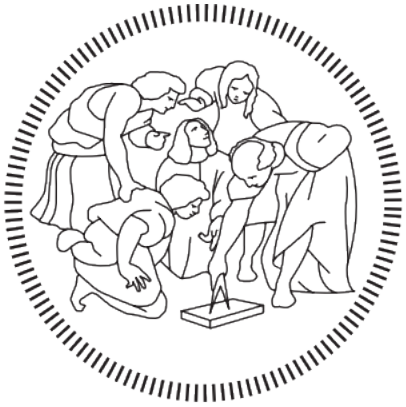

POLITECNICO
MILANO 1863

Bayesian Statistics

Ozone Prediction in Po Valley

A.Y. 2023-2024

Marton Barta - CP: 10884623 - MAT: 216514
Lorenzo Lenzi - CP: 10581949 - MAT: 987119
Natasha Makishima - CP: 10618559 - MAT: 101632
Firas Mtibaa - CP: 10987480 - MAT: 242795
Simon Sebastien Trottier - CP:10988912 - MAT: 246573
Riccardo Villa - CP: 10680712 - MAT: 222757

Prof. Alessandra Guglielmi
Supervisors: Alessandro Carminati and Michela Frigeri

Link to the GitHub Repository : https://github.com/fmtibaa/Bayesian_Project_A1/tree/main

Contents

1. Introduction

The stratospheric ozone layer, serving as Earth's protective shield, is critical in maintaining ecological balance and protecting the biosphere by filtering harmful ultraviolet radiation. This research focuses on Lombardy, a region characterized by its industrial vigor and marked as one of Italy's most important regions, where ozone-level shifts have implications for environmental and public health concerns. The measured ozone level in the troposphere (where humans live) has significant implications on human health and agriculture since ozone acts as a harmful molecule to the living; therefore, understanding its values, and by that, being able to act in a preventing way, could not only have monetary but also health benefits.

Some known factors, such as high temperatures and solar radiation exacerbate ozone concentration, while rain and humidity can mitigate it. However, the variability of wind patterns and geographical elevation complicates the predictability of ozone dispersion.

This research aims to map the ozone levels in Lombardy (Po Valley) in a sophisticated Bayesian statistical framework, drawing on the methodologies pioneered in environmental Bayesian statistics, such as those in *Sahu et al.(2007)*. We integrate a comprehensive array of data sources, including satellite observations and ground-level monitoring stations, to build a robust spatial-temporal model. We will also implement various environmental information, such as humidity, rain, wind direction, etc. obtained from Meteorological Sites that will serve as environmental covariates to enhance our spatiotemporal models.

The stakes are high, as ozone poses a threat not only to respiratory health-implicated in ailments like asthma and stroke but also to the region's agronomy, affecting both yield and economic stability.

2. Exploratory Analysis

2.1. Data Overview and Spatial Distribution

We present an exploratory analysis of an initial dataset of ozone concentration measurements collected from 51 sensors distributed across the region of Lombardy.
Spanning from 2010 to 2017, this dataset contains a total of 5,552,198 data points, each associated with the latitude, longitude, ozone concentration value, and date of measurement.

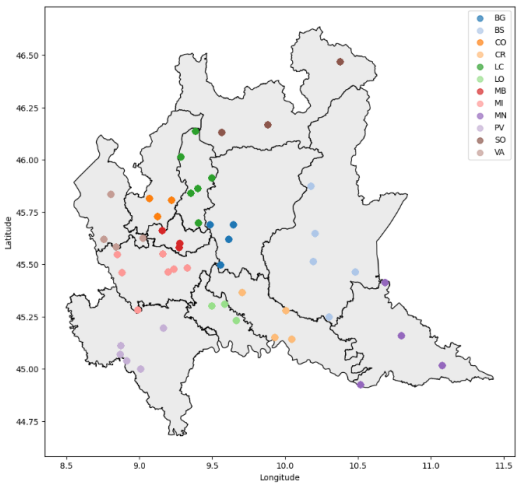

Figure 1: Ozone sensors in Lombardy

The map highlights the geographical spread of the 51 sensors, each representing a monitoring station capturing ozone concentration data.

Along with the sensor locations, we examined the mean ozone levels recorded at each monitoring station. The following map (Figure 2) shows color-coded regions indicating the average ozone concentration across Lombardy. Lighter shades signify higher ozone levels, while darker shades indicate lower concentrations.

