Institut Supérieur de Statistique D'Econométrie

REPUBLIQUE DE COTE D'IVOIRE

Union-Discipline-Travail

MASTER 1

STATISTIQUES – ECONOMETRIE – DATA SCIENCE

MINI PROJET

ANALYSE STATISTIQUES ECONOMETRIQUES

MODÉLISATION ET PRÉVISION À 30 JOURS PAR LA MÉTHODE BOX-JENKINS

Nom: YOBO

Prénom(s): BAYE GUY ANGE HENOC

Enseignant – Encadreur

AKPOSSO DIDIER MARTIAL

AVANS PROPOS

Dans un contexte mondial où la qualité de l'air devient un enjeu de santé publique majeur, la compréhension et la prévision des niveaux de pollution atmosphérique sont essentielles pour anticiper les risques sanitaires et orienter les décisions politiques. La ville de Pékin, en tant que mégapole fortement urbanisée et industrialisée, fait régulièrement face à des niveaux de pollution préoccupants, justifiant la mise en place de systèmes d'analyse et de prévision robustes.

Ce travail s'inscrit dans cette logique en exploitant un **ensemble de données publiques** portant sur la qualité de l'air à Pékin entre **2014 et 2019**. Le jeu de données regroupe des mesures quotidiennes de pollution atmosphérique ainsi que des variables météorologiques telles que la température, la pression, le vent, la pluie et la neige. Ces données, riches et variées, constituent une base solide pour développer un modèle de prévision fiable.

L'objectif principal de cette étude est de réaliser une **prévision des niveaux de pollution de l'air pour les 30 jours à venir**, à l'aide de la **méthodologie de Box & Jenkins**, une approche rigoureuse basée sur les modèles ARIMA (AutoRegressive Integrated Moving Average). Cette méthodologie est particulièrement adaptée à l'analyse des séries chronologiques et permet d'exploiter les dépendances temporelles dans les données historiques.

L'analyse sera réalisée à l'aide du logiciel **[à adapter selon votre choix : R / Python / Excel / Power BI]**, qui offre les outils nécessaires pour l'exploration, la modélisation et la visualisation des données temporelles.

À travers cette étude, nous cherchons non seulement à produire une prévision quantitative, mais également à mieux comprendre les **relations dynamiques entre les conditions météorologiques et les niveaux de pollution**, et à dégager des **enseignements utiles pour la gestion environnementale**.

Table des matières

Liste des figures

Liste des tableaux

INTRODUCTION

Contexte et justification de l'étude

La pollution de l'air est aujourd'hui l'un des enjeux environnementaux et sanitaires majeurs auxquels sont confrontées les grandes métropoles. Pékin, capitale de la Chine, est régulièrement citée parmi les villes les plus touchées par ce phénomène. La croissance urbaine rapide, le trafic routier dense et les conditions météorologiques particulières y contribuent fortement. Dans ce contexte, une analyse rigoureuse de la qualité de l'air et de ses déterminants météorologiques constitue un outil précieux pour anticiper les pics de pollution et orienter les politiques publiques en matière d'environnement et de santé publique.

L'étude s'appuie sur un jeu de données public couvrant la période de 2014 à 2019, offrant un échantillonnage fin et une consolidation journalière des niveaux de pollution ainsi que de multiples variables météorologiques. Ce riche ensemble de données permet de mieux comprendre les interactions entre les conditions atmosphériques et les variations de la pollution au fil du temps.

Problématique

Dans quelle mesure les variables météorologiques influencent-elles les niveaux quotidiens de pollution de l'air à Pékin ? Peut-on prédire efficacement la pollution journalière à partir de ces variables ? Et dans quelle mesure les données de la veille permettent-elles d'anticiper les niveaux du jour ?

Ces interrogations s'inscrivent dans une volonté d'améliorer les modèles prédictifs de qualité de l'air en intégrant des facteurs exogènes (température, pression, précipitations, etc.) ainsi que des dynamiques temporelles (pollution d'hier).

Principaux résultats attendus

L'étude vise principalement à :

- Identifier les facteurs météorologiques ayant une influence significative sur la pollution quotidienne.
- Évaluer la contribution de la pollution de la veille à la prédiction des niveaux du jour.
- Proposer un modèle prédictif fiable et interprétable de la qualité de l'air à Pékin, utilisable dans un cadre de prévention ou d'alerte.

Méthodologie

Pour répondre à cette problématique, plusieurs étapes méthodologiques seront mises en œuvre :

1. **Prétraitement des données**
2. **Analyse exploratoire**
3. **Analyses statistiques et fondements théoriques**
    - **Régression linéaire multiple** pour évaluer l'influence individuelle et conjointe des variables explicatives.
    - **Analyse de corrélation croisée** pour détecter les décalages temporels entre pollution et facteurs météorologiques.
    - **Modèles prédictifs supervisés** (régressions régularisées, arbres de décision ou forêts aléatoires) afin de comparer les performances de prédiction.
    - Validation croisée et évaluation des modèles par des métriques telles que le RMSE, MAE ou R^2.

PARTIE I : ANALYSE DESCRIPTIVE ET EXPLORATOIRE DES DONNEES.

A) APPROCHE METHODOLOGIQUE DES DONNEES

L'approche méthodologique des données englobe l'organisation, la collecte, l'analyse et l'interprétation des données dans le cadre d'une étude ou d'une recherche. Elle repose sur un ensemble de principes, de techniques et de processus visant à traiter les données de manière systématique et rigoureuse, afin d'obtenir des résultats fiables et pertinents.

A.1) Information sur le jeu de donnée

date	pollution_today	dew	temp	press	wnd_spd	snow	rain	pollution_yesterday
02/01/2010	145.958333	-8.5	-5.125	1024.75	24.860000	0.708333	0.0	10.041667
03/01/2010	78.833333	-10.125	-8.541667	1022.7917	70.937917	14.166667	0.0	145.958333
04/01/2010	31.333333	-20.875	-11.5	1029.2917	111.160833	0.0	0.0	78.833333
05/01/2010	42.458333	-24.583333	-14.458333	1033.625	56.92	0.0	0.0	31.333333
06/01/2010	56.416667	-23.708333	-12.541667	1033.75	18.511667	0.0	0.0	42.458333
07/01/2010	69.0	-21.25	-12.5	1034.0833	10.17	0.0	0.0	56.416667
08/01/2010	176.208333	-17.125	-11.708333	1028.0	1.972917	0.0	0.0	69.0
09/01/2010	88.5	-16.333333	-9.125	1029.0417	13.29875	0.0	0.0	176.208333
10/01/2010	57.25	-15.958333	-8.75	1032.5	17.415833	0.0	0.0	88.5
11/01/2010	20.0	-20.708333	-8.708333	1034.3333	41.685833	0.0	0.0	57.25
12/01/2010	20.75	-23.541667	-12.416667	1030.7083	60.378333	0.0	0.0	20.0
13/01/2010	40.208333	-21.958333	-10.0	1030.4583	169.754167	0.0	0.0	20.75
14/01/2010	93.708333	-17.625	-9.5	1025.2083	13.23875	0.0	0.0	40.208333
15/01/2010	45.458333	-17.166667	-7.041667	1036.8333	12.381667	0.0	0.0	93.708333
16/01/2010	177.625	-13.5	-8.416667	1033.0417	2.677917	0.0	0.0	45.458333
17/01/2010	209.208333	-12.083333	-7.25	1028.8333	4.134583	0.0	0.0	177.625
18/01/2010	260.208333	-9.666667	-4.916667	1026.75	4.91625	0.0	0.0	209.208333
19/01/2010	340.75	-3.791667	0.291667	1020.6667	4.788333	0.0	0.0	260.208333
20/01/2010	85.333333	-11.041667	-1.166667	1030.2083	34.48125	0.0	0.0	340.75
21/01/2010	27.041667	-21.166667	-6.125	1036.375	59.070833	0.0	0.0	85.333333
22/01/2010	29.416667	-18.791667	-4.583333	1034.375	93.062083	0.0	0.0	27.041667
23/01/2010	23.965686	-17.708333	-1.916667	1028.0	43.892083	0.0	0.0	29.416667
31/01/2010	39.25	-15.791667	1.333333	1024.2083	62.510417	0.0	0.0	44.291667

Tableau 1 Inffo jeu de donnée

A.2) Détection et traitement des valeurs manquantes et aberrantes / extrêmes

Dans cette section, nous allons identifier visuellement les éventuelles valeurs manquantes ou aberrantes présentes dans notre jeu de données, puis appliquer les traitements appropriés. Ces anomalies peuvent résulter d'erreurs de mesure, de saisie, de calcul, ou encore correspondre à des valeurs extrêmes réelles mais rares.

