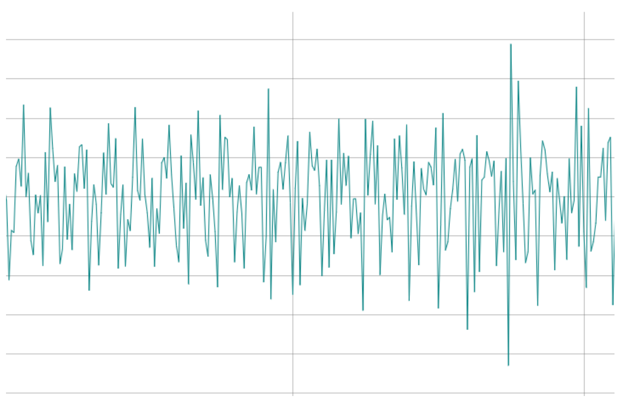
ECOLE NATIONALE DE LA STATISTIQUE
ET DE L'ANALYSE DE L'INFORMATION

Rendu de Série Temporelle

De Janvier 2023 à février 2023



Étudiant :
Patrick Gabin KAMGA

Professeur :
Vincent Lefieux

Table des matières

1 EXERCICE I

1.1 Objectif

Le but de l'exercice est de modéliser la série beer qui désigne la production mensuelle de bière en Australie entre janvier 1956 et février 1991. Cette modélisation sera faite par un modèle SARIMA. Pour se faire, nous allons premièrement visualiser la série, la tronquer si neccessaire. Etudier sa staionnarité et/ou sa saisonalité et retenir le modèle SARIMA le mieux adapté dans ce cas d'étude, enfin effectuer une analyse à posteriori sur les 12 prochains mois.

1.2 Analyse de la Série

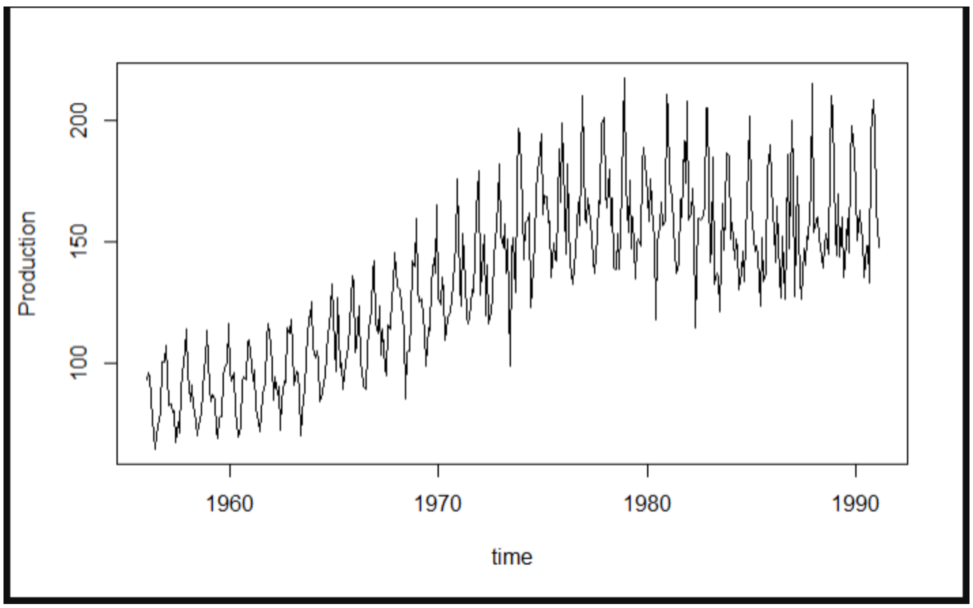


FIGURE 1 – Visualisation de la série

On observe que la serie présente une tendance croissante jusqu'en 1975 et après on a une production constante à partir de 1976.

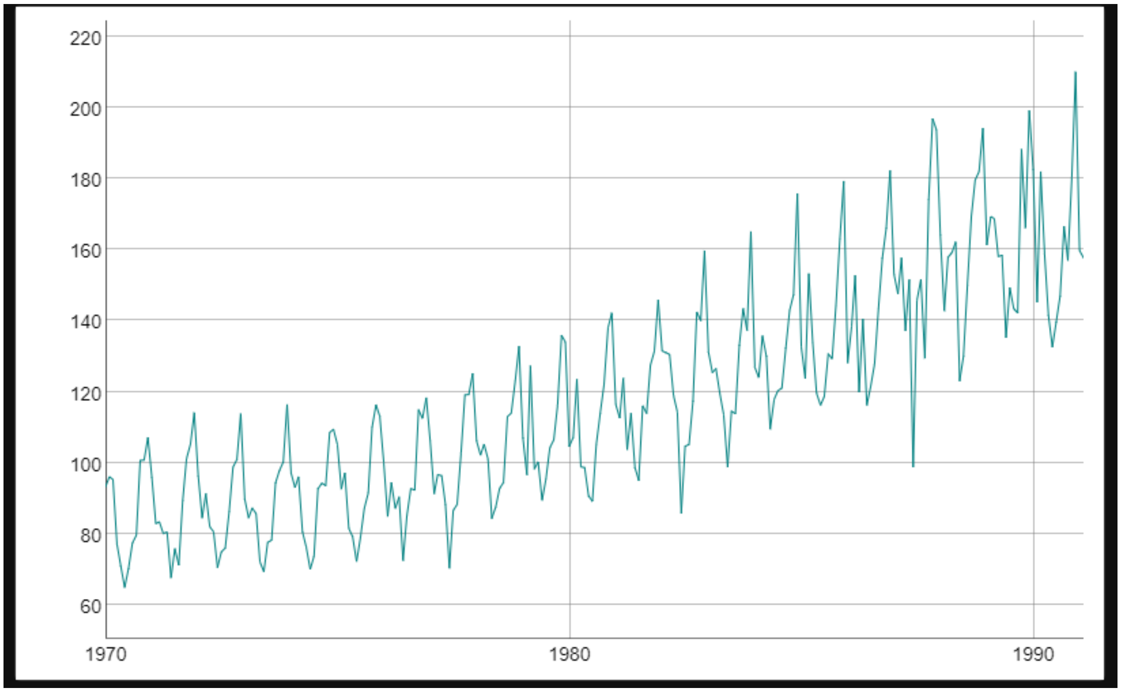


FIGURE 2 – Visualisation de la série tronquée à partir de 1970

En visualisant la serie tronquée à partir de 1970, on constate que la nouvelle série présente toujours cette tendance croissante.
Nous allons tronquer notre série à partir de 1970, puis prendre les 03 denières années comme base test pour l'analyse à postériori. Au vu de la crosssance de la série, bien qu'elle ne soit pas en saisonalité, nous travaillerons plutôt sur le log de la série.
La nouvelle serie n'est pas stationnaire, ainsi le calcul de l'ACF et PACF n'a pas vraiment d'interprétation. On peut tout de même observer l'ACF et se baser dessus pour stationariser la série.

