Study of Air Quality Evolution in Île-de-France

1 State of the Art

1.1 Predictive Algorithms

We conducted an initial literature review on forecasting algorithms that estimate NO2 concentration based on a given set of explanatory variables. For this section, we relied on the articles [1] and [**?**]. The latter article is particularly interesting due to the large number of algorithms explored and compared. The goal of this study is to propose the most accurate forecasting code possible. Among the multitude of algorithms presented, we decided to implement models based on Generalized Additive Models (GAM) and Random Forest after reviewing the article. The first model is a simple additive model that we used naively, with standard splines. Random Forest is a decision tree-based model that is very popular for its flexibility in both classification and regression problems, as well as for its ease of use, which, through a large number of trees, helps prevent overfitting. These two algorithms will reappear in the following section on weather normalization. The third and final predictive algorithm we chose to use is the LSTM algorithm, based on a recurrent neural network that naturally accounts for seasonality. This capability made it particularly relevant for our field of application.

1.2 Weather Normalization

Weather normalization is a technique that, using pollutant concentration data and meteorological data, produces a curve that represents pollutant concentration independent of meteorological influences. This allows us to observe trends that are not attributable to weather effects. The methods implemented in this study are based on references [3], [2], and [**?**]. The first method involves using a generalized additive model, which has the advantage of being an additive model with simple functions that can be explicitly analyzed independently, making it particularly suitable for weather normalization. The difference compared to the referenced article is the choice of the variable to analyze. The last two methods are similar. The second article largely relies on an R module named *rmweather*, initially developed under the name *normalweather* by the author of the first article, who published the functions developed and used during their study. This library is based on Random Forest algorithms that estimate the influence of each parameter on the model's output. In the first article, calculations using methods such as Partial Dependence Plots and Accumulated Local Effects help understand the contributions of meteorological parameters to PM10 concentrations near the studied sites. The Random Forest algorithm is particularly relevant due

to its function, the existence of these methods, and its ease of use, requiring no assumptions about the input data. Evidence of its utility is the use of this library one year later to evaluate the impact of air purification measures implemented by the Chinese government between 2013 and 2017. Their work demonstrated the benefits of the methods used by the government. However, they noted that it was impossible to precisely measure the effectiveness of individual anti-pollution measures, particularly when multiple measures had overlapping effects. We attempted to design quantitative impact evaluation methods through our initial forecasting methods.

2 Methods

2.1 Data

The pollutant concentration data we obtained came from an extraction performed by the client.

The meteorological data we obtained came from MétéoFrance's public database.

2.1.1 Explanatory Variables

Regarding air pollution, we collected only NO2 concentration data, as previously explained.

For meteorological data, we followed the choice of explanatory variables made by [**?**]. Thus, we used the following variables in our project:

1. RR: Daily precipitation amount
2. TN: Minimum daily temperature
3. TX: Maximum daily temperature
4. TM: Average daily temperature
5. FFM: Average wind speed
6. DXY: Wind direction at the time of maximum recorded wind speed

2.1.2 Data Cleaning

Both meteorological and pollutant concentration data contained gaps. We began by addressing this issue. To do so, we selected a measurement station located in the Seine department, in the 13th arrondissement. Additionally, we chose meteorological data from the nearest MétéoFrance station, Parc Montsouris, which provides access to crucial measurements, particularly regarding wind conditions. This station pairing was selected based on the principle of minimizing data gaps in adjacent stations while maximizing the availability of desired parameters. Despite this pre-selection, some data gaps remained. We chose to remove rows containing missing values, as they accounted for less than 3% of our total dataset after pre-selection.

2.2 Synthetic Control

The synthetic control method is a statistical analysis technique primarily used to evaluate the effects of interventions in observational studies, making it ideal for our subject. This method combines multiple control groups to create a synthetic version of the treated group, which is then used for comparison to estimate the effect of an intervention. However, a major issue arises: to construct a synthetic control unit, we need pollutant concentration data from cities similar to Paris where no anti-pollution regulations have been applied. This requirement led us to abandon the idea, as major cities have already implemented pollution control measures.