Les valeurs atypiques, qu'elles soient manquantes ou extrêmes, peuvent fortement perturber les analyses statistiques. Elles ont notamment un impact sur les mesures de tendance centrale (comme la moyenne) et de dispersion (comme l'écart-type), et peuvent fausser les résultats des tests d'hypothèse.

Il est donc essentiel de détecter et de corriger ces valeurs avant toute analyse approfondie, afin de garantir la fiabilité et la robustesse des résultats obtenus.

A.2.1) Visualisation des valeurs manquantes manquantes

Notre jeu de donnée ne contient aucunes valeurs manquantes, ce qui favorise la suite de l'étude statistiques

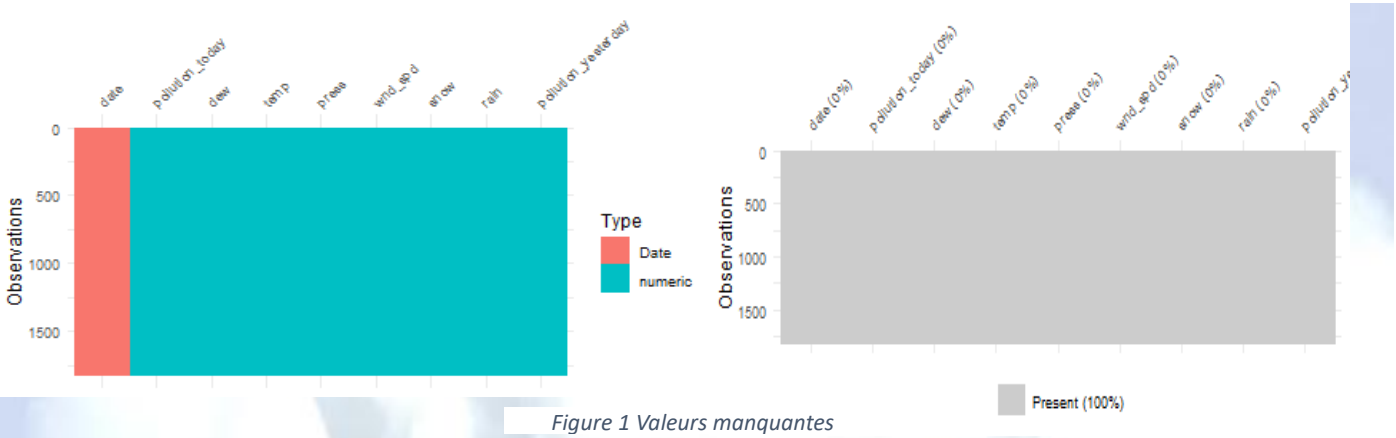
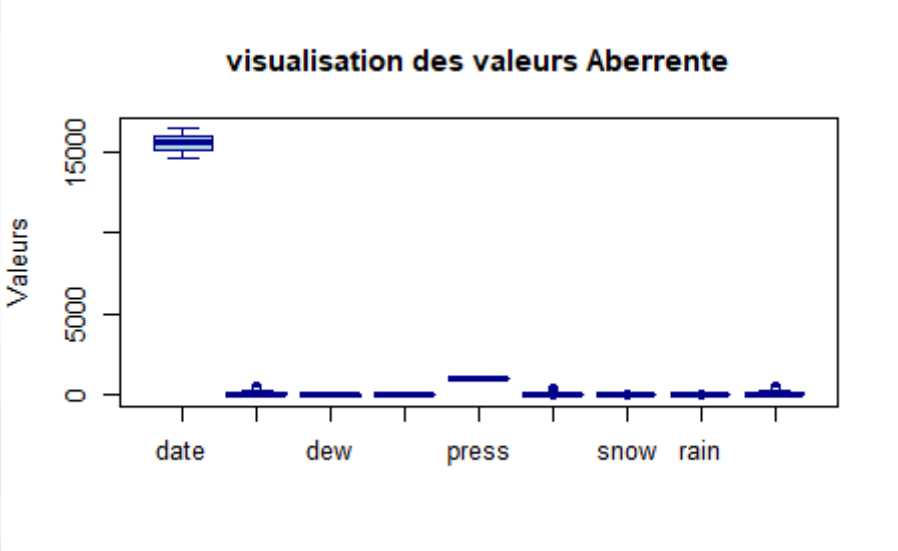


Figure 1 Valeurs manquantes

A.2.2) Visualisation des valeurs aberrantes



B) ETUDE STATISTIQUE DE LA SERIE TEMPORELLE (HOLT-WINTERS)

Le **modèle Holt-Winters**, aussi appelé **exponentielle lissée triple**, est une méthode de prévision utilisée pour les **séries temporelles** présentant à la fois une **tendance** et une **saisonnalité**. Il existe en deux versions : **additive** et **multiplicative**, selon la nature de la saisonnalité (constante ou proportionnelle à la tendance).

Objectif du modèle Holt-Winters

Le but est de **prévoir les valeurs futures** d'une série temporelle en tenant compte de :

- la tendance (croissance ou décroissance)
- la saisonnalité (variations cycliques régulières)
- les résidus (bruit aléatoire)

B.1) Construction de la série

Dans cette étape, nous allons nous concentrer sur les dates et la variable **pollution_today** afin de construire notre série temporelle. Ces deux éléments sont essentiels, car ils nous fourniront les informations nécessaires pour analyser l'évolution de la pollution de l'air au fil du temps. En prenant la date comme axe temporel et **pollution_today** comme variable d'intérêt, nous serons en mesure d'identifier des tendances, des schémas saisonniers ou des motifs récurrents. Ces insights forment la base de toute modélisation en séries temporelles. Cette approche nous permettra de mieux comprendre les variations quotidiennes de la pollution et de poser les bases pour des prévisions futures plus précises.

Date	Pollution_tod
02/01/2010	145.958333
03/01/2010	78.833333
04/01/2010	31.333333
05/01/2010	42.458333
06/01/2010	56.416667
07/01/2010	69.000000
08/01/2010	176.208333
09/01/2010	88.500000
10/01/2010	57.250000
11/01/2010	20.000000
12/01/2010	20.750000
13/01/2010	40.208333
14/01/2010	93.708333
15/01/2010	45.458333
16/01/2010	177.625000
17/01/2010	209.208333
18/01/2010	260.208333
19/01/2010	340.750000
20/01/2010	85.333333
21/01/2010	27.041667
22/01/2010	29.416667
23/01/2010	23.965686
24/01/2010	40.926471
25/01/2010	64.220588
26/01/2010	138.637255
27/01/2010	122.333333
28/01/2010	21.166667
29/01/2010	23.875000
30/01/2010	44.291667
31/01/2010	39.250000
01/02/2010	64.791667
02/02/2010	65.625000
03/02/2010	77.541667
04/02/2010	58.500000
05/02/2010	78.458333

Tableau 2 présentation de la série

Cette série temporelle montre l'évolution des niveaux de pollution pour chaque jour à partir du **2 janvier 2010**. Les valeurs varient considérablement, avec des pics de pollution qui semblent se produire de manière irrégulière. Les valeurs commencent relativement élevées, puis diminuent avant de fluctuer à nouveau. Ce genre de données peut refléter des facteurs saisonniers ou d'autres événements locaux ayant un impact sur la qualité de l'air.