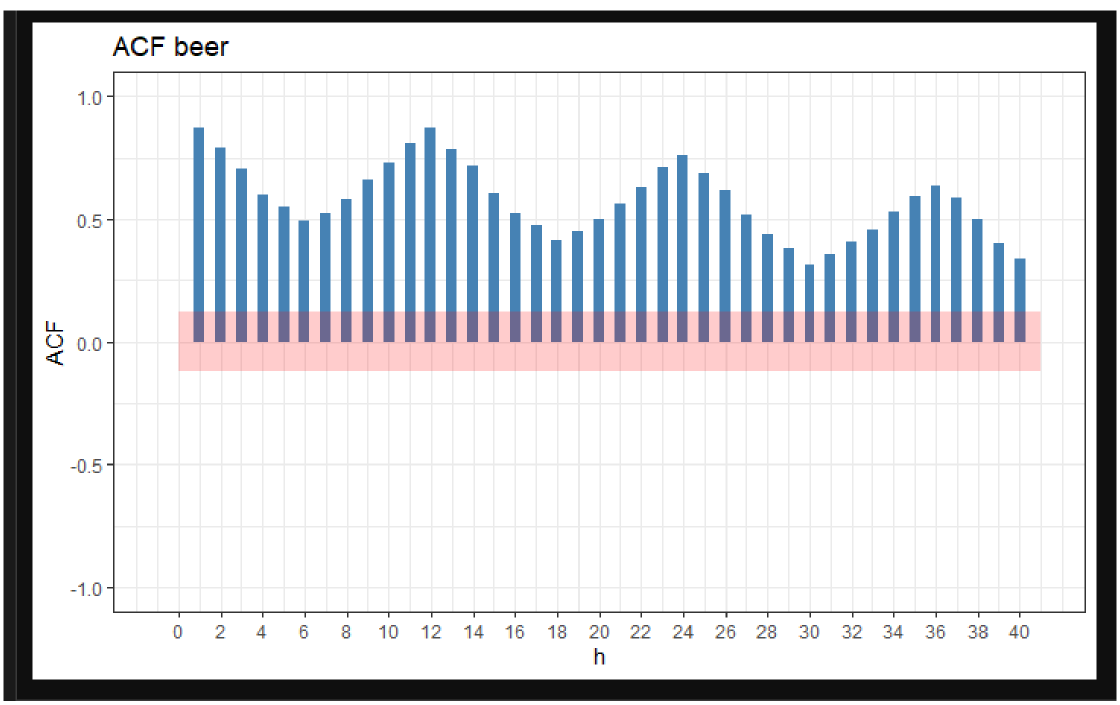


FIGURE 3 – Visualisation de l'ACF

La sortie ACF présente une décroissance lente vers 0, ce qui traduit un problème de non-stationnarité. On décide donc de la différentier de cette façon $(I - B)$.
La réalisation d'une telle différentiation permet d'obtenir une série encore non stationnaire. La sortie ACF presente toujours une décroissance lente vers 0, mais cette fois ci, pour tous les multiples de 12. On décide donc de la différentier de cette façon $(I - B^{12})$.

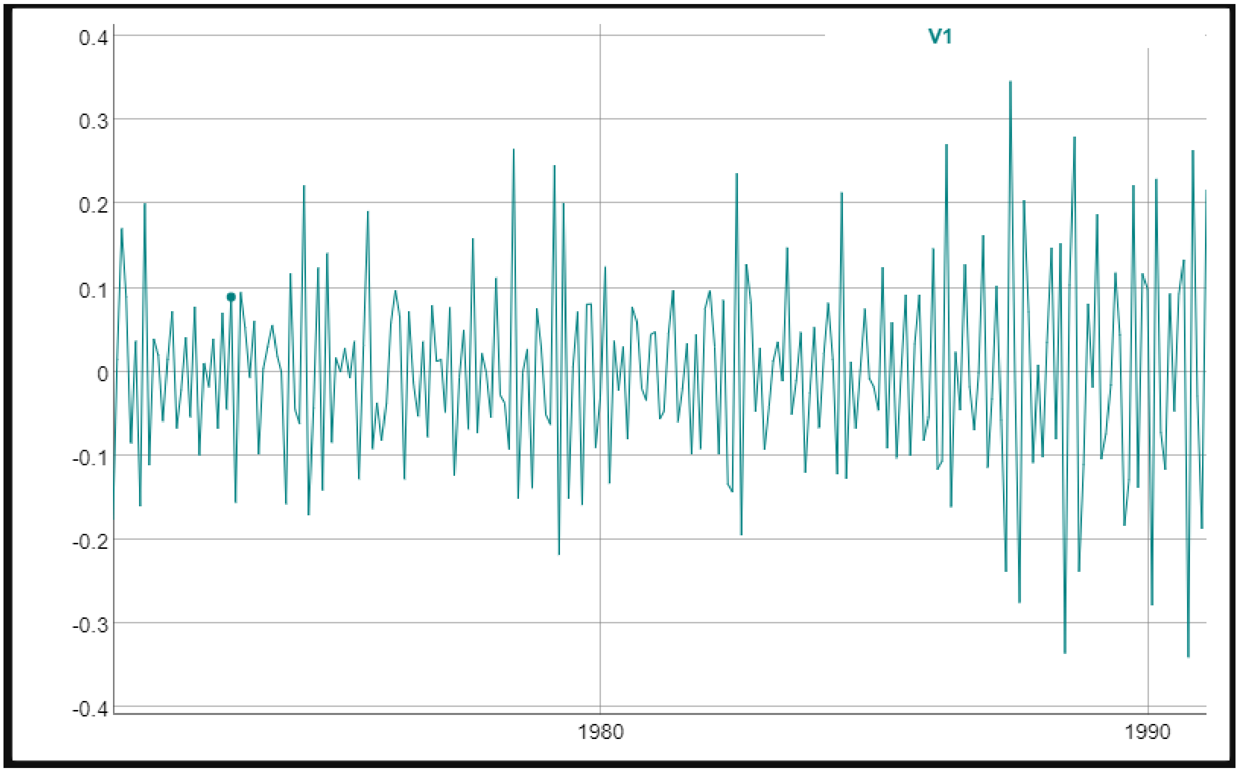


FIGURE 4 – Visualisation de la série différentiée

On observe que les autocorrélation simple décroissent vers 0, ainsi La sortie ACF de la série différenciée semble pouvoir être interprétée comme un autocorrélogramme simple empirique. On identifiera donc un modèle ARMA sur la série $(I - B^{12})(I - B)log(X_t)$, avec X_t la production de Beer.

