Projet de Séries Temporelles Linéaires

ENSAE 2A 2023-2024

Modélisation ARIMA d'un indice de la production industrielle

Fabrication de pesticides et d'autres produits agrochimiques

Julia LU

20 mai 2024

Table des matières

1 Les données

1.1 Présentation de la série

Dans ce projet, nous allons étudier un indice de la production industrielle (IPI) en France dans le secteur de l'industrie chimique. La série choisie est disponible sur le site de l'INSEE (lien de la série) et correspond à l'IPI de la fabrication de pesticides et d'autres produits agrochimiques entre janvier 1990 et février 2024. Cette série a été corrigée des variations saisonnières et des jours ouvrés (CVS-CJO). La série est mensuelle et a pour année de référence 2021, qui est l'année où elle vaut 100 en moyenne. Elle contient 410 observations. Par ailleurs, on a retiré les deux dernières valeurs de la série afin de réaliser les prédictions dans la section 3. La série est représentée sur la figure 1. On remarque que la série semble avoir une tendance non linéaire.

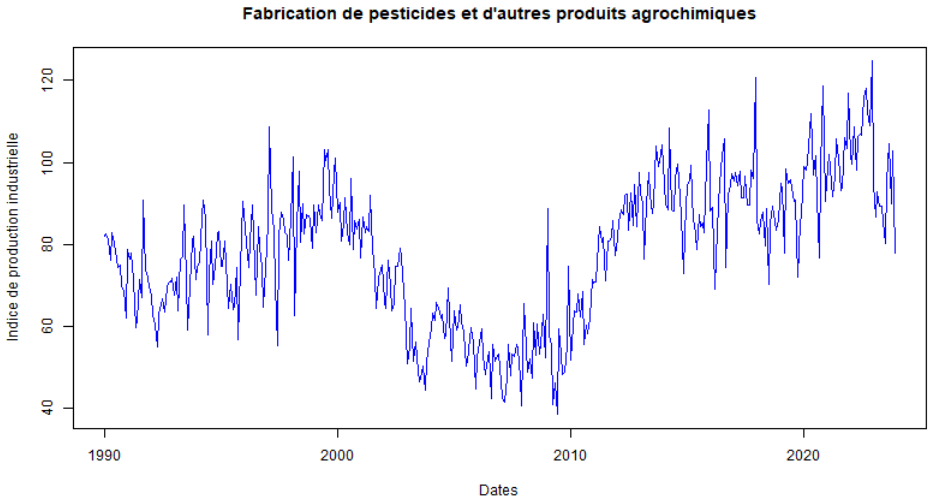


FIGURE 1 – Représentation de la série temporelle étudiée

1.2 Stationnarisation

Avant de pouvoir modéliser la série par un processus ARMA, il faut d'abord vérifier que l'hypothèse de stationnarité est plausible. Si la stationnarité n'est pas vérifiée, il faudra alors transformer la série pour la rendre stationnaire.

1.2.1 Tests de stationnarité de la série initiale

Pour étudier la stationnarité de la série initiale, on va effectuer un test de Dickey-Fuller augmenté (ADF) dont l'hypothèse nulle est la non stationnarité. Mais avant d'effectuer ce test ADF, il convient au préalable de vérifier si la série présente une tendance linéaire et/ou une constante. D'après la représentation graphique (figure 1), la tendance n'est probablement pas linéaire, mais s'il devait y en avoir une, elle serait positive. On régresse la série sur ses dates pour le vérifier (tableau 2 en annexe). Le coefficient associé aux dates est bien

positif et significatif à tous les niveaux usuels (p-valeur < 0.01). La constante est également significative. Ainsi, on effectue le test ADF dans le cas avec constante et tendance non nulles. Cependant, pour que le test ADF soit valide, et donc pouvoir l'interpréter, il faut d'abord s'assurer de l'absence d'autocorrélation des résidus de la régression, par exemple avec des tests portmanteau. On constate donc que le test ADF avec aucun retard n'est pas valide. On ajoute alors des retards jusqu'à obtenir des résidus décorrélés. On obtient ainsi un test ADF valide en considérant 18 retards. La p-valeur de ce test est de $0.6971 > 0.05$, donc le test ADF avec 18 retards ne rejette pas l'hypothèse nulle de non stationnarité de la série, autrement dit, on ne peut pas considérer la série initiale comme stationnaire.

1.2.2 Différenciation et stationnarité de la série transformée

On va donc différencier la série à l'ordre 1 pour la rendre stationnaire. En notant (Y_t) la série initiale, on considère donc la série des différences premières $\Delta Y_t = Y_t - Y_{t-1}$. Comme précédemment, on régresse la série différenciée sur les dates (tableau 3 en annexe). La régression linéaire montre que la série différenciée ne présente pas de tendance ni de constante (p-valeur > 0.05), donc on applique le test ADF dans le cas sans constante ni tendance. Dans ce cas, le test ADF est valide avec 7 retards et la p-valeur est de $0.01 < 0.05$, donc on rejette l'hypothèse de non stationnarité au seuil de 5%. Ainsi, nous pouvons conclure, avec un risque de 5%, que la série différenciée à l'ordre 1 est stationnaire. La série initiale est donc I(1).

1.3 Représentation graphique avant et après transformation

La figure 2 ci-dessous représente la série avant et après différenciation à l'ordre 1.

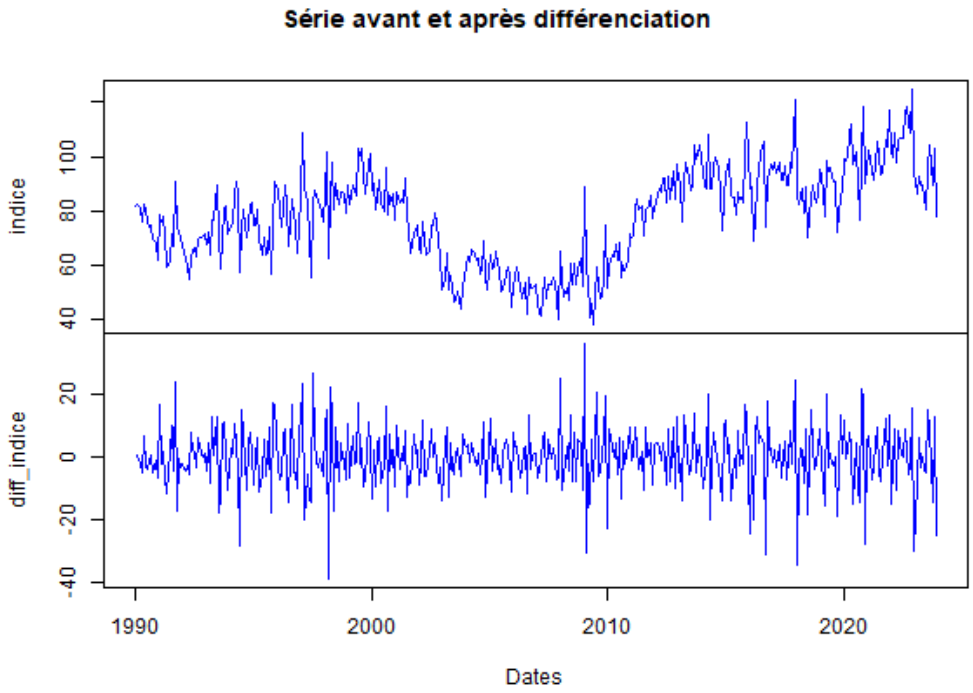


FIGURE 2 – Représentation de la série initiale (haut) et différenciée à l'ordre 1 (bas)

2 Modèles ARMA

2.1 Identification des ordres maximum p_{max} et q_{max}

On a vu que la série différenciée était stationnaire, donc on peut la modéliser par un processus ARMA(p, q) avec $p \leq p_{max}$ et $q \leq q_{max}$. Afin de déterminer les ordres q_{max} et p_{max}, on représente sur la figure 3 les fonctions d'autocorrélation (ACF) et d'autocorrélation partielle (PACF) de la série différenciée.

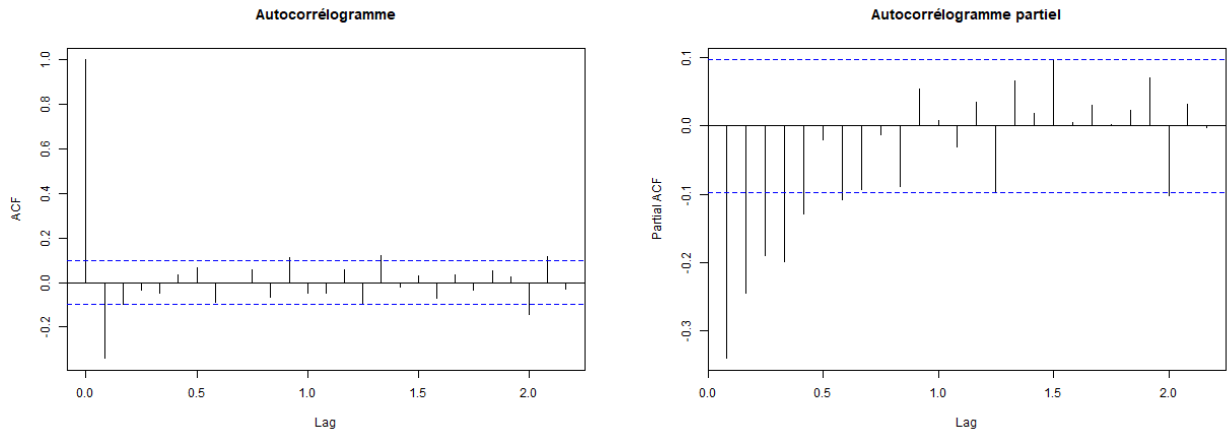


FIGURE 3 – ACF et PACF de la série différenciée

L'ACF permet d'identifier l'ordre q_{max}, tandis que la PACF permet d'identifier p_{max}. Pour cela, on considère les dernières autocorrélations significativement différentes de zéro (ie. dépassant les bornes $\pm 1.96/\sqrt{n}$ de l'intervalle de confiance à 95% d'un test de nullité de l'autocorrélation). D'après l'ACF, le dernier pic significatif (à la limite de la significativité) est celui avec 2 retards, d'où $q_{max} = 2$. D'après la PACF, le dernier pic significatif est celui avec 7 retards, d'où $p_{max} = 7$. On ignore les éventuels pics significatifs pour des ordres supérieurs.