2.3 Formulation of Synthetic Control

2.3.1 Introduction

The Synthetic Control Method (SCM) is a statistical technique used in causal inference to estimate the effect of an intervention in observational studies. It constructs a synthetic version of the treated unit by optimally weighting untreated units from a donor pool. This allows researchers to estimate the counterfactual outcome — what would have happened in the absence of the intervention.

2.3.2 Formal Definition

Let us define the main components of the Synthetic Control framework:

- Y_{it} : Observed outcome variable (e.g., NO2 concentration) for unit i at time t.
- $Y_{it}^{(0)}$: Counterfactual outcome for unit i at time t (i.e., outcome in the absence of intervention).
- $Y_{it}^{(1)}$: Observed outcome after intervention.
- D_{it} : Treatment indicator (1 if unit i is treated at time t, 0 otherwise).
- α_t : Time fixed effects.
- λ_i : Unit fixed effects.
- X_i : Pre-treatment characteristics of unit i.

The observed outcome is:

$$Y_{it} = Y_{it}^{(0)} + \tau_{it} D_{it}$$

where τ_{it} represents the treatment effect.

2.3.3 Construction of the Synthetic Control

Let unit $i = 1$ be the treated unit, and let $\{2, 3, ..., J+1\}$ be the donor pool of untreated units. The goal is to construct a synthetic unit that closely approximates the treated unit before the intervention.

Define the synthetic control as a weighted combination of donor units:

$$\hat{Y}_{1t}^{(0)} = \sum_{j=2}^{J+1} w_j Y_{jt}$$

where:

- $w_j \geq 0$ are weights assigned to each donor unit.
- $\sum_{j=2}^{J+1} w_j = 1$ ensures convexity.

The optimal weights w_j^* are chosen by minimizing the discrepancy between the treated unit and the synthetic control before treatment:

$$\min_{w} \sum_{t \in \mathcal{T}_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} \right)^2$$

subject to:

$$\sum_{j=2}^{J+1} w_j = 1, \quad w_j \geq 0 \quad \forall j.$$

2.3.4 Estimation of Treatment Effect

Once the optimal weights w_j^* are determined, the counterfactual outcome for the treated unit is estimated as:

$$\hat{Y}_{1t}^{(0)} = \sum_{j=2}^{J+1} w_j^* Y_{jt}, \quad \forall t \geq T_0$$

where T_0 is the time of intervention.

The treatment effect is then calculated as:

$$\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^{(0)}, \quad \forall t \geq T_0.$$

2.3.5 Assumptions and Limitations

The Synthetic Control Method relies on the following assumptions:

- **No Interference:** The intervention affects only the treated unit.
- **Convexity:** The synthetic control is a convex combination of donor units.
- **Similarity in Pre-Treatment Trends:** The pre-treatment characteristics of the treated unit must be closely approximated by the synthetic control.

Despite its strengths, SCM has limitations, including sensitivity to donor unit selection and the requirement that no significant unobserved confounders vary over time.

2.3.6 Conclusion

The Synthetic Control Method is a powerful tool for estimating causal effects when randomized experiments are not feasible. By constructing a weighted combination of control units, it provides a data-driven approach to estimate counterfactual outcomes. Its mathematical formulation ensures robustness but requires careful selection of donor units and validation of pre-treatment fit.

2.4 Weather Normalization Using R Module *rmweather*

This study is distinct from the one described in the previous sections, but it aims to achieve similar results using R language libraries such as rmweather, openair, and other libraries mentioned in reference [2]. Daily NO2 concentration weather normalization was performed using Random Forest models. The predictive RF model for a site was used to predict each NO2 concentration 1000 times. For each prediction, explanatory meteorological variables, excluding NO2 concentration, were sampled without replacement and randomly assigned to an observation of the dependent variable (NO2 concentration). The 1000 predictions were then aggregated using the arithmetic mean, representing the "average" meteorological conditions, which constituted the weather-normalized trend. The functions used to apply the weather normalization procedure were reprogrammed based on the R repository skgrange/normalweatherr. Practical details: The chosen number of trees is 300, the number of predictions per row is 1000, the selected node size is 5, and the number of random variables chosen at each split is mtry = 5.

