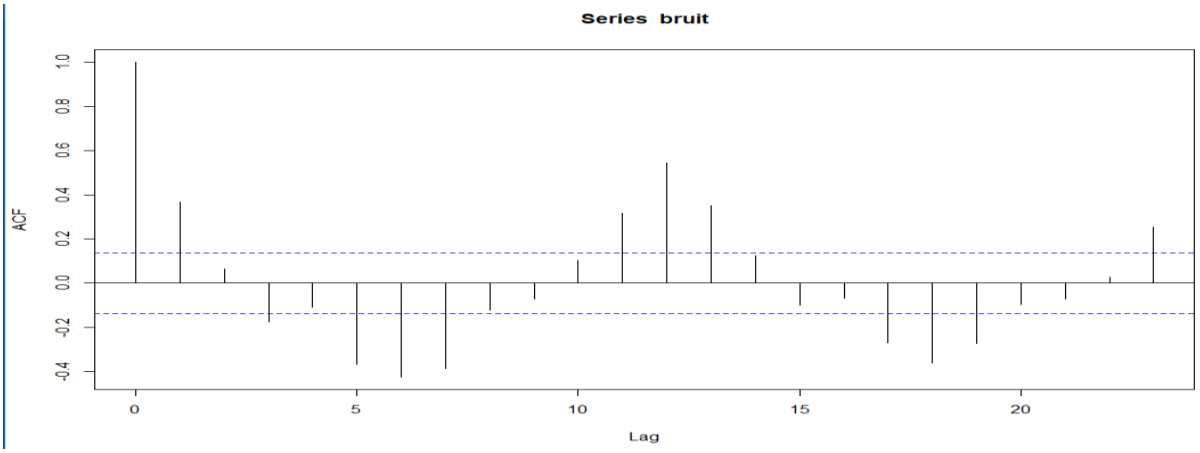
ii. *Test ADF* :
H_0 : ' la trajectoire est issue d'un processus non stationnaire' contre $H_1 = \overline{H_0}$



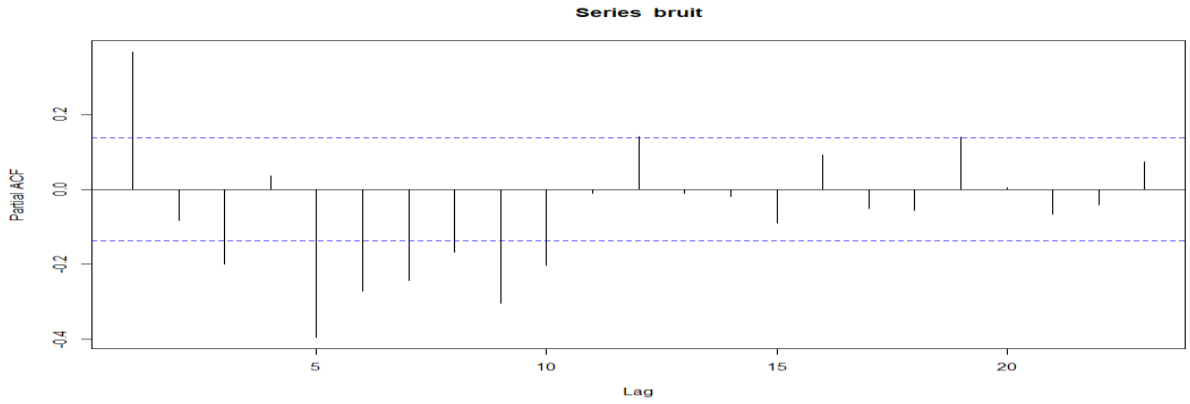
Si on se contente à un seuil de significativité de 5%, alors pour le test KPSS, on ne rejette pas l'hypothèse de stationnarité. Par contre pour le test ADF, on rejette l'hypothèse de non-stationnarité.

Ce résultat nous conduit à conclure que le bruit est issu d'un processus stationnaire.

On s'intéresse maintenant à l'étude des autocorrélations du bruit.



L'ACF montre que ce bruit est plein de corrélations d'où il n'est pas blanc.



La PACF confirme le même résultat. En outre, on remarque que les autocorrélations partielles s'annulent à partir d'un lag égale à 10. Théoriquement, cela nous permet d'adapter le modèle

AR(10) sur le bruit. Pratiquement ce n'est pas possible car on a peu de données vs trop de paramètres.

On s'intéressa dans la suite à construire un modèle ARMA d'ordre p et q du bruit stationnaire.

4. *Construction du modèle ARMA(p,q)* :

On va chercher les modèles ARMA(p,q) les plus adéquats à ce bruit. C'est-à-dire,

$$\Phi(B)(\epsilon_t - m) = \Theta(B)\xi_t$$

Avec $\Phi(B) = I - \phi_1 B - \cdots - \phi_p B^p$ est le polynôme autorégressif.

$\Theta(B) = I + \theta_1 B + \cdots + \theta_q B^q$ est le polynôme moyenne mobile.

(ξ_t) est un bruit blanc de variance non nulle.

$m = E[\epsilon_t] = 0$ (Car le bruit est centré).

Pour ce fait, on va passer par deux étapes :

i. Validation des modèles :

La validation des processus passe par un examen des coefficients estimés (test de Student). De plus, les résidus estimés doivent former un bruit blanc gaussien (tests de Box-Pierce, Ljung-Box et les tests de normalité).

ii. Critères de choix des modèles :

Après avoir validé les modèles, on fait appel aux critères standards (MSE, SCR, MAPE…) et aux critères de l'information (AIC, BIC…).

5. *Application* :

En faisant varier (p,q), nous avons pu sélectionner les modèles suivants.

A. *Validation des modèles* :

a. *Modèle 1* :

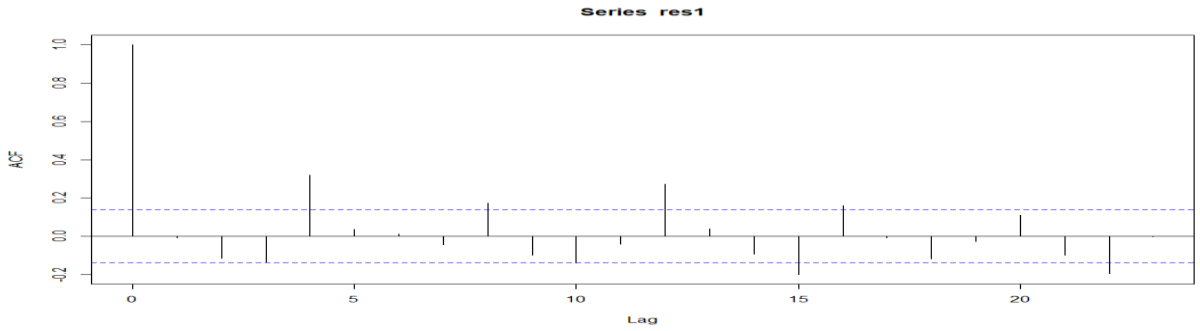
ARMA(2, 2) >> $(I - 1.64B + 0.9B^2)\epsilon_t = (I - 1.77B + 0.79B^2)\xi_t$

i. Significativité des coefficients estimés :

Pour le test de Student, tous les coefficients vérifient $\left|\frac{\hat{c}}{\sigma_{\hat{c}}}\right| > U_{0.95}$ (avec $U_{0.95}$ = 1.96 le quantile d'ordre 0.95 de la loi normale centrée). D'où on peut dire qu'à 5% d'erreur, ils sont tous significatifs.

ii. Examen des résidus estimés :

- L'ACF de ces résidus ne nous permet pas de conclure qu'il s'agit d'un bruit blanc.