2.2 Choix du modèle et validité

On estime donc les coefficients de tous les modèles ARMA(p, q) avec $0 \leq p \leq 7$ et $0 \leq q \leq 2$, et on calcule les critères AIC et BIC pour chacun des modèles. Les résultats sont disponibles dans le tableau 4 et le tableau 5 en annexe. Les deux critères sont minimisés pour $p = 0$ et $q = 2$. On retient donc le modèle ARMA(0,2) ou MA(2). Les coefficients estimés de ce modèle sont affichés dans le tableau 1.

	ma1	ma2
Estimation	-0.57	-0.18
Écart-type	0.05	0.05
p-value	0.00	0.00

TABLE 1 – Coefficients du modèle MA(2)

On constate que les deux coefficients du modèle sont significatifs à tous les niveaux usuels (p-valeur < 0.01), donc le modèle MA(2) est bien ajusté. Il reste donc à vérifier la validité du modèle, c'est-à-dire l'absence d'autocorrelation des résidus.

Pour cela, on réalise des tests de Ljung-Box avec pour hypothèse nulle l'absence d'autocorrélation des résidus d'ordre 1 à k donné. On fait varier le lag k de 1 à 24. Les résultats des différents tests sont rassemblés dans le tableau 6 en annexe. On constate que toutes les p-valeurs sont au dessus de 0.05, donc pour chaque retard $k \in \{1, ..., 24\}$, on ne rejette pas l'hypothèse d'absence d'autocorrélation. Ainsi, le modèle MA(2) est bien valide.

2.3 Écriture du modèle ARIMA pour la série initiale

D'après ce qui précède, la série différenciée suit un modèle MA(2), donc en la notant (X_t) et en utilisant les coefficients estimés plus haut, le modèle vérifié par (X_t) s'écrit :

$$X_t = \epsilon_t - 0.57\epsilon_{t-1} - 0.18\epsilon_{t-2}$$

Or, $X_t = Y_t - Y_{t-1} = (1-L)Y_t$, où L est l'opérateur de lag, et X_t est un processus ARMA(0,2) = MA(2) causal, donc d'après le cours, la série initiale (Y_t) suit un modèle ARIMA(0,1,2). Le modèle vérifié par (Y_t) s'écrit ainsi :

$$Y_t = Y_{t-1} + \epsilon_t - 0.57\epsilon_{t-1} - 0.18\epsilon_{t-2}$$

Par ailleurs, d'après la figure 6 en annexe, les inverses des racines du polynôme MA sont situés à l'intérieur du cercle unité, ce qui signifie que les racines sont bien de module strictement supérieur à 1, donc notre modèle MA(2) est inversible. De plus, il est a fortiori causal puisqu'il n'a pas de partie auto-régressive. Le modèle MA(2) suivi par la série différenciée (X_t) est donc canonique, c'est-à-dire causal et inversible. On peut donc l'utiliser pour réaliser les prédictions.

2.4 Normalité des résidus

Dans le cadre de la prévision, il convient également de vérifier la normalité des résidus. Pour cela, on a représenté en annexe (figure 7) les résidus de notre modèle MA(2) avec leur autocorélogramme et leur distribution. On constate que les résidus se comportent bien comme un bruit blanc (espérance nulle, variance constante et on a vu que les résidus n'étaient pas autocorrélés). De plus, l'histogramme des résidus ressemble à une distribution normale. Pour en être plus sûr, on a tracé sur la figure 8 en annexe le diagramme Quantile-Quantile des résidus. On remarque qu'au centre, les points sont alignés sur la droite, ce qui indique bien une distribution normale. Cependant, on observe des écarts aux extrémités. La distribution n'est donc pas gaussienne sur les queues. Le test de Shapiro Wilk vient confirmer cela. En effet, l'hypothèse nulle de normalité de la distribution est rejetée à tous les niveaux usuels (p-valeur = $0.0037 < 0.01$). Ainsi, les résidus ne sont pas gaussiens bien que leur distribution soit proche de celle d'une loi normale.

3 Prévision

Dans cette partie, on note T la longueur de la série différenciée (X_t), on a $T = 408$, et on suppose que les résidus de la série sont gaussiens, soit $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

3.1 Région de confiance de niveau α

La série différenciée (X_t) suit un modèle MA(2) qui s'écrit :

$$X_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2}$$

Puisque (X_t) suit un modèle MA(2) canonique, les résidus sont l'innovation linéaire de (X_t), ils sont donc orthogonaux à toute fonction du passé linéaire. D'où, en particulier : $EL(\epsilon_{T+1}|X_T, X_{T-1}, ...) = EL(\epsilon_{T+2}|X_T, X_{T-1}, ...) = 0$. Les meilleures prévisions de X_{T+1} et X_{T+2} sachant le passé vérifient donc les équations suivantes :

$$\begin{cases} \hat{X}_{T+1|T} = EL(X_{T+1}|X_T, ...) = -\theta_1 \epsilon_T - \theta_2 \epsilon_{T-1} \\ \hat{X}_{T+2|T} = EL(X_{T+2}|X_T, ...) = -\theta_2 \epsilon_T \end{cases}$$

Posons $X = (X_{T+1}, X_{T+2})^T$ et $\hat{X} = (\hat{X}_{T+1|T}, \hat{X}_{T+2|T})^T$ et calculons la différence $\tilde{X} = X - \hat{X}$.

$$\tilde{X} = \begin{pmatrix} \tilde{X}_{T+1} \\ \tilde{X}_{T+2} \end{pmatrix} = \begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} \\ X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} = \begin{pmatrix} \epsilon_{T+1} \\ \epsilon_{T+2} - \theta_1 \epsilon_{T+1} \end{pmatrix}$$

Comme les résidus sont i.i.d. et gaussiens de variance σ^2, le vecteur $\tilde{X}$ est gaussien et :

$$\mathbb{V}[\tilde{X}_{T+1}] = \sigma^2, \quad \mathbb{V}[\tilde{X}_{T+2}] = \sigma^2(1 + \theta_1^2), \quad \mathrm{Cov}(\tilde{X}_{T+1}, \tilde{X}_{T+2}) = -\theta_1 \sigma^2$$

D'où : $\tilde{X} \sim \mathcal{N}(0, \Sigma)$ avec $\Sigma = \sigma^2 \begin{pmatrix} 1 & -\theta_1 \\ -\theta_1 & 1 + \theta_1^2 \end{pmatrix}$

La matrice Σ est inversible ssi $\det(\Sigma) = \sigma^4 > 0$, ssi $\sigma^2 > 0$, ce qu'on suppose vrai.

De plus, $\tilde{X}^T \Sigma^{-1} \tilde{X} = \frac{\epsilon_{T+1}^2}{\sigma^2} + \frac{\epsilon_{T+2}^2}{\sigma^2}$ qui est la somme de deux variables indépendantes de loi gaussienne centrée réduite, d'où : $\tilde{X}^T \Sigma^{-1} \tilde{X} \sim \chi^2(2)$

On en déduit la région de confiance de niveau α sur les valeurs futures de $(X_{T+1}, X_{T+2})^T$, qui correspond à l'intérieur d'une ellipse :

$$R_\alpha = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \ : \ \begin{pmatrix} x_1 - \hat{X}_{T+1|T} \\ x_2 - \hat{X}_{T+2|T} \end{pmatrix}^T \hat{\Sigma}^{-1} \begin{pmatrix} x_1 - \hat{X}_{T+1|T} \\ x_2 - \hat{X}_{T+2|T} \end{pmatrix} \leq q_{1-\alpha}^{\chi^2(2)} \right\}$$

avec $q_{1-\alpha}^{\chi^2(2)}$ le quantile d'ordre $1 - \alpha$ de la loi $\chi^2(2)$.

On peut également obtenir les intervalles de confiance univariés pour X_{T+1} et X_{T+2} (voir annexe B).

3.2 Hypothèses

Pour obtenir la région de confiance, on a utilisé les hypothèses suivantes :

1. Le modèle MA(2) est parfaitement connu et les coefficients estimés en partie 2.2 sont les vrais coefficients du modèle.
2. Les résidus sont i.i.d. et gaussiens, ie. $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
3. La variance des résidus est connue et strictement positive ($\sigma^2 > 0$).

