

10/01/2025

RAPPORT SAE-SÉRIE TEMPORELLE



SOMMAIRE

01	INTRODUCTION
02	ANALYSE STATISTIQUE
03-06	TENDANCE DE LA SERIE
07	VARIATIONS SAISONNIERES
08	SERIE SAISONNIERE AJUSTEE
09	LES RESIDUS
10-14	PREVISION 2024
15	CONCLUSION
16	SUMMARY (RESUME EN ANGLAIS)

INTRODUCTION



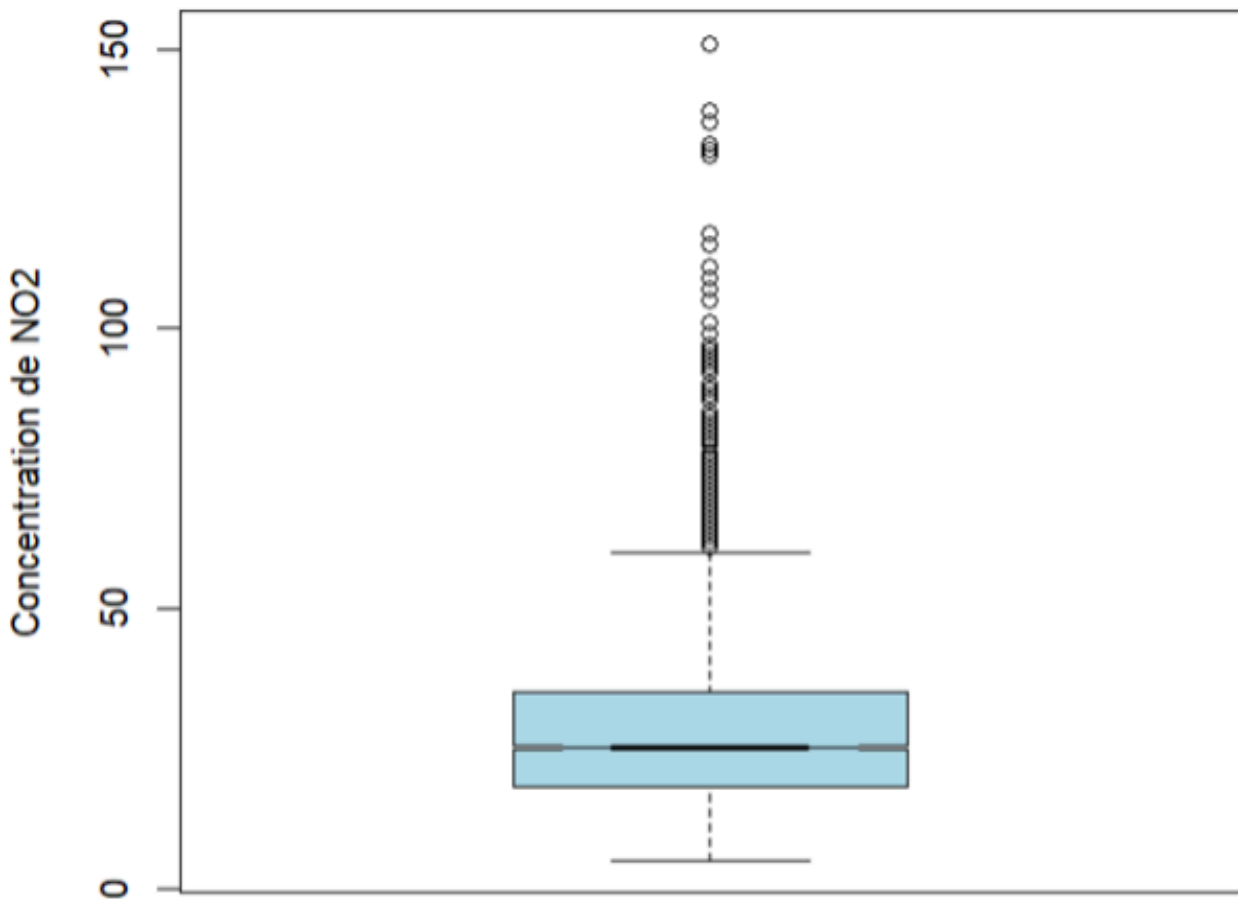
La pollution atmosphérique constitue une problématique majeure des sociétés contemporaines, avec des implications importantes sur la santé publique et l'environnement. Dans le cadre de notre projet de séries temporelles, nous avons choisi d'analyser la concentration de dioxyde d'azote (NO_2) sur une période s'étalant de 2013 à 2023. Cette étude vise à comprendre les variations saisonnières, les tendances générales, et les phénomènes inhabituels observés dans les données, tout en explorant les méthodes statistiques avancées pour mieux appréhender ces dynamiques.

En nous appuyant sur des outils tels que les filtres de moyennes mobiles, les courbes de régression et les modèles ARIMA, nous avons exploré les fluctuations dans les niveaux de NO_2 et identifié des épisodes de pollution marquants, notamment ceux autour des années 2018 et 2020. Notre analyse révèle également des composantes saisonnières récurrentes et des tendances générales d'augmentation, tout en proposant des ajustements pour corriger les variations indésirables.

Ce travail s'inscrit dans une démarche d'approfondissement des techniques de modélisation des séries temporelles, afin d'ouvrir la voie à des prédictions fiables sur l'évolution de la pollution pour les prochaines années.

ANALYSE STATISTIQUE

Boxplot de NO2



```
boxplot(no2, main = "Boxplot de NO2",  
        ylab = "Concentration de NO2", col = "lightblue",  
        border = "black", notch = TRUE)
```