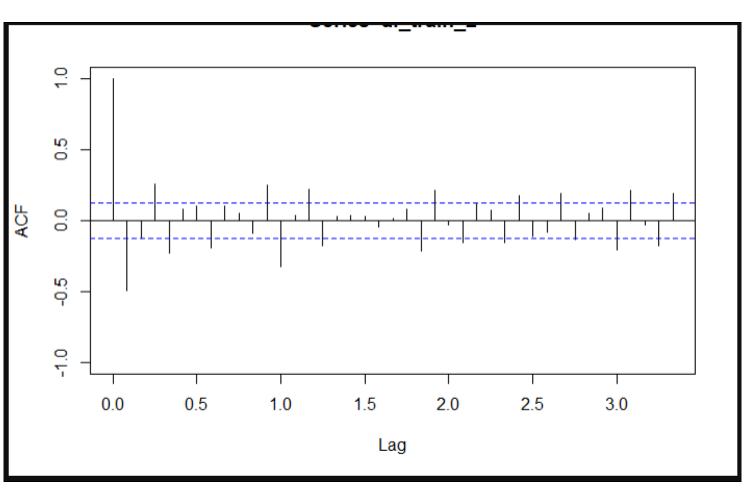


FIGURE 5 – Visualisation de l'ACF de la série différenciée

1.3 Identification des modèles

Nos modèles sont construit sur la série tronquée à partie de 1970.
On estime en premier lieu un modèle $SARIMA(1,0,2)(2,3,1)_{12}$ au vu des autocorrélogrammes empiriques simples et partiels. En réalisant un test de significatifité des différents paramètres on trouve les résultats suivants :

```
ARIMA(1,0,2)(2,3,1)[12]

Coefficients:
         ar1      ma1     ma2     sar1     sar2    sma1
      0.2126  -1.3937  0.3937  -1.1066  -0.5434  -1.000
s.e.  0.3267   0.3110  0.3078   0.0607   0.0643   0.057

sigma^2 = 0.01648:  log likelihood = 90.48
AIC=-166.96   AICc=-166.39   BIC=-143.7

Training set error measures:
                      ME      RMSE        MAE       MPE     MAPE      MASE         ACF1
Training set -0.00251681 0.1166646 0.08769855 -142.9821 503.7898 0.6260565 0.0002806451
[1] "Test statistic"
        ar1         ma1         ma2        sar1        sar2        sma1
  0.6508981  -4.4816813   1.2793348 -18.2262522  -8.4555460 -17.5395864
[1] "p-value"
         ar1          ma1          ma2         sar1         sar2         sma1
5.151123e-01 7.405727e-06 2.007792e-01 0.000000e+00 0.000000e+00 0.000000e+00
[1] "Ljung-Box test for residuals"

        Shapiro-Wilk normality test

data:  model0$residuals
W = 0.98655, p-value = 0.02299
```

FIGURE 6 – Visualisation des résultats du modèle 1

On constate que les p-value des coefficients du modèle sont tous significatifs sauf celui de **ar1**, de plus le test de shapiro testant la normalité des résidus a échoué. Les résidus du modèle 1 contiennent encore assez d'information. Ce qui est aussi visible sur le test de Ljung-Box, testant la blancheur des résidus en fonction des retards. Toutes les p-values sont nulles donc on rejette l'hypothèse de blancheur des résidus au seuil de 5%.

Description: df [8 × 2]

k <dbl>	p_valeur <dbl>
6	0.000
12	0.001
18	0.000
24	0.000
30	0.000
36	0.000
42	0.000
48	0.000

8 rows

FIGURE 7 – Visualisation des résultats du test de blancheur

On va tester un second modèle : $SARIMA(2,0,2)(2,4,2)_{12}$. Tous les paramètres du modèle sont significatifs. Bien que ce modèle échoue au test de shapiro sur la normalité des résidus, il est tout de même valide sur le test de blancheur au moins pour $k = 6, 12$ au seuil de 5%.

```
ARIMA(2,0,2)(2,4,2)[12]

Coefficients:
          ar1      ar2      ma1      ma2     sar1     sar2     sma1    sma2
      -0.7901  -0.2285  -0.3489  -0.5905  -1.0598  -0.5098  -1.8934  0.9510
s.e.   0.3141   0.0746   0.3212   0.3076   0.0663   0.0695   0.0875  0.0869

sigma^2 = 0.02016:  log likelihood = 2.83
AIC=12.35   AICc=13.33   BIC=41.71

Training set error measures:
                       ME      RMSE       MAE       MPE     MAPE      MASE       ACF1
Training set -0.001068646 0.1244076 0.0879666 -219.8526 551.5535 0.6279701 0.00243292
[1] "Test statistic"
       ar1        ar2        ma1        ma2       sar1       sar2       sma1       sma2
 -2.515445  -3.064434  -1.086438  -1.919688 -15.989433  -7.334778 -21.645021  10.944312
[1] "p-value"
         ar1          ar2          ma1          ma2         sar1         sar2         sma1
1.188824e-02 2.180823e-03 2.772854e-01 5.489737e-02 0.000000e+00 2.220446e-13 0.000000e+00
        sma2
0.000000e+00
     Retard p-value
[1,]      6 0.88906
[2,]     12 0.31825
[3,]     18 0.02369
[4,]     24 0.00035
[5,]     30 0.00049
[6,]     36 0.00085

        Shapiro-Wilk normality test

data:  model0$residuals
W = 0.97128, p-value = 8.554e-05
```

