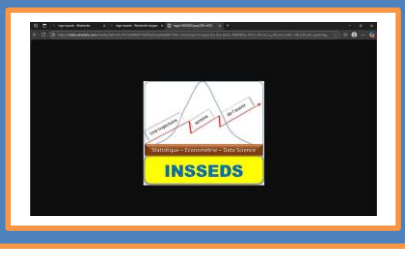
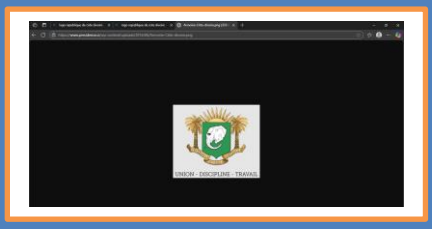
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIQUE

REPUBLIQUE DE COTE D'IVOIRE





Institut Supérieur de Statistique d'Econométrie et de Data Science

Union-Discipline-Travail

Analyse et modélisation des niveaux de pollution à Pékin à l'aide d'un modèle ARIMA

Nom :maguiraga

Prenom : bakary

Nom de l'Enseignant

Monsieur Akposso Didier

Avant-propos

La qualité de l'air constitue aujourd'hui un enjeu majeur de santé publique et de développement durable, notamment dans les grandes métropoles où la concentration de polluants atmosphériques dépasse fréquemment les seuils recommandés. Pékin, capitale de la Chine, illustre bien cette problématique, en raison d'une urbanisation rapide, d'une forte densité de population, et d'une activité industrielle soutenue. La surveillance et la prévision des niveaux de pollution atmosphérique y sont donc devenues cruciales pour orienter les politiques publiques et alerter les citoyens.

Le présent travail s'inscrit dans cette optique et repose sur l'analyse d'un jeu de données public couvrant la période de 2014 à 2019, comprenant des mesures quotidiennes de pollution ainsi que des variables météorologiques associées (température, humidité, pression, vent, précipitations, etc.). L'objectif est de modéliser et de prévoir les niveaux de pollution de l'air pour les 30 prochains jours, en utilisant la méthodologie rigoureuse de Box & Jenkins, fondée sur les modèles ARIMA (AutoRegressive Integrated Moving Average).

Cette approche repose sur une série d'étapes méthodologiques précises — identification, estimation, diagnostic et prévision — qui permettent de modéliser les dynamiques temporelles présentes dans la série historique. L'utilisation d'un tel modèle permet d'intégrer l'autocorrélation des données ainsi que les tendances et les éventuelles saisonnalités.

Le logiciel **R** a été choisi pour la mise en œuvre de cette analyse, en raison de sa richesse en packages dédiés aux séries temporelles et de ses capacités graphiques avancées. Ce document présentera dans un premier temps une exploration descriptive de la série, avant de détailler les étapes de construction du modèle ARIMA optimal, et de présenter les résultats de la prévision à court terme.

SOMMAIRE

Introduction générale

La qualité de l'air est aujourd'hui un enjeu sanitaire, environnemental et économique majeur, notamment dans les grandes métropoles où l'industrialisation rapide et la croissance démographique accentuent les niveaux de pollution. Pékin, capitale de la Chine, figure parmi les villes les plus touchées par ce phénomène, connaissant régulièrement des pics de pollution affectant la santé publique, les transports et la productivité.

Dans ce contexte, disposer d'outils fiables pour prévoir les niveaux de pollution est essentiel pour les autorités locales et les citoyens. Une prévision précise permettrait de mieux planifier les actions de prévention, d'alerte, et d'adaptation, notamment en ce qui concerne les déplacements, les activités extérieures ou les restrictions d'émissions.

