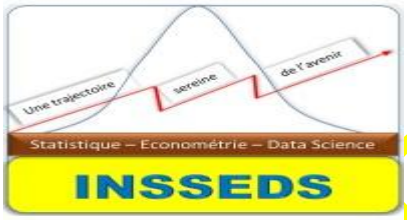
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

REPUBLIQUE DE CÔTE D'IVOIRE

Union – Discipline – Travail



INSTITUT SUPERIEUR DE STATISTIQUE, D'ECONOMETRIE ET DE DATASCIENCE

MASTER 2

STATISTIQUE ECONOMETRIE DATA SCIENCE

MINI PROJET ECONOMETRIE DES SERIES TEMPORELLES

PREVISION DES INDICES DE POLLUTION DE L'AIR POUR LES 30 PREMIERS JOURS DE L'ANNEE 2015

MODELE ECONOMETRIQUE SERIE TEMPORELLE

ANNEE UNIVERSITAIRE : 2024-2025

ETUDIANT : **N'DRI ONESIME**

ENSEIGNANT : **AKPOSSO DIDIER MARTIAL**

AVANT-PROPOS

L'association du théorique au pratique, des connaissances aux compétences et des savoir-faire aux savoirs est la principale tendance récente dans le secteur technique. Dans ce contexte, l'INSSEDS (Institut Supérieur de la Statistique, d'Économétrie et de la Data Science), dans sa formation en master professionnel en statistique, économie et science des données, impose que les divers crédits soient validés en effectuant un mini-projet à la fin de chaque module. Le projet est donc structuré et supervisé de cette manière, visant principalement à faire de chaque élève un participant dynamique, engagé et libre dans la vie active.

Ce document est un rapport de projet Muni axé sur l'économétrie des séries temporelles. Il se divise principalement en trois parties : Prétraitements des données, HOT-WINTER et Méthode BOX-JENKIS.

En règle générale, toutes les analyses et conclusions présentées dans ce rapport relèvent de la responsabilité de l'auteur, qui ne sollicite ni autrui ni l'INSSEDS (Institut Supérieur de Statistique d'Econométrie et de Data Science).

Table des matières

INTRODUCTION

Contexte et justification de l'étude

La qualité de l'air est un sujet qui nous touche tous, car elle a un impact direct sur notre santé et notre bien-être. Dans un monde de plus en plus urbanisé, les niveaux de pollution peuvent atteindre des seuils alarmants, notamment dans les grandes villes. Prévoir ces niveaux de pollution n'est pas seulement une question scientifique, mais aussi une nécessité pour protéger les populations et aider les décideurs à agir en amont. Cette étude vise à apporter une réponse concrète à ce défi en prédisant les niveaux de pollution pour les 30 prochains jours, en s'appuyant sur des données historiques et des méthodes éprouvées.

Problématique

Prévoir la pollution de l'air est une tâche complexe, car elle dépend de nombreux facteurs, notamment les conditions météorologiques comme la température, la pluie, ou la vitesse du vent. Comment pouvons-nous, à partir des données passées, anticiper les niveaux de pollution à venir ? Cette question est au cœur de notre étude. En comprenant les liens entre la météo et la pollution, nous espérons fournir des prévisions fiables pour les semaines à venir.

Principaux résultats attendus

À l'issue de cette étude, nous souhaitons obtenir une prévision précise des niveaux de pollution pour les 30 prochains jours. Ces résultats pourront servir de base pour informer les citoyens et guider les politiques publiques, par exemple en déclenchant des

alertes pollution ou en adaptant les mesures de réduction des émissions. De plus, nous espérons identifier les facteurs météorologiques qui influencent le plus la pollution, ce qui pourrait aider à mieux comprendre ses variations.

Méthodologie

Pour réaliser ces prévisions, nous utiliserons la méthodologie de Box & Jenkins, une approche classique et robuste pour l'analyse des séries temporelles. Cette méthode repose sur des modèles ARIMA (AutoRegressive Integrated Moving Average), qui permettent de capturer les tendances, les saisons et les variations aléatoires dans les données. L'avantage de cette approche est qu'elle est particulièrement adaptée aux données temporelles, comme celles dont nous disposons ici.

Description du jeu de données : dictionnaire des données

Notre jeu de données couvre une période de 5 ans, soit 1825 jours d'observations quotidiennes. Il combine des informations sur la qualité de l'air et les conditions météorologiques. Voici les variables clés que nous analyserons:

- date : La date de l'observation, avec une entrée unique pour chaque jour.
- pollution_today : Le niveau de pollution mesuré le jour même. Il pourrait s'agir d'un indice de qualité de l'air (AQI) ou de la concentration d'un polluant spécifique.

- dew : La température du point de rosée en degrés Celsius, qui indique à quel moment l'air devient saturé en humidité.

- temp : La température de l'air en degrés Celsius, mesurée à un instant donné.

- press : La pression atmosphérique en hPa, un indicateur des conditions météorologiques (pression élevée = temps stable, pression basse = perturbations).

- wnd_spd : La vitesse du vent, qui peut influencer la dispersion des polluants.

- snow : La quantité de neige tombée en millimètres (0 mm signifie aucune neige).

- rain : La quantité de pluie en millimètres (0 mm signifie aucune précipitation).

- pollution_yesterday : Le niveau de pollution mesuré la veille, utile pour analyser les variations jour après jour.

Ces données nous offrent une vue détaillée des interactions entre la météo et la pollution, ce qui constitue une base solide pour nos prévisions. En les exploitant avec soin, nous espérons apporter des réponses utiles et actionnables pour améliorer la qualité de l'air et, par extension, la qualité de vie.

I- Prétraitement des données

Avant d'aborder la deuxième étape consacrée au prétraitement des données, il est important de préciser que quelques ajustements ont déjà été apportés au jeu de données. Ces modifications incluent, entre autres, la conversion des types de données et le renommage des modalités de certaines variables pour une meilleure clarté et cohérence. Tout cela s'est fait dans le logiciel ExcelPower-Query.

Visualisation des données

```
tibble [1,825 × 9] (S3: tbl_df/tbl/data.frame)
   date               : chr [1:1825] "2010-01-02" "2010-01-03"
"2010-01-04" "2010-01-05" ...
   pollution_today    : num [1:1825] 146 79 31 42 56 69 176 88
57 20 ...
   pollution_yesterday: num [1:1825] 10 146 79 31 42 56 69 176
88 57 ...
   dew                : num [1:1825] -8 -10 -21 -25 -24 -21 -17
-16 -16 -21 ...
   temp               : num [1:1825] -5 -9 -12 -14 -13 -12 -12 -
9 -9 -9 ...
   press              : num [1:1825] 1025 1023 1029 1034 1034
...
   wnd_spd            : num [1:1825] 25 71 111 57 19 10 2 13 17
42 ...
   snow               : num [1:1825] 1 14 0 0 0 0 0 0 0 0 ...
   rain               : num [1:1825] 0 0 0 0 0 0 0 0 0 0 ...
```

Ce jeu de données comprend 1825 observations et 9 variables dont 1 sous forme de date et 8 autres sous forme numérique.

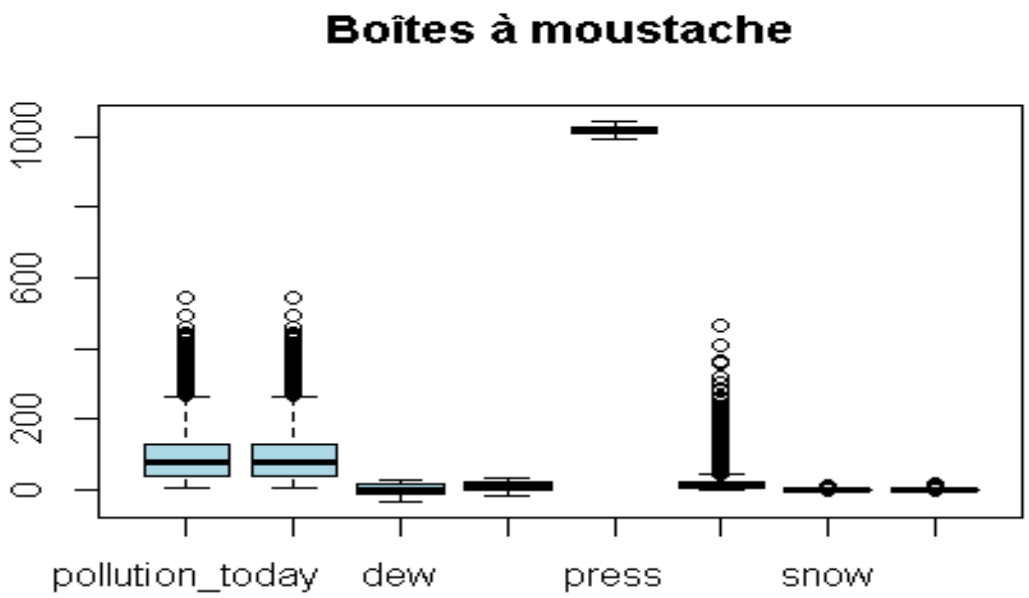
Valeurs manquantes

```
                date     pollution_today pollution_yesterday
dew
                   0                   0                   0
0
                temp               press             wnd_spd
snow
                   0                   0                   0
```

```
0
                    rain
                       0
```

Valeurs abberantes

- **Visualisation**



Ces données représentent des observations réelles, et chaque valeur, même celles qui semblent inhabituelles ou aberrantes, a son importance. Plutôt que de les supprimer, il est préférable de les conserver, car elles peuvent refléter des événements particuliers ou des situations exceptionnelles qui ont un impact sur la qualité de l'air. Les écarter risquerait de fausser notre analyse et de rendre nos prévisions moins fiables. En gardant l'intégralité des données, nous nous assurons que notre étude reste fidèle à la réalité et que nos conclusions soient aussi robustes que possible.