Since no documentation provides an optimal value for mtry, we implemented a Grid Search algorithm, a systematic method for tuning machine learning model hyperparameters to find values that maximize model performance. In practice, to determine the best mtry value:

1. Specify possible values: 2,3,4,5,6
2. Train the model for each of these values
3. Compute the Mean Squared Error (MSE) and explained variance for each mtry value
4. Select the mtry value with the lowest MSE

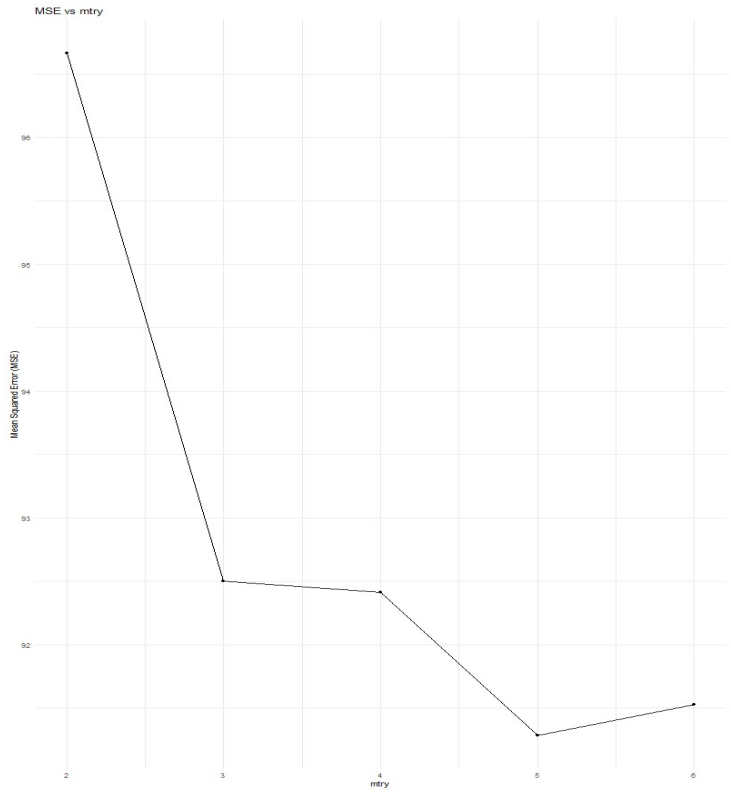

Figure 1: Plot of MSE versus different mtry values for 300 trees

2.5 GAM (Generalized Additive Model)

Generalized Additive Models (GAM) are an extension of Generalized Linear Models (GLM) that allow modeling non-linear relationships between explanatory variables (meteorological data) and the response variable (NO2

concentration). GAMs are particularly useful when relationships between variables are not well represented by simple linear models.

Air pollutant data often exhibit non-linear behavior with meteorological parameters, which prevents researchers from drawing simple conclusions about the parameters of the fitted model [3].

2.5.1 Advantages of GAMs

- **Non-Parametric**: GAMs do not assume a specific form for the relationship between explanatory variables and the response variable. Instead, they use non-parametric smoothing functions to model these relationships.
- **Flexibility**: This approach allows capturing complex non-linear relationships without pre-specifying their form.

2.5.2 Mathematical Formulation

The general formula for a Generalized Additive Model (GAM) is:

$$C_{NO2} = \beta_0 + f_1(TM) + f_2(RR) + f_3(FFM) + f_4(DXY) + \epsilon$$

where:

- C_{NO2} is the NO2 concentration.
- β_0 is the intercept.
- $f_1(TM)$ is a smoothing function of the average temperature (TM).
- $f_2(RR)$ is a smoothing function of precipitation (RR).
- $f_3(FFM)$ is a smoothing function of wind speed (FFM).
- $f_4(DXY)$ is a smoothing function of wind direction (DXY).
- ϵ is the residual error.

2.5.3 Visualization of Individual Effects

One advantage of this approach is that each function $f_i(X_i)$ can be visualized separately to understand the isolated effect of each explanatory variable on the response variable. This makes it possible to interpret how each meteorological variable influences NO2 concentration.