FIGURE 8 – Visualisation des résultats du modèle 2

On estime un autre modèle : $SARIMA(2,0,2)(2,3,2)_{12}$. Ce modèle nous donne des coefficients très significatifs. De plus, la p-value au test de shapiro est de 0.3262 qui est supérieur au seuil de 5%, ce qui signifie que le test de normalité des résidus est validé, donc les résidus de ce modèle suivent bien une distribution normale. Le test de blancheur est tout de même validé pour $k = 6, 12$. Nous utiliserons ce modèle pour faire des prédictions.

```
ARIMA(2,0,2)(2,3,2)[12]

Coefficients:
          ar1      ar2      ma1      ma2     sar1     sar2     sma1    sma2
      -0.7989  -0.2833  -0.3329  -0.5728  -0.7473  -0.3167  -1.9393  1.0000
s.e.   0.1918   0.0714   0.1931   0.1890   0.0741   0.0763   0.0824  0.0824

sigma^2 = 0.009998:  log likelihood = 118.93
AIC=-219.86   AICc=-218.93   BIC=-189.95

Training set error measures:
                       ME       RMSE        MAE       MPE     MAPE      MASE        ACF1
Training set -0.00390307 0.09040334 0.06902551 -91.37459 432.8737 0.4927547 0.003684372
[1] "Test statistic"
       ar1        ar2        ma1        ma2       sar1       sar2       sma1       sma2
 -4.164337  -3.967810  -1.723338  -3.030203 -10.086870  -4.150650 -23.522842  12.135456
[1] "p-value"
         ar1          ar2          ma1          ma2         sar1         sar2         sma1
3.122592e-05 7.253611e-05 8.482738e-02 2.443896e-03 0.000000e+00 3.315331e-05 0.000000e+00
        sma2
0.000000e+00
     Retard p-value
[1,]      6 0.56243
[2,]     12 0.24246
[3,]     18 0.01088
[4,]     24 0.00027
[5,]     30 0.00040
[6,]     36 0.00042

        Shapiro-Wilk normality test

data:  model0$residuals
W = 0.99309, p-value = 0.3262
```

FIGURE 9 – Visualisation des résultats du modèle 3

1.4 Analyse à postériori

Nous choisirons le modèle 3 pour faire notre analyse à postériori. On tronque la série avant 1988 pour entrainer notre modèle et comme test on prendra les 03 dernières années.
On vérifie bien que le modèle 3 est toujours acceptable sur la série tronquée. Ce qui est bien le cas, avec le test de blancheur cette fois réussit pour tous les retards.

```
ARIMA(2,0,2)(2,3,2)[12]

Coefficients:
          ar1      ar2     ma1     ma2     sar1     sar2     sma1    sma2
      -1.1550  -0.9990  1.1485  0.9793  -0.4145  -0.1264  -1.9185  0.9943
s.e.   0.0066   0.0023  0.0486  0.0221   0.0863   0.0869   0.1632  0.1679

sigma^2 = 0.004468:  log likelihood = 197.78
AIC=-377.56   AICc=-376.58   BIC=-348.25

Training set error measures:
                       ME       RMSE        MAE         MPE      MAPE      MASE        ACF1
Training set -0.004383827 0.06004492 0.04411061 -0.09561943 0.9348516 0.7135931 -0.04135107
             ar1       ar2     ma1      ma2      sar1      sar2      sma1     sma2
t.stat -175.4452 -434.8562 23.6295 44.28083 -4.801915 -1.453592 -11.75587 5.920972
p.val     0.0000    0.0000  0.0000  0.00000  0.000002  0.146059   0.00000 0.000000
     Retard p-value
[1,]      6 0.89048
[2,]     12 0.66451
[3,]     18 0.10331
[4,]     24 0.09369
[5,]     30 0.17024
[6,]     36 0.32022
```

Figure 10 – Visualisation des résultats du modèle 3

Tout celà étant fait, On obtient ci-dessous la réprésentation de la production réelle de Beer en Australie de 1989 à 1991, ainsi que la représentation de la production prédite, en plus d'un intervalle de confiance à 95%. On note aussi un *mae* de 18.03 et un *mape* de 8.69. Ce qui n'est pas très mauvais au vu de la moyenne et de l'écart-type de la production de beer.

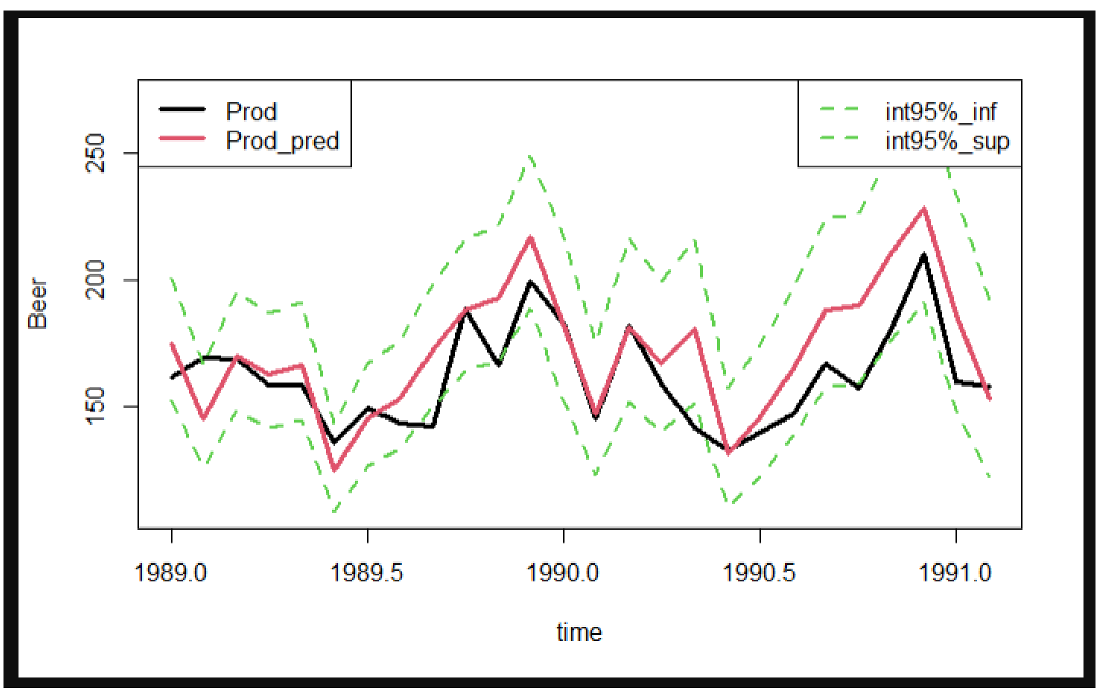


Figure 11 – Visualisation des prédictions

2 EXERCICE 2

2.1 Objectif

L'objectif de cet exercice est de prédire la teneur maximale en ozone *maxO3* à l'orizon d'un jour. On étudiera dans un premier temps les limites de cette prédition avec les modèles de regression, et par la suite on justifiera l'intérêt de plutôt utiliser les modèles modèles de séries temporelles.