3.3 Représentation des prédictions et de la région de confiance

La figure 4 représente la série initiale limitée à partir de l'année 2020, les estimations $\hat{X}_{T+1|T}$ et $\hat{X}_{T+2|T}$ (points rouges) et leur intervalle de confiance à 95% (zones grises). On constate que les vraies valeurs X_{T+1} et X_{T+2} se situent dans les intervalles de confiance, donc nos estimations sont correctes. De plus, on remarque que l'intervalle de confiance est plus large pour la deuxième prédiction.

La figure 5 représente la région de confiance obtenue dans la partie 3.1. Cette région prend la forme d'une ellipse dont le centre est $(\hat{X}_{T+1}, \hat{X}_{T+2}) = (87.59, 91.13)$. Le point bleu correspond aux vraies valeurs $(X_{T+1}, X_{T+2}) = (90.21, 96.95)$ qui ont été retirées de la série initiale au début de l'étude pour pouvoir comparer les prédictions. De même, les vraies valeurs sont situées dans la région de confiance.

Cependant, nous avons vu que ces intervalles de confiance reposent sur l'hypothèse que les résidus sont gaussiens. Or, d'après la partie 2.4, cette hypothèse ne semble pas réaliste.

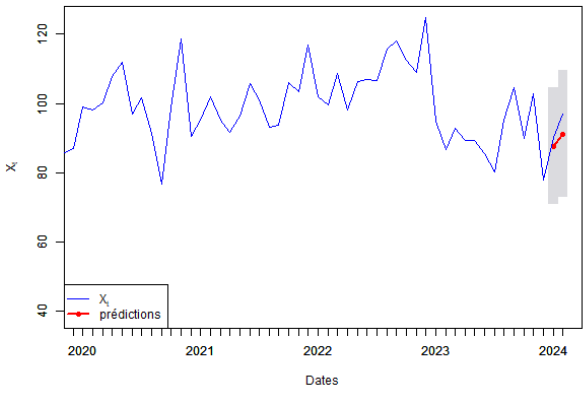


FIGURE 4 – Prédictions de X_{T+1} et X_{T+2} avec leur intervalle de confiance à 95%

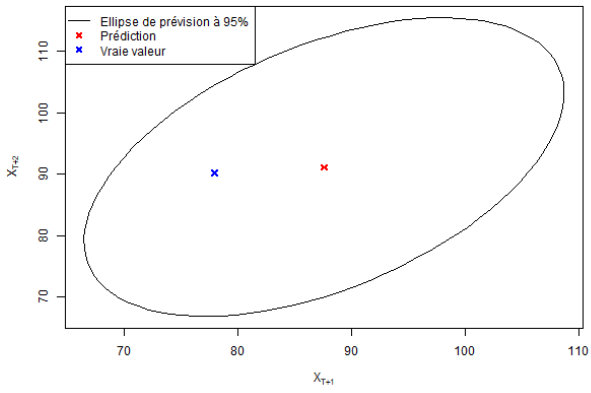


FIGURE 5 – Région de confiance à 95% pour le vecteur (X_{T+1}, X_{T+2})

3.4 Question ouverte

Soit (Y_t) une série stationnaire disponible de $t = 1$ à T. On suppose que Y_{T+1} est disponible plus rapidement que X_{T+1}. D'après le cours, Y_{T+1} améliore la prévision de X_{T+1} si (Y_t) cause instantanément (X_t) au sens de Granger, ie : $\hat{X}_{T+1|\{X_u, Y_u, u \leq T\} \cup \{Y_{T+1}\}} \neq \hat{X}_{T+1|\{X_u, Y_u, u \leq T\}}$

Cette condition peut être testée avec un test de Wald.

Annexes

A Tables et figures

Coefficients	Estimation	Écart-type	t-value	p-value
(Intercept)	-1392.2	159.38	-8.735	<2e-16
dates	0.7326	0.0794	9.226	<2e-16

Table 2 – Résultats de la régression de la série initiale sur les dates

Coefficients	Estimation	Écart-type	t-value	p-value
(Intercept)	-3.1793	101.7909	-0.031	0.9751
dates	0.0016	0.0507	0.031	0.9752

Table 3 – Résultats de la régression de la série différenciée sur les dates

	q=0	q=1	q=2
p=0	3030.77	2926.74	2915.09
p=1	2981.71	2915.14	2916.83
p=2	2958.44	2916.72	2917.02
p=3	2944.68	2918.01	2919.02
p=4	2929.58	2919.18	2921.08
p=5	2924.74	2920.90	2922.69
p=6	2926.58	2922.73	2923.96
p=7	2923.74	2921.49	2923.39

Table 4 – AIC

	q=0	q=1	q=2
p=0	3034.78	2934.76	2927.11
p=1	2989.73	2927.16	2932.87
p=2	2970.47	2932.76	2937.06
p=3	2960.72	2938.05	2943.07
p=4	2949.63	2943.23	2949.14
p=5	2948.79	2948.96	2954.76
p=6	2954.64	2954.80	2960.04
p=7	2955.81	2957.57	2963.48

Table 5 – BIC

lag	p-value
1	
2	0.77
3	0.70
4	0.70
5	0.78
6	0.73
7	0.63
8	0.74
9	0.71
10	0.78
11	0.51
12	0.56
13	0.60
14	0.61
15	0.63
16	0.23
17	0.27
18	0.29
19	0.29
20	0.34
21	0.38
22	0.42
23	0.48
24	0.23

TABLE 6 – Tests d'autocorrélation des résidus pour le modèle MA(2)

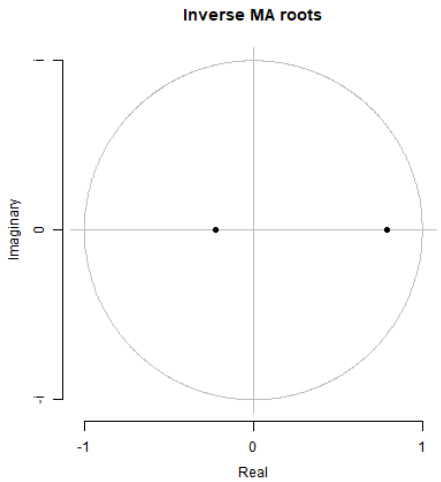


FIGURE 6 – Inverse des racines du modèle MA(2)

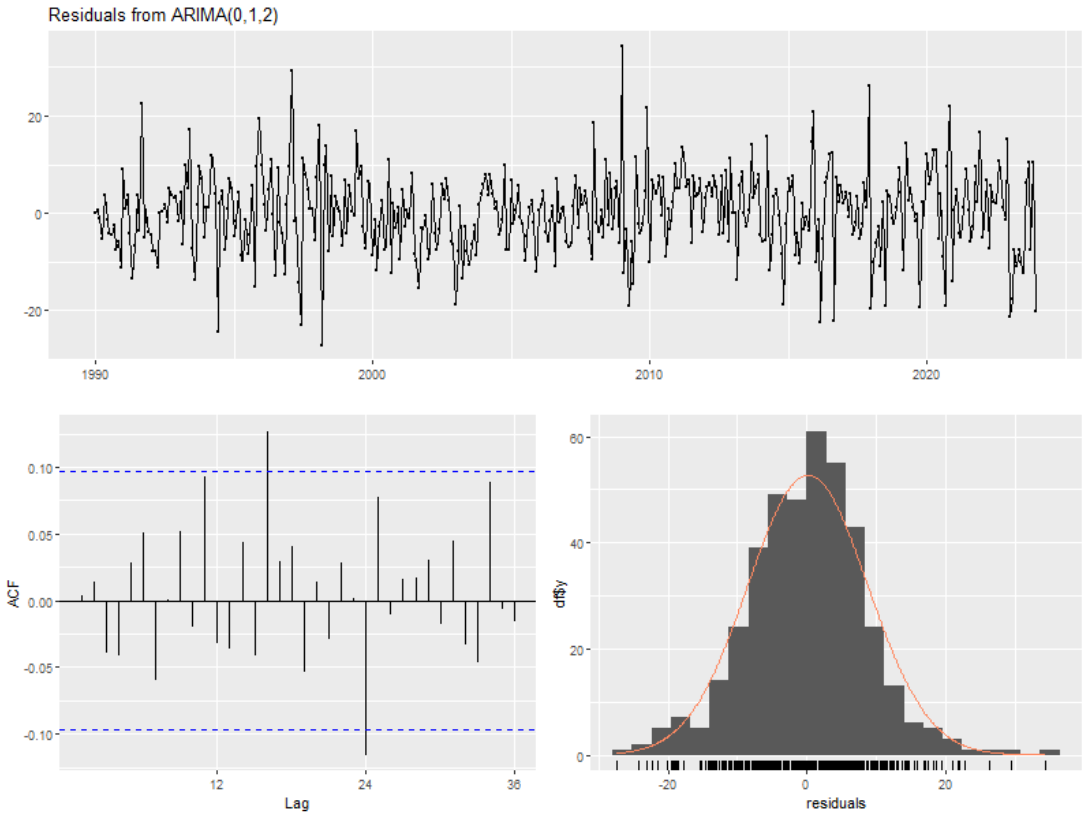


FIGURE 7 – Résidus du modèle MA(2)