2.5.4 General Methodology

Data Preparation The data used includes meteorological measurements and NO2 concentrations. The selected meteorological variables for this study are:

- Average Temperature (TM)
- Precipitation (RR)
- Wind Speed (FFM)
- Wind Direction (DXY)

The data is filtered for the training period (2002-2017) and the testing period (2017-2020).

Modeling with GAM A GAM model is fitted to predict NO2 concentrations based on meteorological variables.

Model Performance The standard performance indicators used are:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

$$RMSE = \sqrt{MSE} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

Final Adjustment and Prediction The final model is fitted to the entire training dataset and used to predict NO2 concentrations for the test period (2016-2020).

2.5.5 Cross-Validation with GAM for Hyperparameter Optimization

Cross-validation is an essential technique in machine learning to assess model performance in a robust and reliable manner. In the context of Generalized Additive Models (GAM), cross-validation not only evaluates model accuracy but also optimizes hyperparameters to improve performance metrics such as the coefficient of determination (R^2), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

The cross-validation process with GAM involves the following steps:

1. **Data Splitting**: The data is divided into multiple subsets called "folds." In this report, we used 5-fold cross-validation (KFold).

2. **Training and Validation**: For each fold, the model is trained on four subsets and tested on the fifth subset. This process is repeated so that each subset is used once as a test set. This ensures model evaluation on different portions of the data, reducing the risk of overfitting.

3. **Performance Score Calculation**: For each fold, we compute performance metrics, including R^2, MSE, and RMSE. These metrics are then averaged across all folds to obtain an overall performance estimate (averaging performed at the end).

2.5.6 Decorrelation of Meteorological Data with GAM

To assess the impact of emission reduction measures and human activities, it is essential to remove the effect of meteorological variables. Generalized Additive Models (GAM) are particularly suited for this task as they flexibly model non-linear relationships between explanatory variables and the response variable.

Method to Isolate Human Emissions

A GAM model is fitted to predict NO2 concentrations based on meteorological variables. The residuals of this model represent variations in NO2 concentrations that cannot be explained by meteorological conditions, mainly attributable to human emissions and other non-meteorological factors.

Final Adjustment and Prediction The final model is fitted to the entire training dataset and used to predict NO2 concentrations for the test period (2016-2020). The obtained residuals are then analyzed to isolate the effect of human emissions.

Results Display The residuals are also displayed to examine variations due to human emissions.

2.5.7 Theil-Sen Algorithm for Decorrelation Evaluation

Introduction The Theil-Sen algorithm is a robust method used to estimate the slope of a linear relationship between two variables. Unlike the Ordinary Least Squares (OLS) estimation, which can be sensitive to outliers (as is the case for some data), the Theil-Sen algorithm is non-parametric and resistant to outliers. This robustness makes it an appropriate choice for evaluating the effectiveness of decorrelation in contexts where data may contain anomalies [2].

Methodology

Slope Estimation The Theil-Sen algorithm calculates the slope of the regression line by considering the median of the slopes of all possible pairs

of data points. For a set of data points (x_i, y_i), the slopes are calculated as follows:

$$\beta_i = \frac{y_j - y_i}{x_j - x_i} \quad for \quad i < j \tag{4}$$

The estimated slope is the median of these slopes β_i.

Application to Decorrelation In the context of meteorological data decorrelation, the Theil-Sen algorithm can be used to assess whether the residuals (after removing meteorological effects) exhibit a significant linear trend. The absence of a trend would indicate good decorrelation, meaning that the remaining variations in the residuals are independent of meteorological variables.

Advantages of the Theil-Sen Algorithm

- **Robustness**: Resistant to outliers.
- **Non-parametric**: Does not make strict assumptions about data distribution.
- **Simplicity**: Easy to implement and interpret.

Multivariate Methodology in the Case of Multiple Explanatory Variables

Computing Slopes for Each Pair of Points For each pair of points (x_i, y_i) and (x_j, y_j) where $i < j$, we compute individual slopes for each explanatory variable. This is done by calculating the difference in the residual values ϵ divided by the difference in values for each explanatory variable X_k (various meteorological variables).