2.2 Prédiction avec les modèle de regression

Les données contiennes 175 lignes avec des valeurs manquantes. Pour cette étude, nous supprimerons ces valeurs manquantes. On prendra comme base d'entrainement des algorithmes, les données allant du 01/04/1995 au 31/12/2001. La base de test quant à elle contiendra les données de l'année 2022.
La visualisation de la matrice de corrélation nous montre quatre groupes de variables relativement correlées entre elles. Le premier groupe est constitué des variables modélisant les températures observées à 6h, 9h, 12h, 15h et 18h. Le second groupe est constitué des varaibles modélisant la nébulosité observée à 12h, 15h et 18h, ainsi que celles modélisant la teneur maximale en ozone observée la veille...
Notons tout de même une forte correlation positive entre la teneur maximale en ozone observée sur la journée et la teneur maximale en ozone observée la veille.

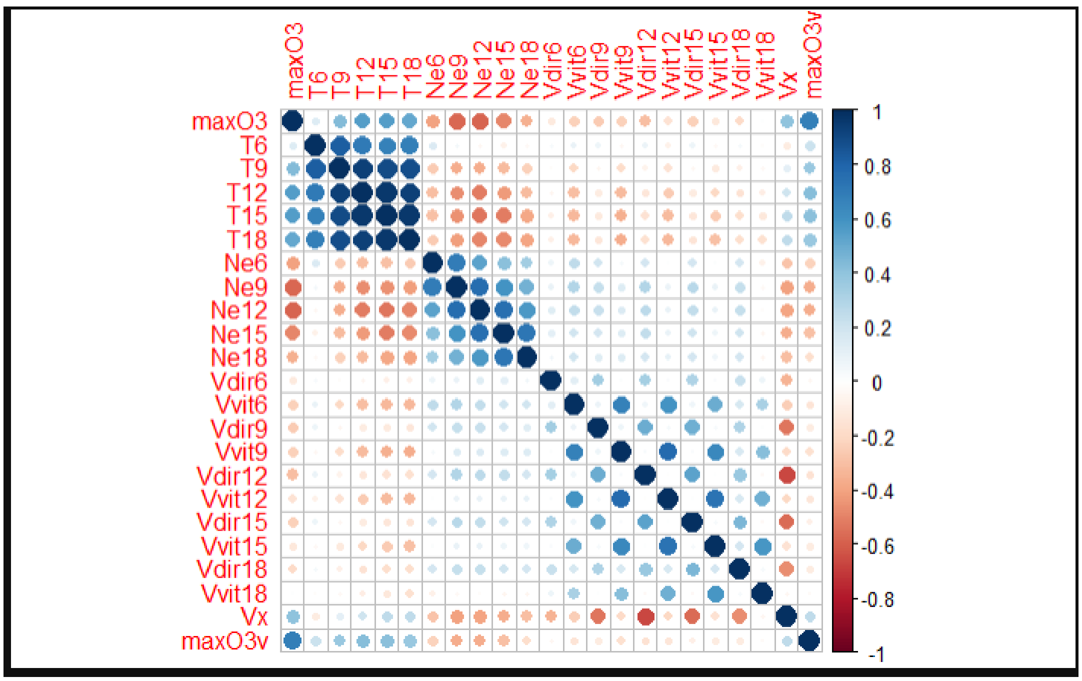


FIGURE 12 – Visualisation matrice de corrélation

Pour mieux visualiser ces groupe de variables correlées entre elles, l'idéal serait de faire un clustering de variable. La distnace entre les variables dépend entièrement de la corrélation entre ces variables. Je prendrais ici :

$$dist(var1, var2) = 1 - |cor(var1, var2)|$$

On peut visualiser ci-dessous, les clusters de variables correlées entre elles au seuil de 0.7.
On voit que les variables modélisant les température observées à 9h, 12h, 15h et 18h ; la nébulosité à 9hr et 12hr, 15hr et 18hr ; et enfin la vitesse du vent à 9hr et 12hr sont le groupe de variable qui sont correllée au seuil de 0.7, donc elles rapportent quasiment la même information. Nous allons donc récupérer juste

une variable représentative de ces variables là, en occurence $T12$ pour le premier groupe, $Ne12$ pour le second groupe, $Ne15$ pour le troisième groupe et $vvit9$ pour le dernier groupe.