B.1.1) Visualisation de la serie

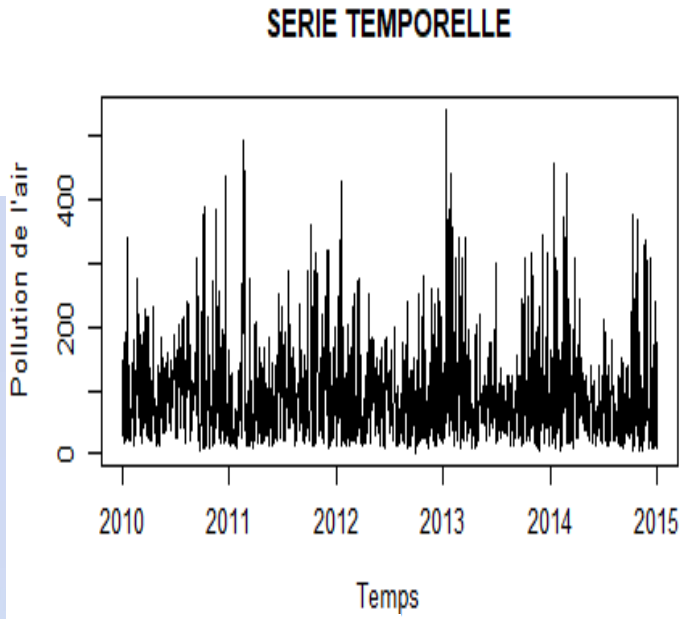


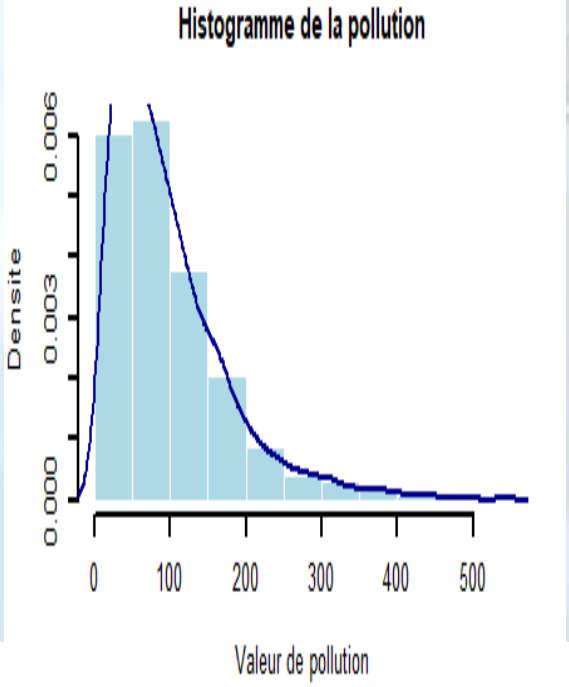
Figure 2 visualisation de la serie

Variabilité marquée
La série présente de fortes fluctuations, avec des pics et des creux importants tout au long de la période observée. Ces variations indiquent une dynamique instable du phénomène.

Absence de tendance globale claire
Aucune tendance évidente à la hausse ou à la baisse ne se dégage sur l'ensemble de la période. Les niveaux de pollution semblent osciller autour d'une moyenne relativement stable.

Présence d'une saisonnalité
Des motifs récurrents apparaissent à intervalles réguliers, suggérant un **comportement saisonnier**. Certains pics semblent revenir chaque année, possiblement liés à des conditions météorologiques ou à des activités humaines saisonnières (chauffage, trafic, etc.).

Épisodes de pollution élevée
À plusieurs reprises, les niveaux de pollution dépassent les **300 unités**, ce qui correspond à des épisodes de pollution intense. Ces valeurs extrêmes méritent une attention particulière en raison de leur impact potentiel sur la santé publique.



- **Distribution asymétrique (skewed)**

La répartition des données n'est pas symétrique : on observe une **asymétrie vers la droite**, avec une concentration importante de valeurs faibles et une longue queue vers les valeurs élevées.

- **Mode autour de 100**

Le **pic de fréquence** se situe aux alentours de 100 unités de pollution, indiquant que ce niveau constitue la valeur la plus couramment observée dans l'échantillon.

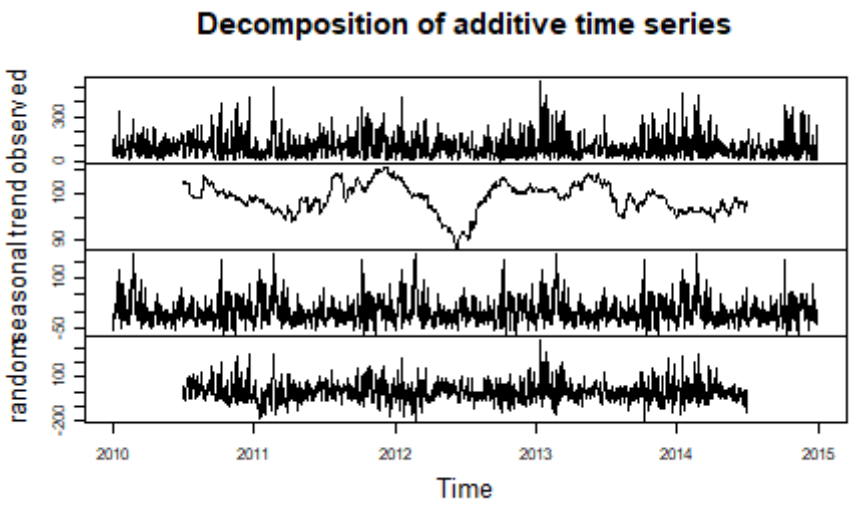
- **Fréquence décroissante pour les valeurs élevées**

À mesure que les valeurs de pollution augmentent au-delà de 200, leur fréquence diminue nettement. Les niveaux supérieurs à 300 sont relativement peu fréquents.

- **Valeurs extrêmes très rares**

Les cas de pollution dépassant les 400 unités sont **exceptionnels**, comme en témoigne la densité quasi nulle dans cette zone. Ces valeurs représentent probablement des **épisodes extrêmes** ponctuels.

B.2) Structure temporelle : tendance et saisonnalité



L'image présente la **décomposition additive** d'une série temporelle en trois composantes distinctes : **la tendance**, **la saisonnalité** et **les résidus**. Cette approche permet de mieux comprendre la structure interne de la série et les facteurs qui influencent son évolution.

Principales observations :

1. **Tendance stable**
   La composante de tendance reste globalement **constante dans le temps**, avec de légères fluctuations autour d'une valeur moyenne proche de zéro. Cela suggère l'absence de dynamique à long terme marquée dans la série.
2. **Saisonnalité bien définie**
   La composante saisonnière est clairement identifiable, avec des **fluctuations régulières et prononcées**. Ce comportement indique une structure cyclique forte, probablement liée à des effets périodiques (par exemple, conditions météorologiques ou activités humaines).
3. **Résidus importants**
   La composante résiduelle présente une **variabilité irrégulière** et de relativement grande amplitude. Cela révèle la présence de **facteurs aléatoires** ou de **phénomènes non modélisés** par la tendance et la saisonnalité.
4. **Bonne séparation des composantes**
   Les trois composantes semblent **bien isolées et synchronisées**, ce qui témoigne de la qualité de la décomposition. Les fluctuations observées dans les résidus n'interfèrent pas avec celles de la tendance ou de la saisonnalité.