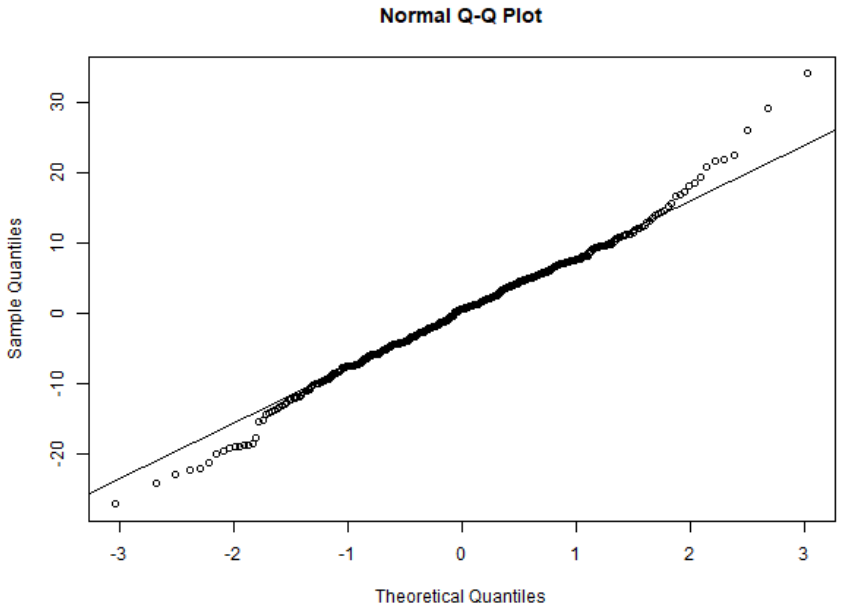


FIGURE 8 – QQplot des résidus du modèle MA(2)

B Prévision et intervalle de confiance

Puisque, par hypothèse, les résidus sont gaussiens, on a vu que :

$$\begin{cases} X_{T+1} - \hat{X}_{T+1|T} \sim \mathcal{N}(0, \sigma^2) \\ X_{T+2} - \hat{X}_{T+2|T} \sim \mathcal{N}(0, \sigma^2(1+\theta_1^2)) \end{cases}$$

On en déduit donc les intervalles de confiance à 95% pour X_{T+1} et X_{T+2} après estimation du coefficient θ_1 et de l'écart-type des résidus $\hat{\sigma}$:

$$IC_{95\%}(X_{T+1}) = \left[\hat{X}_{T+1|T} - 1.96\hat{\sigma} \ ; \ \hat{X}_{T+1|T} + 1.96\hat{\sigma}\right]$$
$$IC_{95\%}(X_{T+2}) = \left[\hat{X}_{T+2|T} - 1.96\hat{\sigma}\sqrt{1+\theta_1^2} \ ; \ \hat{X}_{T+2|T} + 1.96\hat{\sigma}\sqrt{1+\theta_1^2}\right]$$

On obtient ainsi :

$$IC_{95\%}(X_{T+1}) = [70.7 \ ; \ 104.5]$$
$$IC_{95\%}(X_{T+2}) = [71.7 \ ; \ 110.6]$$

C Code R

Projet de Série Temporelles Linéaires

Julia Lu

2024-05-20

Chargement des librairies

```
library(zoo)
library(tseries)
library(forecast)
library(fUnitRoots)
library(ellipse)
library(xtable)
```

Partie I : données

Q1 - Présentation de la série

La série étudiée est l'IPI de la fabrication de pesticides et d'autres produits agrochimiques (https://www.insee.fr/fr/statistiques/serie/010767804)

```
# Importation des données et renommage des colonnes
data <- read.csv("valeurs_mensuelles_pesticides.csv", sep = ";", col.names = c("Dates",
    "Indice", "Codes"))

# On enlève les 3 premières lignes qui ne sont pas des données et on enlève la
# troisième colonne qui n'est pas utile
data <- data[-(1:3), 1:2]
rownames(data) <- NULL  #on réinitialise l'index du DataFrame
# On convertit les valeurs de la colonne 'Indice' en données numériques
data$Indice <- as.numeric(data$Indice)

# On définit les dates de la série
data$Dates[1]  #première date : janvier 1990
```

```
## [1] "1990-01"
```

```
tail(data$Dates, 1)  #dernière date : février 2024
```

```
## [1] "2024-02"
```

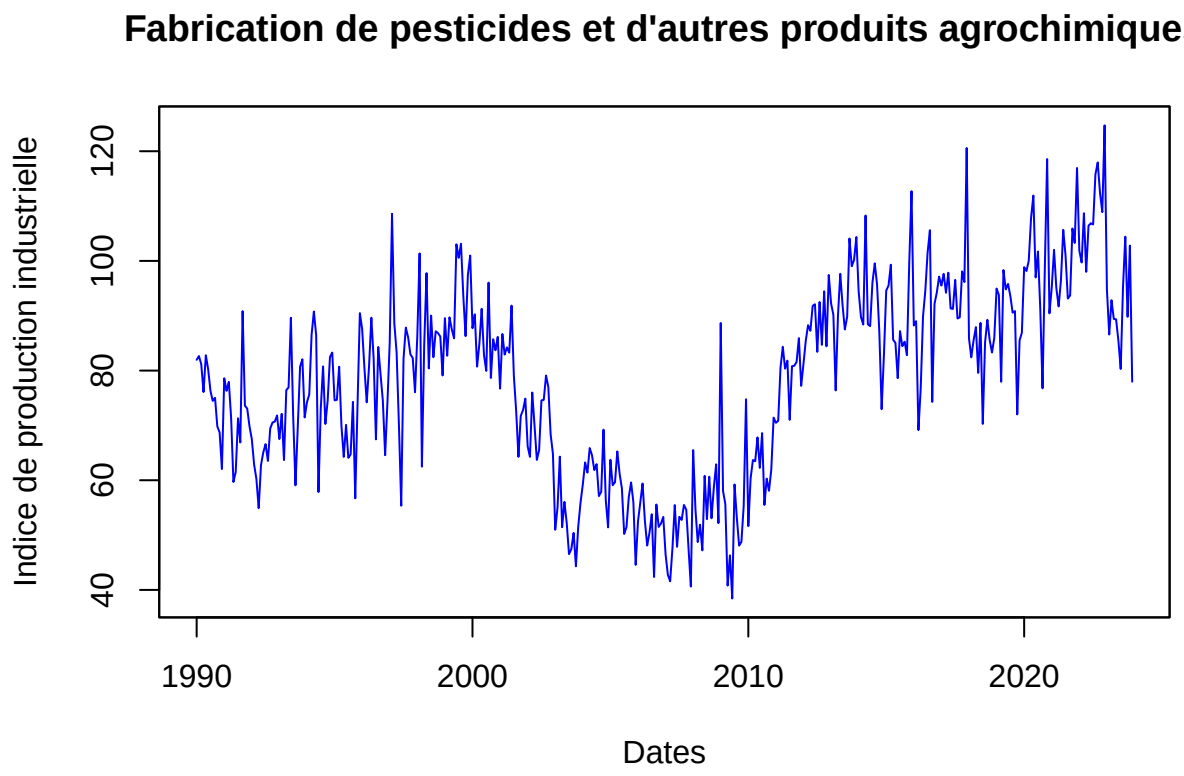
```
dates <- as.yearmon(seq(from = 1990 + 0/12, to = 2024 + 1/12, by = 1/12))

# On créé la série temporelle associée aux valeurs prises par l'indice de
# production
indice.source <- zoo(data$Indice, order.by = dates)

# On supprime les 2 dernières valeurs pour la prédiction
```

```r
indice <- indice.source[1:(length(indice.source) - 2)]
dates2 <- dates[1:(length(dates) - 2)]

# On trace la série
plot(indice, xlab = "Dates", ylab = "Indice de production industrielle", main = "Fabrication de pesticid
    col = "blue")
```



Q2 - Transformation de la série

```r
# On régresse les valeurs de la série sur les dates pour vérifier si la série
# présente une tendance
reg <- lm(indice ~ dates2)
summary(reg)
```

```
##
## Call:
## lm(formula = indice ~ dates2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.514 -10.968   3.196  11.236  37.672
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.392e+03  1.594e+02  -8.735   <2e-16 ***
## dates2       7.327e-01  7.941e-02   9.226   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.74 on 406 degrees of freedom
## Multiple R-squared:  0.1733, Adjusted R-squared:  0.1713
## F-statistic: 85.12 on 1 and 406 DF,  p-value: < 2.2e-16
```

La régression linéaire met en évidence une tendance a priori croissante pour la série (le coefficient associé aux dates est significativement positif) .

```
# On effectue le test de racine unitaire ADF dans le cas avec constante et
# tendance pour déterminer si la série est stationnaire ou non L'hypothèse
# nulle est la non stationnarité (présence de racine unitaire)
adf <- adfTest(indice, lag = 0, type = "ct")
adf
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 0
##   STATISTIC:
##     Dickey-Fuller: -6.7616
##   P VALUE:
##     0.01
##
## Description:
##  Mon May 20 19:12:29 2024 by user: lujul
```

Cependant, on ne peut pas interpréter ce test car on ne sait pas s'il est valide. Pour que le test soit valide, il faut que les résidus de la régression soient non autocorrélés.

```
# On teste donc l'autocorrélation des résidus dans la régression
Qtests <- function(series, k, fitdf = 0) {
    pvals <- apply(matrix(1:k), 1, FUN = function(l) {
        pval <- if (l <= fitdf)
            NA else Box.test(series, lag = l, type = "Ljung-Box", fitdf = fitdf)$p.value
        return(c(lag = l, pval = pval))
    })
    return(t(pvals))
}
Qtests(adf@test$lm$residuals, 24, fitdf = length(adf@test$lm$coefficients))
```

```
##       lag         pval
##  [1,]   1           NA
##  [2,]   2           NA
##  [3,]   3           NA
##  [4,]   4 5.938085e-06
##  [5,]   5 3.041763e-06
##  [6,]   6 2.631102e-07
##  [7,]   7 9.804283e-07
##  [8,]   8 1.069681e-06
##  [9,]   9 1.658078e-07
## [10,]  10 4.518111e-07
## [11,]  11 4.413532e-09
```

```
## [12,]  12 1.067759e-08
## [13,]  13 2.496321e-08
## [14,]  14 4.632029e-09
## [15,]  15 1.049803e-08
## [16,]  16 9.048595e-11
## [17,]  17 1.416562e-10
## [18,]  18 7.852885e-11
## [19,]  19 1.863300e-10
## [20,]  20 1.111194e-10
## [21,]  21 2.253794e-10
## [22,]  22 7.821144e-11
## [23,]  23 6.043954e-11
## [24,]  24 4.929224e-11
```

Toutes les p-valeurs sont en dessous du seuil de 5% donc l'hypothèse nulle d'absence d'autocorrélation des résidus est rejetée. Le test ADF avec aucun retard n'est donc pas valide.