II- ANALYSE DESCRIPTIVE ET PREVISIONS HOT-WINTER

a) Construction de la serie temporelle

Dans cette étape, nous allons nous concentrer sur les dates et la variable pollution_today pour construire notre série temporelle. Ces deux éléments sont essentiels, car ils nous permettront d'analyser l'évolution de la pollution de l'air au fil du temps. En utilisant la date comme axe temporel et pollution_today comme variable d'intérêt, nous pourrons identifier des tendances, des saisonnalités ou des patterns récurrents, ce qui constitue la base de toute modélisation en séries temporelles. Cette approche nous aidera à mieux comprendre comment la pollution évolue jour après jour et à préparer le terrain pour des prévisions futures.

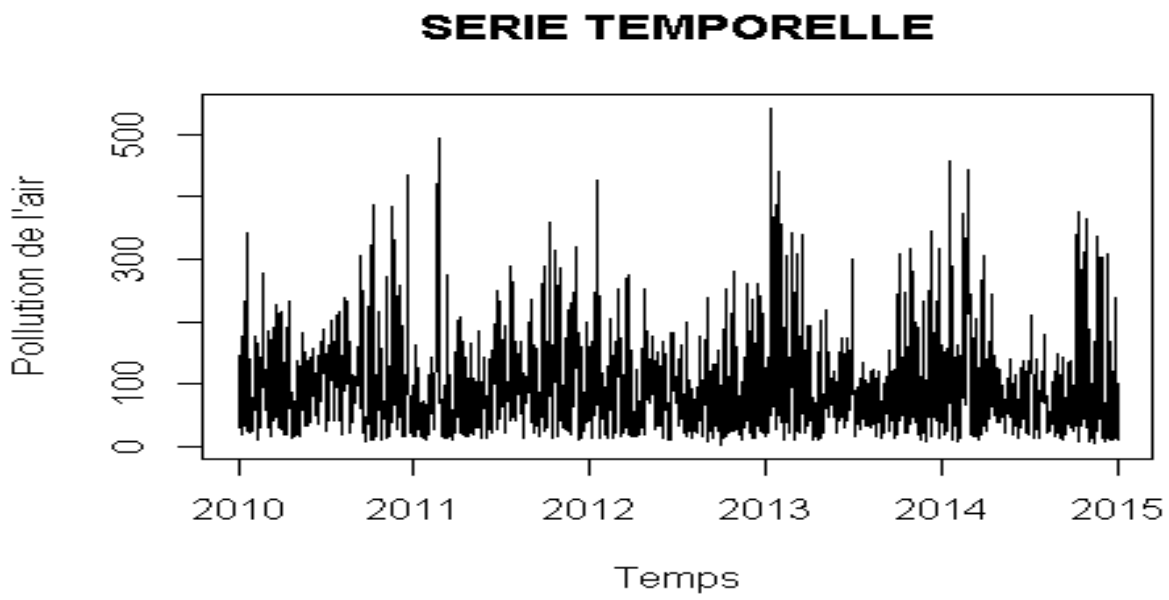
```
146 79 31 42 56 69
```

Les niveaux de pollution de l'air ont commencé à 146 le 2 janvier 2010, ce qui indique que l'air était particulièrement pollué ce jour-là. Heureusement, le lendemain, le 3 janvier, la situation s'est améliorée de manière significative, avec une chute des valeurs à 79. Cette tendance positive s'est poursuivie le 4 janvier, où la pollution a encore diminué pour atteindre 31, offrant une qualité de l'air bien plus respirable. Cependant, à partir du 5 janvier, les choses ont commencé à se dégrader légèrement. La pollution est remontée à 42, puis a continué d'augmenter les jours suivants, passant à 56 le 6 janvier et à 69 le 7 janvier. En résumé, on observe des variations importantes d'un jour à l'autre. Les premiers jours montrent une nette amélioration, mais cette tendance est suivie d'une hausse progressive de la pollution. Ces fluctuations pourraient s'expliquer par plusieurs facteurs, comme les changements

météorologiques, les activités industrielles ou encore l'intensité du trafic. Une analyse plus poussée permettrait de mieux comprendre ces variations et d'identifier les causes précises de ces changements.

b)Graphiques

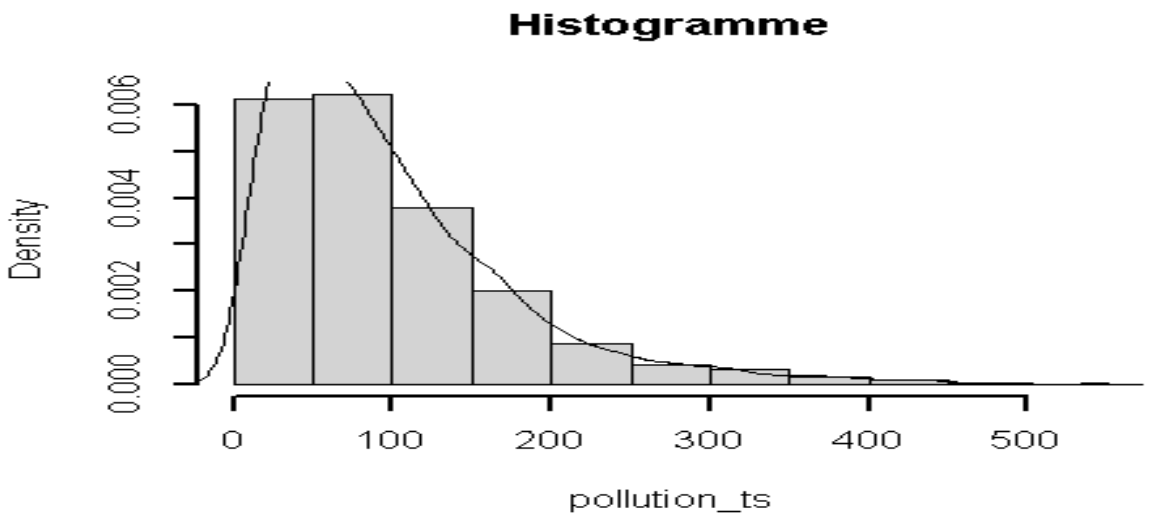
- **Serie temporelle**



En comparant les niveaux de pollution jour après jour, on peut facilement repérer les moments où la qualité de l'air s'améliore ou, au contraire, se dégrade. Par exemple, la forte baisse de pollution observée le 4 janvier 2010, par rapport au 2 janvier, montre une nette amélioration. Cela pourrait s'expliquer par des conditions météo plus favorables ou des efforts ponctuels pour réduire les émissions. Si on regarde les extrêmes – le pic de pollution à 146 le 2 janvier et le niveau le plus bas à 31 le 4 janvier, on voit que la tendance générale est d'abord à la baisse, suivie d'une légère remontée les jours suivants. Cette fluctuation

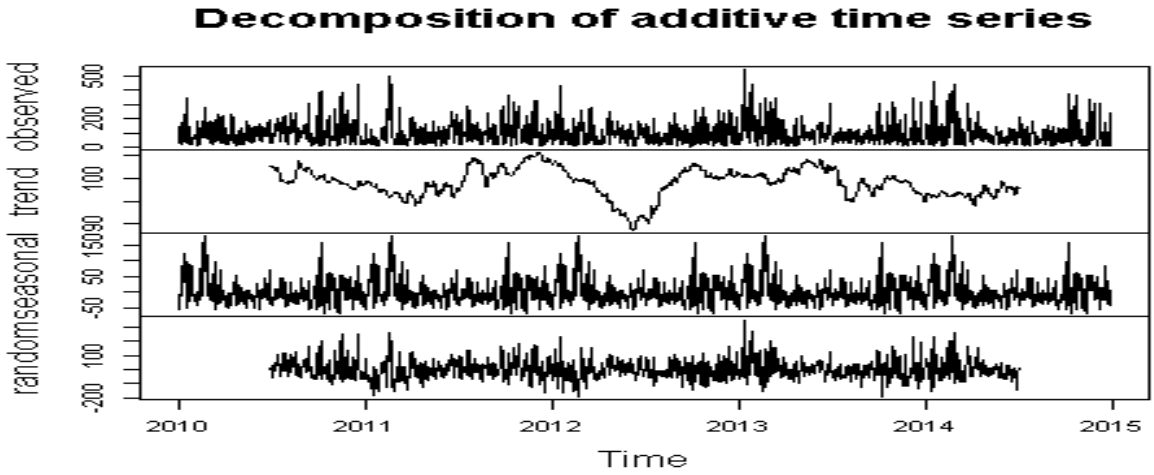
ressemble à un comportement typique d'une série temporelle, où les variations s'additionnent au fil du temps. Une analyse plus approfondie pourrait nous aider à mieux comprendre ces cycles et à anticiper les futurs pics de pollution.