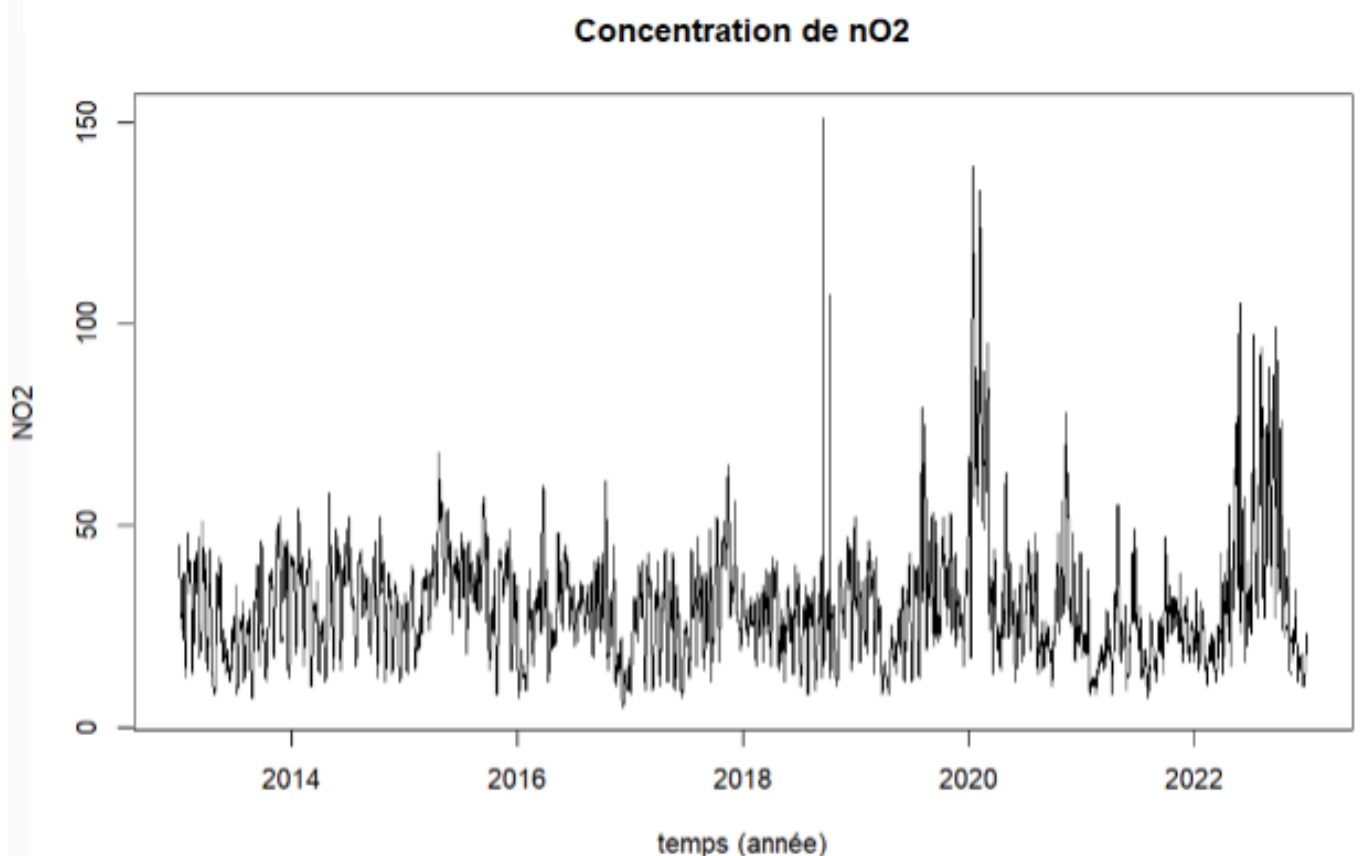
- Le boxplot montre la distribution des concentrations de NO2, avec la majorité des valeurs situées entre environ 20 et 40, ce qui correspond à la boîte principale.
- Il existe de nombreux points au-dessus des moustaches, indiquant une quantité importante de valeurs aberrantes ou concentrations élevées de NO2.
- Cela suggère que la distribution des concentrations de NO2 est asymétrique, avec des épisodes ponctuels de pollution plus élevée.

TENDANCE DE LA SÉRIE

Pour montrer la tendance de cette série, nous allons utiliser un filtre de moyennes mobiles et la courbe de régression des moyennes annuelles.

Tout d'abord, pour poursuivre notre analyse, il faut transformer les données mis à notre disposition en série temporelle à l'aide de la fonction `ts()` sur Rstudio :

```
no2.ts <- ts(no2_data$NO2, start = c(2013,1), frequency = 365, end = c(2023,1))  
no2.ts  
plot(no2.ts, ylab = "NO2", xlab="temps (année)", main = "Concentration de nO2")
```



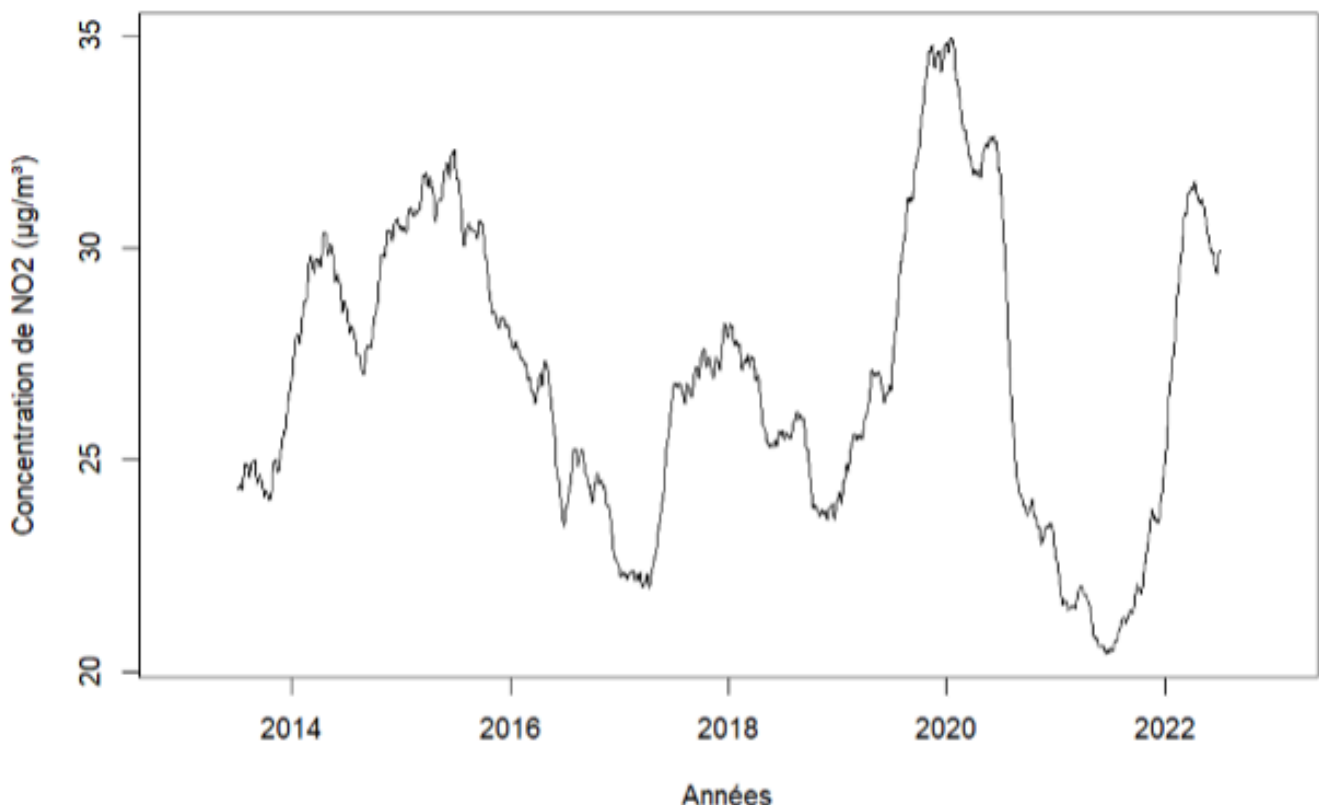
On observe une tendance générale stable avec des variations régulières, mais certains pics notables apparaissent, notamment autour de 2018 et 2020, indiquant des périodes de concentration élevée. On voit que cette série suit un modèle multiplicatif.