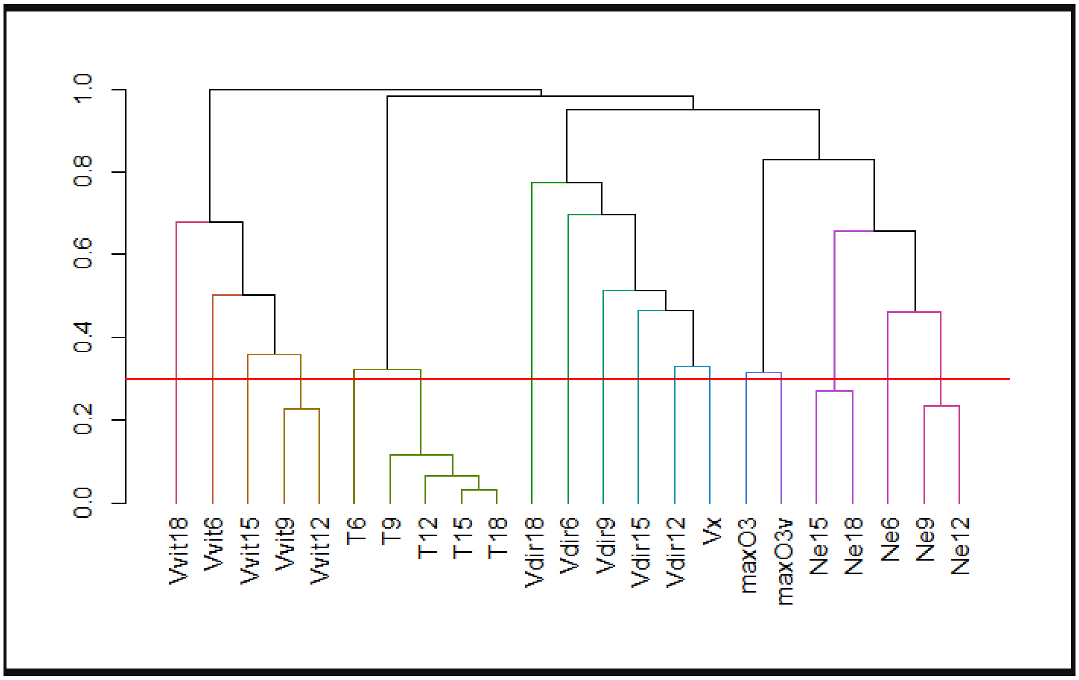


FIGURE 13 – Visualisation cluster de variable

2.2.1 Modèle avec uniquement les variables météorologiques

Nous avons modéliser la teneur maximale en ozone observée sur la journée via un modèle de regression lineaire multiple. En l'appliquant sur le jeux de test, on obtient un *mae* de 12.79 et un *mse* de 253.25.

```
model1 = lm(maxO3~T6+T12+Ne6+Ne12+Ne15+Vdir6+Vvit6+Vdir9+
         Vvit9+Vdir12+Vdir15+Vvit15+Vdir18+Vvit18+Vx,data = df_ozone_na_train[,-23])

summary(model1)

mae = mean(abs(predict(model1,df_ozone_na_test[,-23])-df_ozone_na_test[,1]))

mse = mean((predict(model1,df_ozone_na_test[,-23])-df_ozone_na_test[,1])^2)

mae

## [1] 12.79859

mse

## [1] 253.2521
```

FIGURE 14 – Visualisation Regression lineaire

Comme second modèle, on fait un random forest, avec 500 arbres. On obtient sur ce modèle un *mae* de 11.65 et un *mse* de 219.79. Cela est un peu plus meilleur que celui de la regression lineaire multiple faite précédemment.

```
model2 = randomForest(maxO3~T6+T12+Ne6+Ne12+Ne15+Vdir6+Vvit6+Vdir9+
         Vvit9+Vdir12+Vdir15+Vvit15+Vdir18+Vvit18+Vx,data = df_ozone_na_train[,-23],ntree = 500)

summary(model2)

mae = mean(abs(predict(model2,df_ozone_na_test[,-23])-df_ozone_na_test[,1]))

mse = mean((predict(model2,df_ozone_na_test[,-23])-df_ozone_na_test[,1])^2)

mae

## [1] 11.65642

mse

## [1] 219.7956
```

FIGURE 15 – Visualisation Random Forest

On note aussi que la variable $T12$ qui représente la température observée à 12h ressort comme la variable la plus importante du modèle, suivit par les variables modélisant la nébulosité observée entre 12h et 15h.

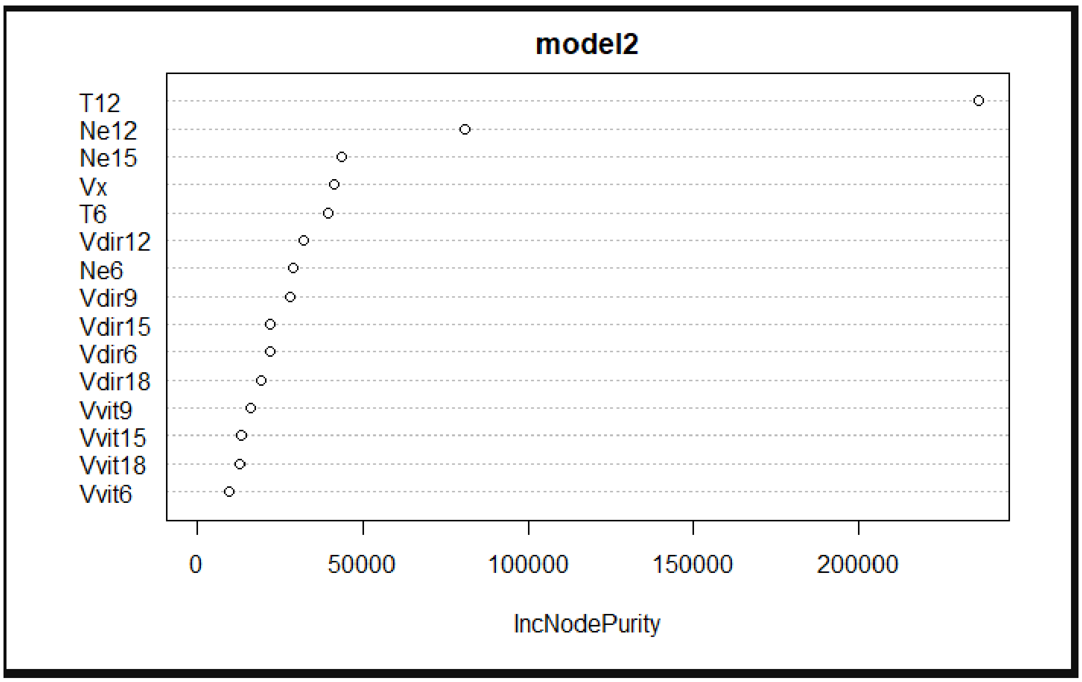


FIGURE 16 – Visualisation Variable Importante

Nous avons modélisé $maxO3$ par un modèle de boosting généralisé. On obtient sur ce modèle un mae de 11.89 et un mse de 222.81. Le modèle de boosting n'et pas meilleur que celui de Random Forest ou du bagging comme cela est présent dans le code r.

```r
set.seed(1)

modele5 = gbm(maxO3~T6+T12+Ne6+Ne12+Ne15+Vdir6+Vvit6+Vdir9+
          Vvit9+Vdir12+Vdir15+Vvit15+Vdir18+Vvit18+Vx, data = df_ozone_na_train[,-23],
              distribution = "gaussian",interaction.depth=6,
              shrinkage = 0.02,n.trees=600,cv.folds=5)

mae = mean(abs(predict(modele5,df_ozone_na_test[,-23])-df_ozone_na_test[,1]))

mse = mean((predict(modele5,df_ozone_na_test[,-23])-df_ozone_na_test[,1])^2)

mae

## [1] 11.89644

mse

## [1] 222.8151

gbm.perf(modele5,method = "cv")
```