B.3) Parametre statistiques

Statistique	Valeur	Interprétation
n	1825	Nombre total d'observations dans la série temporelle
Moyenne (mean)	98.25	Niveau moyen de pollution observé
Écart-type (sd)	76.81	Variabilité des données autour de la moyenne ; indique une dispersion élevée
Médiane	79.17	Valeur centrale : la moitié des observations sont en dessous
Moyenne tronquée (trimmed)	86.79	Moyenne recalculée après exclusion des extrêmes (robuste aux outliers)
MAD (écart médian absolu)	62.70	Mesure robuste de la dispersion, moins sensible aux valeurs extrêmes
Minimum	3.17	Valeur la plus basse de pollution enregistrée
Maximum	541.90	Valeur la plus élevée, indiquant un épisode de forte pollution
Étendue (range)	538.73	Différence entre les valeurs extrêmes ; reflète l'amplitude des données
Asymétrie (skew)	1.62	Distribution asymétrique à droite : les valeurs élevées sont plus fréquentes que les faibles
Aplatissement (kurtosis)	3.44	Distribution plus pointue que la normale, avec des valeurs extrêmes fréquentes
Erreur standard (se)	1.80	Précision de l'estimation de la moyenne ; plus la valeur est basse, plus la moyenne est fiable

B.4) indice de dépendances

- Autocorrelation simple

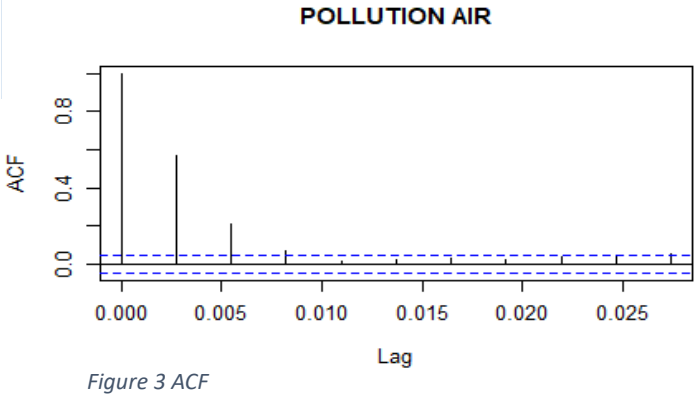


Figure 3 ACF

```
Autocorrelations of series 'Neri_ts', by lag

0.00000 0.00274 0.00548 0.00822 0.01096
0.01370 0.01644 0.01918 0.02192
  1.000   0.569   0.212   0.069   0.012
0.022   0.026   0.025   0.038
0.02466 0.02740
  0.045   0.055
```

❖ Autocorrelation partielle

```
Partial autocorrelations of series 'Neri_ts', by
lag

0.00274 0.00548 0.00822 0.01096 0.01370 0.01644
0.01918 0.02192 0.02466
  0.569  -0.164   0.033  -0.020   0.045  -0.007
0.012   0.028   0.015
0.02740
  0.028
```

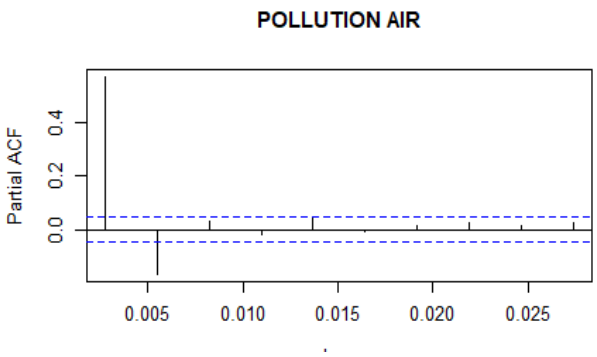


Figure 4 PACF

Interpretation

❖ Analyse de la fonction d'autocorrélation (ACF) de la pollution de l'air

Le graphique présente la **fonction d'autocorrélation (ACF)** appliquée à la série temporelle des niveaux de pollution de l'air, en fonction des décalages temporels (**lags**). Cette analyse permet d'évaluer la dépendance des observations entre elles à différents intervalles de temps.

Principales observations :

Autocorrélation initiale forte
À **lag 0**, la valeur de l'ACF est élevée (environ **0,8**), ce qui indique une **forte dépendance à court terme** : les valeurs de pollution sont fortement corrélées avec les observations immédiatement précédentes.

Décroissance rapide de l'ACF
L'ACF diminue rapidement à mesure que le décalage augmente, ce qui suggère que **l'influence des valeurs passées s'estompe vite**. Cela peut indiquer un processus stationnaire ou à mémoire courte.

Présence d'oscillations amorties
On observe des **oscillations autour de zéro**, dont l'amplitude décroît progressivement. Ce comportement est typique de séries présentant une **composante saisonnière ou cyclique**.

Autocorrélation non significative au-delà d'un certain seuil
À partir d'un **lag d'environ 15**, les valeurs de l'ACF entrent dans la **zone de non-significativité statistique** (généralement définie par les bandes bleues du graphe). Cela indique qu'au-delà de ce seuil, **il n'y a plus de corrélation temporelle significative** entre les observations.

Analyse de la fonction d'autocorrélation partielle (PACF) de la pollution de l'air

Le graphique représente la **fonction d'autocorrélation partielle (PACF)** de la série temporelle des niveaux de pollution de l'air en fonction du décalage (**lag**). Contrairement à l'ACF, la PACF mesure la corrélation entre une observation et ses retards, **en éliminant l'effet des lags intermédiaires**. Elle est particulièrement utile pour identifier l'ordre **AR (autoregressif)** dans les modèles de séries temporelles.

Principales observations :

Corrélation partielle initiale élevée
La PACF présente une valeur significative d'environ **0,8 au premier lag**, ce qui révèle une **forte dépendance immédiate** des observations avec leurs valeurs précédentes.

Chute rapide après le premier lag
Après le premier décalage, les valeurs de la PACF **chutent rapidement vers zéro**, indiquant que les corrélations partielles avec des lags plus lointains sont **faibles ou inexistantes**. Cela suggère qu'un modèle **AR(1)** pourrait suffire à modéliser la dépendance linéaire.

Absence de valeurs significatives au-delà du lag 1
Dès **le deuxième lag**, la majorité des valeurs de PACF tombent **dans l'intervalle de confiance**, et sont donc **statistiquement non significatives**. Cela confirme l'absence de structure autoregressive complexe à long terme.

Pas d'oscillations visibles
Contrairement à la fonction ACF, la PACF **ne présente pas d'oscillations** marquées autour de zéro, ce qui indique une **absence de saisonnalité significative** dans la structure autoregressive de la série..