TENDANCE DE LA SÉRIE

Pour étudier en détail la tendance, on va utiliser un filtre de moyennes mobiles. Pour cela, on commence par mettre au logarithme la série car elle est multiplicative et on veut une série additive pour pouvoir effectuer le `decompose()`. Ensuite, on affiche la courbe de la tendance par moyenne mobile à l'aide d'un `plot()` :

```
#Calcul du logarithme
log_no2 <- log(no2.ts)
dec <- decompose(log_no2, type = "additive")
# Visualisation de la tendance
plot(exp(dec$trend), type = "l",
      main = "Courbe de la tendance par moyenne mobile",
      xlab = "Années", ylab = "Concentration de NO2 (µg/m³)")
```

Courbe de la tendance par moyenne mobile



On observe que la courbe fluctue de manière significative au fil des années, avec plusieurs cycles de hausse et de baisse bien marqués. La concentration de NO2 atteint des sommets autour de 2020, mais diminue rapidement par la suite. Cependant, à partir de 2022, on note une légère reprise vers une tendance à la hausse.

TENDANCE DE LA SÉRIE

Maintenant, nous allons poursuivre notre analyse de la tendance en effectuant une courbe de régression des moyennes annuelles. On commence par calculer les moyennes annuelles en mettant en amont (time(data)) et data en vecteur afin de pouvoir utiliser tapply() puis on crée des séquences pour le tracé de la courbe :

```
# Calculer les moyennes annuelles et tracer la courbe de régression
# des moyennes annuelles
# Mettre en vecteur
x <- as.vector(time(no2.ts))
y <- as.vector(no2.ts)

# Calculer le nombre d'années dans la série temporelle
length(x) / 12

# Créer des classes pour regrouper les données par année
# Ajustement des bornes pour correspondre aux données réelles
xclass <- cut(x, breaks = seq(2013, 2024, by = 1), include.lowest = TRUE, right = FALSE)

# Calculer les moyennes annuelles
meanannual <- tapply(y, xclass, mean, na.rm = TRUE)

# Créer une séquence pour tracer ensuite
tt <- seq(2013, 2023, by = 1) # Ajustement pour correspondre aux moyennes
```

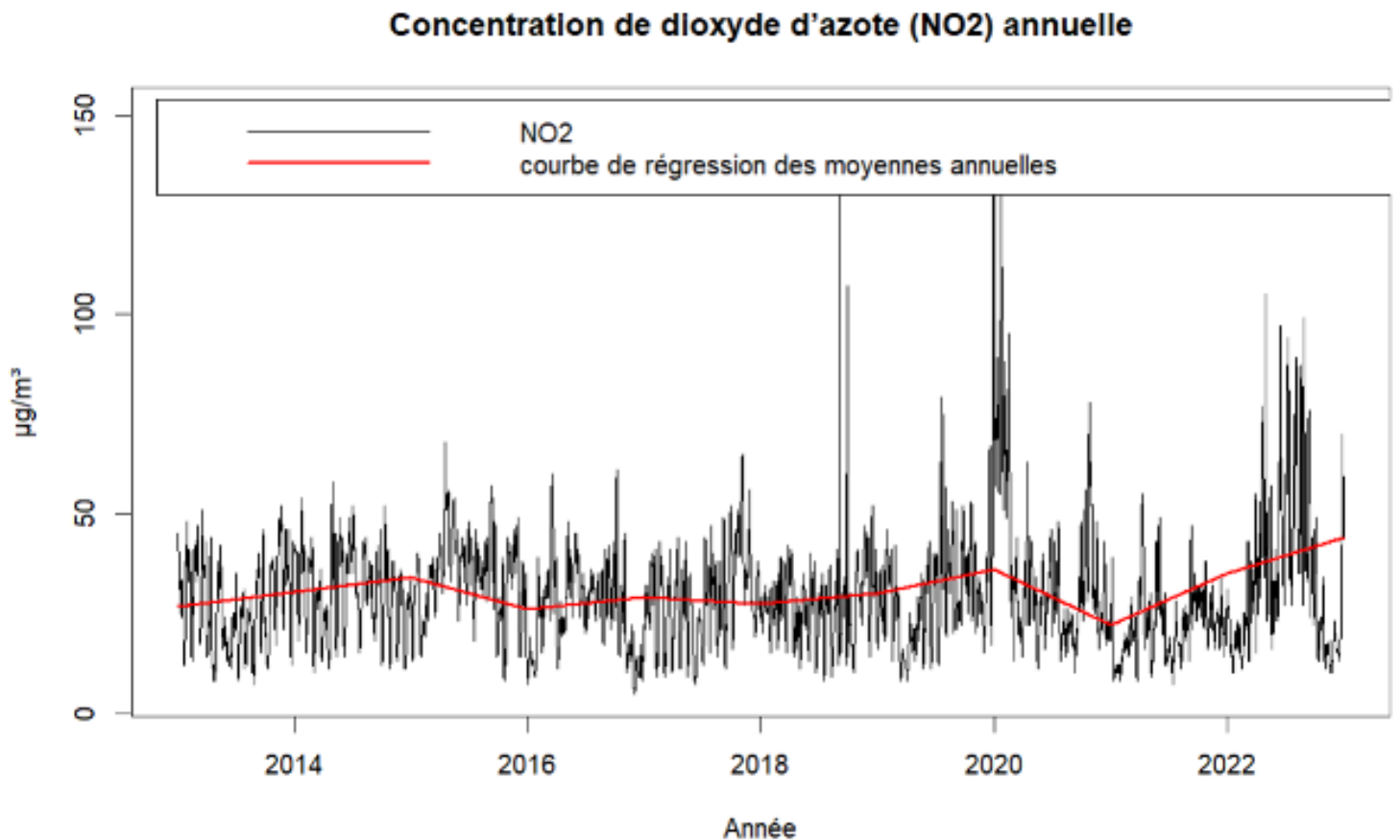
Ensuite on trace la courbe de régression des moyennes annuelles à l'aide d'un plot() ainsi que les légendes comme ci-dessous :

```
# Tracer le graphique principal
plot(x, y, type = "l", main = "Concentration de dioxyde d'azote (NO2) annuelle",
     ylab = "µg/m³", xlab = "Année")

# Ajouter la courbe de régression des moyennes annuelles en rouge
lines(tt, meanannual, col = "red", lwd = 2)

# Ajouter la légende
legend("topleft", lty = c(1, 1), lwd = c(1, 2),
     legend = c("NO2", "courbe de régression des moyennes annuelles"),
     col = c("black", "red"), inset = 0.02)
```