On ajoute alors des lags jusqu'à obtenir des résidus décorrélés avec la fonction ci-dessous (cf TD5).

```r
adfTest_valid <- function(series, kmax, adftype) {
    k <- 0
    noautocorr <- 0
    while (noautocorr == 0) {
        cat(paste0("ADF with ", k, " lags: residuals OK? "))
        adf <- adfTest(series, lags = k, type = adftype)
        pvals <- Qtests(adf@test$lm$residuals, 24, fitdf = length(adf@test$lm$coefficients))[,
            2]
        if (sum(pvals < 0.05, na.rm = T) == 0) {
            noautocorr <- 1
            cat("OK \n")
        } else cat("nope \n")
        k <- k + 1
    }
    return(adf)
}
adf <- adfTest_valid(indice, 24, "ct")
```

```
## ADF with 0 lags: residuals OK? nope
## ADF with 1 lags: residuals OK? nope
## ADF with 2 lags: residuals OK? nope
## ADF with 3 lags: residuals OK? nope
## ADF with 4 lags: residuals OK? nope
## ADF with 5 lags: residuals OK? nope
## ADF with 6 lags: residuals OK? nope
## ADF with 7 lags: residuals OK? nope
## ADF with 8 lags: residuals OK? nope
## ADF with 9 lags: residuals OK? nope
## ADF with 10 lags: residuals OK? nope
## ADF with 11 lags: residuals OK? nope
## ADF with 12 lags: residuals OK? nope
## ADF with 13 lags: residuals OK? nope
## ADF with 14 lags: residuals OK? nope
## ADF with 15 lags: residuals OK? nope
## ADF with 16 lags: residuals OK? nope
## ADF with 17 lags: residuals OK? nope
## ADF with 18 lags: residuals OK? OK
```

```
adf
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 18
##   STATISTIC:
##     Dickey-Fuller: -1.7167
##   P VALUE:
##     0.6971
##
## Description:
##  Mon May 20 19:12:30 2024 by user: lujul
```

On doit considérer 18 lags pour que les résidus ne soient plus autocorrélés. Le test ADF avec 18 lags est donc valide.

pval=0.6971 $>$ 0.05 donc on ne rejette pas l'hypothèse nulle de non stationnarité, ie. la série initiale n'est pas stationnaire, donc la série est au moins I(1).

```
# On va donc considérer la série différenciée à l'ordre 1
diff_indice <- diff(indice, 1)
```

```
# De même, on régresse la série différenciée sur les dates et une constante
reg_diff <- lm(diff_indice ~ dates2[-1])
summary(reg_diff)
```

```
##
## Call:
## lm(formula = diff_indice ~ dates2[-1])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.866  -5.610   0.300   5.905  36.507
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.179263 101.790946  -0.031    0.975
## dates2[-1]    0.001579   0.050717   0.031    0.975
##
## Residual standard error: 10.02 on 405 degrees of freedom
## Multiple R-squared:  2.394e-06,  Adjusted R-squared:  -0.002467
## F-statistic: 0.0009695 on 1 and 405 DF,  p-value: 0.9752
```

Il n'y a pas de tendance ni de constante significatives (p-valeur $>$ 0.05).

On effectue le test ADF dans le cas "nc" (sans constante et sans tendance) en contrôlant pour l'absence d'autocorrélation entre les résidus.

```
adf_diff <- adfTest_valid(diff_indice, 24, "nc")
```

```
## ADF with 0 lags: residuals OK? nope
## ADF with 1 lags: residuals OK? nope
## ADF with 2 lags: residuals OK? nope
## ADF with 3 lags: residuals OK? nope
```

```
## ADF with 4 lags: residuals OK? nope
## ADF with 5 lags: residuals OK? nope
## ADF with 6 lags: residuals OK? nope
## ADF with 7 lags: residuals OK? OK
```

```
adf_diff
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 7
##   STATISTIC:
##     Dickey-Fuller: -10.801
##   P VALUE:
##     0.01
##
## Description:
##  Mon May 20 19:12:30 2024 by user: lujul
```

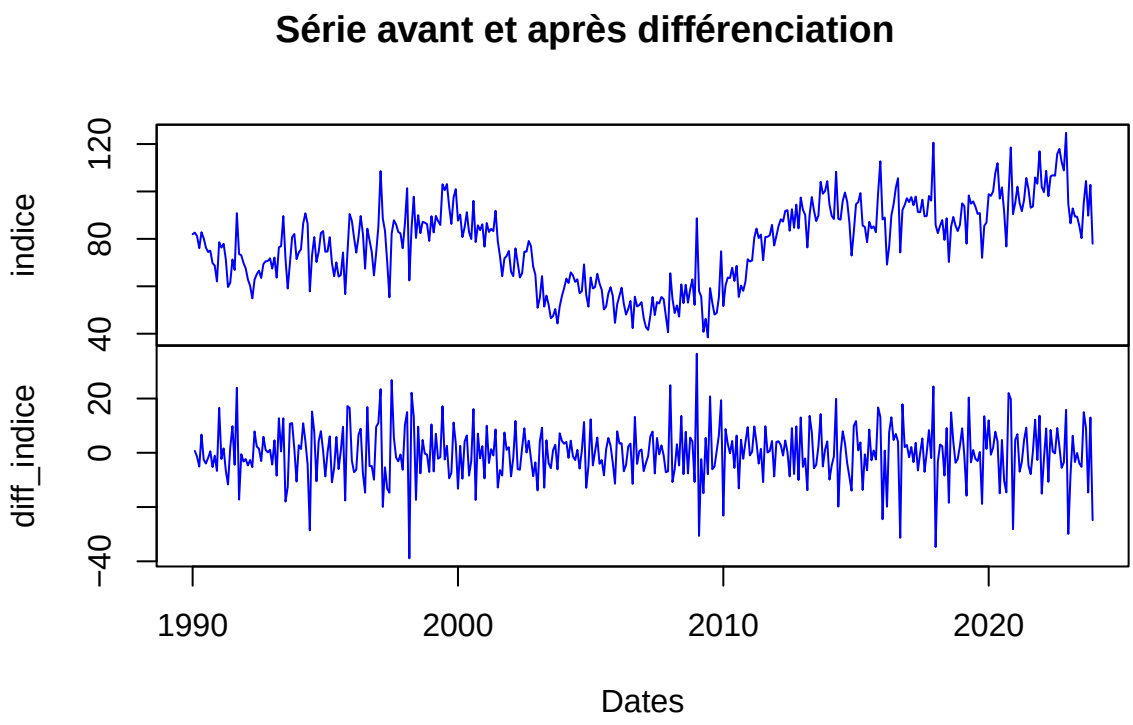
Le test ADF avec 7 lags est valide.

pval=0.01 $<$ 0.05 donc on rejette l'hypothèse de non stationnarité au seuil de 5%

Conclusion : la série différenciée est stationnaire

Q3 - Représentation graphique avant et après transformation

```
plot(cbind(indice, diff_indice), xlab = "Dates", main = "Série avant et après différenciation",
    col = "blue")
```