B.5) Test de normalité

– Graphique

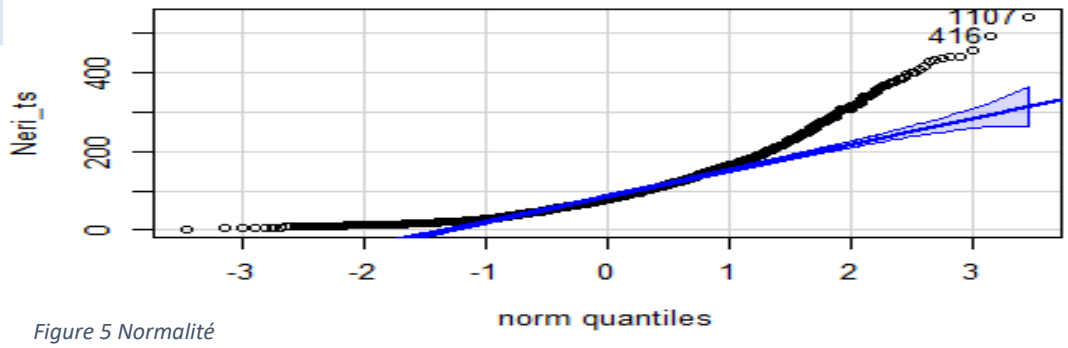


Figure 5 Normalité

```
Shapiro-Wilk normality test

data:  Neri_ts
W = 0.86411, p-value < 2.2e-16
```

H0 : la distribution suit une loi normale

H1 : la distribution ne suit pas une loi normale

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

C) 🔮 Prévision des niveaux de pollution de l'air pour les 30 prochains jours

C.1) Récupération des résidus

```
Time Series:
Start = c(2011, 1)
End = c(2011, 6)
Frequency = 365
[1] -1.982802572 -0.698986060 -0.008380867  0.411525367  0.070415985
[6] -0.023915864
```

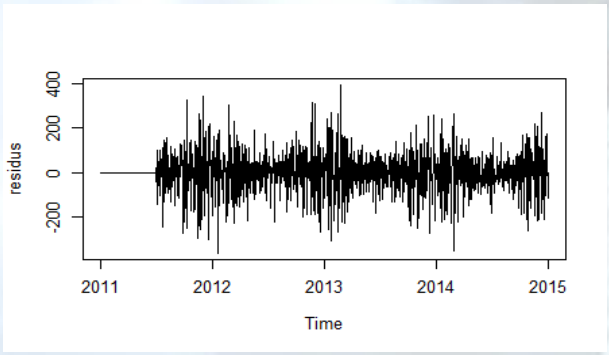
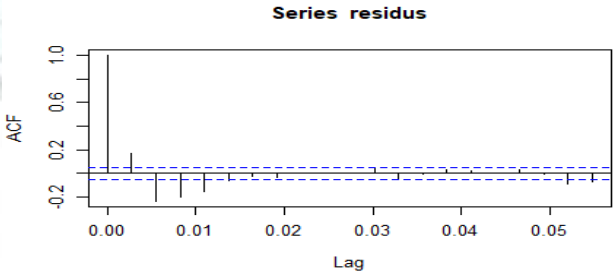


Figure 6 Résidus



- ❖ 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 = 246.28, 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

```
Shapiro-Wilk normality test

data:  residus
W = 0.97518, p-value = 3.555e-15
```

H0 : les résidus suivent une loi normale

H1 : les résidus ne suivent pas une loi normale

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

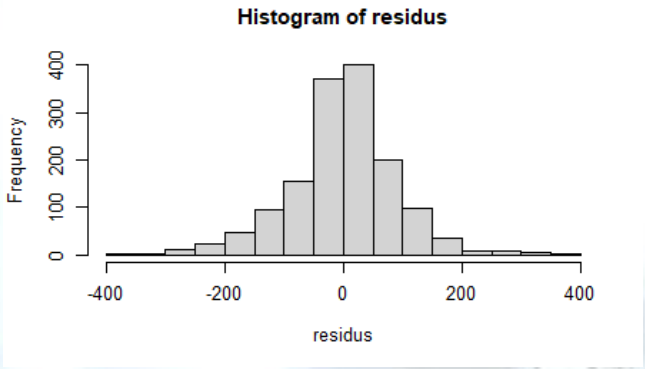


Figure 7 Graphe residus

sont pas des bruits blanc gaussien mais sont centrés

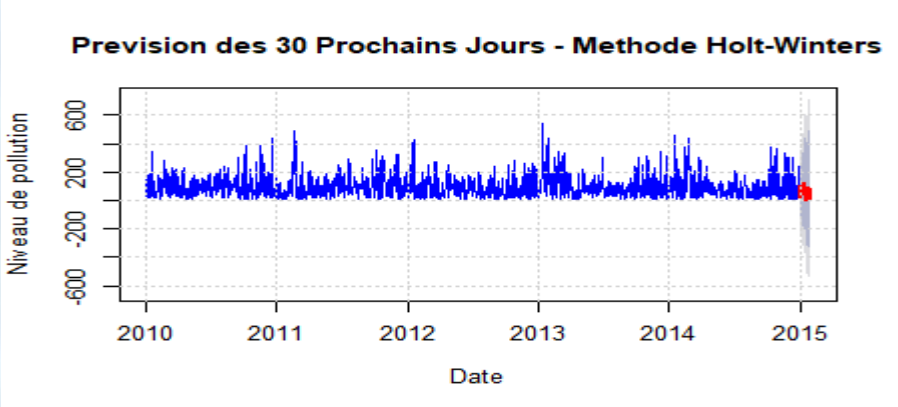


Figure 8 Prévision

PARTIE II : Application de la méthode Box-Jenkins à la modélisation économétrique des séries temporelles

Après avoir appliqué la méthode de Holt-Winters pour capturer la tendance et la saisonnalité de la série, nous passons désormais à une modélisation économétrique plus rigoureuse à travers l'approche Box-Jenkins (ARIMA), afin d'analyser en profondeur la structure dynamique de la série temporelle.

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 for Level Stationarity

data:  Neri_ts
KPSS Level = 0.067898, 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:  Neri_ts
Dickey-Fuller = -10.124, 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:  Neri_ts
Dickey-Fuller Z(alpha) = -703.52, 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

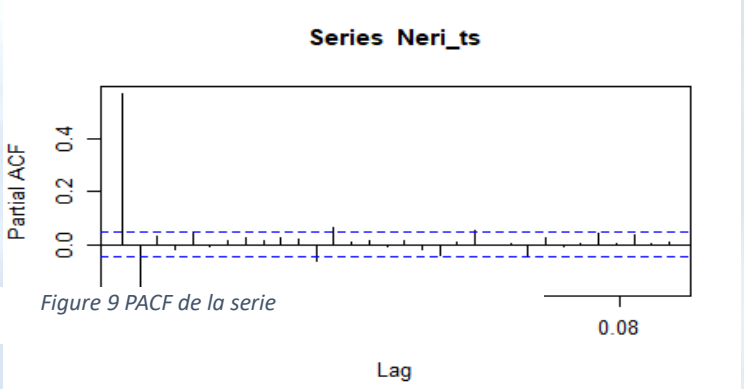
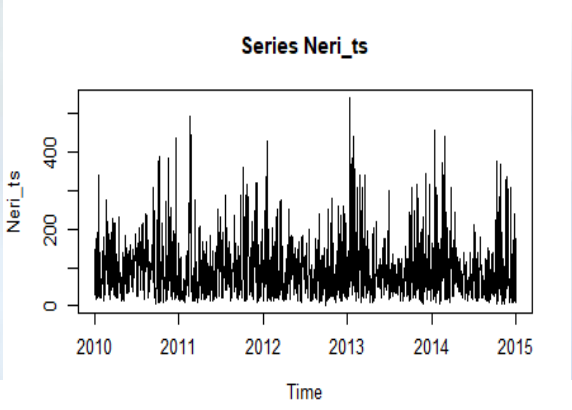
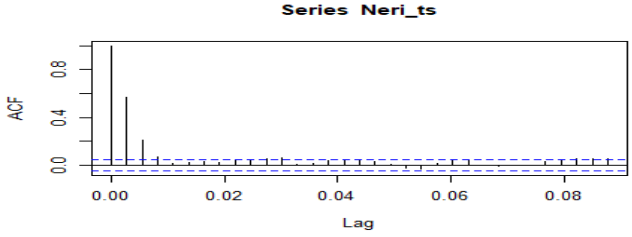