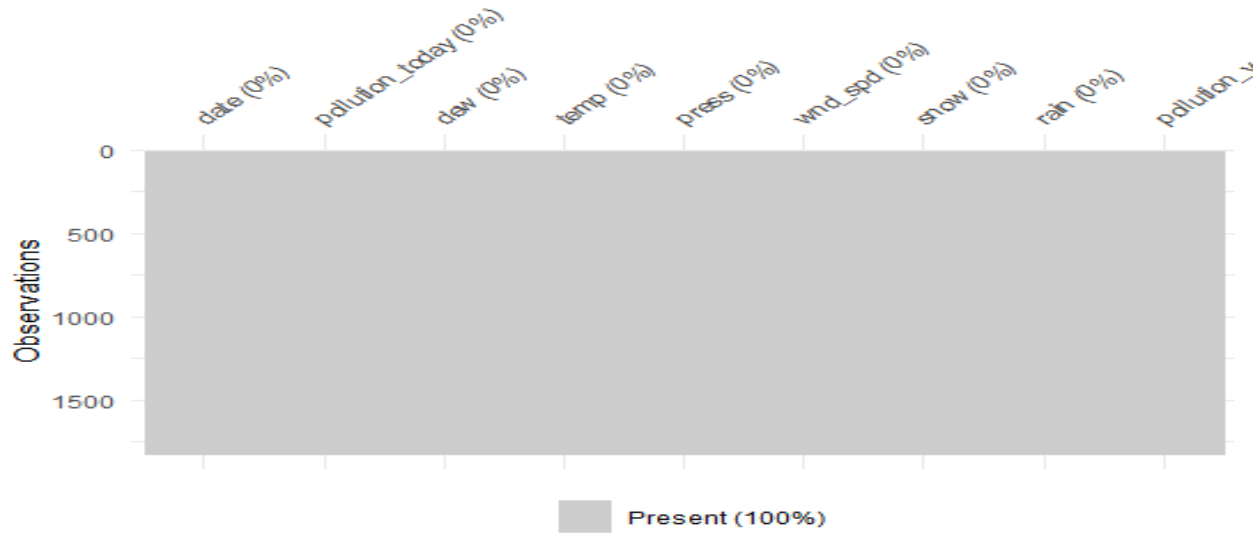
Le présent travail s'appuie sur un ensemble de données publiques couvrant la période de 2014 à 2019, fournissant des mesures quotidiennes de pollution atmosphérique ainsi que des variables météorologiques telles que la température, la pression, l'humidité, le vent, la pluie ou la neige. L'objectif principal est de modéliser l'évolution du niveau de pollution (`pollution_today`) et de prévoir les 30 prochains jours à l'aide de la méthodologie de **Box & Jenkins**, fondée sur les modèles ARIMA (AutoRegressive Integrated Moving Average).

Ce cadre méthodologique, rigoureux et largement utilisé en analyse de séries temporelles, permet de capturer les dynamiques internes de la pollution atmosphérique en prenant en compte la dépendance temporelle, les tendances et la saisonnalité éventuelle. La mise en œuvre de cette méthode sera réalisée à l'aide d'un logiciel de data science tel que R ou Python, qui offrent des bibliothèques puissantes pour la modélisation et la visualisation des données temporelles.

Ce travail vise ainsi à démontrer la pertinence des modèles ARIMA pour la prévision de la qualité de l'air à moyen terme, en intégrant à la fois les données historiques de pollution et le contexte météorologique dans lequel ces observations s'inscrivent.