Donc on passe aux tests de Box-Pierce et Ljung-Box.

- **Remarque** : Il s'agit de deux tests statistiques qui évaluent la corrélation existante entre les résidus. L'hypothèse nulle est H_0 = 'il n'y a pas d'autocorrélation d'ordre ≥ 1 des résidus'.

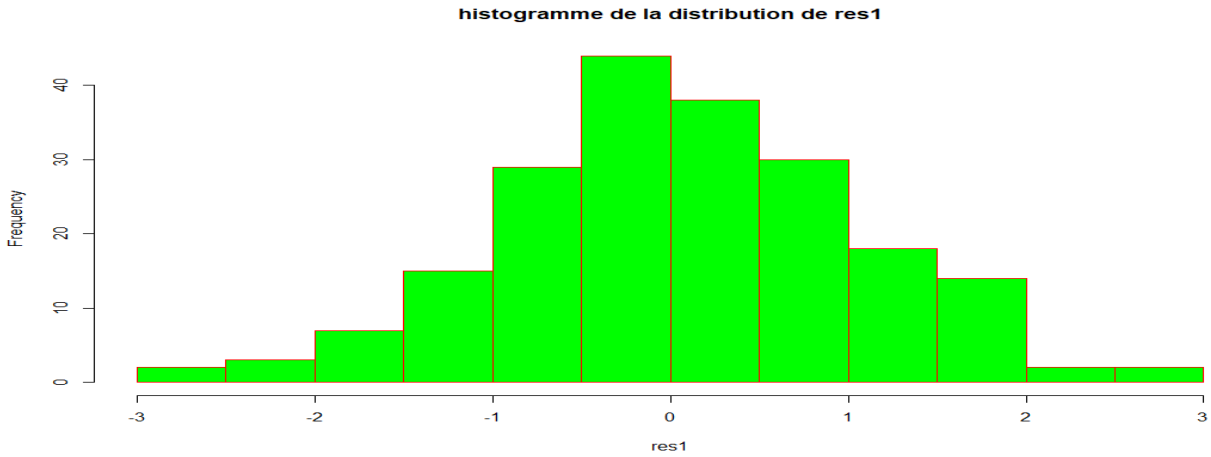
Pour ce modèle, on a un p-val = 0.092 > 5%, Donc on ne rejette pas H_0. Les résidus estimés forment un bruit blanc avec une certitude de 95%.

- Ensuite, on va étudier la normalité de ces résidus.
- Test de Shapiro-Wilk :

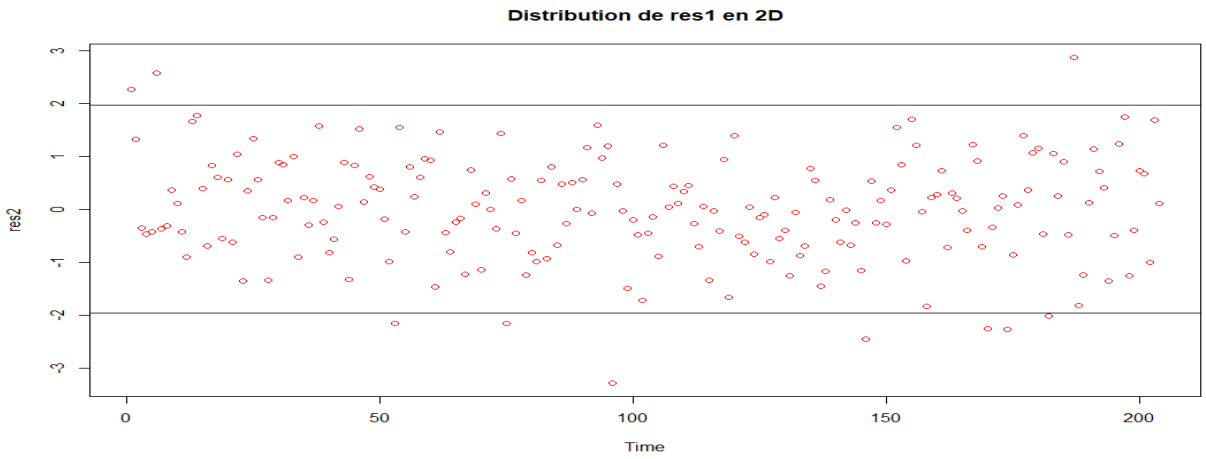
- **Remarque** : Le test de Shapiro-Wilk teste l'hypothèse nulle H_0 : 'l'échantillon est issu d'une population normalement distribuée' contre $H_1 = \overline{H_0}$. Il a été publié en 1965 par Samuel Sanford Shapiro et Martin Wilk.

Le p-val de ce test est de 0.89. On ne rejette pas l'hypothèse nulle.

- Distribution des résidus estimés :



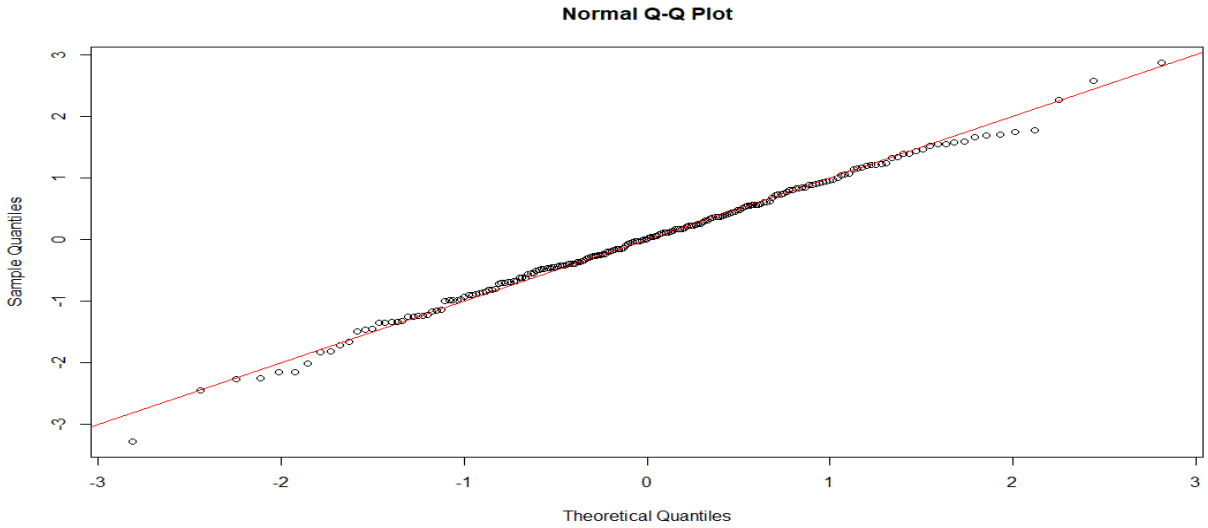
La forme de la distribution des résidus estimés représentée sous forme d'histogramme ressemble beaucoup à la distribution d'une loi normale centrée.