Figure 9 PACF de la serie





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 = Neri_ts, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.5691    98.2408
s.e.  0.0192     3.4271

sigma^2 estimated as 3987:  log likelihood = -10155.01,  aic = 20316.01
```

- Le modèle ARIMA (0,0,3)

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

Coefficients:
         ma1     ma2     ma3  intercept
      0.6706  0.2580  0.0986    98.2498
s.e.  0.0233  0.0276  0.0233     2.9494

sigma^2 estimated as 3867:  log likelihood = -10127.18,  aic = 20264.35
```

- Le modèle ARIMA (1,0,3)

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

Coefficients:
          ar1     ma1     ma2     ma3  intercept
      -0.0276  0.6978  0.2758  0.1039    98.1215
s.e.   0.2451  0.2439  0.1609  0.0530     2.9416

sigma^2 estimated as 3867:  log likelihood = -10127.17,  aic = 20266.33
```

- BILAN DES 3 MODELES

	df <dbl>	AIC <dbl>
mod2	5	20264.35
mod3	6	20266.33
mod1	3	20316.01

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: Neri_ts
ARIMA(1,0,1) with non-zero mean

Coefficients:
         ar1     ma1     mean
      0.3768  0.2924  98.2292
s.e.  0.0364  0.0378   3.0196

sigma^2 = 3879:  log likelihood = -10128.59
AIC=20265.19   AICc=20265.21   BIC=20287.23
```

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 <dbl>	AIC <dbl>
mod2	5	20264.35
mod.auto	4	20265.19
mod1	3	20316.01

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

• Bruit blanc des résidus

```
        Box-Pierce test

data:  res1
X-squared = 16.047, df = 1, p-value = 6.18e-05


        Box-Pierce test

data:  res2
X-squared = 0.00010645, df = 1, p-value = 0.9918


        Box-Pierce test

data:  res_mod.auto
X-squared = 0.00032849, df = 1, p-value = 0.9855
```

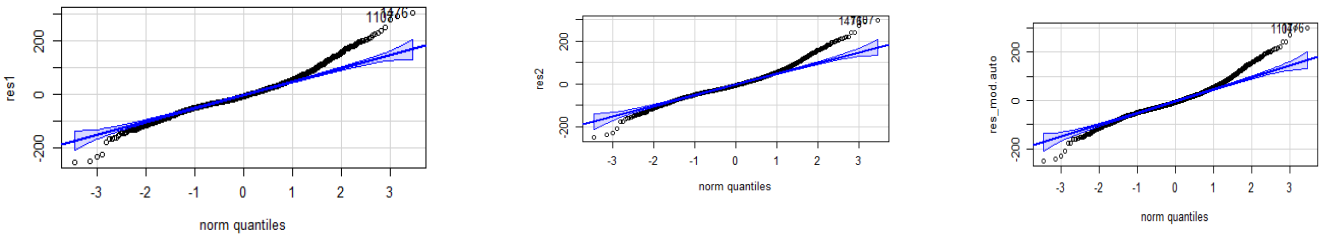


Figure 10 Normalité residuelle

- Normalité des résidus

```
Shapiro-Wilk normality test

data:  res1
W = 0.96445, p-value < 2.2e-16
```

- Centralité des résidus

Indicateurs	Tests	Modèle 1	Modèle 2	Modele au
AIC		20265,65	20266,26	20317,1(
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 nongaussien et noncentré	Les résidus suivent un processus bruit blanc non-gaussien non-centré	Les résidu suivent un processus bruit blan non-gaussi et non-cen

```
[1] -0.03030707
[1] -0.02358461
[1] -0.01472122
```