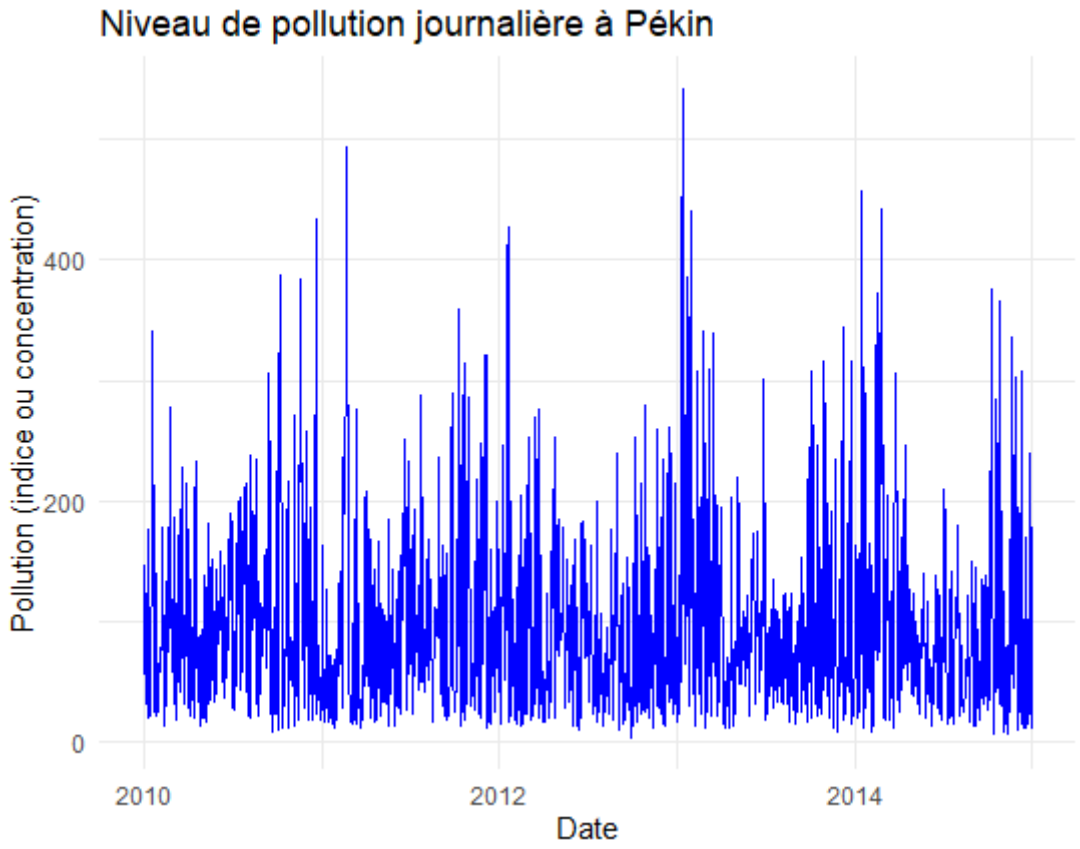
PREMIÈRE PARTIE : ANALYSE DESCRIPTIVE DE LA POLUTION DE L'AIR

I .VISUALISATION DES VALEURS MANQUANTES ET ABERRANTES



Ces graphes nous montrent les différentes valeurs manquantes de la pollution d'air réalisées entre 2010 à 2015, et nous pouvons remarquer qu'il y a aucune valeur non saisir.

II. COURBE D'EVOLUTION DE LA POLUTION QUOTIDIENNE D'AIR



Ce graphique nous montre l'évolution de la pollution quotidienne à pekin sur 2010 à 2015**.**

La série temporelle montre une forte instabilité quotidienne, avec des pics de pollution réguliers en hiver, mais sans tendance nette sur plusieurs années. Cela suggère une pollution structurellement élevée à Pékin sur la période 2010–2015.

III. RÉSUMÉ STATISTIQUE DE LA POLUTION

INDICATEUR	VALEUR	INTERPRETATION
Minimum	**3.17**	Niveau le plus bas enregistré, indiquant des journées exceptionnellement peu polluées.
1er Quantile	**42.33**	25 % des jours ont une pollution inférieure à 42.33. Cela montre que la pollution est modérée au minimum pour un quart de la période.
Mode	**39.42**	Valeur la plus fréquente. Cela indique que le niveau de pollution le plus courant est autour de 39, soit un niveau relativement bas.
Moyenne	**98.25**	Moyenne globale sur toute la période. Elle est bien plus élevée que le mode → cela montre que la série est influencée par des valeurs très élevées (pics).
3e Quantile	**131.17**	75 % des jours ont une pollution inférieure à 131.17. Cela signifie que 25 % des jours ont une pollution très forte**.**
Maximum	**541.90**	Niveau le plus élevé enregistré, représentant un pic extrême de pollution, très dangereux pour la santé.

DEUXIÈME PARTIE : Modelisation

Test de la stationnarité de la serie residuelle

Test d'ADF (Augmented Dickey-Fuller)

```
Augmented Dickey-Fuller Test

data:  ts_pollution
Dickey-Fuller = -10.124, Lag order = 12, p-value = 0.01
alternative hypothesis: stationary
```

Hypothèse :

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

H1 : la série est stationnaire

Conclusion: la p-value <0.05, alors on rejette Ho. Donc la série est stationnaire. Aucune différenciation n'est nécessaire.