- **Histogramme**



Les indices de pollution de l'air se situent dans la classe [0,100]

c) Tendance et composante saisonnière



Composant observé : Ce compartiment représente les valeurs réelles de la série temporelle “pollution_ts” mesurées au fil du temps. On observe des fluctuations avec des pics et des creux, reflétant les variations quotidiennes ou périodiques des niveaux de pollution. Ces variations peuvent être influencées par des facteurs tels que les conditions météorologiques, les activités humaines ou des événements spécifiques.

Tendance : La tendance montre l’évolution générale des niveaux de pollution sur une période plus longue. Dans ce cas, on peut voir une tendance à la hausse, indiquant que les niveaux de pollution ont globalement augmenté au fil du temps, malgré des variations ponctuelles. Cette augmentation pourrait être liée à des facteurs comme l’urbanisation, l’industrialisation ou l’augmentation du trafic.

Composant saisonnier : Ce compartiment met en évidence les variations périodiques qui se répètent à intervalles réguliers, comme des cycles annuels ou mensuels. Les fluctuations saisonnières dans les niveaux de pollution peuvent être dues à des facteurs tels que les changements de température, les conditions météorologiques ou des événements récurrents (par exemple, l’utilisation accrue de chauffage en hiver). Ces variations montrent que la pollution de l’air n’est pas constante et peut varier en fonction de la période de l’année.

Composant aléatoire : Ce compartiment capture les variations imprévisibles qui ne peuvent pas être expliquées par la tendance ou les variations saisonnières. Il représente le “bruit” dans les données, qui peut être causé par des événements exceptionnels (comme des incendies ou des tempêtes) ou des erreurs de mesure. Ce composant montre que, malgré les tendances et les

cycles saisonniers, il existe toujours une part d'incertitude et de variabilité dans les niveaux de pollution.

d)Indice statistique

```
minimum
 [1] 3

 maximum
 [1] 542

 mode
 [1] 23

 mediane
 [1] 79

 moyenne
 [1] 98.2389

 quantile
   0%  25%  50%  75% 100%
    3   42   79  131  542

 coefficient_variation
 [1] 78.18636

 variance
 [1] 5899.687

 ecart_type
 [1] 76.80942

 coefficient_assymetrie
 [1] 1.620844

 interpretation_skewness
 [1] "distribution etalee a droite"

 coefficent_applatissement
 [1] 6.445042

 interpretation_kurtosis
 [1] "distribution leptokurtique"
```

- **Indice de tendance centrale**

MOYENNE : 98.2389 la moyenne de l'indice de pollution de l'air aujourd'hui est 98

25% (42) : 25 % des indices de pollution de l'air aujourd'hui est moins de 42

50% (79) : 50% des indices de pollution de l'air aujourd'hui est moins de 79

75% (131) : 75% des indices de pollution de l'air aujourd'hui est moins de 131

- **Indice de dispersion**

ECART-TYPE : 77, La plupart des indices de pollution de l'air aujourd'hui est compris entre 21 et 175

- **Indice de forme**

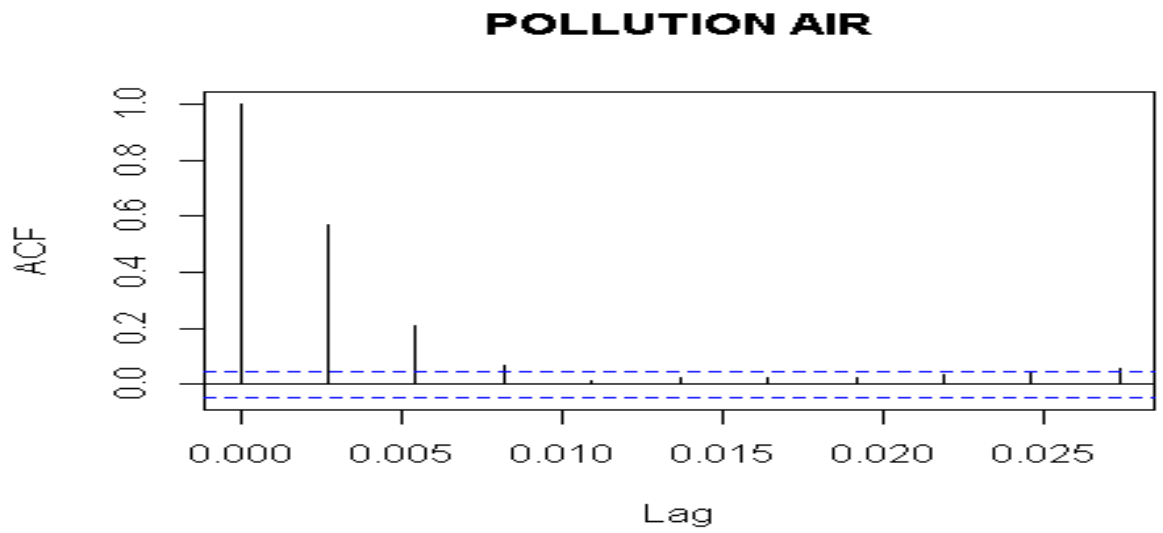
Skewness : 1.620844 distribution étalée à droite. Cela peut indiquer que la majorité des valeurs sont concentrées à gauche de la moyenne et que quelques valeurs plus petites tirent la moyenne vers la droite.

Kurtosis : 6.445042 Une kurtosis de 6.445042 indique que la distribution de la variable pollution_today est leptokurtique. Cela signifie que la courbe de la distribution est plus pointue que celle d'une distribution normale. En d'autres termes, la distribution présente moins de valeurs extrêmes que ce que l'on pourrait observer dans une distribution normale.

- **Indice de dependance**

  - **Autocorrelation simple**

```
 0.00000 0.00274 0.00548 0.00822 0.01096 0.01370 0.01644 0.01918
0.02192 0.02466
   1.000   0.569   0.212   0.069   0.012   0.022   0.027   0.025
0.038   0.045
 0.02740
   0.055
```



Le graphique ainsi obtenu est un corrélogramme. On peut constater une forte auto-corrélation

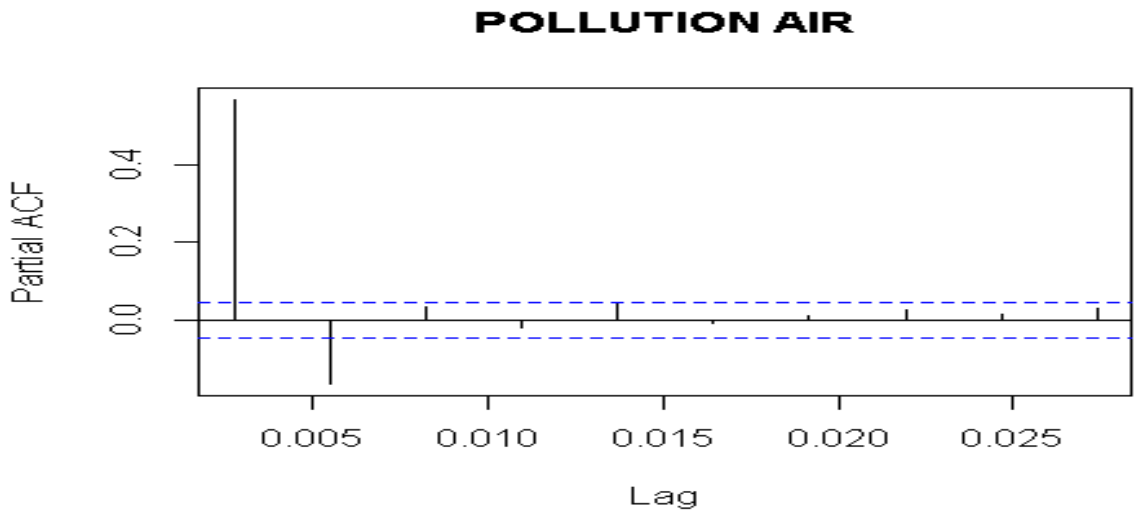
- d'ordre 1 (0.569)
- d'ordre 2 (0.212)

  - **Autocorrelation partielle**

L'autocorrélation partielle (PACF) permet de quantifier la dépendance linéaire entre deux réalisations successives mais conditionnellement aux réalisations intermédiaires.

```
 0.00274 0.00548 0.00822 0.01096 0.01370 0.01644 0.01918 0.02192
```