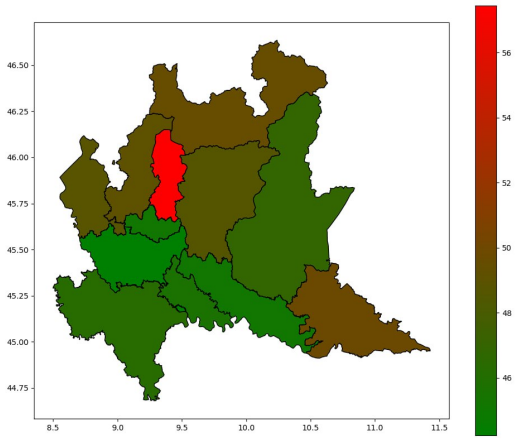

Figure 2: Mean ozone levels in Lombardy (2010-2017)

2.2. Temporal Trends

We explore the temporal trends in ozone concentration over the study period spanning from 2010 to 2017. Analysis of yearly variations provides valuable insights into the long-term dynamics of air quality and helps identify trends and anomalies.

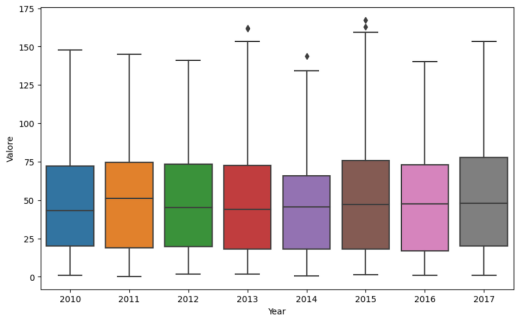

Figure 3: Boxplot of ozone levels over the years (2010-2017)

Figure 3 displays boxplots illustrating the distribution of ozone levels recorded each year from 2010 to 2017. At first sight, the annual analysis via boxplots reveals a stable trend in ozone concentrations, suggesting a persistent equilibrium in the regional ozone levels. While seasonal trends are expected within individual years, the lack of conspicuous patterns across multiple years indicates a relatively steady state of ozone concentration within the dataset. Further analysis may be warranted to elucidate underlying factors influencing ozone dynamics within the region.

This observation underscores the complexity of atmospheric dynamics influencing ozone dispersion and motivates a deeper Bayesian statistical examination to uncover underlying patterns and relationships.

2.3. Meteorological Correlations

An analysis of the interactions between ozone levels and meteorological variables is conducted to uncover the atmospheric influences on ozone dynamics. To provide a comprehensive analysis, we augmented the original dataset with meteorological variables, including precipitation, wind speed, temperature, sunshine duration, and others.
By integrating these meteorological covariates with ozone concentration data, we aim to explore potential correlations and dependencies between atmospheric conditions and ozone levels.

This investigation is pivotal in refining our model, ensuring it accurately reflects the environmental factors affecting ozone concentration. As later seen, this analysis and our initial hypothesis will be the key driving

motivation to select certain covariates that would be compared to a so-called "full-model" later.
We constructed a correlation matrix Figure 4 to visualize the pairwise correlations between ozone concentration and various meteorological variables. The correlation matrix provides insights into the strength and direction of relationships between variables, facilitating the identification of significant associations and potential drivers of ozone dynamics. By examining the correlation coefficients, we can discern which meteorological covariates exhibit the strongest associations with ozone levels and infer their impact on air quality dynamics.

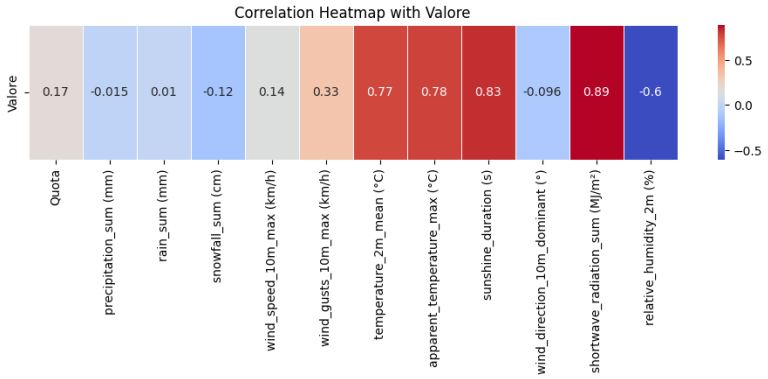

Figure 4: Ozone Value Correlation Heatmap

2.4. Handling Missing Data

The presence of missing data necessitates the careful implementation of an imputation technique to maintain the integrity of the model. Our approach leverages both temporal and spatial dependence, ensuring a robust dataset for analysis. The filling was done by taking the mean of the previous and following week at certain locations. However, as later explained, it was necessary to reduce the number of stations, therefore, we used the linear interpolation for a limited number of stations only.

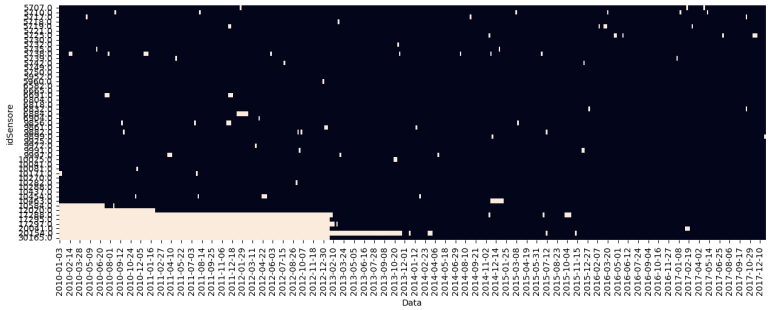

Figure 5: Missing Values by Sensor and Time

As we see in Figure 6, some of the locations started recording their "Valore" later in time, therefore, those locations were automatically considered in the validation set, so no significant data imputation was required.