- VISUALISATIONS DES RESIDUS DU MODELE 2

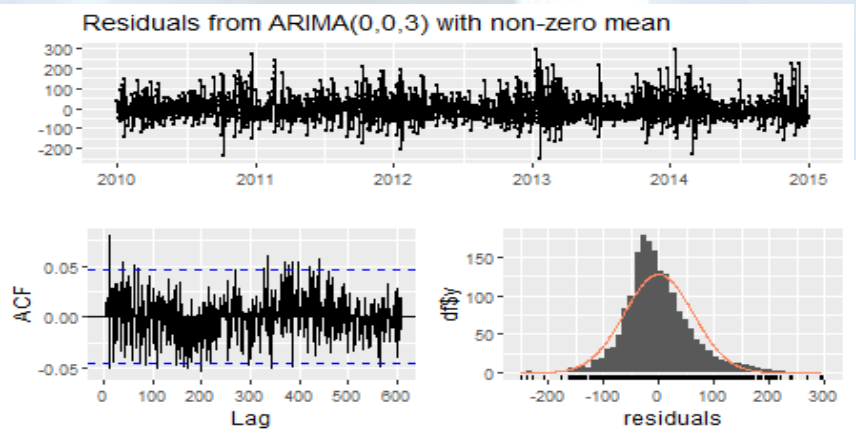


Figure 11 Residual

```
Ljung-Box test

data:  Residuals from ARIMA(0,0,3) with non-zero mean
Q* = 396.91, df = 362, p-value = 0.09982

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
2015.0000	51.26299	-28.4271085	130.9531	-70.61250	173.1385
2015.0027	79.58658	-16.3615408	175.5347	-67.15341	226.3266
2015.0055	93.99278	-4.1327363	192.1183	-56.07725	244.0628
2015.0082	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0110	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0137	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0164	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0192	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0219	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0247	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0274	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0301	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0329	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0356	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0384	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0411	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0438	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0466	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0493	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0521	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0548	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0575	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0603	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0630	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0658	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0685	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0712	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0740	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0767	98.24981	-0.1895393	196.6892	-52.30018	248.7998
2015.0795	98.24981	-0.1895393	196.6892	-52.30018	248.7998

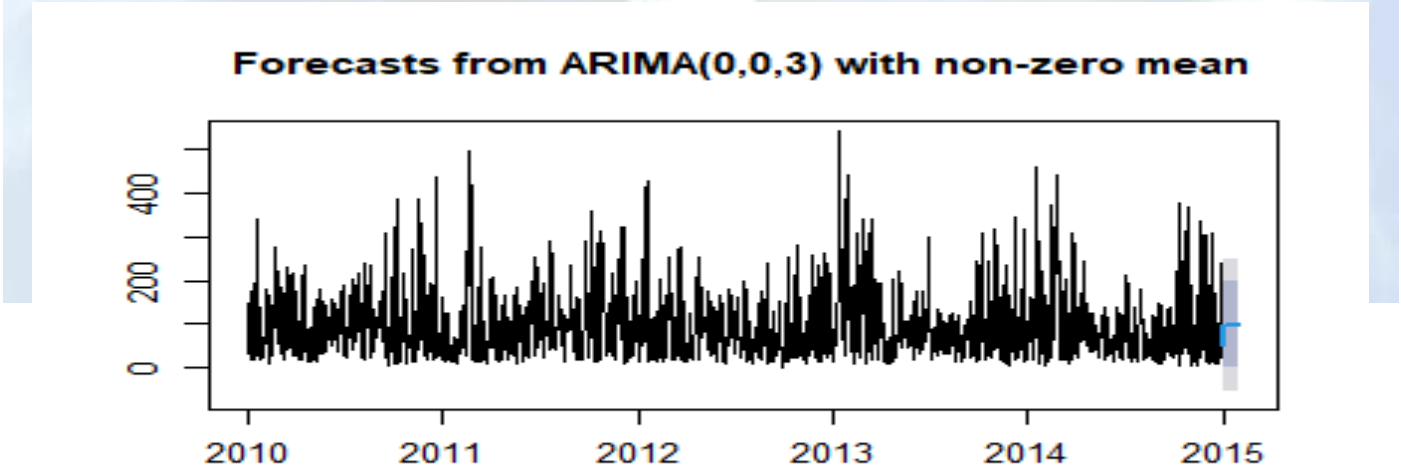
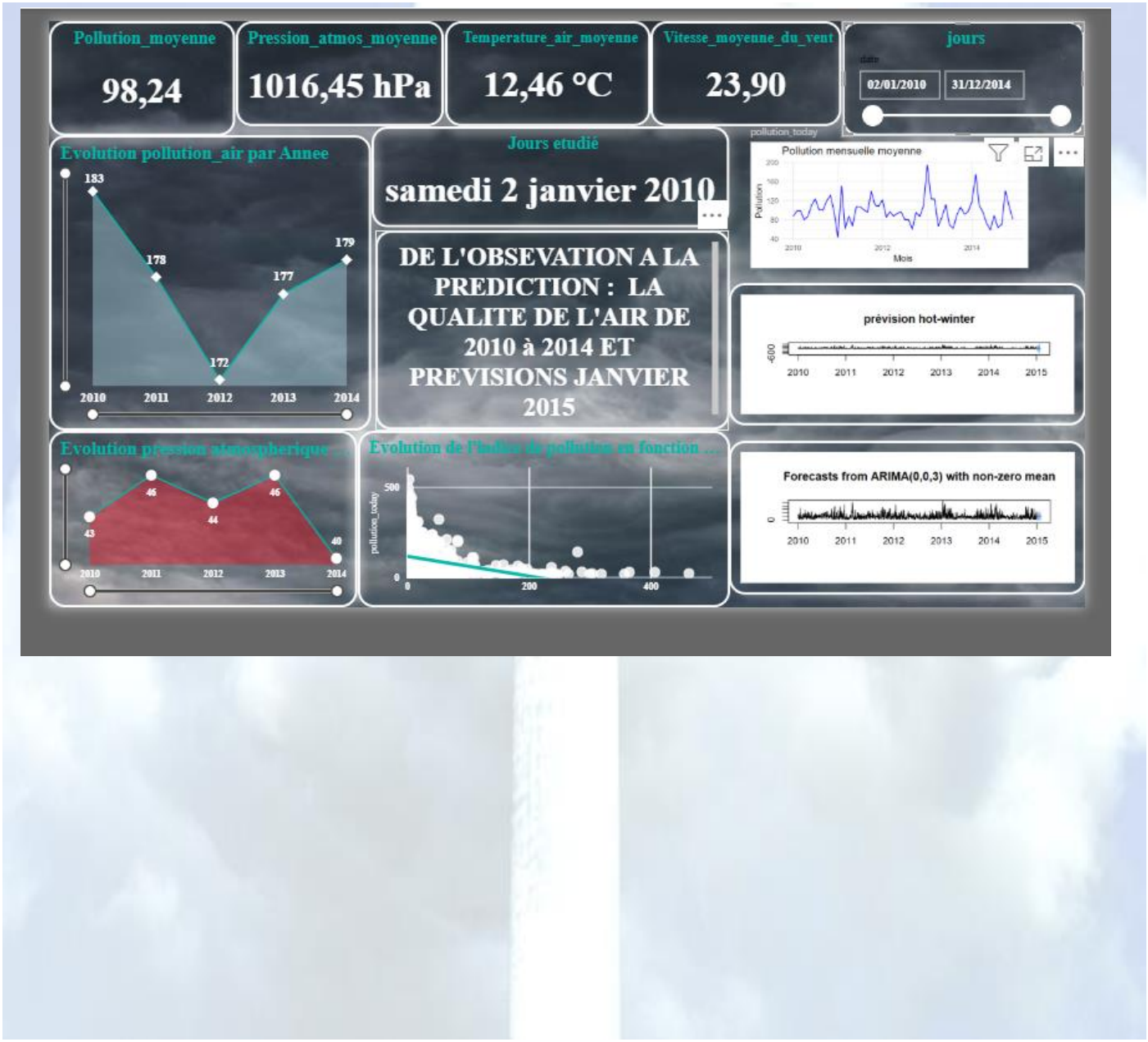


Figure 12 PREDICTION

Dashboard


Pollution_moyenne
98,24
Pression_atmos_moyenne
1016,45 hPa
Temperature_air_moyenne
12,46 °C
Vitesse_moyenne_du_vent
23,90
jours
02/01/2010
31/12/2014
Evolution pollution_air par Annee
183
178
172
177
179
2010
2011
2012
2013
2014
Jours etudié
samedi 2 janvier 2010
DE L'OBSEVATION A LA PREDICTION : LA QUALITE DE L'AIR DE 2010 à 2014 ET PREVISIONS JANVIER 2015
Pollution mensuelle moyenne
Pollution
Mois
prévision hot-winter
Evolution pression atmospherique
Evolution de l'indice de pollution en fonction
pollution_today
Forecasts from ARIMA(0,0,3) with non-zero mean

Conclusion

Rappel de la problématique

Cette étude visait à répondre à une question centrale : **dans quelle mesure les variables météorologiques et les données de pollution de la veille permettent-elles de prédire les niveaux quotidiens de pollution de l'air à Pékin** ? Face à la complexité des dynamiques atmosphériques et à l'impact croissant de la pollution sur la santé publique, il était essentiel d'évaluer la pertinence de ces facteurs dans la modélisation de la qualité de l'air.

Résumé des principaux résultats obtenus

L'analyse statistique a révélé plusieurs résultats significatifs :

- **La température, la pression atmosphérique et la vitesse du vent** apparaissent comme des variables explicatives importantes, influençant notablement les niveaux de pollution.
- La **pollution mesurée la veille** s'est avérée être un prédicteur fort, soulignant l'inertie et la persistance du phénomène d'un jour à l'autre.
- Les **modèles de régression linéaire** ont permis d'expliquer une partie significative de la variance des niveaux de pollution, tandis que les modèles plus complexes (comme les forêts aléatoires) ont montré de meilleures performances prédictives, au prix d'une interprétabilité moindre.

Recommandations

Sur la base des résultats obtenus, plusieurs recommandations peuvent être formulées :

- **Renforcer la surveillance météorologique et environnementale** pour améliorer la précision des prédictions à court terme.
- **Utiliser la modélisation prédictive dans les systèmes d'alerte** afin de prévenir les populations des épisodes de pollution à venir.
- Intégrer ces modèles dans des **politiques publiques de gestion de la qualité de l'air**, notamment en adaptant certaines activités urbaines ou industrielles lors de conditions favorables à l'accumulation des polluants.

Limites de l'analyse

Cette étude présente toutefois certaines limites :

- Le dataset ne précise pas **l'unité exacte de mesure de la pollution**, ce qui limite l'interprétation physique des résultats.
- Les **données sont consolidées à l'échelle journalière**, ce qui empêche une analyse fine des variations intra-journalières.
- Certains **facteurs externes non pris en compte**, comme le trafic routier, les émissions industrielles ou les politiques environnementales, pourraient jouer un rôle significatif.

Perspectives et autres analyses pouvant améliorer les résultats

Pour affiner les prédictions et approfondir la compréhension du phénomène, plusieurs pistes sont envisageables :

- **Intégrer des données exogènes supplémentaires**, comme les données de trafic, les émissions industrielles, ou la topographie de la ville.
- Explorer des **modèles temporels avancés** tels que les réseaux de neurones récurrents (RNN) ou les modèles ARIMA pour capturer les dynamiques temporelles plus fines.
- Étendre l'analyse à **d'autres villes ou régions** pour comparer les effets contextuels et climatiques sur la pollution.
- Réaliser une **analyse saisonnière approfondie**, afin de détecter des variations spécifiques à certaines périodes de l'année (hiver, mousson, etc.).

Sources du code R