En faisant une représentation graphique des résidus en 2D, on remarque que 95% des points sont situés dans l'intervalle [-1.96, 1.96]. Ce dernier correspond à l'intervalle de confiance de la loi normale centrée.

- Diagramme Quantile-Quantile :

- **Remarque** : Le diagramme Q-Q est un outil graphique permettant d'évaluer la pertinence d'ajustement d'une distribution donnée à un modèle théorique. Il compare la position de certains quantiles dans la population observée avec leur position dans la population théorique.



Pour ce modèle, on remarque que les valeurs des quantiles des résidus estimés sont très proches des quantiles de la loi normale centrée.

➢ Tout cela nous ramène à dire que les résidus estimés suivent une distribution gaussienne centrée réduite.

- **Conclusion** : Le modèle ARMA(2, 2) est bien valide.

b. *4 Autres modèles* :

En faisant les mêmes études sur d'autres modèles, nous avons retenu :

ARMA(2, 4) >> $(I + 0.49B - 0.45B^2)\epsilon_t = (I + 0.71B - 0.57B^2 - 0.76B^3 - 0.37B^4)\xi_t$

ARMA(1, 3) >> $(I - 0{,}41B\ )\epsilon_t = (I - 0.22B - 0.30B^2 - 0.47B^3)\xi_t$

ARMA(1, 2) >> $(I + 0.72B)\epsilon_t = (I + 1.17B + 0.49B^2)\xi_t$

ARMA(3, 2) >> $(I - 0.35B - 0.57B^2 + 0.6B^3)\epsilon_t = (I - 0.17B - 0.82B^2)\xi_t$

B. *Choix des modèles* :

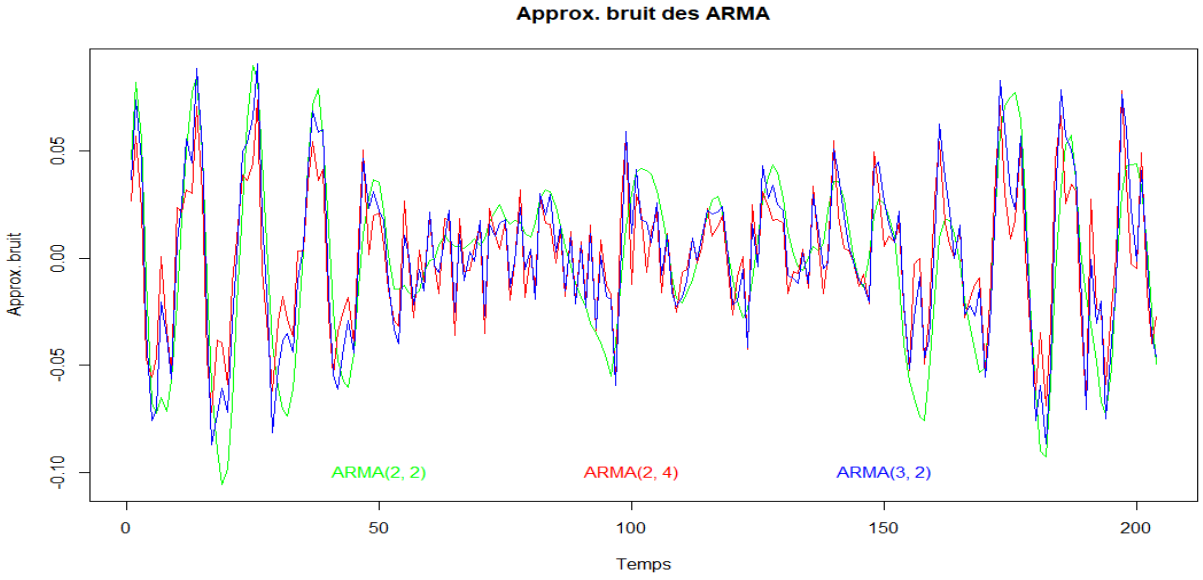
Remarque : Dans notre cas, il est possible d'augmenter la vraisemblance du modèle en faisant varier les paramètres p et q. L'AIC permet de pénaliser les modèles en fonction du nombre de paramètres d'où l'intérêt de l'utiliser comme critère de validation.

D'autre part, on va choisir l'erreur moyenne quadratique MSE pour comparer les valeurs calculées aux valeurs observées.

En se basant sur ces deux critères, nous allons sélectionner les modèles qui ont les valeurs les plus petites d'AIC et de MSE.

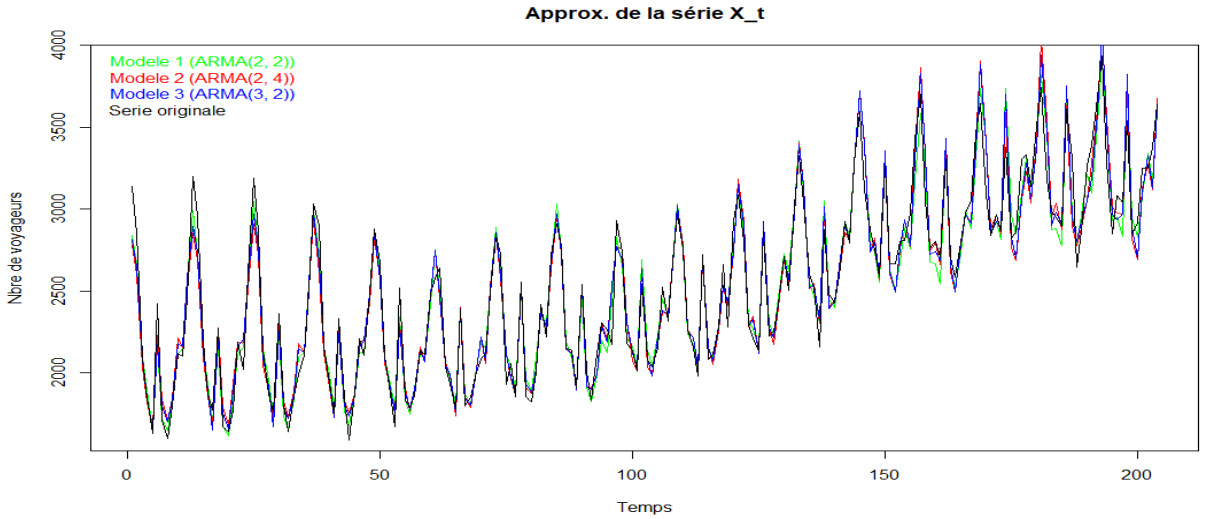
Modèle	AIC	MSE
ARMA(2, 2)	-714.85	0.04
ARMA(3, 2)	-682.26	0.043
ARMA(2, 4)	-664.37	0.045
ARMA(1, 3)	-663	0.046
ARMA(1, 2)	-626.74	0.051