2.5. Takeaways

The preliminary analysis not only provides a foundational understanding of the dataset but also highlights the significance of incorporating meteorological variables into our spatial-temporal model. These initial exploration steps set the stage for more in-depth investigations using advanced modeling techniques.
The correlation analysis between ozone concentration and meteorological covariates revealed significant associations, highlighting the importance of incorporating atmospheric conditions into the spatial-temporal model, and their potential impact on ozone dynamics; this is essential for developing accurate and reliable predictive models that can capture spatial and temporal patterns.

The Bayesian statistics framework offers a powerful approach for detecting long-term trends in ozone concentrations while accounting for uncertainties inherent in the data.

The data in the original scale were modified to work with the weekly average ozone level instead of the daily values. They underwent a two-step transformation process to facilitate the modeling. Initially, a logarithmic scale transformation was applied to the dataset to mitigate skewness and heteroscedasticity, enhancing the normality of the distributions. Then, the transformed data were standardized to ensure comparability and reduce the influence of scale differences among variables.

3. Proposed Model

In the following section, we formalize the statistical framework underpinning our analysis. We progress by defining key notations, followed by outlining the assumptions and structure of our utilized Bayesian model.

3.1. Model Notation

The following notations are pivotal to our model formulation:

- $O_l(s_i, w)$ denotes the true ozone level at location s_i for week w of year l.
- $Z_l(s_i, w)$ represents the observed ozone level at location s_i for week w of year l.
- $x_{lj}(s_i, w)$ signifies the measured value of meteorological variable j at location s_i for week w of year l.
- $\delta_{lj}(s_i, w) = x_{lj}(s_i, w) - x_{lj}(s_i, w-1)$ encapsulates the week-over-week change in meteorological variable j at location s_i for year l.

3.2. Model Assumptions

Our model is predicated on the following assumptions:

- The observed ozone level, $Z_l(s_i, w)$, is the sum of the true ozone level, $O_l(s_i, w)$, and an error term, $\epsilon_l(s_i, w)$, where $\epsilon_l(s_i, w) \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ for all l, i, and w.
- The week-over-week changes in variables, $\delta_{lj}(s_i, w)$, are spatially correlated yet temporally independent.

3.3. Model Specification

We aim to construct an autoregressive model to obtain the dynamics of ozone levels:

$$O_l(\mathbf{s}, w) = \rho O_l(\mathbf{s}, w-1) + \xi_l \mathbf{1} + \delta_l(\mathbf{s}, w)\boldsymbol{\beta} + \eta_l(\mathbf{s}, w), \tag{1}$$

where:

- $\rho \in (0, 1)$ is the autoregressive coefficient, capturing temporal dependence.
- ξ_l represents the global intercept for year l.
- $\delta_l(\mathbf{s}, w)\boldsymbol{\beta}$ accounts for adjustments based on weather parameters.
- $\eta_l(\mathbf{s}, w)$ denotes a spatially correlated error term.

The model initiates with:

$$O_l(\mathbf{s}, 1) = \mu_l \mathbf{1} + \gamma_l(\mathbf{s}), \tag{2}$$

where $\gamma_l(\mathbf{s})$ reflects regional effects, and μ_l the global effect for the initial week of year l.

3.4. Prior Selection

To fully specify our Bayesian framework, we establish priors for each model component as follows:

- $\rho \sim N(0, 1)I(0 < \rho < 1)$ to enforce the constraint on the autoregressive coefficient.
- $\xi_1, \ldots, \xi_L \overset{iid}{\sim} N(0, 1)$ for the global intercepts across years.
- $\eta_1(\mathbf{s}, 1), ..., \eta_L(\mathbf{s}, 1), ..., \eta_1(\mathbf{s}, W), ..., \eta_L(\mathbf{s}, W) | \sigma_\eta, \phi_\eta \overset{iid}{\sim} N(0, \Sigma_\eta)$, with $\Sigma_\eta(i, j) = \sigma_\eta^2 exp(-\phi_\eta d_{ij})$:
  For the spatially correlated error terms and regional effects, we adopt Gaussian processes with exponential covariance functions to model spatial dependencies.
- $\mu_1, ..., \mu_L \overset{iid}{\sim} N(0, 1)$
- $\gamma_1(\mathbf{s}), ..., \gamma_L(\mathbf{s}) | \sigma_\gamma, \phi_\gamma \overset{iid}{\sim} N(0, \Sigma_\gamma)$, with $\Sigma_\gamma(i, j) = \sigma_\gamma^2 exp(-\phi_\gamma d_{ij})$
- $\boldsymbol{\beta} \sim N(0, I_p)$ for the regression coefficients associated with weather adjustments.
- $\frac{1}{\sigma_\epsilon^2}, \frac{1}{\sigma_\eta^2}, \frac{1}{\sigma_\gamma^2} \overset{iid}{\sim} Gamma(a, b)$, where $a = 2$ and $b = 1$:
  Inverse-Gamma distributions are selected for the precision parameters of error terms, promoting robust inference by accommodating uncertainty.