FIGURE 17 – Visualisation Boosting

Le dernier modèle, sans doute aussi le meilleur de nos modèle est le xgboost. Il nous donne un *mae* de 11.52 et un *mse* de 219.63.

```r
set.seed(0)

data_train = df_ozone_na_train[,c("maxO3","T6","T12","Ne6","Ne12","Ne15","Vdir6","Vvit6","Vdir9",
          "Vvit9","Vdir12","Vdir15","Vvit15","Vdir18","Vvit18","Vx")]

data_test = df_ozone_na_test[,c("maxO3","T6","T12","Ne6","Ne12","Ne15","Vdir6","Vvit6","Vdir9",
          "Vvit9","Vdir12","Vdir15","Vvit15","Vdir18","Vvit18","Vx")]

train_x = data.matrix(data_train[,-1])

train_y = df_ozone_na_train$maxO3


test_x = data.matrix(data_test[,-1])

test_y = df_ozone_na_test$maxO3

xgb_train = xgb.DMatrix(data = train_x, label = train_y)

xgb_test = xgb.DMatrix(data = test_x, label = test_y)

#define watchlist and param

watchlist = list(train=xgb_train, test=xgb_test)

param <- list(max_depth = 2, eta = 0.2, verbose = 0, nthread = 2,
              objective = "reg:squarederror", eval_metric = "mae",lambda=0.1)

#fit XGBoost model and display training and testing data at each round

model = xgb.train(param,data = xgb_train, nrounds = 68, watchlist=watchlist)

## mae

# 11.526850

##mse

# 219.63
```

FIGURE 18 – Visualisation Boosting

2.2.2 Modèle avec les variables météorologiques et le pic d'ozone de la veille

Nous allons entrainer les modèles précédents en y ajoutant désormais, la teneur maximale en ozone observée la veille *maxO3v*. Dans ce cas de figure c'est le modèlme de boosting généralisé qui produit de meilleurs résultats. On obtient un *mae* de 9.23 et un *mse* de 141.56. Les résultats des autres modèles sont visibles dans le code r.

```
set.seed(1)

modele5 = gbm(maxO3~T6+T12+Ne6+Ne12+Ne15+Vdir6+Vvit6+Vdir9+
          Vvit9+Vdir12+Vdir15+Vvit15+Vdir18+Vvit18+Vx+maxO3v, data = df_ozone_na_train,
              distribution = "laplace",interaction.depth=4,
              shrinkage = 0.02,n.trees=600,cv.folds=5)

mae = mean(abs(predict(modele5,df_ozone_na_test)-df_ozone_na_test[,1]))

mse = mean((predict(modele5,df_ozone_na_test)-df_ozone_na_test[,1])^2)

mae

## [1] 9.232696

mse

## [1] 141.5671
```

FIGURE 19 – Visualisation Boosting

2.3 Intérêt de Série temporelle

Jusqu'ici, avec tous ces modèles, on n'obtient pas vraiment de meilleurs résultats pour la prédiction de teneur maximale en ozone observée sur la journée. En outre, il est clair que si l'on permute deux quelconques lignes de notre fichier, alors tout le fichier devient erroné. L'aspect temporelle n'est ce fait pas négligeable ici. De plus, puisque nous cherchons à modéliser la prédiction de la teneur maximale en ozone observée sur la journée connaissant la teneur maximale en ozone observée la veille, il nous suffira de modéliser uniquement cette dernière.
En effet, si l'on veut par exemple prédire la teneur maximale en ozone observée sur la journée du 15/04/1995, il nous suffira de prédire la teneur maximale en ozone observée la veille du 16/04/1995. Il y a donc un grand intérêt d'utiliser un modèle de série temporelle (utilisant le pic d'ozone de la veille) afin de prédire $maxO3$.
Nous visualisons la variable $maxO3v$, elle n'a pas de tendance mais pas n'est stationnaire. Son ACF décroit lentement vers 0. Après une différence, son ACF peut être interpreté comme une autocorrélation simple empirique.