For each explanatory variable k, we have:

$$\beta_{i,j}^k = \frac{y_j - y_i}{x_{j,k} - x_{i,k}} \tag{5}$$

Combining the Slopes For each explanatory variable k, we take the median of the computed slopes $\beta_{i,j}^k$ across all pairs of points (i, j).

The final slope for the explanatory variable k is:

$$\hat{\beta}^k = median(\beta_{i,j}^k) \tag{6}$$

Intercept Estimation The intercept can be estimated by taking the median of the differences between observed values and predicted values using the estimated slopes for each explanatory variable:

$$\hat{\beta}_0 = median\left(y_i - \sum_{k=1}^{p} \hat{\beta}^k x_{i,k}\right) \tag{7}$$

Final Model The final multivariate regression model using the Theil-Sen estimator is given by:

$$\hat{y} = \hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}^k X_k \tag{8}$$

Conclusion The Theil-Sen algorithm for multiple variables allows robust estimation of linear relationships in the presence of multiple explanatory variables, using medians to ensure robustness against outliers. This makes it a powerful tool for air quality analysis, where data can be influenced by numerous meteorological and anthropogenic factors.

2.6 Results - GAM

2.6.1 Visualization of Individual Effects

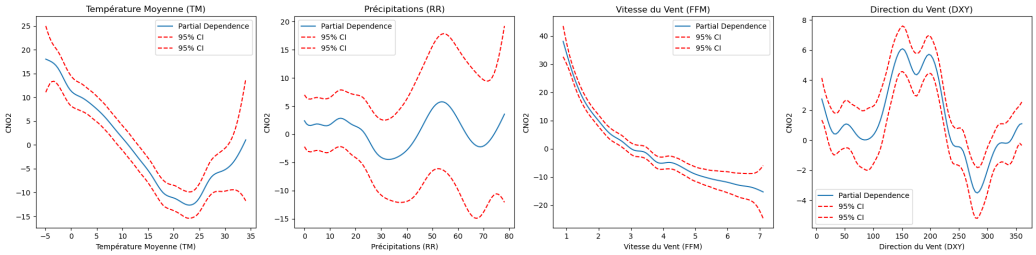

Figure 2: Partial dependence plots for the variables TM, RR, FFM, and DXY.

Explanation of the Graphs

Average Temperature (TM)

Interpretation: We observe that NO2 concentration significantly decreases as temperature increases up to about 25°C, then starts to rise again. This indicates that temperature has a non-linear effect on NO2 concentration.

Precipitation (RR)

Interpretation: The relationship between precipitation and NO2 concentration appears to be complex, with cyclic variations. This may indicate that precipitation has a variable effect on NO2 concentration, with periods of decrease and increase.

Wind Speed (FFM)

Interpretation: We observe that NO2 concentration significantly decreases as wind speed increases. Higher wind speed helps disperse pollutants, thereby reducing NO2 concentration.

Wind Direction (DXY)

Interpretation: The relationship between wind direction and NO2 concentration is also complex, with multiple peaks and troughs. This may indicate that certain wind directions are associated with stronger pollution sources or more effective pollutant dispersion.