3.5. Specification of δ

• $\forall(l,w) \in [1,L] \times [1,W], \delta_l(\mathbf{s},w)|A, \boldsymbol{\phi} \overset{iid}{\sim} N(0,\Sigma_\delta)$
with $A = (\mathbf{a}_1, ..., \mathbf{a}_p), \Sigma_\delta(i,j) = \sum_{k=1}^{p} \exp(-\phi_k d_{ij}) \mathbf{a}_k^T \mathbf{a}_k$
• $\phi_1, ..., \phi_p \overset{iid}{\sim} U(0.001, 0.1)$
• $\forall(i,j) \in [1,p]^2, a_{i,j} \overset{iid}{\sim} N(0,1)$
This model framework leverages the power of Bayesian statistics to accommodate the complexity inherent in environmental data, facilitating nuanced insights into the spatiotemporal dynamics of ozone levels.

4. Simulation Strategy

In this section, we delineate the simulation strategy adopted to refine our Bayesian spatial-temporal model for predicting ozone levels in the Lombardy region. The aim of simulation in our study is twofold: first, to enhance the predictive accuracy of our model; second, to assess the robustness and reliability of our model's predictions (Chapter 5).

4.1. A reduced model to select the covariates

At first, a reduced model was deployed to determine which covariates were relevant. In order to do so, we decided not to model δ as explained in the initial model because the its specification is only used when we have a relevant number of sensor on which we don't have observation for the values of the covariates and this is not the case in our dataset. We will only use the δ values of the chosen covariates of the sensor and the time on which we will do the prediction part. Here is the reduced model with its priors :

$$O_l(\mathbf{s},w) = \rho O_l(\mathbf{s},w-1) + \xi_l \mathbf{1} + \delta_l(\mathbf{s},w)\boldsymbol{\beta} + \eta_l(\mathbf{s},w)$$

The initial condition is:

$$O_l(\mathbf{s},1) = \mu_l \mathbf{1} + \gamma_l(\mathbf{s})$$

The priors are :

- $\rho \sim N(0,1)I(0<\rho<1)$
- $\xi_1, ..., \xi_L \overset{iid}{\sim} N(0,1)$
- $\eta_1(\mathbf{s},1), ..., \eta_L(\mathbf{s},1), ..., \eta_1(\mathbf{s},W), ..., \eta_L(\mathbf{s},W)|\sigma_\eta, \phi_\eta \overset{iid}{\sim} N(0,\Sigma_\eta)$, with $\Sigma_\eta(i,j) = \sigma_\eta^2 exp(-\phi_\eta d_{ij})$
- $\mu_1, ..., \mu_L \overset{iid}{\sim} N(0,1)$
- $\gamma_1(\mathbf{s}), ..., \gamma_L(\mathbf{s})|\sigma_\gamma, \phi_\gamma \overset{iid}{\sim} N(0,\Sigma_\gamma)$, with $\Sigma_\gamma(i,j) = \sigma_\gamma^2 exp(-\phi_\gamma d_{ij})$
- $\boldsymbol{\beta} \sim N(0, I_p)$
- $\frac{1}{\sigma_\epsilon^2}, \frac{1}{\sigma_\eta^2}, \frac{1}{\sigma_\gamma^2} \overset{iid}{\sim} Gamma(a,b)$, where $a=2$ and $b=1$

This was done in order to improve the efficiency of the sampling. The sampling part is processed using a stan implementation in which the data corresponds to the weekly ozone levels recorded in our 42 stations for only 1 year and δ corresponds to the covariates (weather parameters) we want to test.

In order to select the most relevant covariates, we tried different approaches.
At first, leveraging insights from exploratory analysis and prior research, we identified key meteorological variables significantly correlated (see in Figure 4) with ozone levels. These included "Temperature_2m_mean", "Apparent_temperature_max", "Sunshine_duration", and "Shortwave_radiation_sum". The reason for this was so that we would obtain different model that we could compare using the model comparison tool provided by arviz.