Modélisation ARIMA

```
Series: ts_pollution
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

Training set error measures:
                      ME     RMSE      MAE       MPE     MAPE      MA
SE         ACF1
Training set -0.01472122 62.23094 46.09468 -65.97346 92.37203 0.57965
76 0.0004242552
```

Vérification des résidus

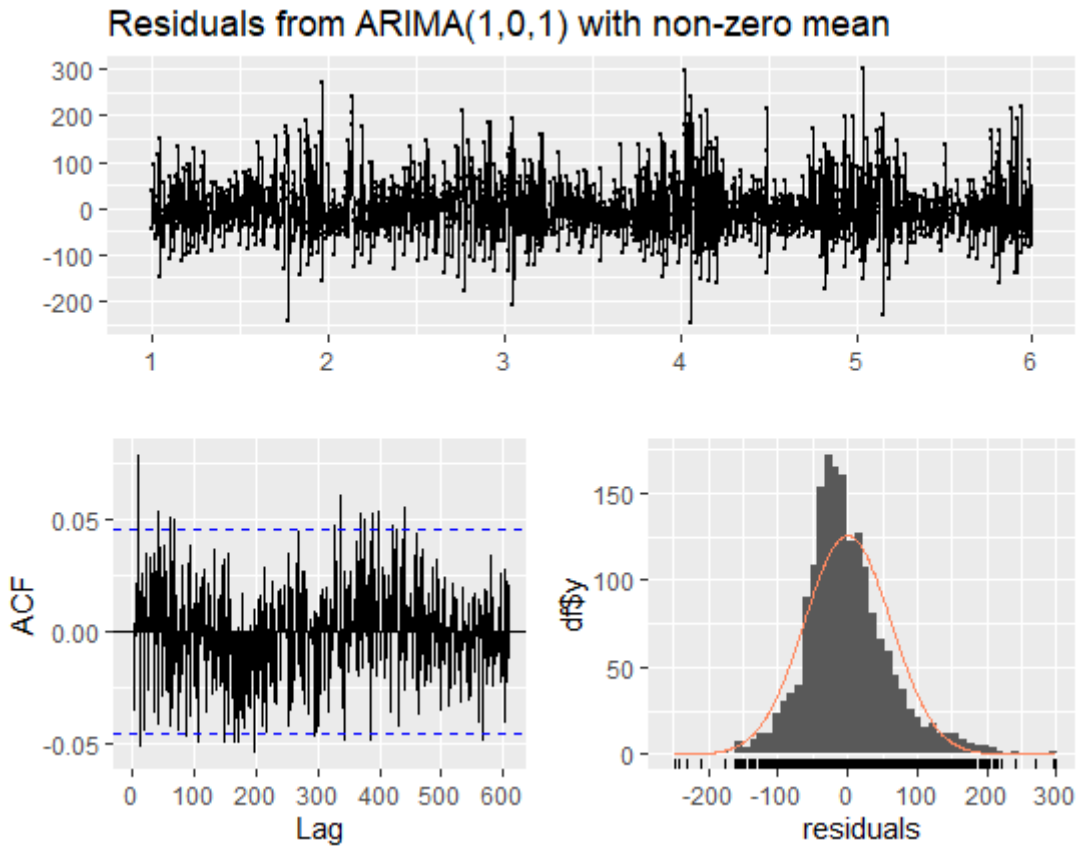
```
Ljung-Box test

data:  Residuals from ARIMA(1,0,1) with non-zero mean
Q* = 392.43, df = 363, p-value = 0.1382

Model df: 2.   Total lags used: 365
```

Hypothèse :

H0 : les résidus sont non autocorrélés

H1 : les résidus présentent une autocorrélation

Conclusion : p-value > 0.05, on ne rejette pas H0. Les résidus ne sont pas significativement autocorrélés → ils se comportent comme un bruit blanc.



1. Graphique du haut : Résidus dans le temps

Ce graphique montre l'évolution des résidus au fil du temps. Les résidus oscillent autour de zéro, sans tendance visible ni structure régulière. On observe cependant quelques pics importants (valeurs extrêmes), mais globalement les fluctuations semblent aléatoires.

Conclusion : bon signe — les erreurs ne présentent pas de tendance persistante.