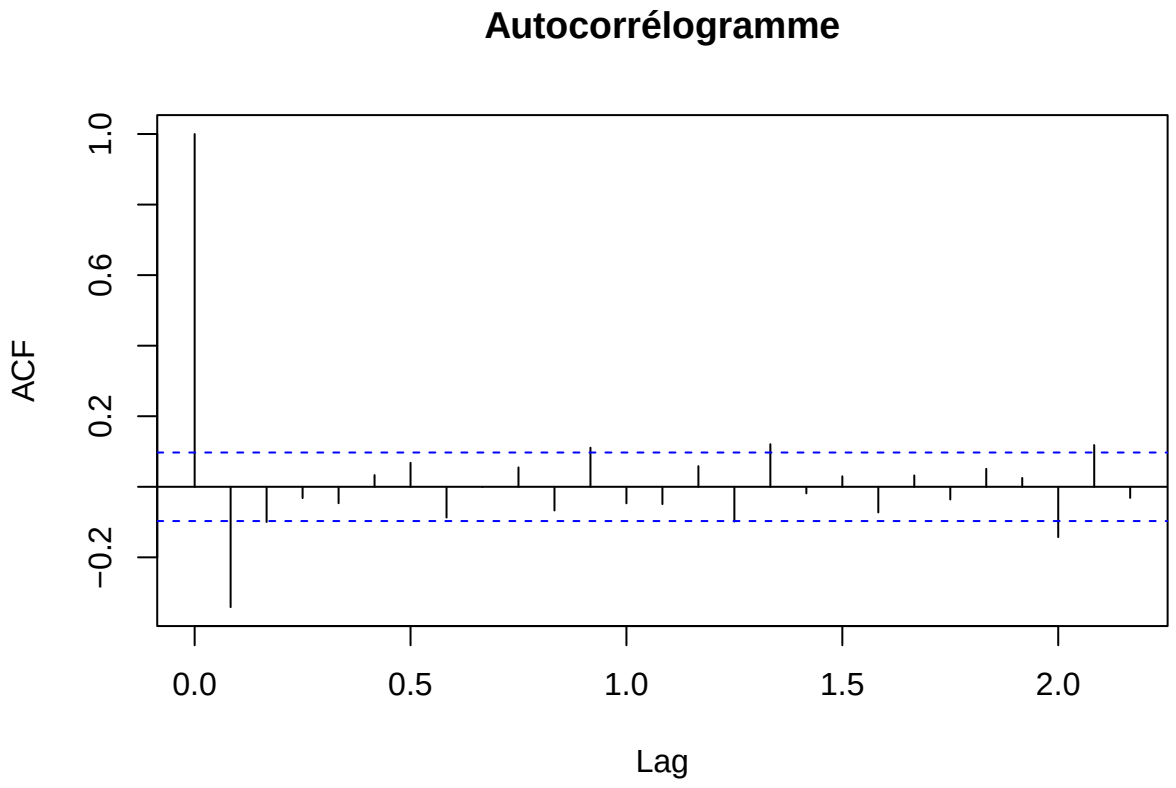


Partie 2 : Modèles ARMA

Q4 - Choix du modèle ARMA(p,q) pour la série différenciée

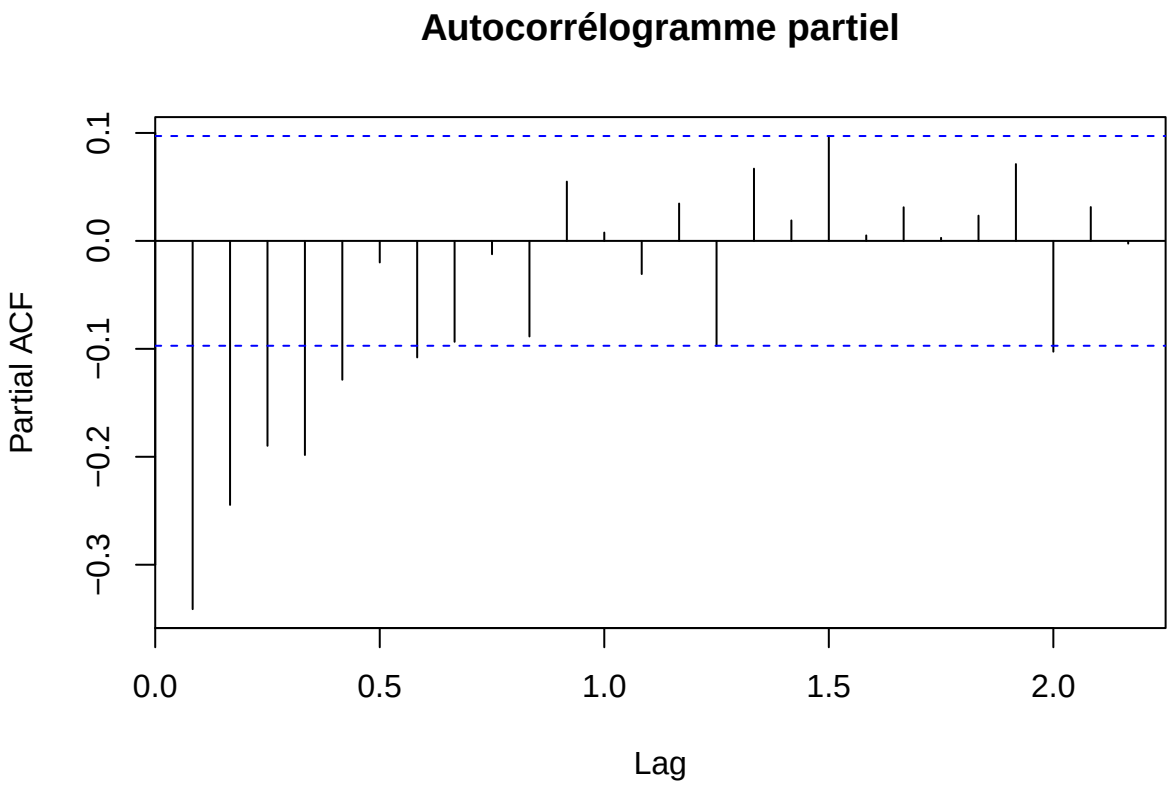
On détermine d'abord qmax et pmax grâce à l'acf et la pacf.

```
# On trace l'autocorrélogramme de la série différenciée (on regarde les
# autocorrélations jusqu'à deux ans de retard)
acf(diff_indice, main = "Autocorrélogramme")
```



L'ACF ne présente plus de pics significativement différents de zéro au delà de 2 retards. On fait le choix d'ignorer les éventuels pics significatifs pour des retards supérieurs à 10. On a donc qmax=2.

```
# On trace l'autocorrélogramme partiel de la série différenciée (on regarde les
# autocorrélations jusqu'à deux ans de retard)
pacf(diff_indice, main = "Autocorrélogramme partiel")
```



La PACF ne présente plus de pics significatifs au delà de 7 retards. On fait le choix d'ignorer les éventuels pics significatifs pour des retards supérieurs à 10. On a donc pmax=7.

```
# qmax and pmax
qmax <- 2
pmax <- 7
```

```
# On calcule le AIC/BIC pour chaque combinaison possible de p<=pmax et q<=qmax
pqs <- expand.grid(0:pmax, 0:qmax)
mat <- matrix(NA, nrow = pmax + 1, ncol = qmax + 1)
rownames(mat) <- paste0("p=", 0:pmax)
colnames(mat) <- paste0("q=", 0:qmax)
AICs <- mat
BICs <- mat
for (row in 1:dim(pqs)[1]) {
    p <- pqs[row, 1]
    q <- pqs[row, 2]
    estim <- try(arima(diff_indice, c(p, 0, q), include.mean = F))
    AICs[p + 1, q + 1] <- if (class(estim) == "try-error")
        NA else estim$aic
    BICs[p + 1, q + 1] <- if (class(estim) == "try-error")
        NA else BIC(estim)
}
```

```
AICs
```

```
##           q=0      q=1      q=2
## p=0 3030.769 2926.741 2915.086
```

```
## p=1 2981.711 2915.136 2916.833
## p=2 2958.443 2916.722 2917.021
## p=3 2944.682 2918.006 2919.015
## p=4 2929.584 2919.180 2921.082
## p=5 2924.740 2920.900 2922.691
## p=6 2926.582 2922.734 2923.957
## p=7 2923.735 2921.489 2923.392
```

```
AICs == min(AICs)
```

```
##       q=0   q=1   q=2
## p=0 FALSE FALSE  TRUE
## p=1 FALSE FALSE FALSE
## p=2 FALSE FALSE FALSE
## p=3 FALSE FALSE FALSE
## p=4 FALSE FALSE FALSE
## p=5 FALSE FALSE FALSE
## p=6 FALSE FALSE FALSE
## p=7 FALSE FALSE FALSE
```

Le modèle ARMA(0,2) = MA(2) minimise le critère AIC donc on garde ce modèle.

```
arima012 <- arima(indice, c(0, 1, 2), include.mean = F)
```

```
BICs
```

```
##          q=0      q=1      q=2
## p=0 3034.778 2934.759 2927.112
## p=1 2989.729 2927.162 2932.869
## p=2 2970.469 2932.757 2937.065
## p=3 2960.717 2938.050 2943.068
## p=4 2949.628 2943.233 2949.143
## p=5 2948.792 2948.962 2954.761
## p=6 2954.643 2954.804 2960.037
## p=7 2955.806 2957.568 2963.481
```

```
BICs == min(BICs)
```

```
##       q=0   q=1   q=2
## p=0 FALSE FALSE  TRUE
## p=1 FALSE FALSE FALSE
## p=2 FALSE FALSE FALSE
## p=3 FALSE FALSE FALSE
## p=4 FALSE FALSE FALSE
## p=5 FALSE FALSE FALSE
## p=6 FALSE FALSE FALSE
## p=7 FALSE FALSE FALSE
```

Le modèle MA(2) minimise également le critère BIC.

A présent, on va donc étudier l'ajustement et la validité du modèle MA(2).

```
# Fonction de test des significations individuelles des coefficients (cf. TD4)
signif <- function(estim) {
    coef <- estim$coef
    se <- sqrt(diag(estim$var.coef))
    t <- coef/se
    pval <- (1 - pnorm(abs(t))) * 2
    return(rbind(coef, se, pval))
```

```
}
signif(arima012)
```

```
##              ma1           ma2
## coef -0.57012120 -0.1758634512
## se    0.04817568  0.0466969694
## pval  0.00000000  0.0001658459
```

Les deux coefficients du modèle sont significatifs (pval<0.05) donc le modèle ARMA(0,2) est bien ajusté.