Another approach was to build a model with all the available covariates, and then to keep only the ones whose associated parameter in the posterior distribution of β is relevant. By relevant it is meant that 0 doesn't belong to the posterior 95% confidence interval of the associated parameter.
By using this method, we obtained the following confidence intervals for the β :

Precipitation_sum (mm)	(-0.599079, 1.17779)
Rain_sum (mm)	(-1.14169, 0.609446)
Snowfall_sum (cm)	(-0.240372, 0.14824)
Wind_speed_10m_max (km/h)	(-0.0158849, 0.0385158)
Temperature_2m_mean (°C)	(-0.0573153, 0.413884)
Apparent_temperature_max (°C)	(-1.13284, -0.622585)
Sunshine_duration (s)	(0.0111516, 0.177066)
Wind_gusts_10m_max (km/h)	(0.0147535, 0.0742081)
Wind_direction_10m_dominant (°)	(-0.0230213, 0.00380585)
Shortwave_radiation_sum (MJ/m^2)	(-0.313512, 0.0424594)
Relative_humidity_2m (%)	(-0.170991, -0.0886358)

Table 1: Confidence intervals

As depicted in the table, the most relevant covariates seem to be "Apparent_temperature_max", "Sunshine_duration", "Wind_gusts_10m_max", and "Relative_humidity".

In this part, we noticed that the time needed to complete the MCMC was quite long so we had some concerns about the final model which we knew more computationally demanding and we started thinking of solutions. The first solution that came to our mind was to reduce the number of stations since the stations not selected could be useful later to validate or not the accuracy of the model, and the second one was to reduce the time period, since 7 years with weekly data is a lot to process. However, we could manage it in different ways : either we reduced the time period from 7 years to 4 for instance, or we could also decide to take monthly data instead of weekly (or 2 weeks by 2 weeks). We chose the first option since the second one could imply a loss of consistency in our autoregressive model since successive month are less correlated than successive weeks.

4.2. The full model

Taking into account the covariates selection and computational cost issues we talked about before, we finally decided to run the model by considering a time period that goes from January 1, 2010, to December 31, 2010. The covariates selected for this model were : "Apparent_temperature_max", "Sunshine_duration", "Wind_gusts_10m_max", and "Relative_humidity_2m".
In this final model, we decided to use a subset of the available sensor stations. Out of the total 51 stations initially present in our dataset, 9 stations were set aside as validation set. Consequently, the final model was trained and evaluated using data from the remaining 42 stations.

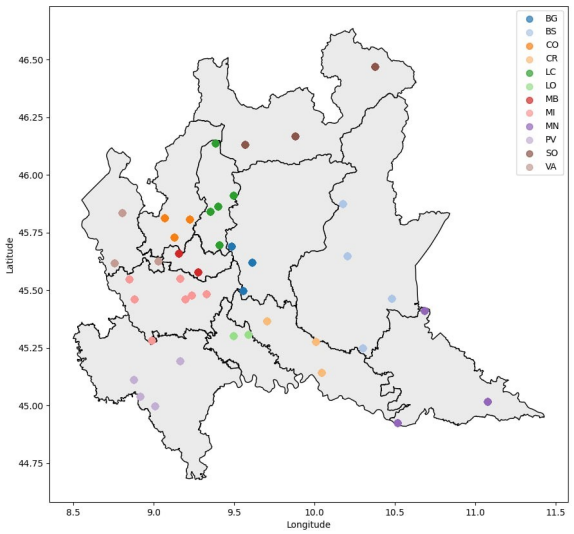

Figure 6: Final stations

5. Posterior Analysis

With the covariate selection finalized, we proceeded to tune model parameters to optimize predictive performance.

5.1. Model Checking

In the model checking phase, we used the validation set with the 9 stations that we previously set aside, to select the optimal values for spatial and temporal correlation parameters, denoted as ϕ_γ and ϕ_η respectively by computing the Validation Mean Squared Error (VMSE). As in *Sahu et al.(2007)* We conducted a two-dimensional grid search ranging from 0.004 to 0.05.

The results can be observed in the following table.

	0.004	0.005	0.01	0.05
0.004	60118.8443	79119.0544	77433.0772	50169.4225
0.005	77124.6015	69765.9095	69530.8637	53181.7744
0.01	84454.4041	79119.0544	53792.1207	57358.8004
0.05	73032.5067	78414.9593	56566.3592	71742.3862

Table 2: VMSE

In this table the values in the first row correspond to the values for the ϕ_η, while the values in the first column correspond to the ones for the ϕ_γ. It follows that the pair of values that provides the minimum VMSE is ($\phi_\eta = 0.05$, $\phi_\gamma = 0.004$), and those were the chosen ones for our model.

5.2. Model Comparison

Once we have obtained the optimal values for ϕ_γ and ϕ_η, we decided to compare models with different choices for the covariates, to verify if the method used previously provided the best result, and to see if we could obtain smaller values for the VMSE.
The first model was the one with the covariates selected based on their confidence interval, as in section 4.2, that provided a VMSE = 50169.4225.
The second model was constructed by selecting covariates according to the correlation map in figure 4, where we identified and included the most strongly correlated variables, which were "Temperature_2m_mean", "Apparent_temperature_max", "Sunshine_duration", and "Shortwave_radiation_sum". This second model resulted in a VMSE = 167018.2634.
The covariates for the third model were taken from *Sahu et al.(2007)* resulting in : "Apparent_temperature_max", "Wind_speed_10m_max", and "Relative_humidity_2m". In this third model, by running different simulations, we noticed that swapping the values of ϕ_γ and ϕ_η resulted in a significantly improved VMSE of 20.2133.