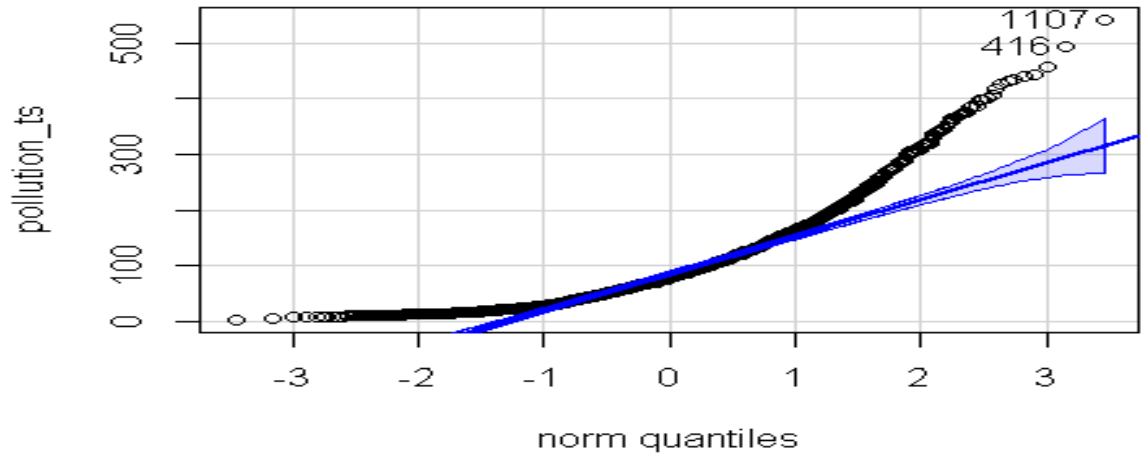
```
0.02466 0.02740
   0.569  -0.164   0.032  -0.020   0.045  -0.007   0.012   0.028
0.015   0.029
```



L'autocorrélation observée au décalage 0.00822 était un effet résiduel de l'autocorrélation pour un décalage de 0.00274

e) Test de normalité

– Graphique



– **Test**

H0 : la distribution suit une loi normale

H1 : la distribution ne suit pas une loi normale

```
  Shapiro-Wilk normality test

data:  pollution_ts
W = 0.8641, p-value < 2.2e-16
```

P-value < 0,05, on rejette H0 et on conclut que la distribution ne suit pas une loi normale.

f) Prévision des indices d'air de pollution pour les 30 prochains jours

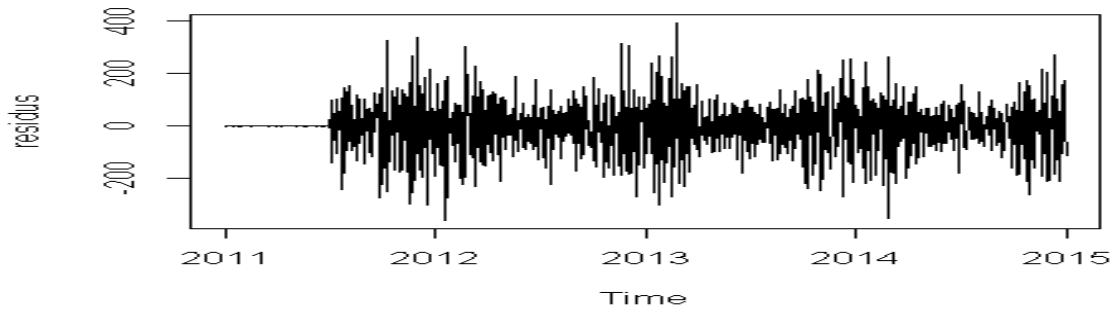
– **Validation du modèle de prévision**

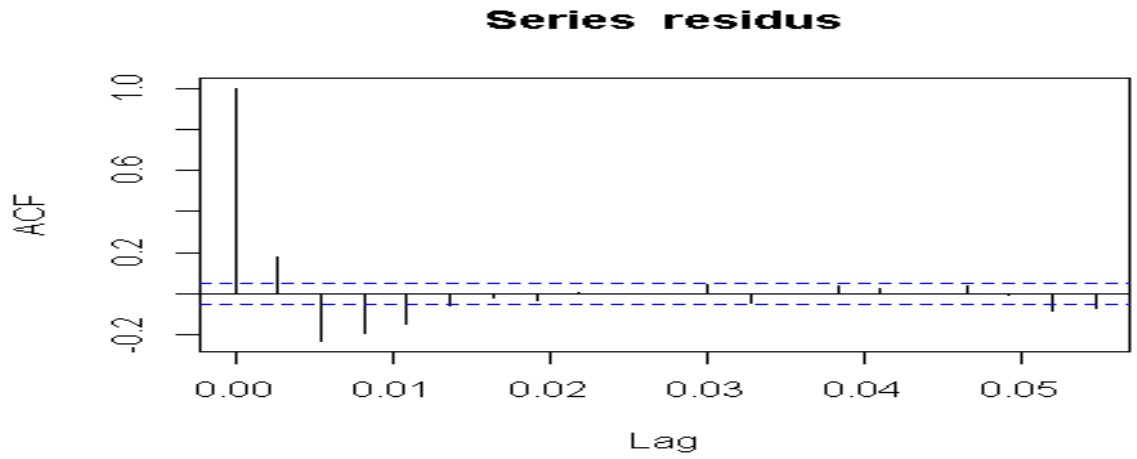
- **Récupération des résidus**

Visualisations les 6 premières résidus:

```
 [1] -1.975201448 -0.699281151 -0.009042258 0.409185146
0.070760306
 [6] -0.025224766
```

- **Graphique des résidus**





TEST

Box-Ljung test

H0 : la série est un bruit blanc

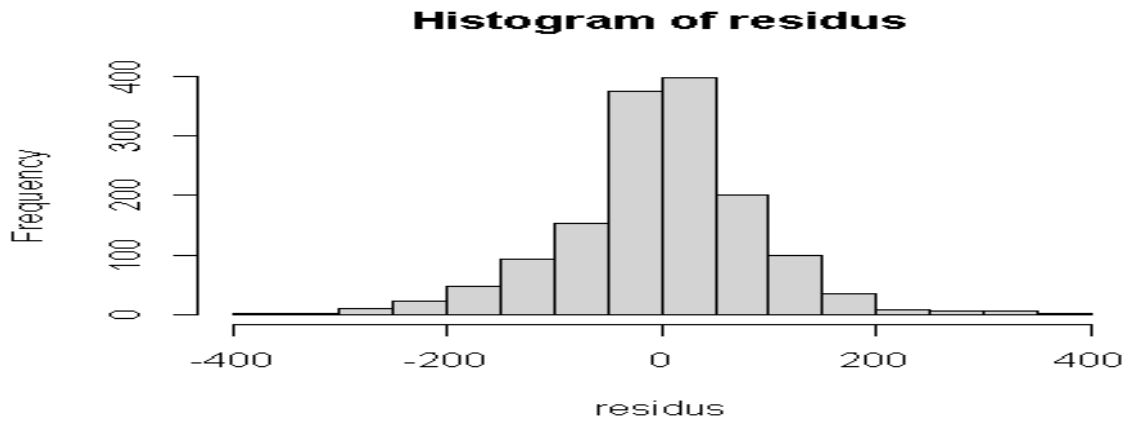
H1 : la série n'est pas un bruit blanc

```
 Box-Ljung test

data:  residus
X-squared = 246.68, df = 20, p-value < 2.2e-16
```

p-value < 0.05 donc on rejette H0 et on conclut que la série n'est pas un bruit blanc

Shapiro-Wilk normality test

Pour vérifier si les erreurs de prévision sont normalement réparties avec le zéro moyen, nous pouvons tracer un histogramme des erreurs de prévision.



On peut aussi faire un test de Shapiro Wilk

H0 : les résidus suivent une loi normale

H1 : les résidus ne suivent pas une loi normale

```
 Shapiro-Wilk normality test

data:  residus
W = 0.97512, p-value = 3.406e-15
```

Conclusion : la p-value < 0.05 donc on rejette H0 et on conclut les résidus ne suit pas une loi normale

Moyenne des résidus

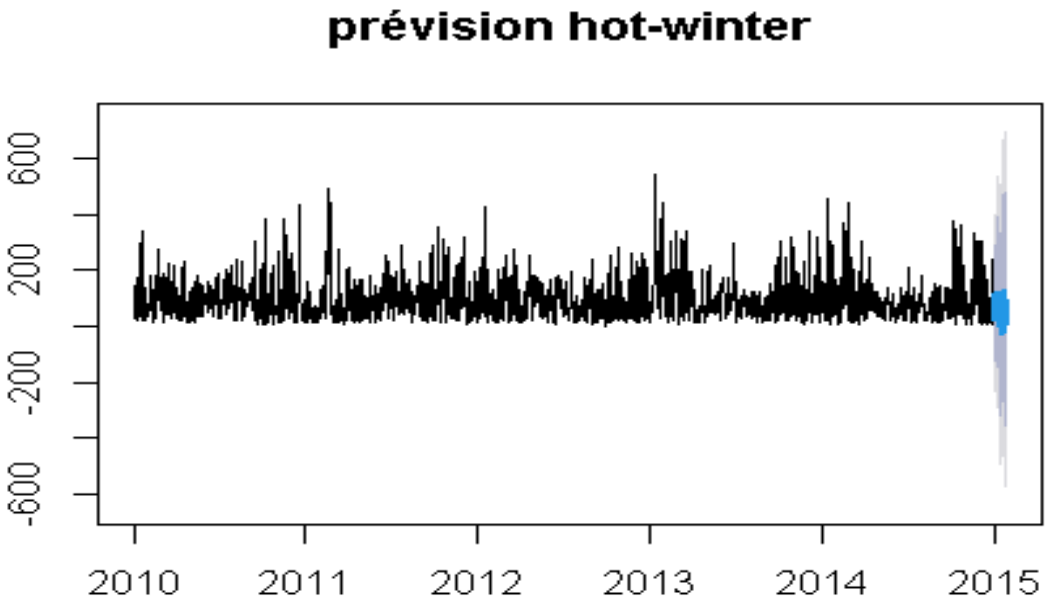
```
[1] 0.03315528
```

Les résidus de la série temporelle ne sont pas des bruits blanc-gaussien mais sont centrés