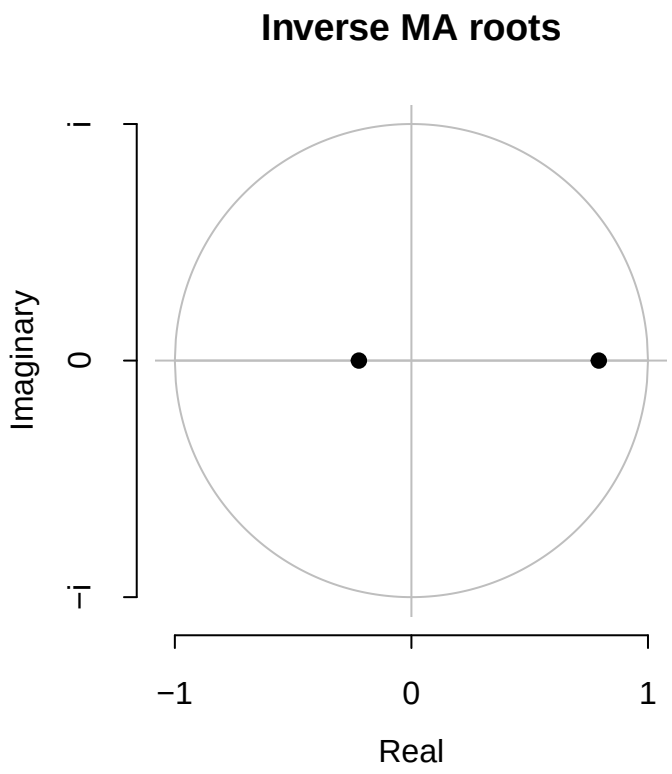
```
# Tests d'absence d'autocorrélation des résidus
qtest_arima012 <- Qtests(arima012$residuals, 24, length(arima012$coef) - 1)
qtest_arima012
```

```
##       lag      pval
##  [1,]   1        NA
##  [2,]   2 0.7706713
##  [3,]   3 0.6962715
##  [4,]   4 0.7018840
##  [5,]   5 0.7838566
##  [6,]   6 0.7270789
##  [7,]   7 0.6349692
##  [8,]   8 0.7435963
##  [9,]   9 0.7115108
## [10,]  10 0.7810497
## [11,]  11 0.5122521
## [12,]  12 0.5637225
## [13,]  13 0.5985435
## [14,]  14 0.6089081
## [15,]  15 0.6269542
## [16,]  16 0.2325623
## [17,]  17 0.2700481
## [18,]  18 0.2914494
## [19,]  19 0.2853131
## [20,]  20 0.3383075
## [21,]  21 0.3770663
## [22,]  22 0.4168084
## [23,]  23 0.4777087
## [24,]  24 0.2346000
```

Toutes les p-valeurs sont au dessus du seuil de 5% donc le modèle MA(2) est valide (résidus non autocorrélés).

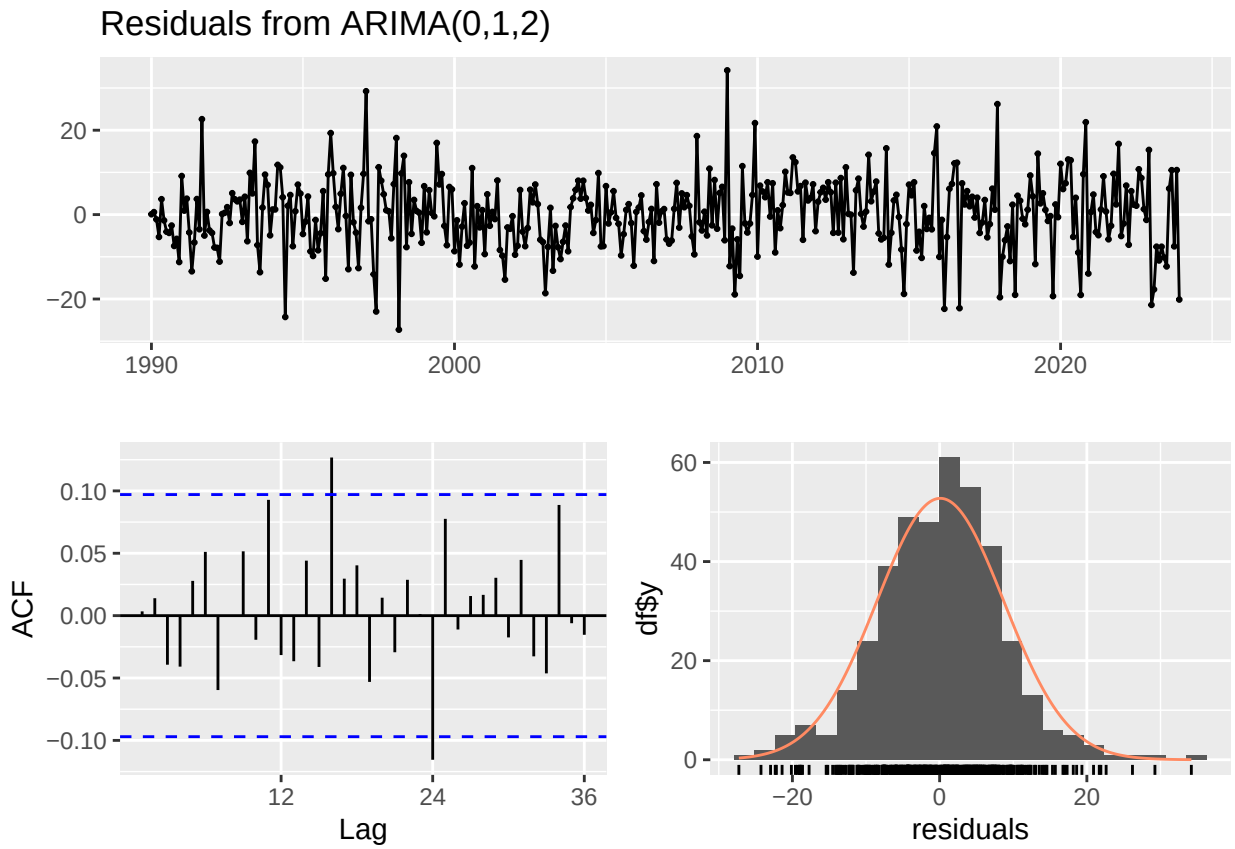
Conclusion : le modèle MA(2) est bien ajusté et valide.

```
# On plot l'inverse des racines
plot(arima012)
```



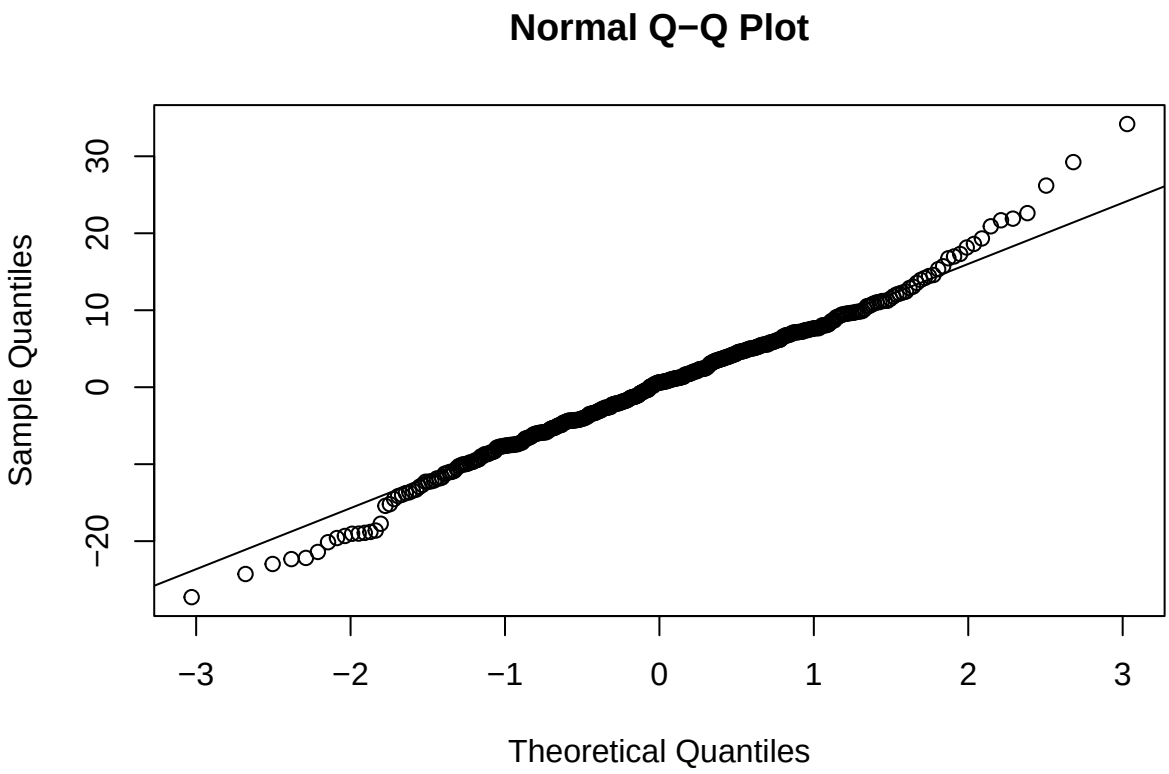
Toutes les racines du polynôme MA ont un module > 1. Le modèle MA(2) est bien inversible donc canonique.

```
# On étudie également la normalité des résidus de notre modèle
checkresiduals(arima012)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,2)
## Q* = 27.517, df = 22, p-value = 0.1922
##
## Model df: 2.   Total lags used: 24
```

```
# Diagramme Quantile-Quantile
qqnorm(arima012$residuals)
qqline(arima012$residuals)
```




```
# Test de normalité de Shapiro Wilk (H0 : les résidus sont gaussiens)
shapiro.test(arima012$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  arima012$residuals
## W = 0.98903, p-value = 0.003743
```

pval=0.0037 < 0.01 donc on rejette H_0.