We further explore model comparison in the following section dedicated to prediction.

6. Simulation Study

In this section, we present a comprehensive graphical overview detailing the origins of VMSE (Validation Mean Squared Error) values. We have visualized the prediction values in logarithmic scale against the actual values, providing a temporal superposition of both the real and predicted values for the models under comparison. This approach facilitates a nuanced examination of the disparities in VMSE values and affords insight into the positioning of predicted and actual ozone levels.

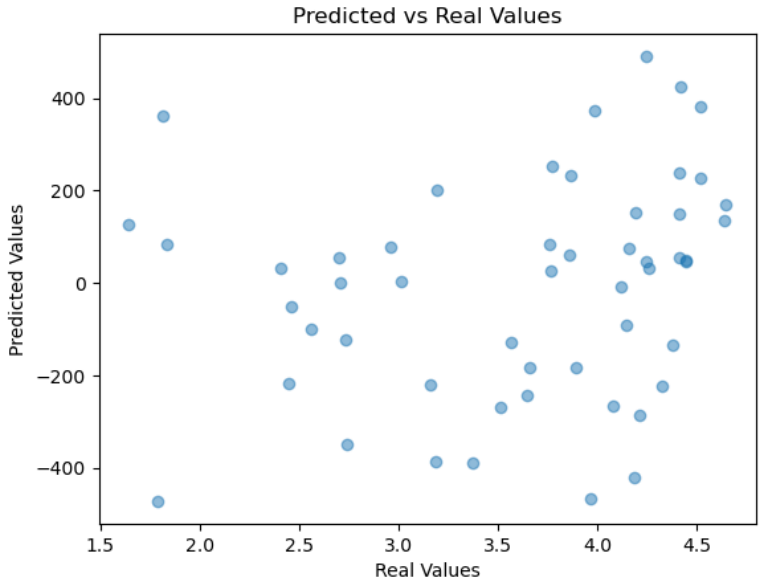

Figure 7: Model 1

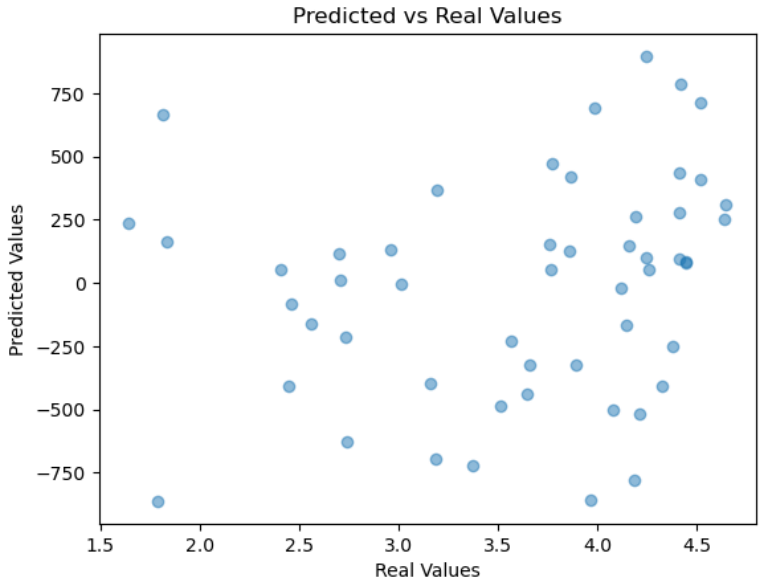

Figure 8: Model 2

The initial two plots conspicuously illustrate a substantial disparity between the predicted values and their actual counterparts, underscoring the unsatisfactory predictive capabilities of these two models. Furthermore, it is noteworthy that while the overall patterns in the cloud points remain consistent, there exists a discernible difference in the range of values. This variation, resembling a scaling effect, may be attributed to alterations in covariates, thereby providing a potential explanation for the observed discrepancy.