– Méthode Hot-winter

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
01-01-2015	25.337660	-88.89550	139.5708	-149.3669	200.0422
02-01-2015	105.993551	-31.88545	243.8725	-104.8742	316.8613
03-01-2015	79.574881	-78.45050	237.6003	-162.1041	321.2538
04-01-2015	116.048911	-59.83009	291.9279	-152.9348	385.0326
05-01-2015	61.993856	-130.08640	254.0741	-231.7676	355.7553
06-01-2015	83.554532	-123.46293	290.5720	-233.0514	400.1604

07-01-2015	76.483022	-144.46410	297.4301	-261.4264	414.3925
08-01-2015	36.038586	-198.01063	270.0878	-321.9088	393.9860
09-01-2015	51.439951	-195.01580	297.8957	-325.4816	428.3615
10-01-2015	73.993744	-184.27325	332.2607	-320.9915	468.9790
11-01-2015	123.609702	-145.95150	393.1709	-288.6486	535.8680
12-01-2015	108.512864	-171.88800	388.9137	-320.3233	537.3490
13-01-2015	89.665413	-201.17140	380.5022	-355.1311	534.4619
14-01-2015	37.727346	-263.18369	338.6384	-422.4764	497.9311
15-01-2015	29.755419	-280.90332	340.4142	-445.3561	504.8670
16-01-2015	124.536404	-195.57335	444.6462	-365.0292	614.1020
17-01-2015	1.704107	-327.58552	330.9937	-501.9009	505.3091
18-01-2015	53.289764	-284.93067	391.5102	-463.9738	570.5533
19-01-2015	14.554595	-332.36682	361.4760	-516.0159	545.1251
20-01-2015	-31.239444	-386.64888	324.1700	-574.7913	512.3124
21-01-2015	40.206294	-323.49313	403.9057	-516.0240	596.4366
22-01-2015	99.462832	-272.34179	471.2675	-469.1633	668.0889
23-01-2015	129.446732	-250.29012	509.1836	-451.3107	710.2041
24-01-2015	-21.241802	-408.74855	366.2649	-613.8822	571.3986
25-01-2015	3.764889	-391.35899	398.8888	-600.5249	608.0547
26-01-2015	57.118964	-345.47796	459.7159	-558.5999	672.8378
27-01-2015	92.955660	-316.97810	502.8894	-533.9839	719.8952
28-01-2015	58.017908	-359.12366	475.1595	-579.9451	695.9809
29-01-2015	91.509945	-332.71699	515.7369	-557.2892	740.3091
30-01-2015	8.157668	-423.03822	439.3536	-651.2995	667.6149



Bien que les résidus de la série temporelle ne soient pas des bruits blanc-gaussien mais sont centrés…la méthode de prévision selon le HOT-WINTER offre des prévisions avec des intervalles de confiance à des seuils de 80% et 95%.

III- MODELISATION ECONOMETRIQUE SERIE TEMPORELLE (METHODE BOX-JENKINS)

Etant donné que la statistique descriptive de la série ait déjà été fait plus haut nous passons à la méthode BOX-JENKINS qui se déroule en trois étapes qui sont : L'identification, Estimation et la prévision. Mais avant faisons un test pour vérifier s'il y a saisonnalité (Test de Kruskal-Wallis).

H0 : il n'y a pas de saisonnalité

H1: il y a saisonnalité

```
 Kruskal-Wallis rank sum test

data:  pollution_today by date
Kruskal-Wallis chi-squared = 1824, df = 1824, p-value = 0.4956
```

Une p-value sensiblement supérieure ou égale à 0.05 indique que nous ne pouvons pas rejeter H0, ce qui signifie qu'il n'y a pas de saisonnalité

A. IDENTIFICATION

Vérification de la stationnarité de la série

Pour cela il existe une battérie de test mais les plus connus sont : kpss, adf, pp.

- **Kpss-test (Kwiatkowski-Phillips-Schmidt-Shin)**

H0 : la série est stationnaire

H1 : la série n'est pas stationnaire

```
 KPSS Test for Level Stationarity

data:  pollution_ts
KPSS Level = 0.06781, Truncation lag parameter = 8, p-value = 0.1
```

p-value > 0.05 donc on ne peut rejeter H0 et on conclut que la série est stationnaire.

- **Adf-test (Augmented Dickey-Fuller)**

H0 : présence de racine unitaire donc la série n'est pas stationnaire

H1 : la série est stationnaire

NB : présence de racine unitaire signifie que la variable est intégrée d'ordre 1

```
 Augmented Dickey-Fuller Test

data :  pollution_ts
Dickey-Fuller = -10.127, Lag order = 12, p-value = 0.01
alternative hypothesis : stationary
```

p-value < 0.05 donc on rejette H0 et on conclut que la série est stationnaire.

- **pp-test (Phillips-Perron)**

H0 : présence de racine unitaire donc la série n'est pas stationnaire

H1 : la série est stationnaire

NB : présence de racine unitaire signifie que la variable est intégrée d'ordre 1

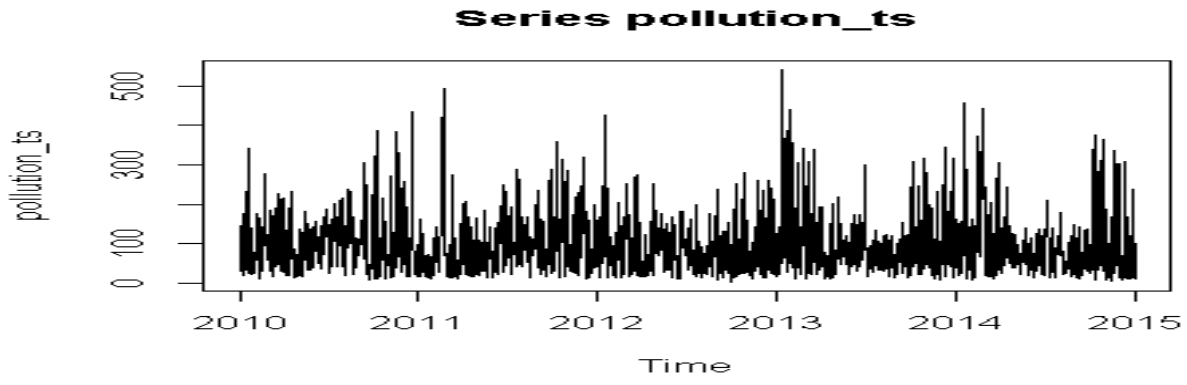
```
  Phillips-Perron Unit Root Test

 data:  pollution_ts
 Dickey-Fuller Z(alpha) = -704.18, Truncation lag parameter = 8,
p-value
 = 0.01
 alternative hypothesis: stationary
```

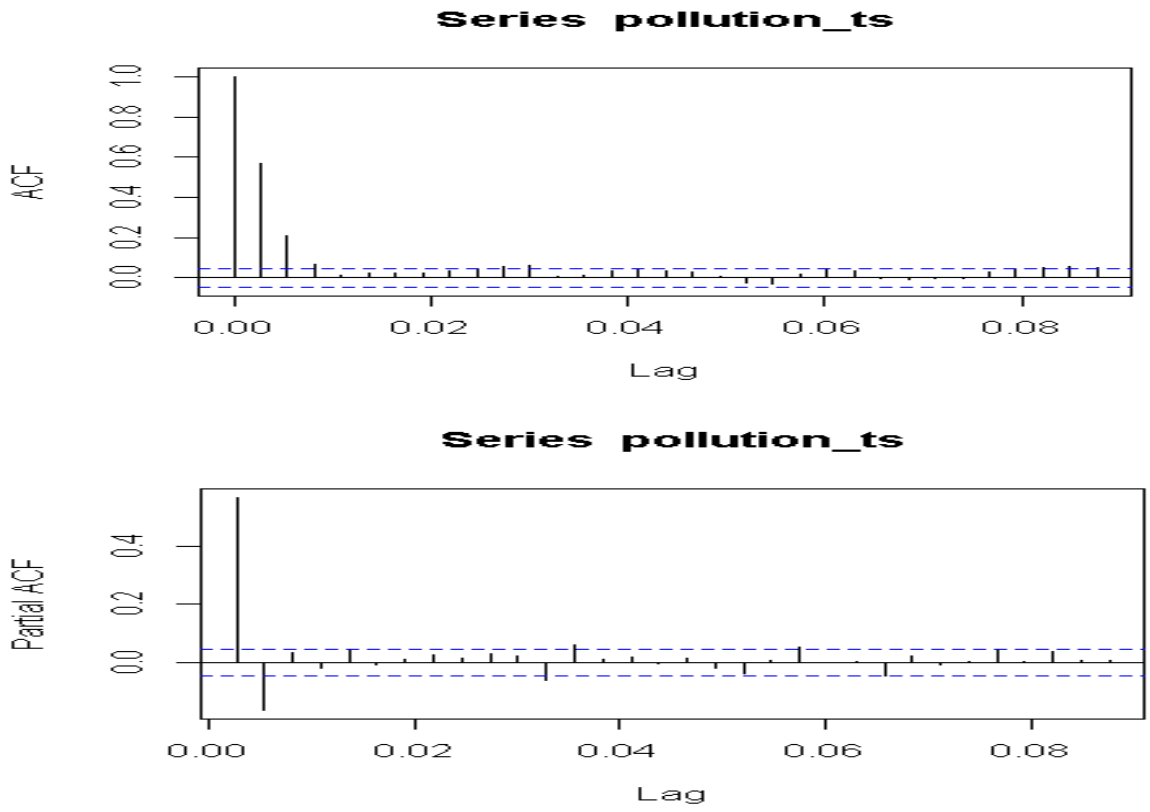
p-value < 0.05 donc on rejette H0 et on conclut que la série est stationnaire