On retient les modèles colorés en vert.



Le graphe ci-dessus représente les approximations du bruit pour les 3 modèles retenus.

C. Approximations de (X_t) :



Visuellement, on remarque que les courbes de la série originale et ses approximations sont presque confondues (On rappelle que pour la série originale on a enlevé les 6 premières et les 6 dernières valeurs).

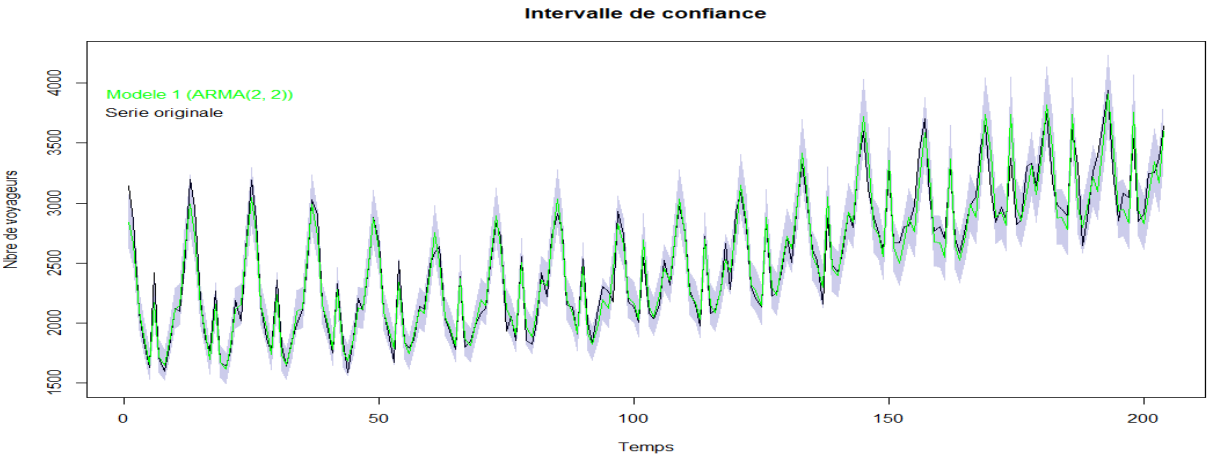
Pour avoir confirmation de ce constat, on va regarder la RMSE (racine carré de l'erreur moyenne quadratique) et le MAPE (l'erreur absolue moyenne en pourcentage) de chaque approximation.

Approximation issue du modèle	RMSE	MAPE
1	111.32	3.18%
2	122.94	3.44%
3	118.88	3.33%

Le Tableau ci-dessus confirme que l'erreur entre la série originale et ses approximations est raisonnable ce qui justifie le choix des modèles.

- **Remarque** : A l'aide de la commande auto.arima qu'on applique sur le bruit, R nous préconise le modèle ARMA(2, 2).

Ci-dessous, on a une représentation graphique de la série de base et son approximation issue du modèle ARMA(2, 2).



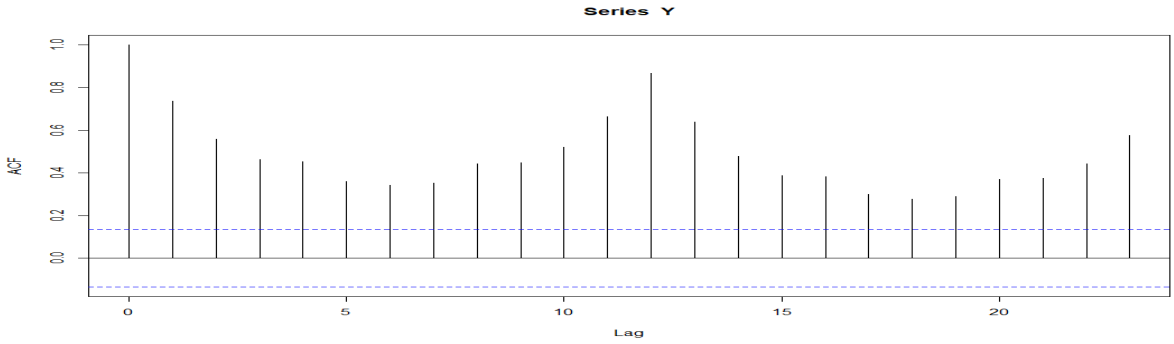
On remarque que la courbe de la série originale est bien dans l'intervalle de confiance de son approximation.

III. **Chapitre 2 : Modélisation SARIMA**

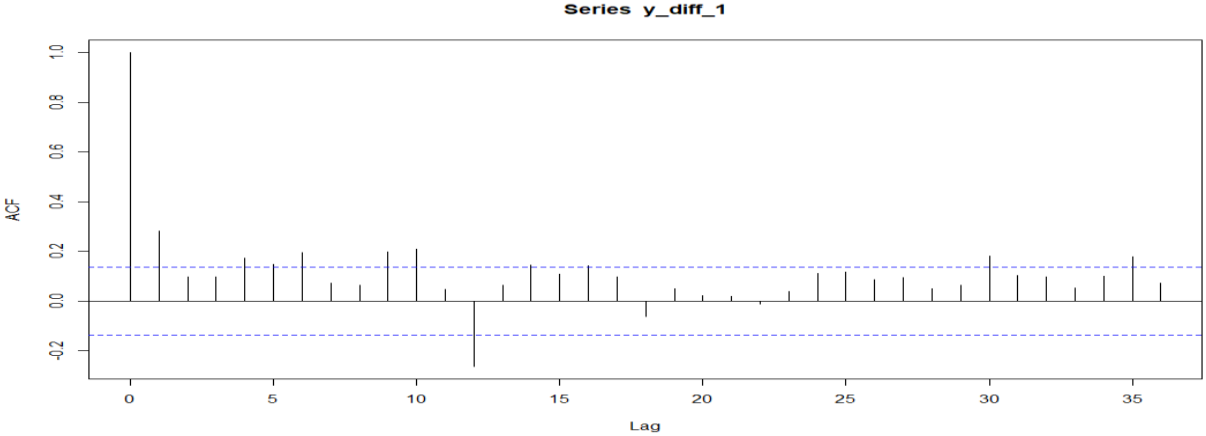
Dans ce chapitre, nous allons chercher un modèle SARIMA qui modélise au mieux notre série.