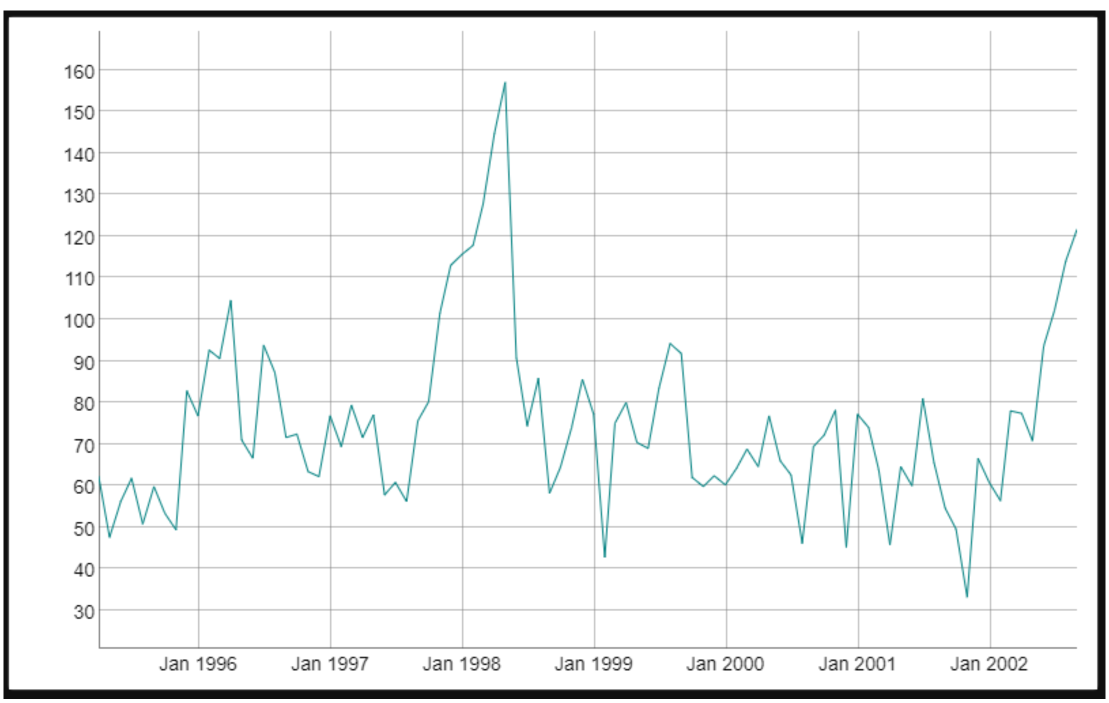


FIGURE 20 – Visualisation ACF ozone

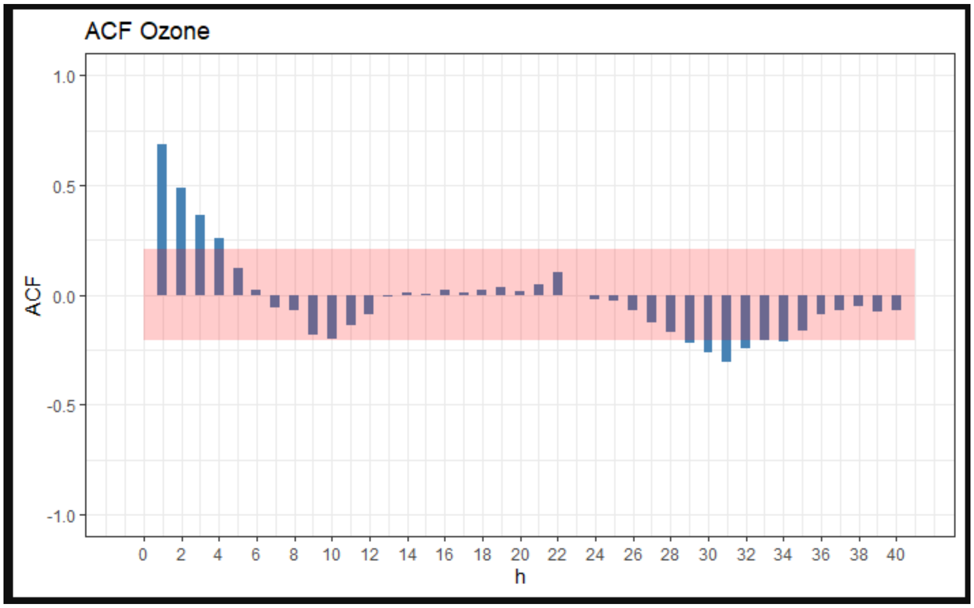


Figure 21 – Visualisation ACF ozone

Au vu des autocorrélation simple et partielle, on a modélisé la série avec un modèle $SARIMA(1,0,1)(0,1,1)_{12}$. Les p-values des paramètres du modèle sont tous très significatifs. Le test de blacheur des résidus pour les retards de 6 à 42 est aussi réussi. Le test de shapiro sur la normalité des résidus est validé. Ce modèle peut donc être utiliser pour faire notre analyse à postériori.

```
arima(x = df_ozone_diff, order = c(1, 0, 1), seasonal = list(order = c(0, 1,
    1), period = 12), include.mean = TRUE, method = "CSS-ML")

Coefficients:
         ar1      ma1     sma1
      0.6932  -0.8999  -1.0000
s.e.  0.2068   0.1788   0.2145

sigma^2 estimated as 256.5:  log likelihood = -335.28,  aic = 678.55

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE        ACF1
Training set -1.057894 14.89766 11.22135 81.87472 146.0394 0.5615142 -0.03629719
[1] "Test statistic"
      ar1       ma1      sma1
 3.352821 -5.032373 -4.661087
[1] "p-value"
         ar1          ma1         sma1
7.999240e-04 4.844456e-07 3.145442e-06
     Retard p-value
[1,]      6 0.98318
[2,]     12 0.81904
[3,]     18 0.97928
[4,]     24 0.86127
[5,]     30 0.96002
[6,]     36 0.97874

        Shapiro-Wilk normality test

data:  modele_ozone$residuals
W = 0.98417, p-value = 0.3534
```

Figure 22 – Visualisation $SARIMA(1,0,1)(0,1,1)_{12}$

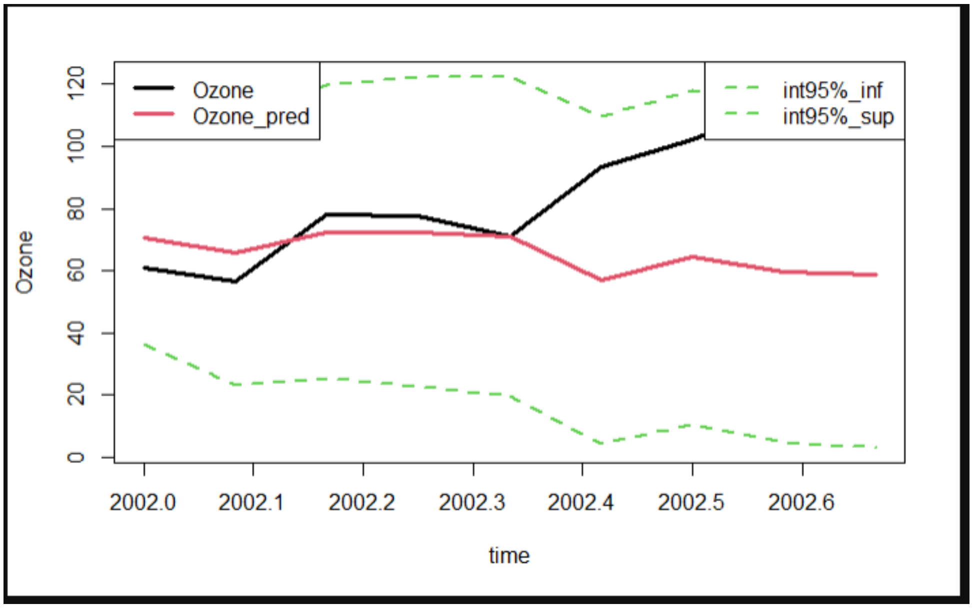


FIGURE 23 – Visualisation Prédition