En somme, la série temporelle pollution_ts est stationnaire donc pas besoin de la stationnariser et procéder à une modélisation ARMA (p, q)

- **Détermination des combinaisons d'auto régression(p) et de moyenne mobile (q)**

- **Graphiques**





D'après les autocorrélations simples et partiels il s'agit d'un modèle : ARMA (1,3)

Voici donc les modèles ARMA possibles pour la série pollution_ts : ARMA (1,0) ARMA (0,3)

Voici donc les modèles possibles ARIMA pour la série initiale pollution_ts : ARIMA (1,0,0) ARIMA (0,0,3)

On procède à la deuxième phase de la méthode BOX-JENKIS qui est celle de l'estimation.

B. ESTIMATION

Estimation des modèles par la fonction arima

Le modèle ARIMA (1,0,0)

```
Call:
arima(x = pollution_ts, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.5687    98.2346
s.e.  0.0192     3.4255

sigma^2 estimated as 3989:  log likelihood = -10155.55,  aic =
20317.1
```

Le modèle ARIMA (0,0,3)

```
Call:
arima(x = pollution_ts, order = c(0, 0, 3))

Coefficients:
         ma1     ma2     ma3  intercept
      0.6700  0.2575  0.0983    98.2436
s.e.  0.0233  0.0276  0.0233     2.9486

sigma^2 estimated as 3869:  log likelihood = -10127.83,  aic =
20265.65
```

Le modèle ARIMA (1,0,3)

```
Call:
arima(x = pollution_ts, order = c(1, 0, 3))

Coefficients:
          ar1     ma1     ma2     ma3  intercept
      -0.0253  0.6950  0.2739  0.1032    98.1219
s.e.   0.2450  0.2439  0.1608  0.0530     2.9414

sigma^2 estimated as 3869:  log likelihood = -10127.81,  aic =
20267.63
```

- **BILAN DES 3 MODELES**

```
      df      AIC
mod2   5 20265.65
mod3   6 20267.63
mod1   3 20317.10
```

Pour des raisons de AIC on va retenir le modele 2. Par contre pour des raisons de parcimonie, on va préférer le modèle 1 parce qu'il a moins de paramètres à estimer

Estimation automatique des modèles par la fonction auto.arima () du package forecast

```
Series: pollution_ts
 ARIMA(1,0,1) with non-zero mean

 Coefficients:
          ar1     ma1     mean
       0.3766  0.2920  98.2230
 s.e.  0.0364  0.0378   3.0188

 sigma^2 = 3882:  log likelihood = -10129.23
 AIC=20266.46   AICc=20266.48   BIC=20288.5
```

Le modèle proposé automatique est le modèle avec le plus petit AIC est le modèle ARIMA (1,0,1)

Nous allons mettre en compétition les trois modèles :

```
           df      AIC
 mod2       5 20265.65
 mod.auto   4 20266.46
 mod1       3 20317.10
```

TESTS DE VALIDATION DES MODELES : Test sur les résidus en détail

- **Bruit blanc des résidus**

```
 Box-Pierce test

data:  res1
X-squared = 15.972, df = 1, p-value = 6.428e-05
```

```
 Box-Pierce test

data:  res2
X-squared = 8.8323e-05, df = 1, p-value = 0.9925
```

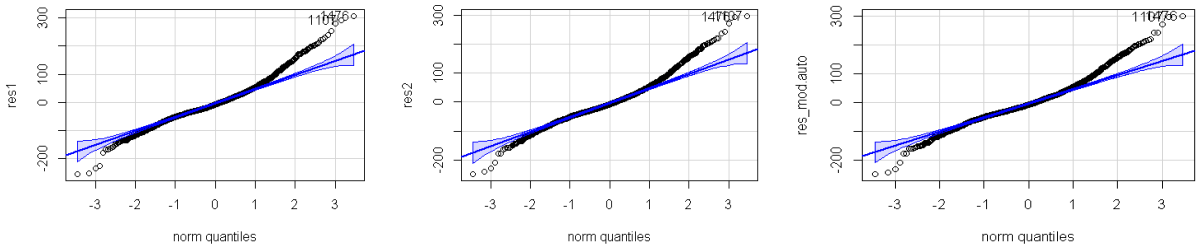
```
 Box-Pierce test

data:  res_mod.auto
X-squared = 0.00032164, df = 1, p-value = 0.9857
```

- **Normalité des résidus**

```
 Shapiro-Wilk normality test

data:  res1
W = 0.9644, p-value < 2.2e-16
```

```
 Shapiro-Wilk normality test

data:  res2
W = 0.96348, p-value < 2.2e-16
```

```
 Shapiro-Wilk normality test

data:  res_mod.auto
W = 0.96288, p-value < 2.2e-16
```



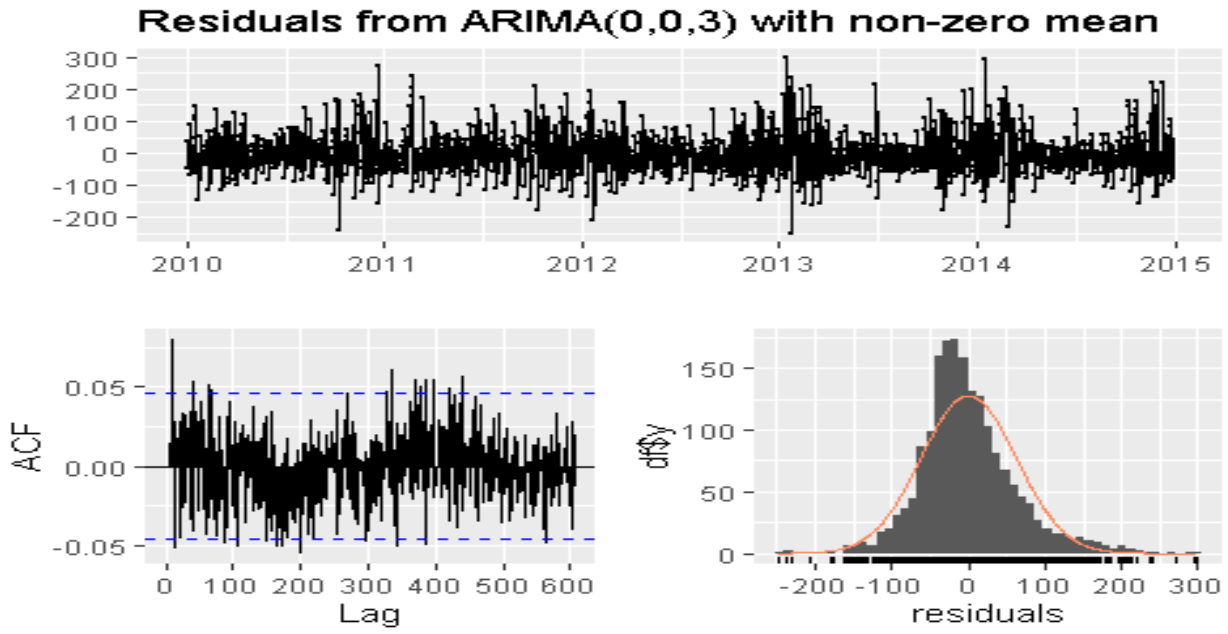
- **Centralité des résidus**

```
-0.03029916

-0.02356272

-0.01471253
```

Indicateurs	Tests	Modèle 1	Modèle 2	Modele auto
AIC		20265,65	20266,26	20317,10
Bruit blanc	Box.test()	NON	OUI	OUI
Normalité des résidus	shapiro.test() jarque.bera.test(x) qqPlot()	NON	NON	NON
Moyenne des résidus égale à 0	mean()	NON	NON	NON
CONCLUSION		Les résidus suivent un processus non-bruit blanc non-gaussien et non-centré	Les résidus suivent un processus bruit blanc non-gaussien non-centré	Les résidus suivent un processus bruit blanc non-gaussien et non-centré