Comme on l'a vu dans le chapitre 1, la variance de (X_t) est très élevée d'où l'intérêt d'étudier la série $(Y_t) = (\log(X_t))$.

1. *Stationarisation de la série* :



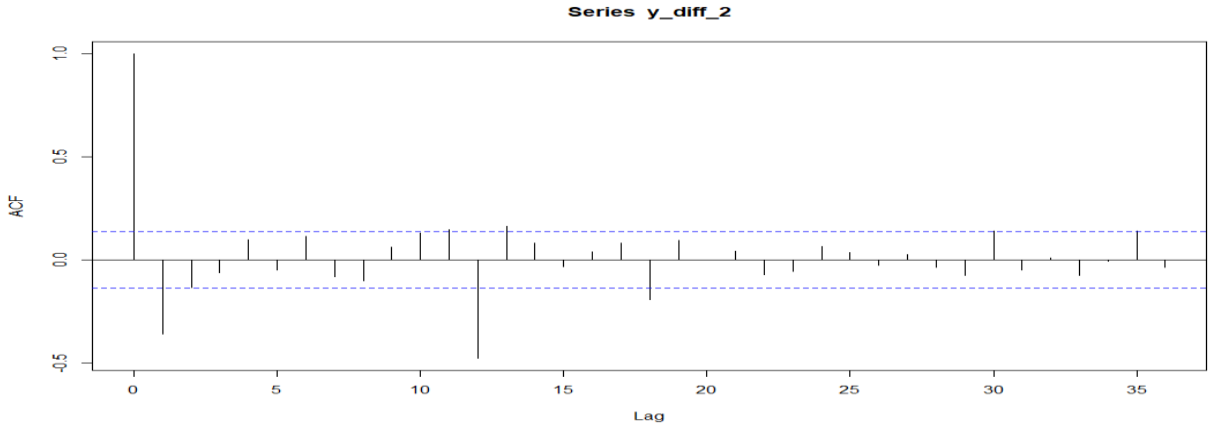
La sortie de l'ACF de (Y_t) nous démontre que la série n'est pas stationnaire. De plus, on remarque une décroissance lente vers 0 pour les multiples de 12. Donc on effectue une différentiation de $(I - B^{12})$.



La série différenciée semble être stationnaire. Pour s'assurer de ce constat, on fait les tests de stationnarité :

- KPSS : $p - val = 0.01 < 0.05$, donc on rejette l'hypothèse de stationnarité avec un risque d'erreur de 5%.
- ADF : $p - val = 0.01 < 0.05$, là on rejette l'hypothèse de non stationnarité avec un risque d'erreur de 5%.

A l'issu de ces deux tests, on ne peut pas affirmer que la série est stationnaire. Cette fois-ci, on va effectuer une deuxième différentiation de $(I - B)$.

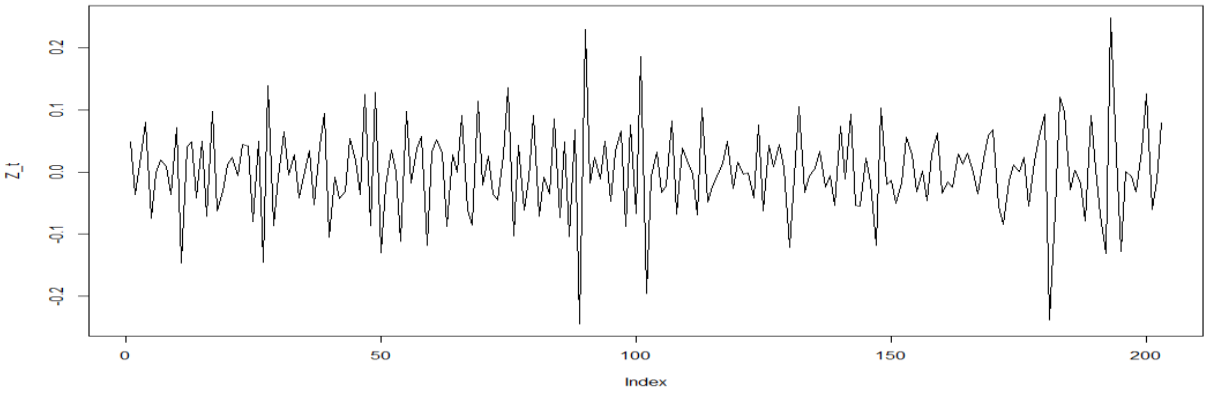


En testant la stationnarité de cette nouvelle série, on obtient :

- KPSS : $p - val = 0.1 > 0.05$
- ADF : $p - val = 0.01 < 0.05$

Les deux tests affirment à 5% d'erreur que la série doublement différentiée est stationnaire.

On identifiera donc un modèle ARMA sur la série $(Z_t) = \big((I - B)(I - B^{12})Y_t\big)$



2. *Validation des modèles* :

On va essayer d'adapter (Z_t) à des modèles ARMA ayant l'écriture suivante :

$$\Phi_s(B^s)\Phi(B)Z_t = \Theta_s(B^s)\Theta(B)\epsilon_t$$

- $\Phi_s(B^s) = I - \alpha_1 B^s - \cdots - \alpha_p B^{sp}$, $\Theta_s(B^s) = I + \beta_1 B^s + \cdots + \beta_q B^{sq}$ sont les nouveaux opérateurs.
- $\Phi(B) = I - \phi_1 B - \cdots - \phi_p B^p$, $\Theta(B) = I + \theta_1 B + \cdots + \theta_q B^q$ sont les polynômes autorégressif et moyenne mobile.
- (ϵ_t) est un bruit blanc de variance non nulle.
- On rappelle que la moyenne de (Z_t) est nulle.

Comme on l'a fait dans la modélisation additive, on va utiliser les critères suivants pour valider nos modèles :

- Significativité des coefficients estimés.
- Les résidus estimés doivent former un bruit blanc gaussien.