2.6.2 Prediction for 2016-2020 with GAM

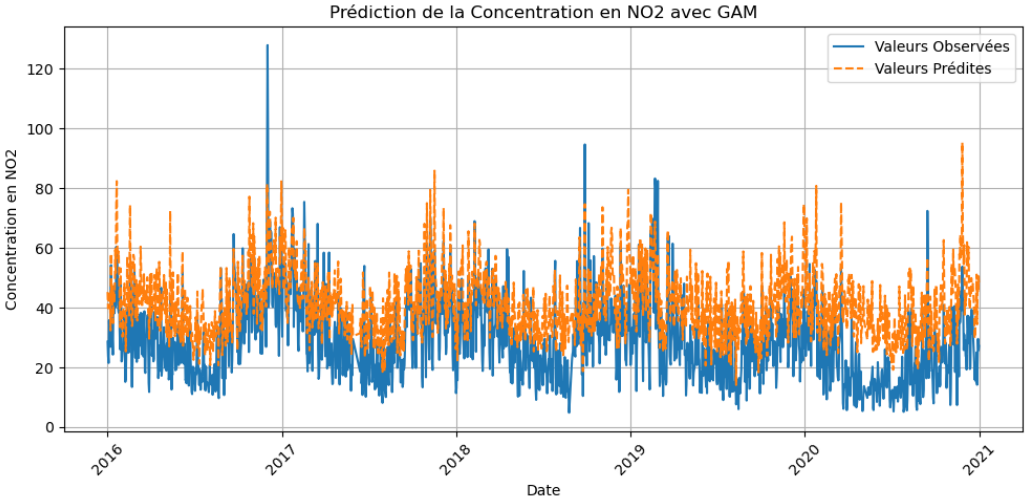

Figure 3: Prediction of NO2 Concentration with GAM (2016-2020)

Graph Explanation Figure 3 presents the NO2 concentration predictions using the GAM model over the period from 2002 to 2020. The observed values (in blue) and the predicted values (in orange dotted line) are displayed. The model was trained on data from 2002 to 2016 and then tested on data from 2016 to 2020.

- Validation R^2 : 0.404
- Validation MSE : 187.836
- Validation RMSE : 13.705
- Test R^2 : 0.202
- Test MSE : 232.093
- Test RMSE : 15.235

Again, the results show good performance on validation data but poorer performance on test data. This indicates that the model could benefit from better generalization.

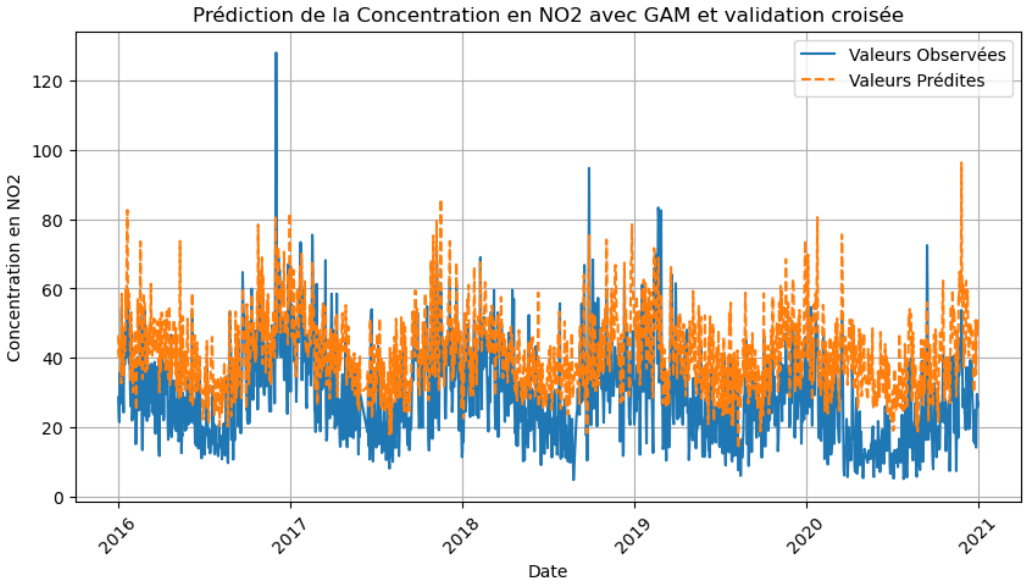

Figure 4: Prediction of NO2 Concentration with GAM and Cross-Validation (2016-2020)

Graph Explanation Figure 4 shows the prediction of NO2 concentration using the GAM model with cross-validation over the period from 2016 to 2020. The observed values (in blue) are compared to the predicted values (in orange dotted line). The model was trained using data from 1995 to 2016 and then tested on data from 2016 to 2020.

- Validation R^2 : 0.393
- Validation MSE : 187.836
- Validation RMSE : 13.705
- Test R^2 : 0.180
- Test MSE : 232.093
- Test RMSE : 15.235

The results show that the model has moderate performance on validation data with an R^2 of 0.393 but does not generalize well on test data, as indicated by the negative R^2 of 0.180.