On va préférer ici le modèle 2 par principe de parcimonie

VISUALISATIONS DES RESIDUS DU MODELE 2

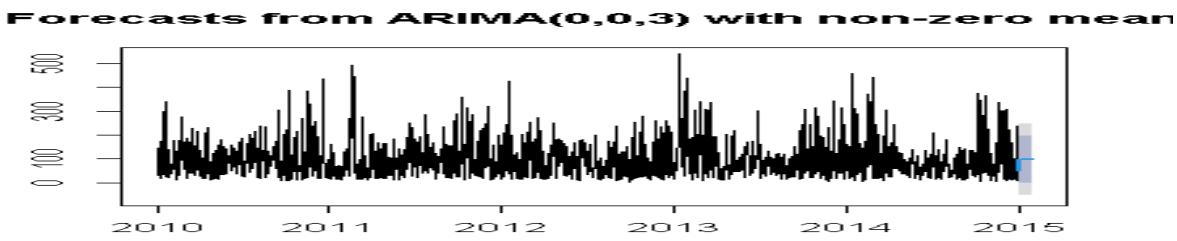


```
 Ljung-Box test

data:  Residuals from ARIMA(0,0,3) with non-zero mean
Q* = 396.62, df = 362, p-value = 0.1016

Model df: 3.   Total lags used: 365
```

Ainsi valider les résidus du modèle 2, On peut passer à la dernière étape de la méthode de BOX-JENKINS qui est la Prévision.

C.PREVISION

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
01-01-2015	51.31127	-28.4071530	131.0297	-70.60754	173.2301
02-01-2015	79.60294	-16.3525186	175.5584	-67.14827	226.3542
03-01-2015	93.99844	-4.1288682	192.1257	-56.07433	244.0712
04-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
05-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
06-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
07-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
08-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
09-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
10-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
11-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
12-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
13-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
14-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
15-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
16-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
17-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
18-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
19-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
20-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
21-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
22-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
23-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
24-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
25-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
26-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
27-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
28-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
29-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944
30-01-2015	98.24360	-0.1962746	196.6835	-52.30720	248.7944



Bien que les résidus du modèle 2 soient des bruits blanc- non gaussien et non centrés…la méthode de prévision selon BOX-JENKINS offre des prévisions avec des intervalles de confiance à des seuils de 80% et 95%.

CONCLUSION GENERALE

Cette étude s'est concentrée sur l'analyse et la prévision des niveaux de pollution de l'air sur une période de 5 ans, allant de 2010 à 2014 (soit 1825 jours), en s'appuyant sur des données météorologiques et de qualité de l'air. Deux méthodes de prévision ont été utilisées : celle de Box-Jenkins et celle de Holt-Winters. Bien que ces méthodes aient montré des résultats significatifs, certaines limites liées aux résidus des modèles ont été identifiées. Voici les principales conclusions tirées de cette analyse :

Pour la méthode de Holt-Winters, les résidus de la série temporelle ne suivent pas une distribution gaussienne (bruit blanc), mais ils sont centrés. Cela signifie que le modèle parvient à capturer une partie importante de la structure des données, même s'il reste des pistes d'amélioration pour mieux modéliser les résidus. En ce qui concerne la méthode de Box-Jenkins, les résidus du modèle 2 sont également non gaussiens et non centrés, ce qui suggère que le modèle ne capture pas entièrement la variabilité des données. Malgré cela, il fournit des prévisions fiables, accompagnées d'intervalles de confiance robustes.

Les deux méthodes ont permis de générer des prévisions pour les 30 prochains jours, avec des intervalles de confiance à 80 % et 95 %. Ces intervalles sont essentiels, car ils donnent une idée de l'incertitude associée aux prévisions, un élément crucial pour la prise de décision. La méthode de Box-Jenkins, en particulier, a démontré une forte capacité à modéliser les tendances et les variations saisonnières des niveaux de pollution, même si les résidus ne sont pas parfaitement gaussiens. Cela met en lumière l'importance de bien comprendre les spécificités des données avant de choisir un modèle de prévision. Les résultats de cette

étude sont particulièrement utiles pour anticiper les épisodes de pollution et mettre en place des mesures préventives. Les variables météorologiques, comme la température, la pression atmosphérique, la vitesse du vent et les précipitations, jouent un rôle clé dans la dispersion ou l'accumulation des polluants. Leur inclusion dans les modèles de prévision est donc justifiée. Par ailleurs, les intervalles de confiance fournis par les deux méthodes permettent aux décideurs d'évaluer plus finement les risques liés aux niveaux de pollution prévus.

Recommandations

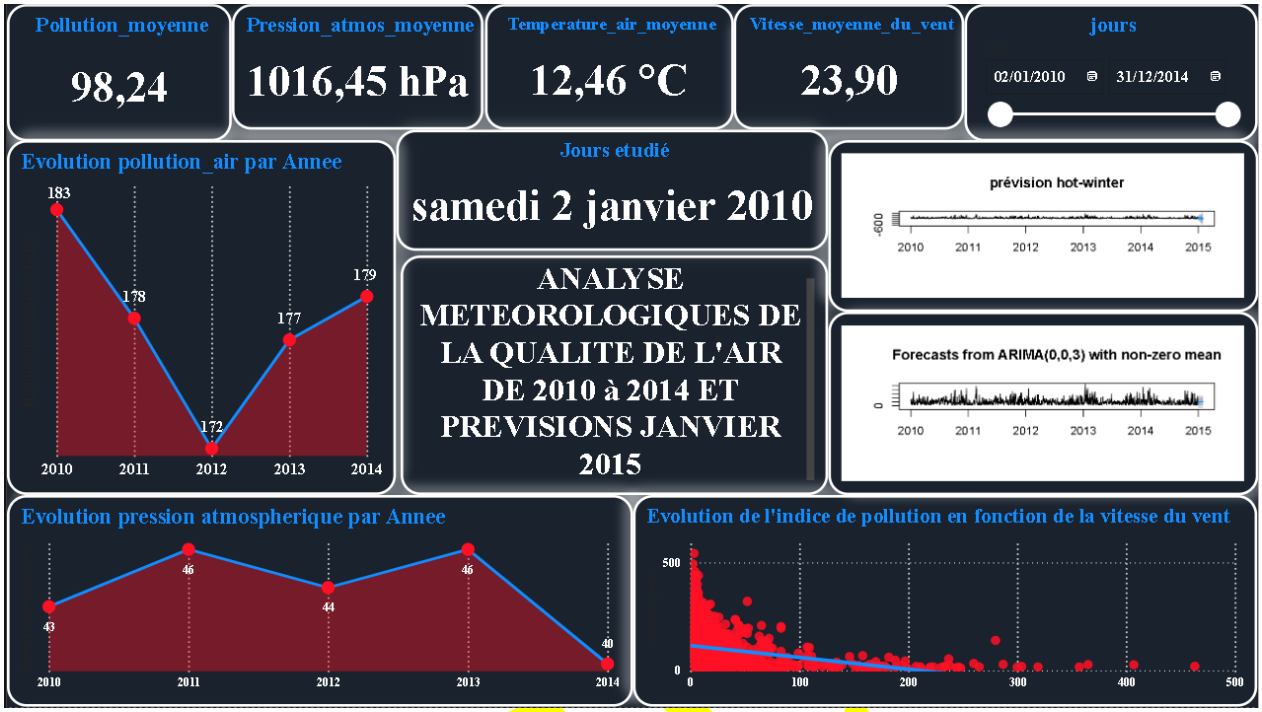
- Explorer d'autres modèles : Il serait intéressant de tester d'autres approches, comme les modèles à mémoire longue (ARFIMA) ou des modèles hybrides combinant des techniques statistiques et d'apprentissage automatique. Ces méthodes pourraient mieux capturer les particularités des résidus et affiner les prévisions.
- Ajuster les modèles existants : Les modèles actuels pourraient être améliorés en appliquant des transformations aux données ou en utilisant des techniques de rééchantillonnage pour mieux gérer les résidus non gaussiens et non centrés.
- Enrichir les données : L'ajout de variables supplémentaires, comme les émissions de polluants, les données sur le trafic routier, les activités industrielles ou même des informations satellitaires, pourrait apporter une plus grande précision aux prévisions.
- Valider les modèles sur des données récentes : Pour s'assurer de la robustesse des modèles, il serait utile de les

tester sur des données récentes et indépendantes. Cela permettrait de vérifier leur performance en conditions réelles et d'identifier d'éventuelles limites.

- Mieux communiquer les résultats : Les intervalles de confiance, qui reflètent l'incertitude des prévisions, devraient être clairement expliqués aux décideurs et au public. Une communication transparente est essentielle pour une prise de décision éclairée.
- Surveillance continue : Maintenir une surveillance continue des niveaux de pollution de l'air et des conditions météorologiques pour ajuster les prévisions et les mesures en temps réel.
- Réduction des émissions : Mettre en œuvre des politiques pour réduire les émissions de polluants atmosphériques provenant des sources principales, telles que les véhicules, les industries et les centrales électriques.
- Promotion des transports durables : Encourager l'utilisation des transports en commun, du covoiturage, du vélo et de la marche pour réduire les émissions de gaz d'échappement.
- Sensibilisation du public : Informer et sensibiliser le public aux effets de la pollution de l'air sur la santé et aux mesures qu'ils peuvent prendre pour réduire leur exposition.