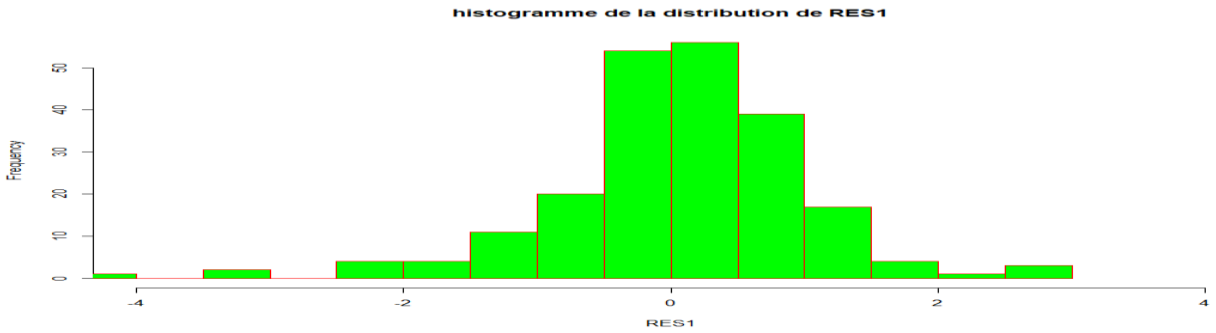
A. *Modèle 1* :

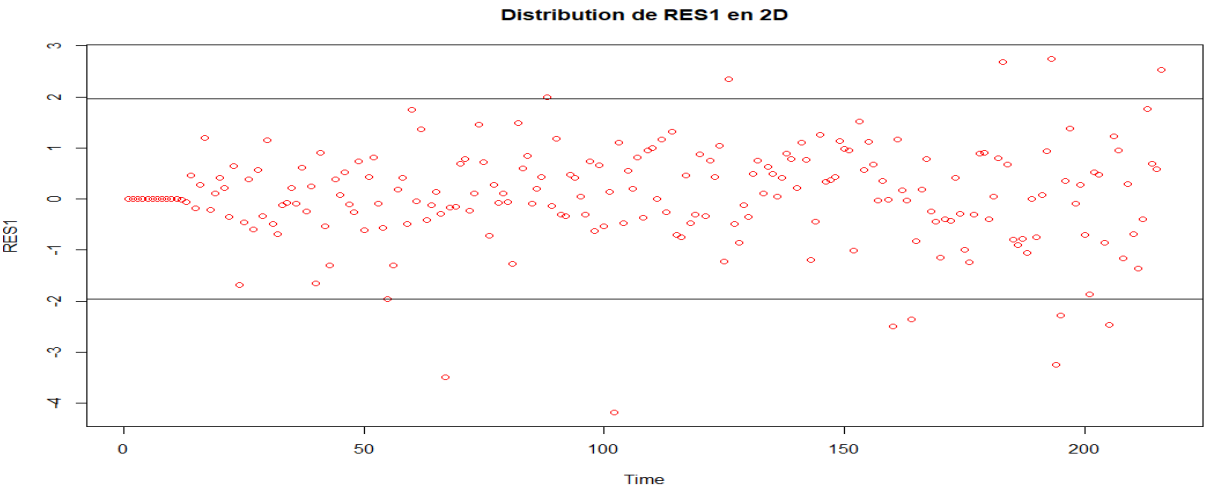
$$\text{SARIMA}(1,1,1)(0,1,1)_{12} >> (I - 0.16B)Z_t = (I - 0.49B^{12})(I - 0.89B)\epsilon_t$$

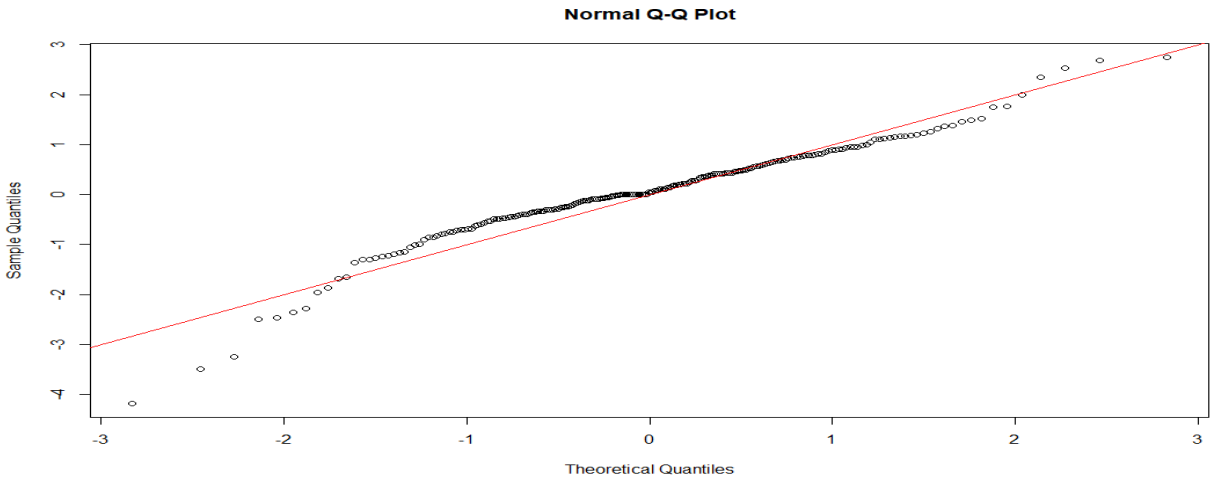
En faisant le test de Student, on constate que les coefficients estimés sont tous significatifs.

Ensuite, on passe aux tests de Box-Pierce et Ljung-Box. Ces derniers nous indiquent que les résidus estimés forment un bruit blanc.

Finalement, on va vérifier si ces résidus sont normalement distribués. Le test de Shapiro-wilk nous indique que ce n'est pas le cas ($p - val < 5\%$). Par contre, on constate clairement d'après l'histogramme et le graphe de distribution ainsi le diagramme Quantile-Quantile qu'il s'agit d'une distribution normale.







D'où, on retient le modèle.