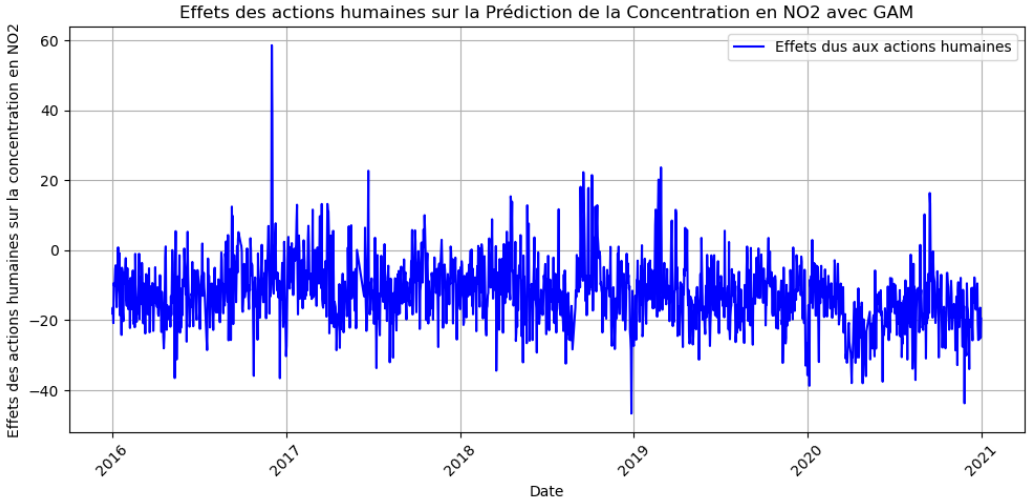

Figure 5: Effects of Human Actions on NO2 Concentration Prediction with GAM (2016-2020)

Graph Explanation Figure 5 shows the effects of human actions on NO2 concentration prediction for the period from 2016 to 2020. These effects are calculated by subtracting the predicted values from the observed values and then smoothing the residuals. These effects are represented in blue.

The variations in NO2 concentrations not explained by meteorological variables are primarily attributed to human emissions. This graph highlights the impact of human activities on pollution levels.

These results demonstrate the importance of considering meteorological variables in NO2 concentration modeling while emphasizing that human activities play a significant role in observed variations.

2.6.3 Decorrelation Results with Theil-Sen

Figure 6 shows the effects of human actions on NO2 concentration prediction obtained with the GAM model, along with the trend estimated using the multivariate Theil-Sen algorithm.

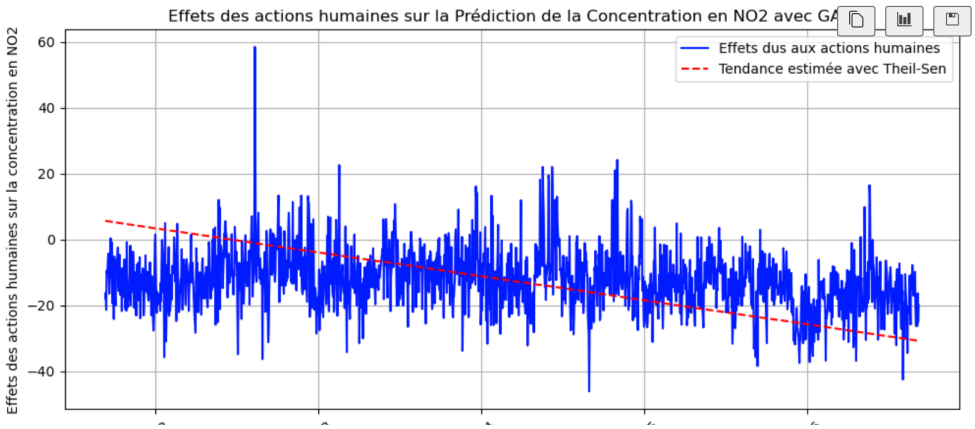

Figure 6: Effects of Human Actions on NO2 Concentration Prediction with GAM and Theil-Sen Estimated Trend

The estimated slope with the multivariate Theil-Sen algorithm is **-0.0203**, and the intercept is **5.6557**. A slope close to zero indicates no significant trend, meaning that the effects of human actions on NO2 concentration are well decorrelated from meteorological variables. This confirms that the GAM model has successfully isolated human action effects, thereby reducing the influence of meteorological variables in the residuals.