D'apprès le diagramme Q-Q et le test de Shapiro Wilk, l'hypothèse de normalité des résidus n'est pas vérifiée.

Partie 3 : Prévision

Dans cette partie, on suppose que les résidus sont gaussiens.

```
# On prédit les valeurs de la série aux horizons T+1 et T+2 (janvier et février
# 2024) et les régions de confiance à 95% associées
predictions <- predict(arima012, n.ahead = 2)
```

```
# On récupère l'écart-type de nos résidus et le coefficient theta1
sigma <- sd(arima012$residuals)  #sigma=8.6
theta1 <- arima012$coef[1]  #theta1=-0.57
```

```
# Calcul des bornes de l'intervalle de confiance à 95% des prédictions
borne_inf_1 <- predictions$pred[1] - 1.96 * sigma
borne_inf_1
```

```
## [1] 70.70049
```

```
borne_sup_1 <- predictions$pred[1] + 1.96 * sigma
borne_sup_1
```

```
## [1] 104.486
```

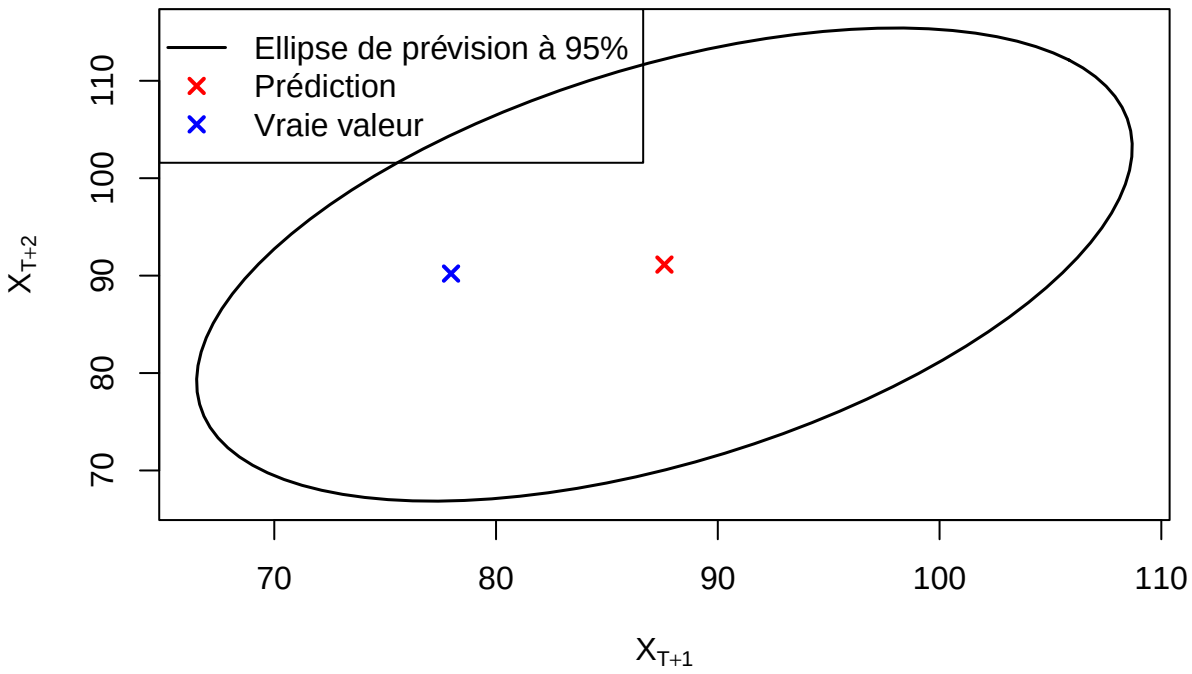
```
borne_inf_2 <- predictions$pred[2] - 1.96 * sigma * sqrt(1 + theta1^2)
borne_inf_2
```

```
##     ma1
## 71.6891
```

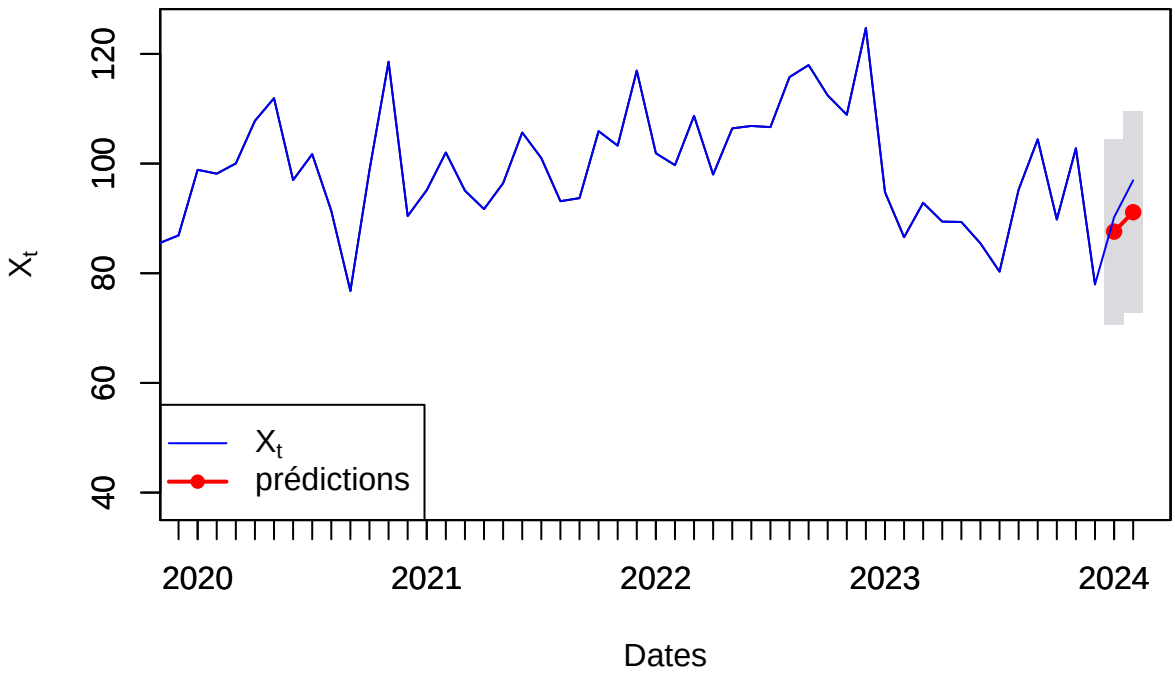
```
borne_sup_2 <- predictions$pred[2] + 1.96 * sigma * sqrt(1 + theta1^2)
borne_sup_2
```

```
##      ma1
## 110.5797
```

```
# Tracé de la région de confiance
Sigma <- matrix(data = c(sigma^2, -theta1 * sigma^2, -theta1 * sigma^2, sigma^2 *
    (1 + theta1^2)), nrow = 2, byrow = TRUE)  #matrice de variance-covariance
ell <- ellipse(x = Sigma, centre = c(predictions$pred[1], predictions$pred[2]), t = sqrt(qchisq(0.95,
    2)))
plot(ell, type = "l", xlab = expression(X[T + 1]), ylab = expression(X[T + 2]), lwd = 1.5)
points(x = predictions$pred[1], y = predictions$pred[2], pch = 4, lwd = 2, col = "red")  #prédiction
points(x = as.vector(indice.source)[length(indice.source) - 2], y = as.vector(indice.source)[length(indi
    1], pch = 4, lwd = 2, col = "blue")  #vraie valeur
lgd_labels <- c("Ellipse de prévision à 95%", "Prédiction", "Vraie valeur")
legend("topleft", legend = lgd_labels, col = c("black", "red", "blue"), lty = c(1,
    0, 0), pch = c(NA, 4, 4), lwd = c(1.5, 2, 2))
```




```r
# Représentation des prédictions et de leur intervalle de confiance à 95% ainsi
# que de la série initiale
forecast <- forecast(arima012, h = 2, level = 0.95)
plot(forecast, fcol = "red", xlim = c(2020, 2024 + 1/12), main = "")
lines(x = c(2024, 2024 + 1/12), y = forecast$mean, col = "red", lwd = 2)
par(new = T)
plot(indice.source, col = "blue", xlim = c(2020, 2024 + 1/12), xlab = "Dates", ylab = expression(X[t]))
legend("bottomleft", legend = c(expression(X[t]), "prédictions"), col = c("blue",
    "red"), lty = 1, pch = c(NA, 16), lwd = c(1, 2))
```



Les points rouges correspondent prédictions et les zones en gris clair représentent les intervalles de confiance à 95%.