Pour plus de détails, vous pouvez consulter les lignes directrices de l'OMS sur la qualité de l'air (https://www.who.int/fr/news-room/questions-and-answers/item/who-global-air-quality-guidelines)

POWER BI



CODE R et POWER-QUERY

- **Power-Query**

```
let

    Source =
Csv.Document(File.Contents("C:\Users\HP\Downloads\INSSE
DS\cours\ECONOMETRIE\Mini projet économétrie Serie
temporelle\air_pollution.csv"),[Delimiter=",",
Columns=9, Encoding=1252, QuoteStyle=QuoteStyle.None]),

    #"En-têtes promus" = Table.PromoteHeaders(Source,
[PromoteAllScalars=true]),

    #"Type modifié" = Table.TransformColumnTypes(#"En-
têtes promus",{{"date", type text}, {"pollution_today",
type text}, {"dew", type text}, {"temp", type text},
{"press", type text}, {"wnd_spd", type text}, {"snow",
type text}, {"rain", type text},
{"pollution_yesterday", type text}}),
```

```
    #"Valeur remplacée" = Table.ReplaceValue(#"Type
modifié",".","",Replacer.ReplaceText,{"pollution_today
"}),

    #"Type modifié1" =
Table.TransformColumnTypes(#"Valeur
remplacée",{{"pollution_today", Int64.Type}}),

    #"Valeur remplacée1" = Table.ReplaceValue(#"Type
modifié1",".","",Replacer.ReplaceText,{"dew"}),

    #"Type modifié2" =
Table.TransformColumnTypes(#"Valeur
remplacée1",{{"dew", Int64.Type}}),

    #"Valeur remplacée2" = Table.ReplaceValue(#"Type
modifié2",".","",Replacer.ReplaceText,{"temp"}),

    #"Type modifié3" =
Table.TransformColumnTypes(#"Valeur
remplacée2",{{"temp", Int64.Type}}),

    #"Valeur remplacée3" = Table.ReplaceValue(#"Type
modifié3",".","",Replacer.ReplaceText,{"press",
"wnd_spd", "snow", "rain", "pollution_yesterday"}),

    #"Type modifié4" =
Table.TransformColumnTypes(#"Valeur
remplacée3",{{"press", Int64.Type}, {"wnd_spd",
Int64.Type}, {"snow", Int64.Type}, {"rain",
Int64.Type}, {"pollution_yesterday", Int64.Type}}),

    #"Colonnes permutées" = Table.ReorderColumns(#"Type
modifié4",{"date", "pollution_today",
"pollution_yesterday", "dew", "temp", "press",
"wnd_spd", "snow", "rain"})

in

    #"Colonnes permutées"
```

- **R**

I- Pretraitement des données

Visualisation des données

```
library(readxl)

air_pollution1 <-
read_excel("C:/Users/HP/Downloads/INSSEDS/cours/ECONOMETRIE/Mini
projet économétrie Serie temporelle/air_pollution1.xlsx")

str(air_pollution1)
```

Valeurs manquantes

```
colSums(is.na(air_pollution1))
```

Valeurs abberantes

Visualisations

```
library(akposso)

afficher_boites_a_moustache(air_pollution1)
```

II- ANALYSE DESCRIPTIVE ET PREVISIONS HOT-WINTER

a) Construction de la serie temporelle

```
library(dplyr)

pollution <- air_pollution1 %>%
select(date,pollution_today)

pollution$date <- as.Date(pollution$date)
```

Créer une série temporelle de la colonne 'airpollution'

```
pollution_ts  <-  ts(pollution$pollution_today,  start  =  c(2010,
01), frequency = 365)

head(pollution_ts)
```

b) Graphiques

Visualiser la série temporelle

```
plot(pollution_ts, col = "black", main = "SERIE TEMPORELLE", xlab
= "Temps", ylab ="Pollution de l'air")
```

Histogramme

```
hist(pollution_ts, main = "Histogramme", prob=TRUE,lwd = 2)
lines(density(pollution_ts,na.rm = FALSE))
```

c) Tendance et composante saisonnière

```
decomposition_add=decompose(pollution_ts, type = "add")
plot(decomposition_add)
```

d) Indice statistique

```
library(onesime)
onesime_qt_resume(pollution_ts)
```

Autocorrélation simple

```
acf(pollution_ts,lag.max=10,plot = FALSE, main="POLLUTION AIR")
acf(pollution_ts,lag.max=10,plot = TRUE, main="POLLUTION AIR")
```

Autocorrélation partielle

```
pacf(pollution_ts,lag.max=10,plot = FALSE, main="POLLUTION AIR")
pacf(pollution_ts,lag.max=10,plot = TRUE, main="POLLUTION AIR")
```

e) Test de normalité

Graphique

```
library(car)
qqPlot(pollution_ts)
```

Test

```
shapiro.test(pollution_ts)
```

f) Prévision des indices d'air de pollution pour les 30 prochains jours

Validation du modèle de prévision

Récupération des résidus

```
xlisse <- HoltWinters(pollution_ts)
residus <- residuals(xlisse)
head(residus)
```

Graphique des résidus

```
plot(residus)
acf(residus, lag.max=20, na.action = na.pass)
```

TEST

```
Box.test(residus, lag=20, type="Ljung-Box")
```

Shapiro-Wilk normality test

```
hist(residus)
shapiro.test(residus)
```

Moyenne des résidus

```
mean(residus)
```

Méthode Hot-winter

```
library(tseries)
library(forecast)
xlisse <- HoltWinters(pollution_ts)
```

Faire une prévision pour les 30 prochains jours

```
prevision <- forecast(xlisse, h = 30)
forecast(xlisse,h = 30)
```

Visualiser la prévision pour les 30 prochains jours

```
plot(prevision,main = "prévision hot-winter")
```

III- MODELISATION ECONOMETRIQUE SERIE TEMPORELLE (METHODE BOX-JENKINS)

Test de Kruskal-Wallis pour la saisonnalité

```
test_result <- kruskal.test(pollution_today ~ date,
data = pollution)

print(test_result)
```

A) IDENTIFICATION

A-1) Vérification de la stationnarité de la série

- kpss

```
kpss.test(pollution_ts)
```

- adf

```
adf.test(pollution_ts)
```

- pp

```
pp.test(pollution_ts)
```

A-2) Détermination des combinaisons d'auto régression(p) et de moyenne mobile (q)

Graphiques

```
plot(pollution_ts,main="Series pollution_ts")

acf(pollution_ts)

pacf(pollution_ts)
```

B) ESTIMATION

Estimation des modèles par la fonction arima

le modèle ARIMA(1,0,0)

```
mod1 <- arima (pollution_ts, order=c(1,0,0))

mod1
```

le modèle ARIMA(0,0,3)

```
mod2 <- arima (pollution_ts, order=c(0,0,3))
mod2
```

le modèle ARIMA(1,0,3)

```
mod3 <- arima (pollution_ts, order=c(1,0,3))
mod3
```

BILAN des 3 MODELES

Choix du meilleur modele par le critère AIC minimum

```
sc.AIC = AIC(mod1,mod2,mod3)
sort.score <- function(x, score = c("bic", "aic")){
 if (score == "aic"){
 x[with(x, order(AIC)),]
 } else if (score == "bic") {
 x[with(x, order(BIC)),]
 } else {
 warning('score = "x" only accepts valid arguments
("aic","bic")')
 }
}
sort.score(sc.AIC, score ="aic")
```

Estimation automatique des modèles par la fonction auto.arima() du package forecast

```
auto.arima(pollution_ts)


mod.auto<-arima(pollution_ts,order=c(1,0,1))


sc.AIC = AIC(mod1,mod2,mod.auto)
```

```
sort.score <- function(x, score = c("bic", "aic")){
 if (score == "aic"){
 x[with(x, order(AIC)),]
 } else if (score == "bic") {
 x[with(x, order(BIC)),]
 } else {
 warning('score = "x" only accepts valid arguments
("aic","bic")')
 }
}
sort.score(sc.AIC, score ="aic")
```

TESTS DE VALIDATION DES MODELES : Test sur les résidus en détail

```
res1 <- residuals(mod1)
res2 <- residuals(mod2)
res_mod.auto <- residuals(mod.auto)
```

Bruit blanc des résidus

```
Box.test(res1)
Box.test(res2)
Box.test(res_mod.auto)
```

Normalité des résidus

```
shapiro.test(res1)
shapiro.test(res2)
shapiro.test(res_mod.auto)
library(car)
qqPlot(res1)
```

```
qqPlot(res2)

qqPlot(res_mod.auto)
```

Centralité des résidus

```
mean(res1)

mean(res2)

mean(res_mod.auto)
```

VISUALISATIONS DES RESIDUS DU MODELE 2

```
checkresiduals(mod2)
```

C) PREVISION

```
library(forecast)

mod2 <- arima(pollution_ts, order=c(0,0,3))

prediction <- forecast(mod2,h=30) # pour les 30
prochains jours

prediction

plot(prediction)
```