TENDANCE DE LA SÉRIE

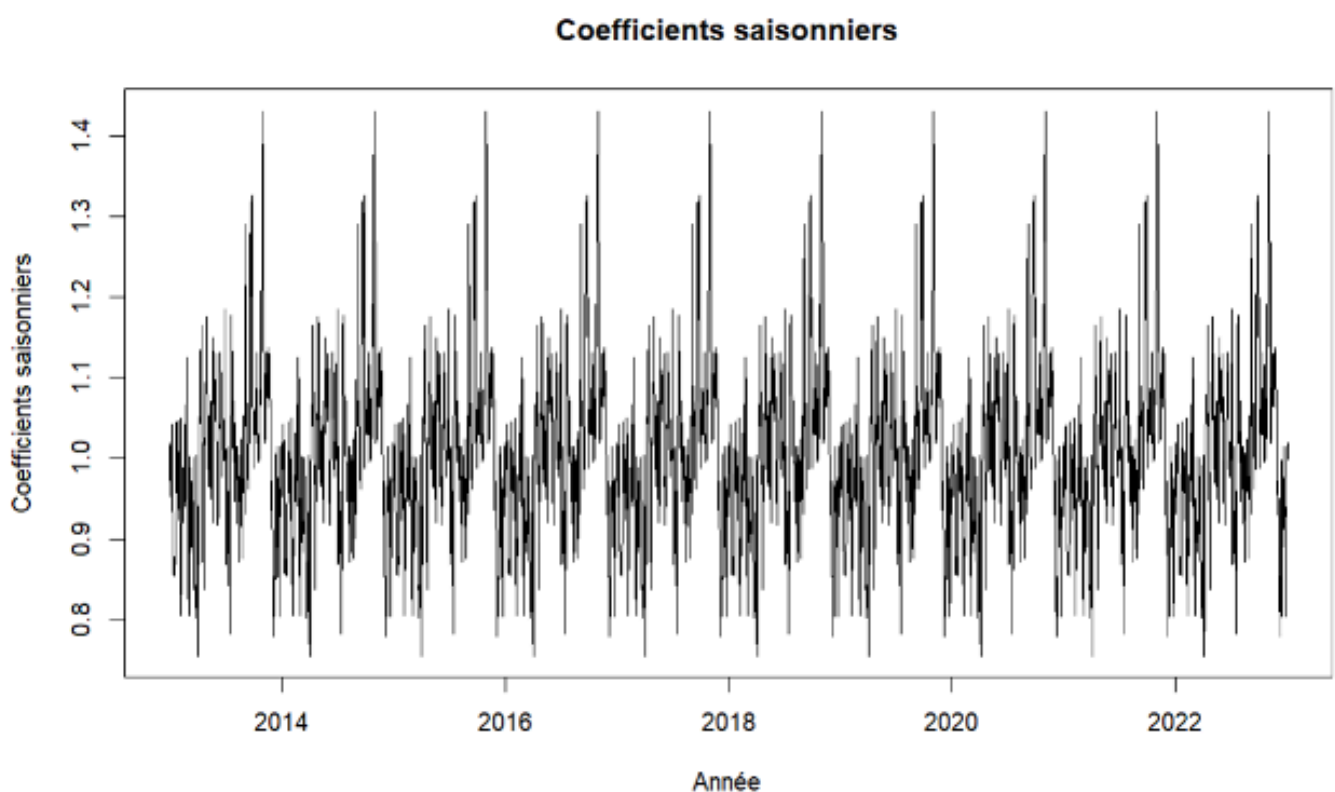


Le graphique montre l'évolution annuelle de la concentration de dioxyde d'azote (NO₂) entre 2013 et 2023, avec une tendance générale légèrement croissante. La courbe de régression des moyennes annuelles (en rouge) indique une augmentation progressive des niveaux moyens de NO₂, malgré d'importantes fluctuations et des pics marqués, notamment autour de 2020.

VARIATIONS SAISONNIÈRES

Les variations saisonnières se réfèrent aux fluctuations régulières et prévisibles qui se produisent à des intervalles de temps spécifiques, souvent liées à des saisons, des mois, des trimestres ou d'autres périodes régulières. Nous allons donc calculer et tracer les coefficients saisonniers :

```
# Calcul des coefficients saisonniers
coefficients_saisonniers <- exp(dec$seasonal)
# Tracer les coefficients saisonniers
plot(coefficients_saisonniers, type = "l", main = "Coefficients saisonniers",
      xlab = "Année", ylab = "Coefficients saisonniers")
```



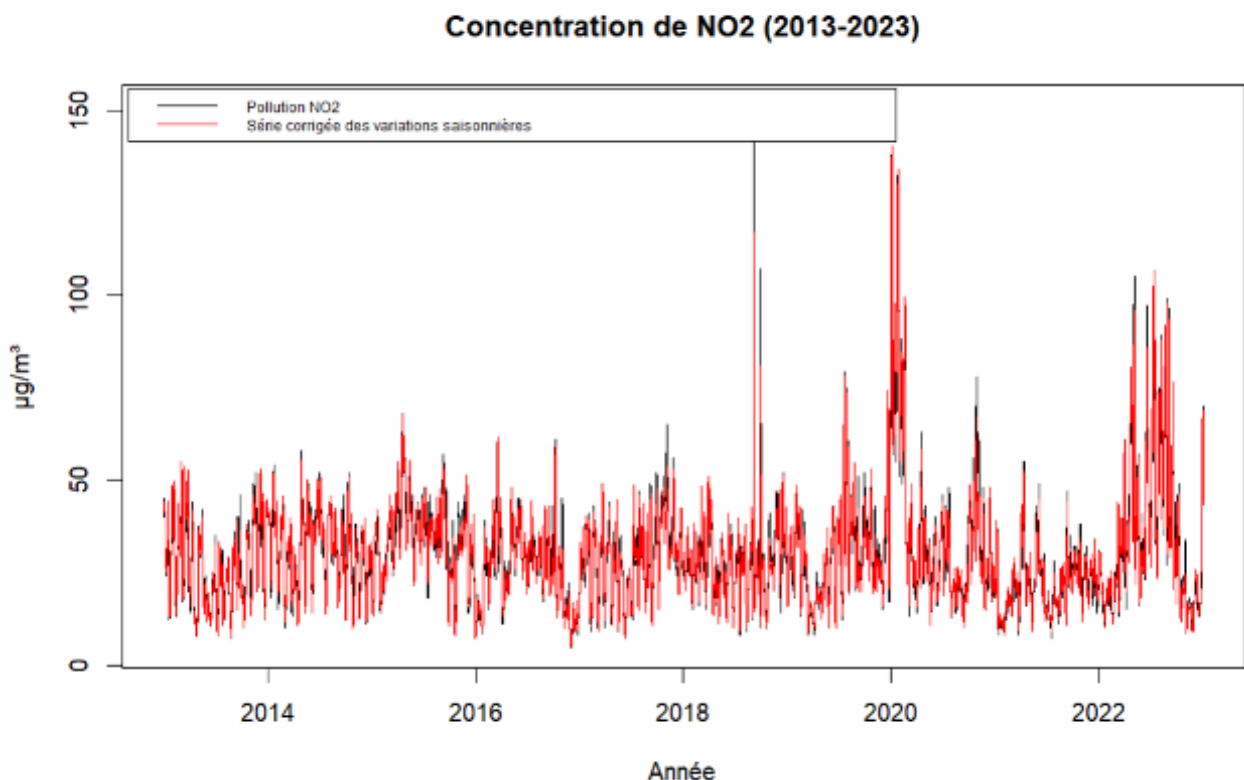
La courbe des coefficients saisonniers ci-dessus est périodique avec des intervalles réguliers qui correspondent aux cycles annuels. Elle met en évidence des variations saisonnières marquées, reflétées par une amplitude significative. Ces variations indiquent que la concentration de dioxyde d'azote (NO_2) fluctue de manière récurrente selon les saisons.