```
# I- Pretraitement des données

## Visualisation des données

library(readxl)

library(readr)

data <- read_csv("C:/Users/yoboh/OneDrive/Bureau/INSSEDS/MASTER 1&2/MINI
  PROJET/PROJETS/6_ECONOMETRIE/AIR_POLLUTION/Data BASE/air_pollution.csv")

print(data)

str(data)

summary(data)


# Valeurs manquantes

library(visdat)

vis_dat(data)

vis_miss(data)

sum(is.na(data))


# Valeurs aberrante

boxplot(data,

      main = "Distribution de la variable",

      ylab = "Valeurs",

      col = "lightblue",

      border = "darkblue")


# II) ANALYSE DESCRIPTIVE ET PREVISIONS HOTWINTER

## a) Construction de la serie temporelle

library(dplyr)

Neri <- data %>%

select(date,pollution_today)

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

```
##Créer une série temporelle de la colonne 'airpollution'
Neri_ts <- ts(Neri$pollution_today, start = c(2010,01), frequency = 365)
head(Neri_ts)
print(Neri_ts)

## b) Graphiques
###Visualiser la série temporelle
plot(Neri_ts, col = "black", main = "SERIE TEMPORELLE", xlab
    = "Temps", ylab ="Pollution de l'air")


### Regroupement par mois
library(dplyr)
library(lubridate)

Neri_mensuel <- Neri %>%
 mutate(mois = floor_date(date, "month")) %>%   # crée une colonne 'mois'
 group_by(mois) %>%
 summarise(pollution_moyenne = mean(pollution_today, na.rm = TRUE)) %>%
 ungroup()
print(Neri_mensuel)

library(ggplot2)
ggplot(Neri_mensuel, aes(x = mois, y = pollution_moyenne)) +
 geom_line(color = "blue") +
 labs(title = "Pollution mensuelle moyenne", x = "Mois", y = "Pollution") +
 theme_minimal()

##############################################################################################"##
 ######
library(dplyr)
library(lubridate)
library(ggplot2)
```

```
library(zoo)

# Calcul de la moyenne mobile et des bandes (avec une fenêtre de 3 mois, par exemple)
Neri_mensuel3 <- Neri %>%
  mutate(mois = floor_date(date, "month")) %>%
  group_by(mois) %>%
  summarise(pollution_moyenne = mean(pollution_today, na.rm = TRUE)) %>%
  arrange(mois) %>%
  mutate(
    moyenne_mobile = rollmean(pollution_moyenne, k = 3, fill = NA, align = "right"),
    ecart_type = rollapply(pollution_moyenne, width = 3, FUN = sd, fill = NA, align = "right"),
    bande_sup = moyenne_mobile + ecart_type,
    bande_inf = moyenne_mobile - ecart_type
  )

# Graphique avec ggplot2
ggplot(Neri_mensuel, aes(x = mois)) +
  geom_line(aes(y = pollution_moyenne), color = "blue", alpha = 0.4) +  # ligne pollution réelle
  geom_line(aes(y = moyenne_mobile), color = "darkgreen") +           # moyenne mobile
  geom_ribbon(aes(ymin = bande_inf, ymax = bande_sup), fill = "lightgreen", alpha = 0.3) +  # bandes
  labs(title = "Pollution mensuelle avec bandes type trading",
       x = "Mois", y = "Pollution") +
  theme_minimal()
##########################################################################################"

# Charger les bibliothèques
library(dplyr)
library(lubridate)
library(plotly)

# Exemple de données (remplace ceci par ton propre import si tu lis depuis un fichier)
# Neri <- read.csv("chemin/vers/ton_fichier.csv")
```

```
# ou tu peux avoir déjà Neri chargé en mémoire comme dans ton exemple


# Étape 1 : Créer les composantes Open, High, Low, Close par mois
Neri_candle <- Neri %>%
 mutate(mois = floor_date(date, "month")) %>%
 group_by(mois) %>%
 summarise(
  Open = first(pollution_today),
  High = max(pollution_today, na.rm = TRUE),
  Low = min(pollution_today, na.rm = TRUE),
  Close = last(pollution_today)
 ) %>%
 ungroup()

# Étape 2 : Créer le graphique en chandelier avec plotly
fig <- plot_ly(data = Neri_candle, x = ~mois, type = "candlestick",
          open = ~Open, close = ~Close,
          high = ~High, low = ~Low) %>%
 layout(title = "Pollution mensuelle - Chandelier japonais",
       xaxis = list(title = "Mois"),
       yaxis = list(title = "Niveau de pollution"))

# Afficher le graphique interactif
fig

##############################################################################


# Histogramme avec courbe de densite
hist(Neri_ts,
    main = "Histogramme de la pollution",  # Titre plus descriptif
```

```
    xlab = "Valeur de pollution",        # Ajout d'un label pour l'axe x
    ylab = "Densite",                    # Ajout d'un label pour l'axe y
    prob = TRUE,                         # Pour avoir une densite (pas des frequences brutes)
    col = "lightblue",                   # Couleur des barres
    border = "white",                    # Bordure blanche
    lwd = 2)                             # Largeur de ligne des bordures de barres


# Ajouter la courbe de densité
lines(density(Neri_ts, na.rm = TRUE),
     col = "darkblue",
     lwd = 2)


#c) Tendance et composante saisonnière
decomposition_add=decompose(Neri_ts, type = "add")
plot(decomposition_add)


# d) Indice statistique
library(psych)
describe(Neri_ts)


### Autocorrélation simple
acf(Neri_ts,lag.max=10,plot = FALSE, main="POLLUTION AIR")
acf(Neri_ts,lag.max=10,plot = TRUE, main="POLLUTION AIR")


### Autocorrélation partielle
pacf(Neri_ts,lag.max=10,plot = FALSE, main="POLLUTION AIR")
pacf(Neri_ts,lag.max=10,plot = TRUE, main="POLLUTION AIR")


# e) Test de normalité Graphique
library(car)
qqPlot(Neri_ts)
### Test
```

```
shapiro.test(Neri_ts)


#f) Prévision des indices d'air de pollution pour les 30 prochains jours
### Validation du modèle de prévision


### Récupération des résidus
xlisse <- HoltWinters(Neri_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(Neri_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 = data)
print(test_result)

#A) IDENTIFICATION
##A-1) Vérification de la stationnarité de la série - kpss
kpss.test(Neri_ts)

##- adf
adf.test(Neri_ts)

##- pp
pp.test(Neri_ts)

# A-2) Détermination des combinaisons d'auto régression(p) et de moyenne mobile (q)
###Graphiques
plot(Neri_ts,main="Series Neri_ts")
acf(Neri_ts)
pacf(Neri_ts)

#B) ESTIMATION
##Estimation des modèles par la fonction arima le modèle ARIMA(1,0,0)
mod1 <- arima (Neri_ts, order=c(1,0,0))
mod1

## le modèle ARIMA(0,0,3)
```

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

## le modèle ARIMA(1,0,3)
mod3 <- arima (Neri_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(Neri_ts)

mod.auto<-arima(Neri_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(Neri_ts, order=c(0,0,3))
prediction <- forecast(mod2,h=30) # pour les 30 prochains jours

#prediction
plot(prediction)
```