Luca Ordronneau

Air Quality in Catalonia

Ocean Data Challenge

Discord :
lucanew#3793

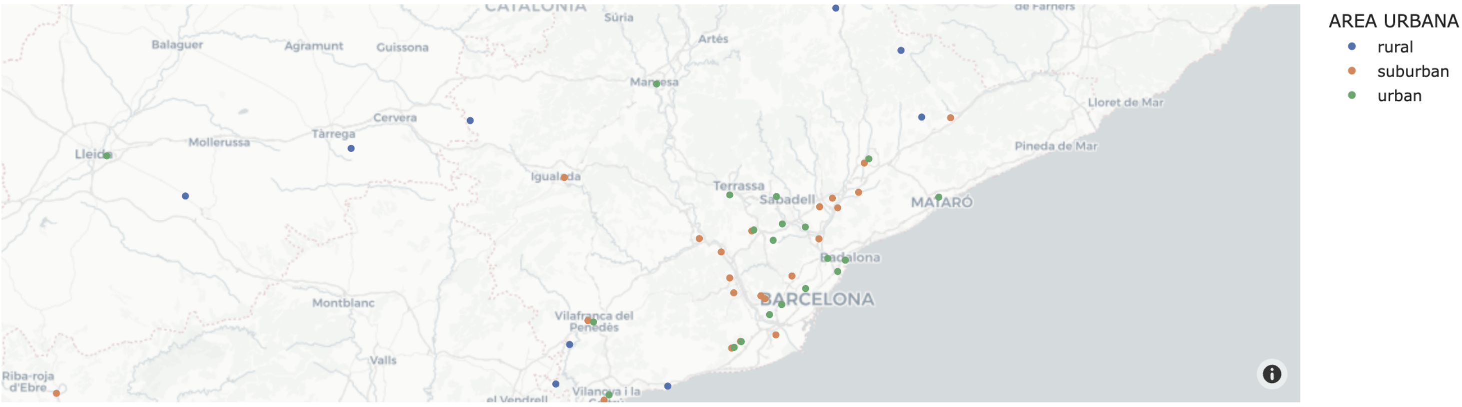
GitHub :
lucaordronneau

Version of
February 12, 2023

Introduction

Air pollution is a serious issue that affects millions of people worldwide, and it is essential to analyze its evolution to inform strategies to reduce its impact. Data scientists can play a significant role in this process by providing comprehensive analytics and predictive algorithms based on open data. This report describes a challenge that aims to analyze the evolution of air pollutants in Catalonia over the past three decades and develop algorithms to predict air pollutant concentrations. The dataset contains hourly measurements of air pollution from 1991 to the present, collected by the automatic measuring stations of the Air Pollution Monitoring and Forecasting Network. In this report, we present a global analysis of the air quality in Catalonia, results from predictive algorithms that were built to predict the concentration of pollutants in the air using Ocean Protocol's Compute-to-Data technology, and provide recommendations based on our findings. The results of this study can inform policies to mitigate the effects of air pollution in Catalonia and beyond.

Exploring Urbanization in Catalonia: A Map of Rural, Suburban, and Urban Areas



Methodology

Data Preparation

In order to analyse the dataset, I had to performs various data preparation steps. The data preparation removes rows that do not have the molecular weight to convert ppm into µg/m3, converts all others units to µg/m3, and transforms the data from wide to long format. It also fills missing values with the mean of the hour of the same month and year from the same station (CODI EOI).

Features

The process of global analysis involves manipulating data to create two-dimensional plots for extracting various information. In my work of developing algorithms to predict pollutant concentrations, I had to create specific features and the insights from global analysis were invaluable in this process. The following are the steps I took for feature creation:

- Select the pollutant of interest, in this case NO, from the input DataFrame. (present from 1991 until today and has a high correlation with other polluants).
- Create a pivot table to calculate the mean concentration of all pollutants associated with each type of urban area (i.e., urban, suburban, rural) for each hour/month.
- Create a new feature for each type of urban area that represents the mean concentration of NO associated with that area for each hour/month.
- Create a pivot table to calculate the mean altitude for all pollutants for each hour/month.
- Create a new feature that represents the mean altitude associated with NO for each month.
- Create cyclic features for the hour, day, month (sin and cosine) to capture the cyclical pattern of hourly, daily and monthly changes.