B. *4 autres Modèles* :

En vérifiant les mêmes critères sur d'autres modèles, nous avons pu retenir :

SARIMA$(0,1,1)(0,1,0)_{12}$ >> $Z_t = (I - 0.90B\ )\epsilon_t$
SARIMA$(0,1,2)(1,1,0)_{12}$ >> $(I + 0.52B^{12})Z_t = (I - 0.69B - 0.16B^2)\epsilon_t$
SARIMA$(3,1,0)(1,1,0)_{12}$ >> $(I + 0.51B^{12})(I + 0.64B + 0.42B^2 + 0.33B^3)Z_t = \epsilon_t$
SARIMA$(3,1,0)(1,1,1)_{12}$ >> $(I + 0.29B^{12})(I + 0.61B + 0.40B^2 + 0.32B^3)Z_t = (I - 0.29B^{12})\epsilon_t$

3. *Choix du modèle ARMA* :

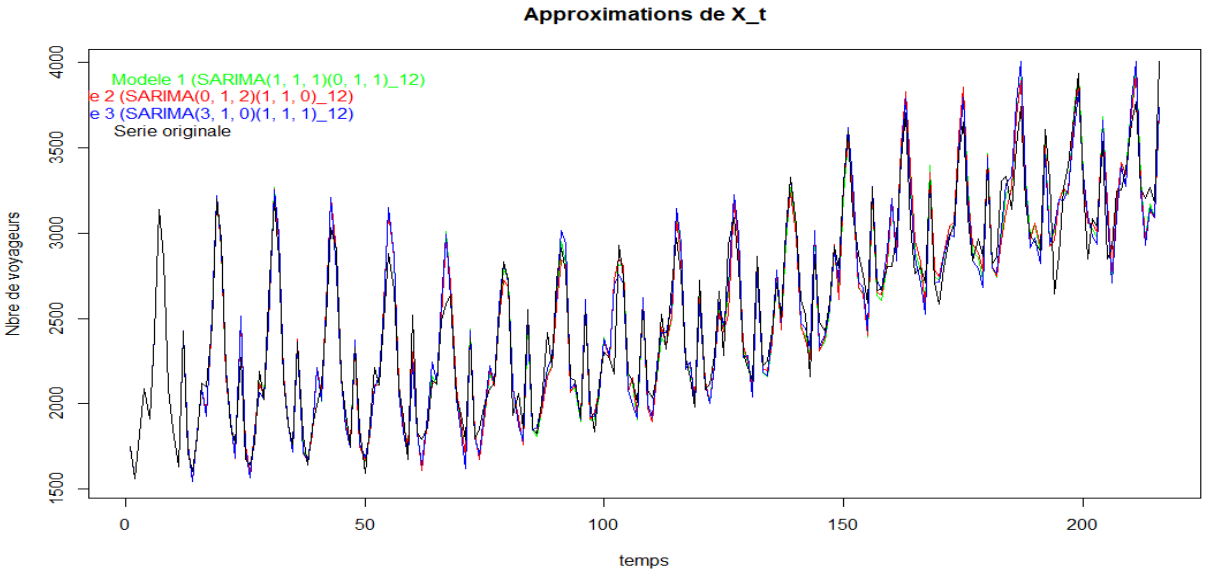
On commence par extraire les approximations de (Z_t) qu'on note $(\widehat{Z_t})$ pour les 5 modèles sélectionnés.

Ensuite, on va comparer $\left(e^{\widehat{Z_t}+\frac{\hat{\sigma}^2}{2}}\right)$ de chaque modèle à la série originale (X_t) en se basant sur les critères RMSE/ MAPE.

Modèle	RMSE	MAPE
SARIMA$(1,1,1)(0,1,1)_{12}$	125.80	3.53%
SARIMA$(0,1,1)(0,1,0)_{12}$	144.98	3.82%
SARIMA$(0,1,2)(1,1,0)_{12}$	126.47	3.53%
SARIMA$(3,1,0)(1,1,0)_{12}$	128.63	3.74%
SARIMA$(3,1,0)(1,1,1)_{12}$	127.25	3.67%

On finit par sélectionner les 3 modèles ayant la RMSE/ MAPE la plus petite, c'est à dire :

- SARIMA$(1,1,1)(0,1,1)_{12}$
- SARIMA$(0,1,2)(1,1,0)_{12}$
- SARIMA$(3,1,0)(1,1,1)_{12}$



IV. **Chapitre 3 : Prédiction**

Nous allons calculer le pouvoir prédictif des modèles SARIMA sélectionnés dans le chapitre 2. Ensuite, nous allons faire la prévision avec le meilleur modèle.

1. *Le pouvoir prédictif* :

- **Définition** : Le pouvoir prédictif d'un modèle est un critère en général très apprécié afin de comparer un ensemble de modélisations pour un même jeu de données. Une stratégie pour l'évaluer consiste à retirer une certaine quantité de mesure en fin de série, à estimer le modèle sur la série tronquée, puis à reprédire les valeurs supprimées. On obtient un écart (en un sens à déterminer) entre réalité et prédiction pour chacun des modèles.

Calculons le pouvoir prédictif de nos modèles SARIMA :

- On commence par supprimer les 12 dernières valeurs, qui définissent une période, de notre série originale (X_t).
- On adapte les modèles SARIMA sur la nouvelle série $(Y'_t) = (\log(X_t))_{t=1,\dots,204}$.
- On utilise la fonction FORCAST de R pour construire les prédicteurs $\left(\widehat{Y'_t}\right)$ permettant de prédire les 12 dernières valeurs de (Y'_t).

Il n'est pas judicieux de poser $\left(\widehat{X'_t}\right) = \left(e^{\widehat{Y'_t}}\right)$.

>> De manière informelle, vu que (Y'_t) se comporte comme un SARIMA, alors on peut l'écrire sous la forme $Y'_t = C_{t-1} + \epsilon_t$ où C_{t-1} contient une information accessible à l'instant $t-1$ et (ϵ_t) est un bruit additif.

>> Sous l'hypothèse $\epsilon_t \sim N(0, \sigma^2)$, on a $e^{\epsilon_t} \sim LN(0, \sigma^2)$ qui n'est pas centré.

>> On doit donc corriger notre prévision par le facteur $E[e^{\epsilon_t}] = e^{\frac{\sigma^2}{2}}$ que l'on estime naturellement en injectant $\hat{\sigma}^2$ à la place de σ^2.

- On en déduit les prédicteurs $\left(\widehat{X'_t}\right) = \left(e^{\widehat{Y'_t} + \frac{\hat{\sigma}^2}{2}}\right)$ de la série tronquée $(X_t)_{t=1,\dots,204}$
- Finalement, on compare les valeurs prédites pour $t \in \{205, \dots, 216\}$ avec les valeurs supprimées de (X_t).