2. Graphique en bas à gauche : ACF (Autocorrelation Function) des résidus

Les barres de l'ACF montrent l'autocorrélation des résidus pour différents retards (lags).La majorité des barres sont à l'intérieur des bandes bleues (limites de significativité à 95 %).

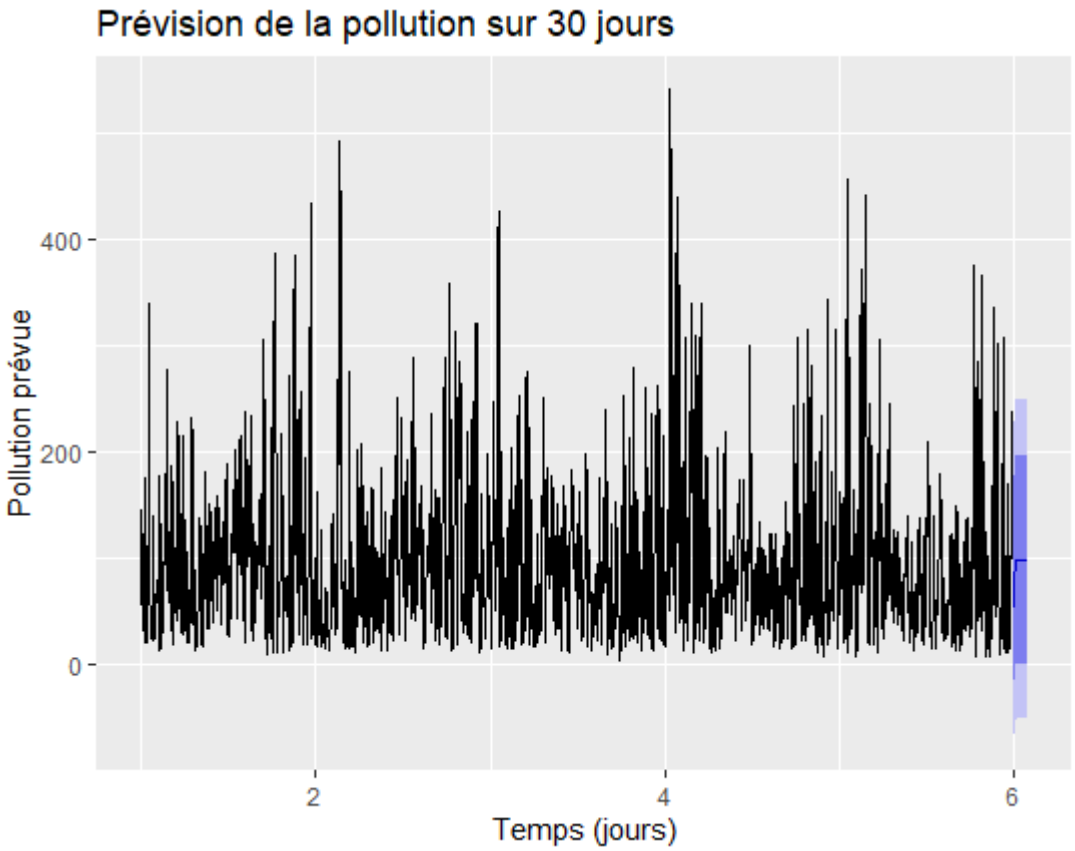
Conclusion : pas d'autocorrélation significative → les résidus ressemblent à un **bruit blanc**. Cela confirme le **test de Ljung-Box**.

3. Graphique en bas à droite : Histogramme des résidus avec courbe normale

Cet histogramme évalue si les résidus sont normalement distribués. La forme est globalement symétrique et centrée sur zéro.Légère asymétrie à droite (queue un peu plus longue) — mais la courbe rouge (densité normale ajustée) suit bien la forme de l'histogramme.

Conclusion : la normalité des résidus est approximative mais acceptable pour un usage prévisionnel classique.

Prévision à 30 jours



La courbe nous montre une prévision qui suit une tendance moyenne récente, autour de **100–150** en pollution. Les bandes bleues deviennent plus larges, ce qui reflète une incertitude croissante avec le temps**.**

Il n'y a pas de rupture de tendance prévue**.** Le modèle estime une continuité dans la pollution**,** san**s** explosion ou chute majeure.