SÉRIE SAISONNIÈRE AJUSTÉE

Nous voulons maintenant ajuster la série saisonnière car on veut améliorer la précision de l'analyse. En l'ajustant, on peut corriger les effets saisonniers indésirables dans nos données. Voici le code qui a permis de générer la série ajustée :

```
# Série corrigée des variations saisonnières
CVS <- y / exp(dec$seasonal)

# Tracer la série temporelle et la série corrigée
plot(x, y, type = "l", main = "Concentration de NO2 (2013-2023)",
     ylab = "µg/m³", xlab = "Année")
lines(x, CVS, type = "l", col = "red")
legend("topleft", lty = c(1, 1), col = c("black", "red"),
      legend = c("Pollution NO2", "Série corrigée des variations saisonnières"),
      inset = 0.006, cex = 0.6)
```



La série corrigée des variations saisonnières, illustrée en rouge, est très proche de la série de base représentée en noir. Cela indique que les variations saisonnières n'ont qu'un impact limité sur les données brutes. La similitude entre les deux courbes suggère que les fluctuations observées sont principalement dues à des facteurs non saisonniers, et que la série de base reflète déjà bien les tendances sous-jacentes de la pollution au NO₂.

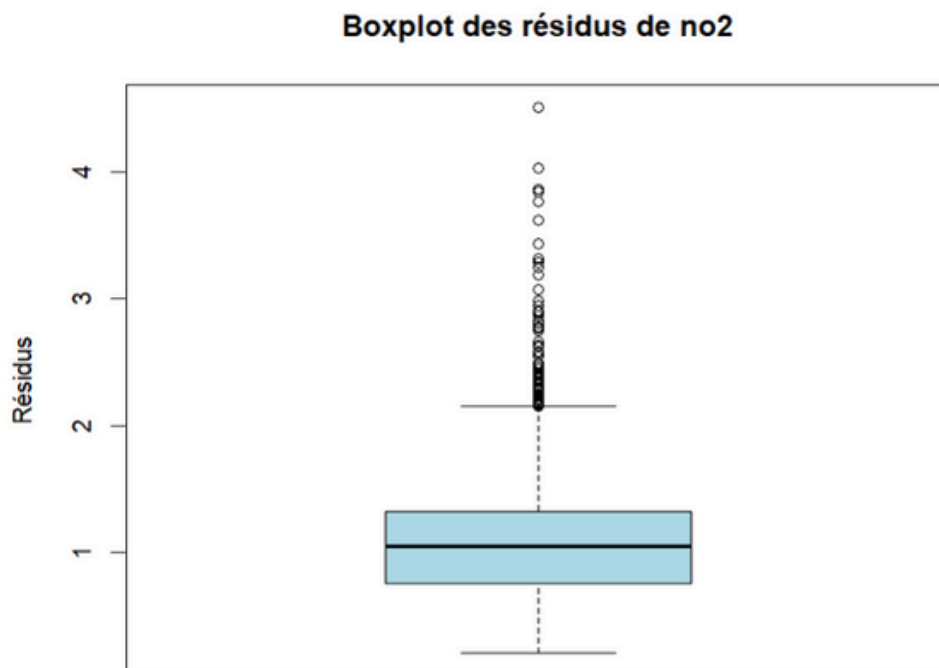
LES RÉSIDUS

A présent, on veut tracer le boxplot des résidus. On utilise le code ci-dessous :

```
#####Les résidus de no2|
#résidus :
residus = y/(exp(dec$trend)*exp(dec$seasonal))

#boxplot des résidus :
boxplot(residus,main = "Boxplot des résidus de no2",
        ylab = "Résidus", col = "lightblue", border = "black")
```

On obtient la boîte à moustache suivante :



```
> summary(residus)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
0.2175  0.7630  1.0526  1.0839  1.3201  4.5093   364
```

Les résultats de la boîte à moustaches montrent que la médiane des résidus est d'environ 1.05, avec des valeurs allant de 0.22 à 4.51. La majorité des résidus se situe entre 0.76 et 1.32 (1er et 3e quartiles), tandis que la moyenne est légèrement supérieure à la médiane (1.08), suggérant la présence de valeurs extrêmes. Les 364 valeurs manquantes (NA's) pourraient toutefois limiter l'interprétation des résultats.

PRÉVISION POUR LE RESTE DE L'ANNÉE 2024

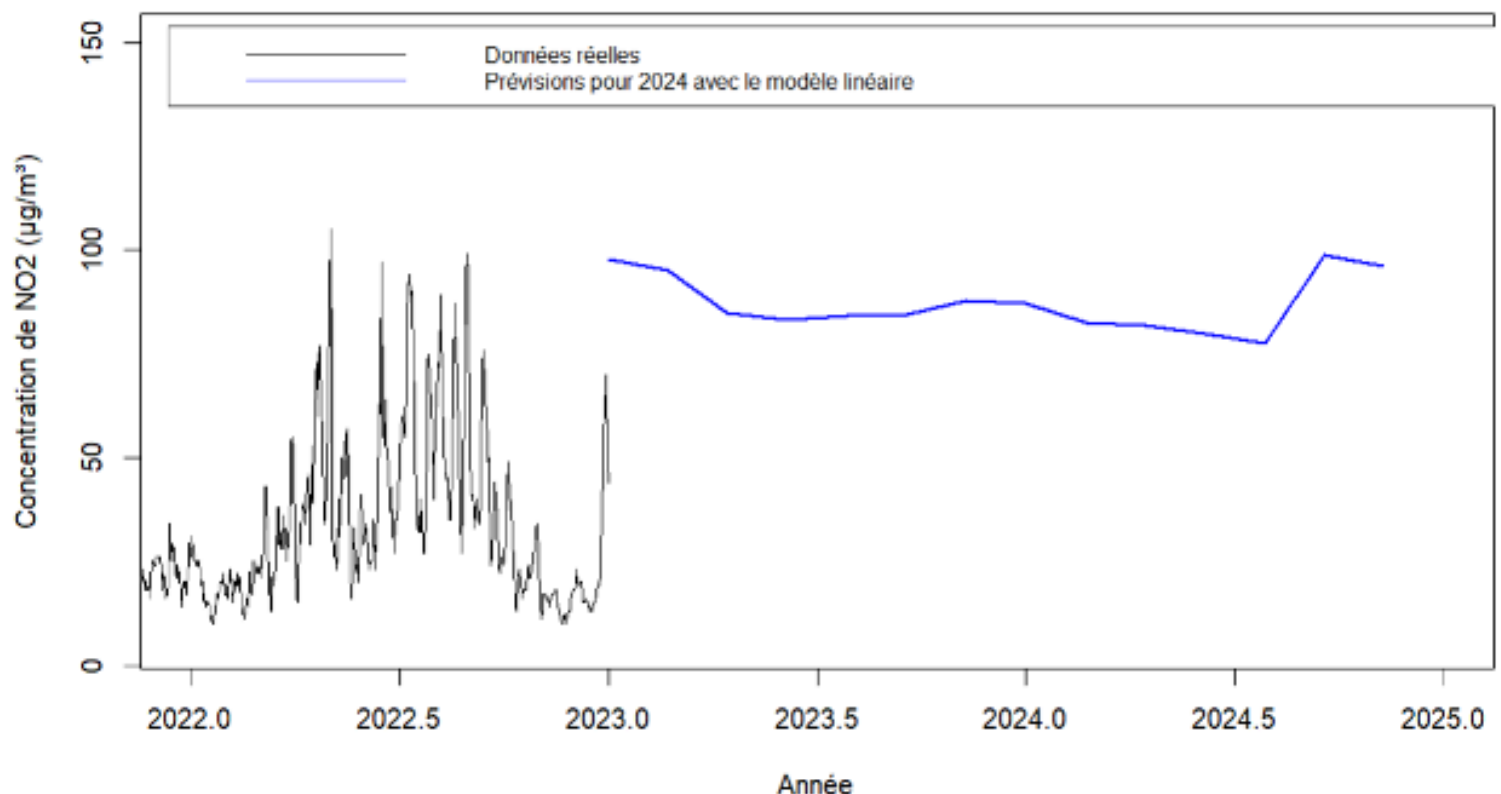
Nous cherchons à prévoir les productions d'énergie hydroélectrique pour le reste de l'année 2024 à travers 3 méthodes différentes :