Dans le tableau ci-dessous, on s'est basé sur le MAPE pour calculer le pouvoir prédictif de chaque modèle :

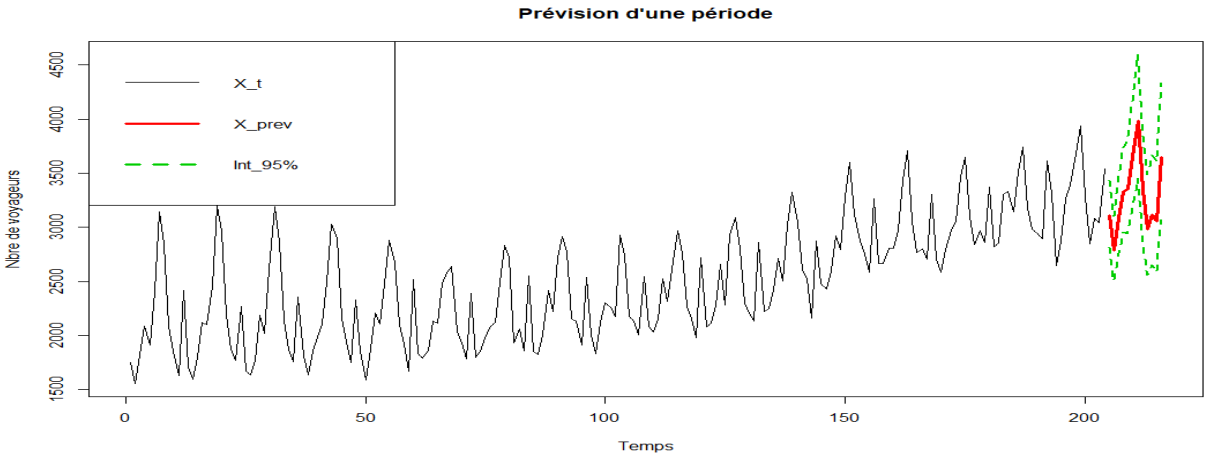
Modèle	MAPE	MAPE_N
SARIMA$(1,1,1)(0,1,1)_{12}$	4.84%	4.83%
SARIMA$(0,1,2)(1,1,0)_{12}$	4.74%	4.73%
SARIMA$(3,1,0)(1,1,1)_{12}$	4.65%	4.67%

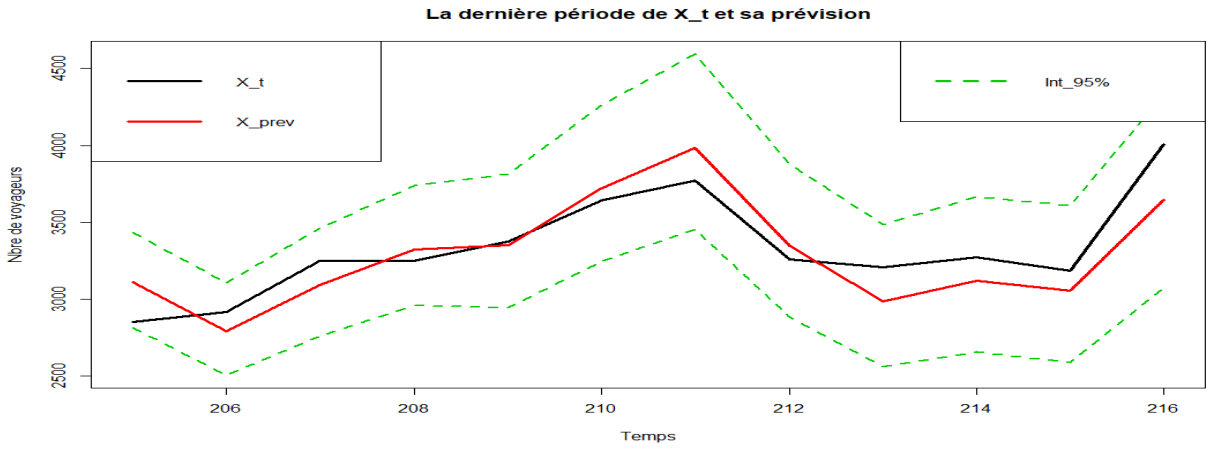
MAPE_N est simplement le MAPE des prédicteurs $\left(\widehat{X'_t}\right) = (e^{\widehat{Y'_t}})$, qui ne tiennent pas compte de la correction de la prévision par le facteur $e^{\frac{\hat{\sigma}^2}{2}}$.

>> On constate que MAPE_N est très proche MAPE car le facteur correctif $e^{\frac{\hat{\sigma}^2}{2}} \sim 1$ pour les 3 modèles.

On retient le modèle SARIMA $(3,1,0)(1,1,1)_{12}$ qui a le pouvoir prédictif le plus grand.

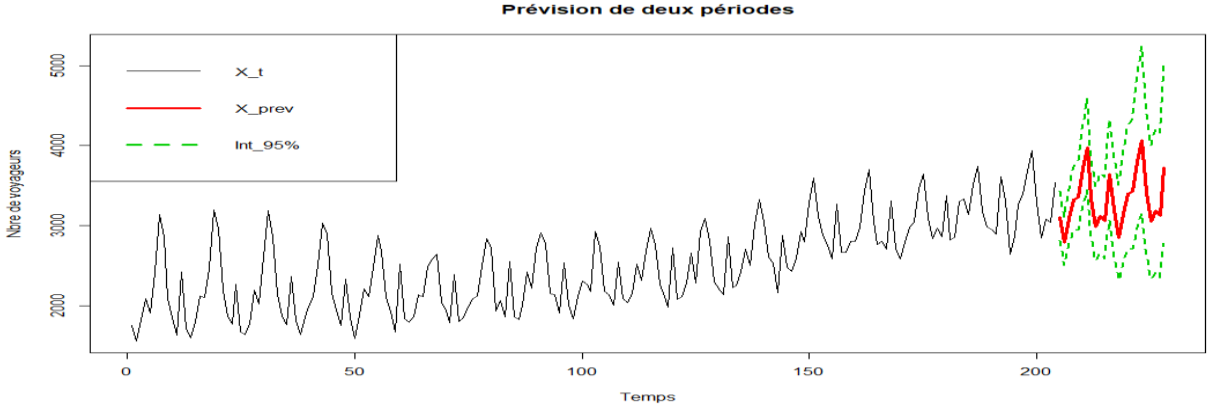
Les graphes ci-dessous représentent la prévision des 12 dernières valeurs de la série (X_t) .





On constate que la réalisation de (X_t) pendant la dernière période est bien dans l'intervalle de prévision à 95% (basé sur les données antérieures à 205).

On continue par prédire une 2ème période :



- **Remarque** : Autant qu'on prédit à long terme, la précision de la prévision diminue. Ce qui explique la dispersion de l'intervalle de prévision

V. **Chapitre 4 : Conclusion**

La plupart des professionnels analysent les données de façon relativement simpliste.

Pourtant, toute corrélation dans les données risque d'accroître le taux de fausses alertes. Il peut s'avérer opportun d'essayer de modéliser les données à l'aide d'une technique de modélisation de série chronologique telle que les ARIMA et les SARIMA.

Utiliser correctement l'approche ARIMA / SARIMA peut fournir un très bon ajustement aux données et offrir d'excellentes prévisions de comportement futur, ce qui est important dans un monde incertain.

Reste à savoir que l'identification d'une série chronologique n'est pas automatique ni unique. Le Data-Scientist intervient pour guider les choix, poser et arbitrer les compromis, trier les solutions en fonction de l'utilisation qu'il compte faire du modèle.