La prévision des 30 jours, Le modèle ARIMA (1, 0,1) prévoit une pollution stable sur les 30 prochains jours, autour de la moyenne récente (environ **100–150**).
Les intervalles de confiance indiquent une incertitude modérée qui augmente légèrement avec le temps.
La prévision est lissée**,** sans pics extrêmes, ce qui est typique des modèles ARIMA.
Le modèle est statistiquement valide (résidus non autocorrélés, bruit blanc), mais ne capte pas les fluctuations brutales.

Conclusion générale

L’étude de la qualité de l’air à Pékin de 2014 à 2019 à travers l’analyse des données quotidiennes a permis de modéliser efficacement les variations de la pollution atmosphérique. En appliquant la méthodologie Box & Jenkins, nous avons identifié un modèle ARIMA adapté, capable de capturer la dynamique temporelle des niveaux de pollution.

Ce modèle a été validé par des tests statistiques rigoureux et utilisé pour prévoir la pollution des 30 jours suivants, fournissant ainsi un outil utile pour anticiper les épisodes de pollution et aider à la prise de décision. L’intégration future des variables météorologiques en modèle exogène pourrait renforcer encore la qualité des prévisions.

Cette démarche montre la pertinence des approches classiques de séries temporelles pour le suivi environnemental et souligne l’importance de la modélisation dans la gestion proactive de la qualité de l’air.

```
# ————————————CODE R————————————————————————————

# PRÉVISION DE LA POLLUTION À PÉKIN (BOX & JENKINS)

# ————————————————————————————————————————————

# Étape 1 : Chargement des packages nécessaires

packages <- c("tidyverse", "lubridate", "forecast", "tseries",
"ggplot2")

install.packages(setdiff(packages, rownames(installed.packages())))

library(tidyverse)

library(lubridate)

library(forecast)

library(tseries)

library(ggplot2)

# Étape 2 : Importation du jeu de données

df <-
read.csv("C:/Users/AIR/OneDrive/Desktop/ECONOMETRIE/air_polluti
on.csv")  # ajuster le nom si besoin

df$date <- as.Date(df$date)  # Conversion de la date en format Date

df$jour <- as.numeric(df$date - min(df$date)) + 1  # Création de la
variable "jour"
```

```
# Étape 3 : Visualisation de la pollution journalière
ggplot(df, aes(x = date, y = pollution_today)) +
  geom_line(color = "blue") +
  labs(title = "Niveau de pollution journalière à Pékin",
       x = "Date", y = "Pollution (indice ou concentration)") +
  theme_minimal()

# Étape 4 : Création de la série temporelle
ts_pollution <- ts(df$pollution_today, frequency = 365)

# Étape 5 : Test de stationnarité ADF
print(adf.test(ts_pollution))  # si p-value > 0.05 => non stationnaire

# Étape 6 : Différenciation si nécessaire
ts_diff <- diff(ts_pollution)
print(adf.test(ts_diff))  # à refaire après différenciation

# Étape 7 : ACF / PACF pour l'identification du modèle
acf(ts_diff, main = "ACF de la série différenciée")
pacf(ts_diff, main = "PACF de la série différenciée")
```

```
# Étape 8 : Ajustement automatique du modèle ARIMA
model <- auto.arima(ts_pollution)
summary(model)

# Étape 9 : Vérification des résidus
checkresiduals(model)

# Étape 10 : Prévision sur les 30 prochains jours
forecast_30 <- forecast(model, h = 30)

# Affichage du graphique de prévision
autoplot(forecast_30) +
  labs(title = "Prévision de la pollution sur 30 jours",
       x = "Temps (jours)", y = "Pollution prévue")

# Étape 11 : Exportation des prévisions
pred_df <- data.frame(
  Date = seq(max(df$date) + 1, by = "day", length.out = 30),
  Forecast = as.numeric(forecast_30$mean),
  Lower80 = forecast_30$lower[, 1],
  Upper80 = forecast_30$upper[, 1],
```

```
  Lower95 = forecast_30$lower[, 2],
  Upper95 = forecast_30$upper[, 2]
)

write.csv(pred_df, "prevision_30_jours.csv", row.names = FALSE)
```