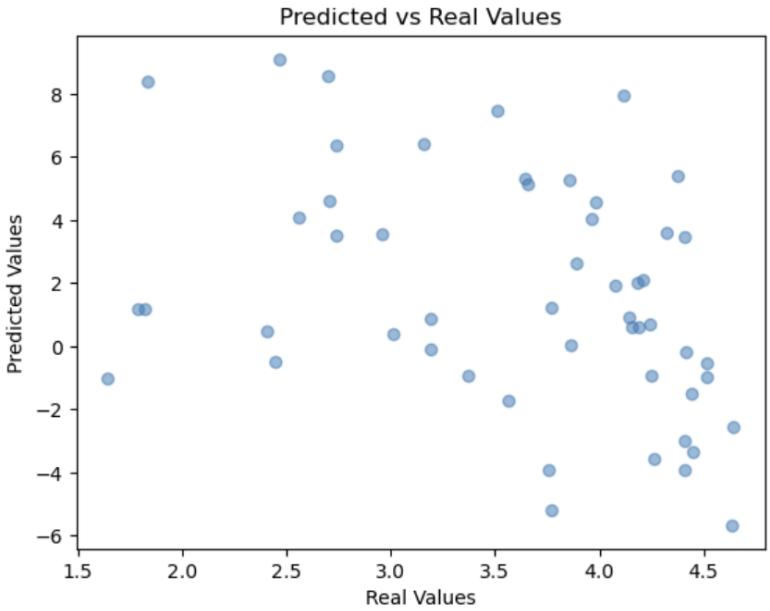

Figure 9: Model 3

Despite its imperfections, the third plot unequivocally demonstrates the superior performance of the model proposed in the paper compared to the other two counterparts. Notably, there is a noticeable contraction in the range of predicted values, albeit not aligning perfectly with the range of real values. Nonetheless, this model successfully predicts certain values within this range.

The following plots show where the real values fall inside the credible interval of the predictions.

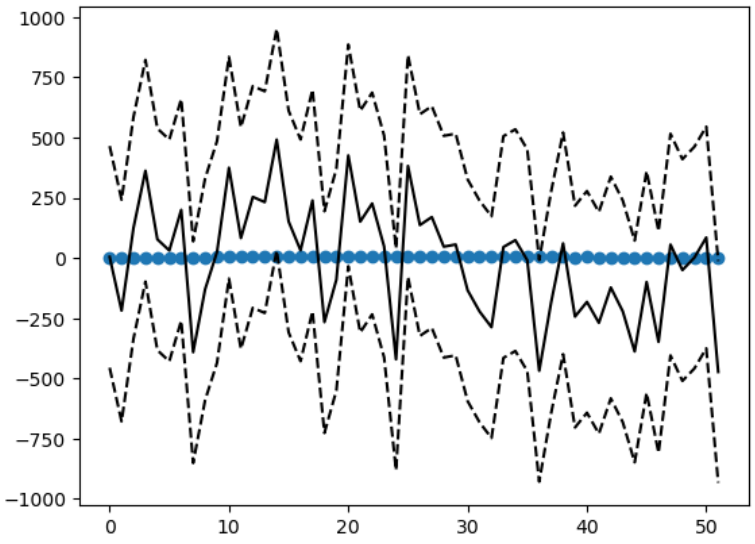

Figure 10: Model 1

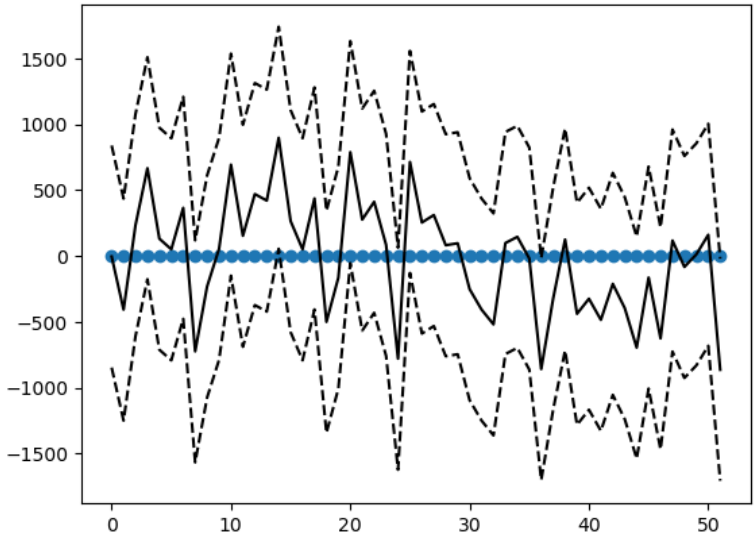

Figure 11: Model 2

The first two plots are really not informative: due to the high size of the interval, we are not even able to see any difference for the different observations to be compared.