Tout d'abord le Modèle Linéaire :

Nous utilisons un modèle linéaire pour estimer la tendance des données avec ce code;

```
# Représentation graphique des données et des prévisions
plot(no2.ts, main = "Prévisions des concentrations de NO2 avec un modèle linéaire",
     xlab = "Année", ylab = "Concentration de NO2 (µg/m³)", xlim = c(2022, 2025))
lines(x_fut, prediction_lin, col = "blue", lwd = 2)
legend("topleft", legend = c("Données réelles", "Prévisions pour 2024 avec le modèle linéaire"),
     lty = c(1, 1), col = c("black", "blue"), cex = 0.8, inset = 0.02)
```

Prévisions des concentrations de NO2 avec un modèle linéaire



PRÉVISION POUR LE RESTE DE L'ANNÉE 2024

Puis avec le modèle Arima:

Nous utilisons un modèle Arima pour estimer la tendance des données avec ce code;

```
# Définir la série temporelle (exemple pour des données de 2022 à 2024 avec fréquence quotidienne)
no2.ts <- ts(no2, start = c(2022, 1), end = c(2024, 8), frequency = 365)

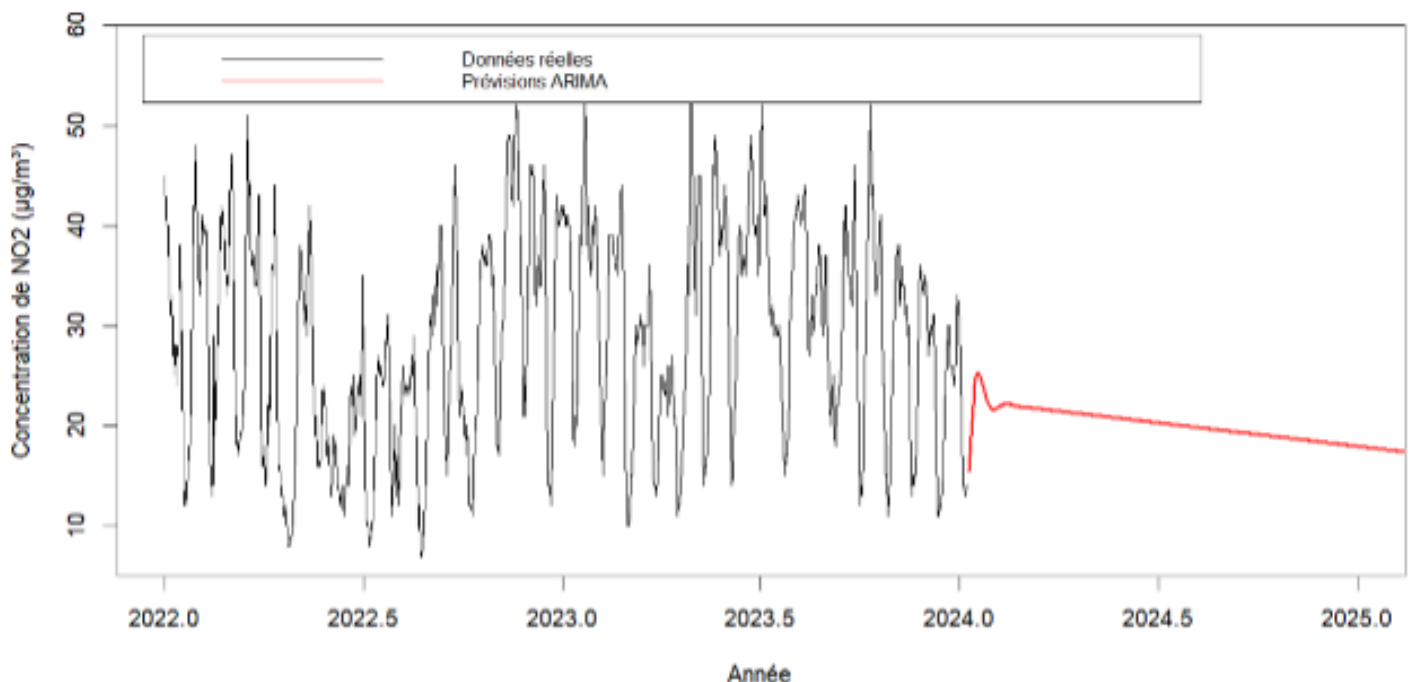
# Diviser les données en ensemble d'entraînement (jusqu'au 31 août 2024)
train_data <- window(no2.ts, end = c(2024, 243)) # Jusqu'au 243e jour de 2024

# Ajustement du modèle ARIMA
arima_model <- auto.arima(train_data, seasonal = TRUE)
summary(arima_model) # Résumé du modèle ARIMA

# Prévisions jusqu'à la fin de 2025
arima_forecast <- forecast(arima_model, h = 365 + 122) # Prévisions pour 487 jours (reste de 2024 + 2025)

# Représentation graphique des données et des prévisions
plot(no2.ts, main = "Prévisions des concentrations de NO2 avec un modèle ARIMA",
     xlab = "Année", ylab = "Concentration de NO2 (µg/m³)", xlim = c(2022, 2025))
lines(arima_forecast$mean, col = "red", lwd = 2) # Prévisions
legend("topleft", legend = c("Données réelles", "Prévisions ARIMA"),
      lty = c(1, 1), col = c("black", "red"), cex = 0.8, inset = 0.02)
```

Prévisions des concentrations de NO2 avec un modèle ARIMA



PRÉVISION POUR LE RESTE DE L'ANNÉE 2024

Et enfin avec le modèle Holt-Winters:

Nous utilisons un modèle Holt-Winters pour estimer la tendance des données avec ce code;

```
# 3. Méthode 3 : Lissage exponentiel (Holt-Winters)
# Définir la série temporelle (exemple pour des données de 2022 à 2024 avec fréquence quotidienne)
no2.ts <- ts(no2, start = c(2022, 1), end = c(2024, 8), frequency = 365)

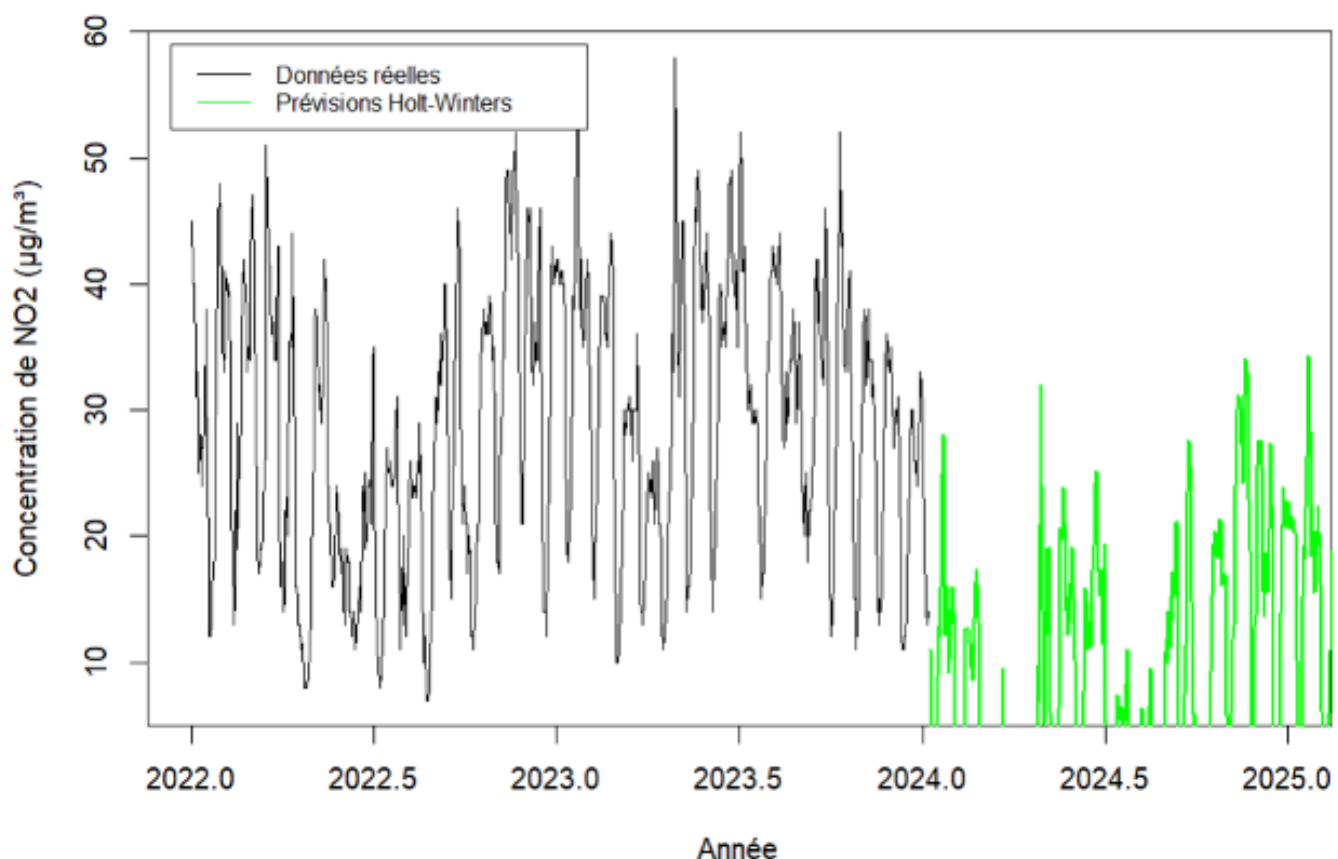
# Diviser les données en ensemble d'entraînement (jusqu'au 31 août 2024)
train_data <- window(no2.ts, end = c(2024, 243)) # Jusqu'au 243e jour de 2024

# Ajustement du modèle Holt-Winters
hw_model <- Holtwinters(train_data) # Ajustement du modèle
summary(hw_model) # Résumé du modèle Holt-Winters

# Prévisions jusqu'à la fin de 2025
hw_forecast <- forecast(hw_model, h = 365 + 122) # Prévisions pour 487 jours (reste de 2024 + 2025)

# Représentation graphique des données et des prévisions
plot(no2.ts, main = "Prévisions des concentrations de NO2 avec un modèle Holt-Winters",
      xlab = "Année", ylab = "Concentration de NO2 (µg/m³)", xlim = c(2022, 2025))
lines(hw_forecast$mean, col = "green", lwd = 2) # Prévisions
legend("topleft", legend = c("Données réelles", "Prévisions Holt-Winters"),
      lty = c(1, 1), col = c("black", "green"), cex = 0.8, inset = 0.02)
```