The absence of a visible trend in the residuals means that the remaining variations are mainly due to human actions, validating the effectiveness of the decorrelation method used.

2.6.4 Prediction of NO2 Concentration with GAM During the COVID Period (2020-2023)

Figure 7 presents the prediction of NO2 concentration using the GAM model over the COVID period from 2020 to 2023. The observed and predicted values are compared to evaluate the model's performance.

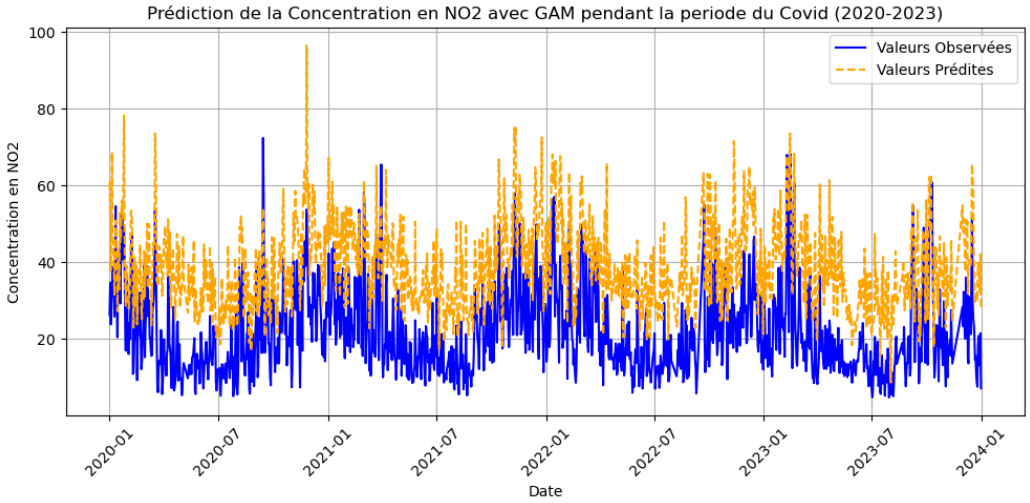

Figure 7: Prediction of NO2 Concentration with GAM During the COVID Period (2020-2023)

The performance metrics for this prediction are as follows:

- $\mathbf{R^2}$: 0.128
- **MSE** : 18.5758
- **MAPE** : 0.933

The negative coefficient of determination (R^2) indicates that the model fails to capture the variance in test data, meaning that the predictions are less accurate than the mean of observed values.

The MSE and RMSE show high prediction errors, reflecting the model's difficulty in properly fitting the test data.

The MAPE (Mean Absolute Percentage Error) of 1.0485 indicates an average absolute error of more than 100

These results suggest that the GAM model used here fails to correctly predict NO2 concentration for the COVID period. This may be due to significant changes in pollution-influencing factors during this period, such as variations in human activities, lockdown measures, and other contextual factors not captured by the meteorological variables used in the model.

References

[1] Yue-Shan Chang, Hsin-Ta Chiao, Satheesh Abimannan, Yo-Ping Huang, Yi-Ting Tsai, and Kuan-Ming Lin. An LSTM-based aggregated model for air pollution forecasting. *Atmospheric Pollution Research*, 11(8):1451–1463, 2020.

[2] Stuart K. Grange, David C. Carslaw, Alastair C. Lewis, Eirini Boleti, and Christoph Hueglin. Random forest meteorological normalisation models for Swiss PM10 trend analysis. *Atmospheric Chemistry and Physics*, 18(9):6223–6239, 2018.

[3] Said Munir, Gulnur Coskuner, Majeed S. Jassim, Yusuf A. Aina, Asad Ali, and Martin Mayfield. Changes in air quality associated with mobility trends and meteorological conditions during COVID-19 lockdown in Northern England, UK. *Atmosphere*, 12(4):504, 2021.

[4] Guillaume Staerman and Amaury Durand. Prédiction de la pollution de l'air à Londres, 2018.

[5] Tuan V. Vu, Zongbo Shi, Jing Cheng, Qiang Zhang, Kebin He, Shuxiao Wang, and Roy M. Harrison. Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. *Atmospheric Chemistry and Physics*, 19(17):11303–11314,