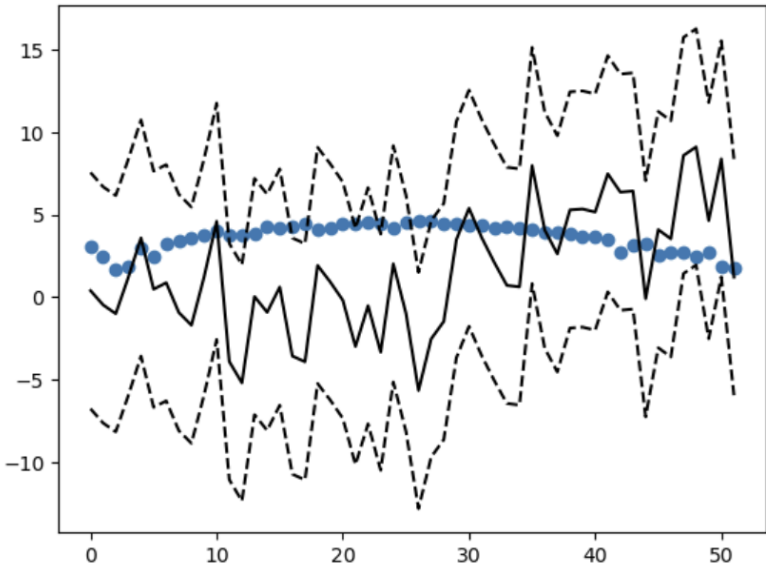

Figure 12: Model 3

The third plot presents a noteworthy observation as temporal ordering along the x-axis reveals a consistent rise in ozone level concentrations during the summer months. Regrettably, our predictive models fail to capture this prominent seasonal variation, highlighting an area for potential improvement in future iterations.

7. Discussion

Parameter	Mean	MCSE	Standard Deviation	95% CI
ρ	0.6465	0.0010	0.0174	(0.6183, 0.6751)
β_1	-0.7731	0.0095	0.0976	(-0.9237, -0.6056)
β_2	0.0335	0.0019	0.0337	(-0.0214, 0.0886)
β_3	0.0542	0.0006	0.0143	(0.0308, 0.0782)
β_4	-0.1207	0.0012	0.0212	(-0.1564, -0.0866)
$1/\sigma_\eta^2$	1.9846	0.0186	0.2635	(1.5901, 2.4512)
$1/\sigma_\epsilon^2$	4.1152	0.0028	0.0834	(3.9825 , 4.2571)

Table 3: Estimation of the parameters for model 1

$\{\beta_1,\beta_2,\beta_3,\beta_4\}$ = {"Apparent_temperature_max", "Sunshine_duration", "Wind_gusts_10m_max", and "Relative_humidity_2m"}

Parameter	Mean	MCSE	Standard Deviation	95% CI
ρ	0.6351	0.0007	0.0175	(0.6064, 0.6640)
β_1	0.5206	0.0052	0.1273	(0.3138, 0.7299)
β_2	-1.2528	0.0058	0.1356	(-1.4828, -1.0299)
β_3	0.0618	0.0020	0.0484	(-0.0181, 0.1395)
β_4	0.0532	0.0076	0.0972	(-0.1047, 0.2119)
$1/\sigma_\eta^2$	2.1825	0.0240	0.3013	(1.7477, 2.7254)
$1/\sigma_\epsilon^2$	4.1114	0.0030	0.0823	(3.9753 , 4.2448)

Table 4: Estimation of the parameters for model 2

$\{\beta_1,\beta_2,\beta_3,\beta_4\}$ = {"Temperature_2m_mean", "Apparent_temperature_max", "Sunshine_duration", "Shortwave_radiation_sum"}

Parameter	Mean	MCSE	Standard Deviation	95% CI
ρ	0.6493	0.0010	0.0176	(0.6205, 0.6784)
β_1	-0.8590	0.0134	0.0977	(-1.0189, -0.6937)
β_2	0.0315	0.0006	0.0132	(0.0105, 0.0533)
β_3	-0.1203	0.0012	0.0213	(-0.1572, -0.0855)
$1/\sigma_\eta^2$	0.3007	0.0117	0.0730	(0.2070 , 0.4408)
$1/\sigma_\epsilon^2$	3.9786	0.0081	0.0858	(3.8414 , 4.1195)

Table 5: Estimation of the parameters for model 3

$\{\beta_1,\beta_2,\beta_3\}$ = {Apparent_temperature_max", "Wind_speed_10m_max", and "Relative_humidity_2m"}

As we can see from the tables, the values of ρ suggest that there exists a strong correlation between ozone concentrations in successive weeks, as it was expected.
By looking at the estimates of σ_η^2 and σ_ϵ^2, we can see that more variation is explained by the spatiotemporal effects than by the pure error process.

References

[1] Open meteo. `https://open-meteo.com/en/docs/historical-weather-api`.

[2] Hussein A Al-Amery and Osama T Al-Taai. The ozone effect on shortwave solar radiation in the atmosphere over iraq. In *AIP Conference Proceedings*, volume 2290. AIP Publishing, 2020.

[3] Amandine Chevalier, François Gheusi, Robert Delmas, Carlos Ordóñez, Claire Sarrat, Régina Zbinden, Valérie Thouret, Gilles Athier, and J-M Cousin. Influence of altitude on ozone levels and variability in the lower troposphere: a ground-based study for western europe over the period 2001–2004. *Atmospheric Chemistry and Physics*, 7(16):4311–4326, 2007.

[4] Sujit K Sahu, Alan E Gelfand, and David M Holland. High-resolution space–time ozone modeling for assessing trends. *Journal of the American Statistical Association*, 102(480):1221–1234, 2007.