Prévisions des concentrations de NO2 avec un modèle Holt-Winters



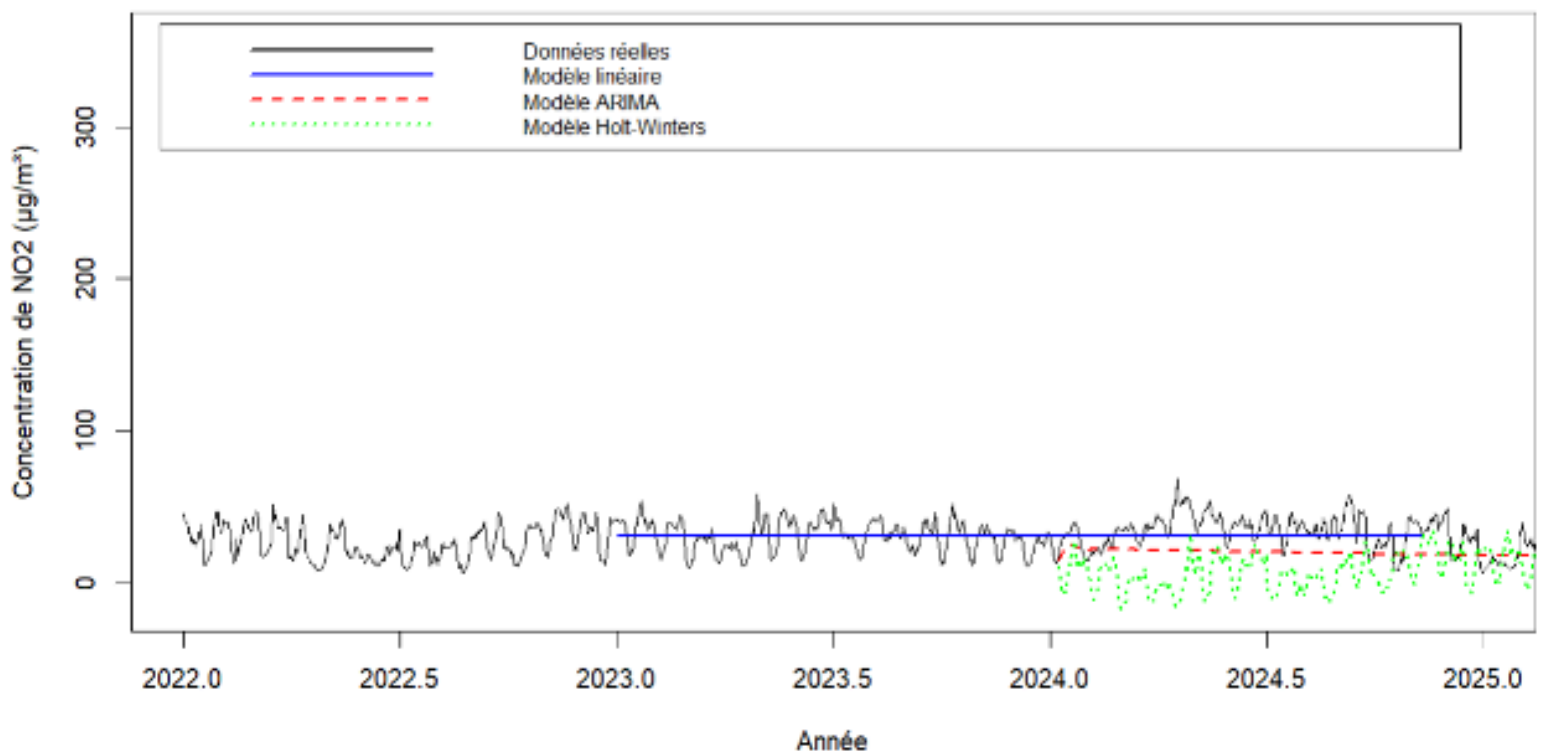
PRÉVISION POUR LE RESTE DE L'ANNÉE 2024

Puis nous avons comparé les modèles de prévision des concentrations de NO₂ à l'aide de ce code:

```
# --- Représentation graphique ---  
no2.ts <- ts(no2, start = c(2022, 1), frequency = 365)  
plot(no2.ts, main = "Comparaison des modèles de prévision des concentrations de NO2",  
      xlab = "Année", ylab = "Concentration de NO2 (µg/m³)", xlim = c(2022, 2025),  
      ylim = range(no2.ts, tend_lin_fut, arima_forecast$mean, hw_forecast$mean))  
lines(x_fut, tend_lin_fut, col = "blue", lwd = 2, lty = 1)  
lines(arima_forecast$mean, col = "red", lwd = 2, lty = 2)  
lines(hw_forecast$mean, col = "green", lwd = 2, lty = 3)
```

On voit que la courbe de prévision qui se rapproche le plus de celle des données réelles (en noir) est la verte (Holt Winters).

Comparaison des modèles de prévision des concentrations de NO₂



PRÉVISION POUR LE RESTE DE L'ANNÉE 2024

A l'aide des prévisions 2024, nous n'avons pas pu conclure sur la méthode la plus efficace. Alors, nous allons calculer l'erreur quadratique moyenne pour chaque méthode afin de trouver celle qui est la plus efficace grâce au code suivant :

```
#erreur quadratique moyenne de prévision pour chacune des méthodes :  
  
# Extraire les données de septembre à décembre 2024  
no2.ts <- ts(no2, start = c(2024, 1), frequency = 365)  
  
data_test <- window(no2.ts, start = c(2024, 9), end = c(2024, 12))  
# 1. Modèle ARIMA  
prev_arima <- forecast(arima_model, h = length(data_test))$mean # Prévisions ARIMA  
mse_arima <- mean((data_test - prev_arima)^2, na.rm = TRUE) # Calcul de MSE pour ARIMA  
  
# 2. Modèle Holt-Winters  
prev_hw <- forecast(hw_model, h = length(data_test))$mean # Prévisions Holt-Winters  
mse_hw <- mean((data_test - prev_hw)^2, na.rm = TRUE) # Calcul de MSE pour Holt-Winters  
  
# 3. Modèle Linéaire  
# Prévisions du modèle linéaire (ajuster la longueur pour correspondre à data_test)  
prev_lin <- tend_lin_fut[1:length(data_test)] # Prévisions linéaires ajustées  
mse_lin <- mean((data_test - prev_lin)^2, na.rm = TRUE) # Calcul de MSE pour le modèle linéaire
```

Bien que les erreurs quadratiques moyennes (MSE) suggèrent que le modèle linéaire est le plus précis en termes d'ajustement global, le modèle Holt-Winters semble être le plus efficace visuellement pour capturer la structure des données (variations saisonnières et dynamiques). Cela montre que le choix du modèle dépend de l'objectif : minimiser l'erreur globale ou bien refléter les tendances sous-jacentes des données.

```
> cat("Erreur Quadratique Moyenne (MSE) :\n")  
Erreur Quadratique Moyenne (MSE) :  
> cat("Modèle ARIMA : ", mse_arima, "\n")  
Modèle ARIMA : 70.44259  
> cat("Modèle Holt-Winters : ", mse_hw, "\n")  
Modèle Holt-Winters : 744.4633  
> cat("Modèle Linéaire : ", mse_lin, "\n")  
Modèle Linéaire : 23.65005
```


CONCLUSION



À travers cette étude, nous avons exploré la concentration de dioxyde d'azote (NO_2) en utilisant les outils et méthodologies des séries temporelles. L'analyse de nos données a permis de mettre en évidence des tendances significatives, notamment une légère augmentation des niveaux moyens de NO_2 sur la période étudiée, accompagnée de fluctuations saisonnières bien marquées. Ces résultats soulignent l'importance de facteurs environnementaux et humains dans les variations des concentrations de pollution.

Les épisodes de pic observés autour des années 2018 et 2020 reflètent des phénomènes ponctuels nécessitant une attention particulière, tandis que les méthodes employées, telles que les filtres de moyennes mobiles et les modèles ARIMA, ont démontré leur pertinence dans la modélisation et la prévision. Cependant, les limitations inhérentes aux données, notamment les valeurs manquantes et les résidus extrêmes, montrent qu'une approche complémentaire pourrait affiner nos résultats.

Ce travail ouvre des perspectives intéressantes pour l'approfondissement des recherches sur la pollution atmosphérique et sur l'utilisation d'approches statistiques pour informer les politiques publiques et les stratégies environnementales. Enfin, il constitue une base solide pour l'analyse future de données environnementales plus complexes et interconnectées.

SUMMARY

This project focuses on analyzing nitrogen dioxide (NO₂) pollution over the period from 2013 to 2023 using time series methods. Key findings highlight an overall increasing trend in NO₂ concentration, periodic seasonal variations, and notable pollution spikes, particularly around 2018 and 2020. Techniques such as moving averages, annual regression trends, and ARIMA modeling were utilized to explore and predict NO₂ levels effectively.

While the results offer valuable insights into the dynamics of air pollution, challenges such as missing data and extreme residuals were noted, underscoring the need for additional methods to enhance accuracy. This study lays the groundwork for further research into environmental data analysis and its application in policymaking and public health.