Model : SARIMAX

I used SARIMAX model to forecast the concentration of pollutant NO for the next 24 months and between 15-02-2023 and 28-02-2023. SARIMAX is a popular time series forecasting model that is useful for modeling and predicting data that exhibits seasonal or periodic patterns.

SARIMAX allows the modeling of both the trend and seasonality in the data. Additionally, it is useful for modeling time series data with a known seasonal pattern, such as monthly or quarterly data. The model also allows for the inclusion of exogenous variables (features). These exogenous variables can be useful for improving the accuracy of the forecast by capturing external factors that may influence the concentration of the pollutant.

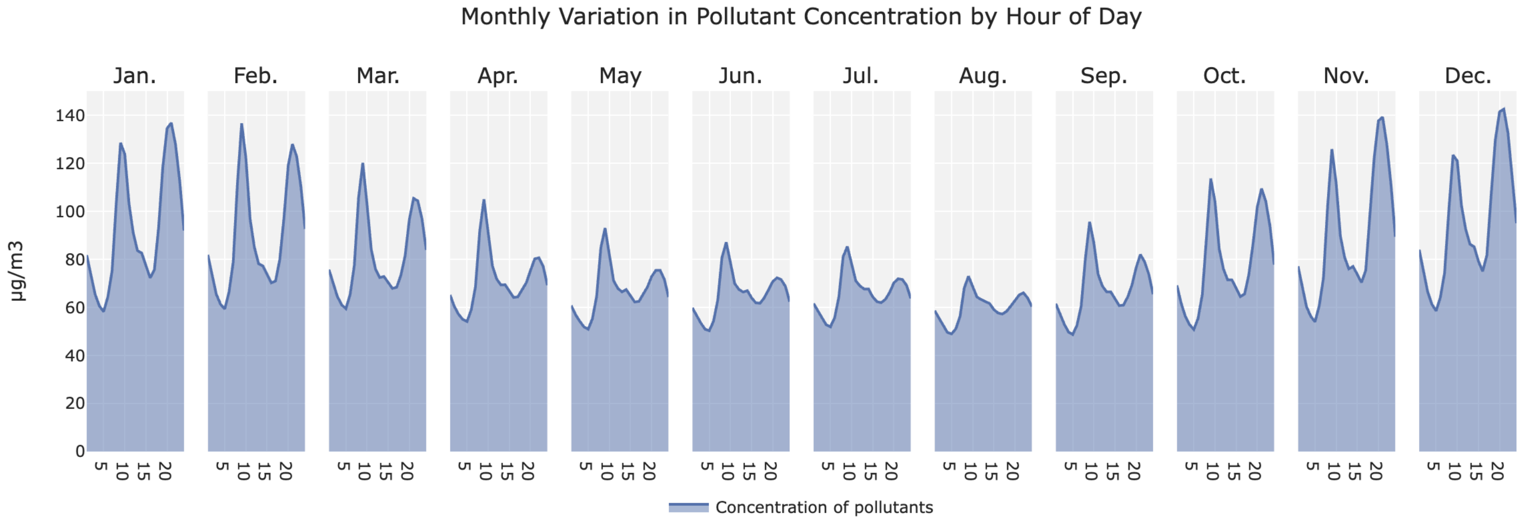
Global Analysis and Predictions results

Best/Worst hours, months in terms of pollution

On average, the best hours for all pollutants were found to be between 2am and 6am with an average concentration of 58.07 µg/m3. The worst hours, with an average concentration of 99.85 µg/m3, were identified as 9am, 10am, 8pm, 9pm, and 10pm. This periodicity is likely due to increased traffic and activity during typical working hours.

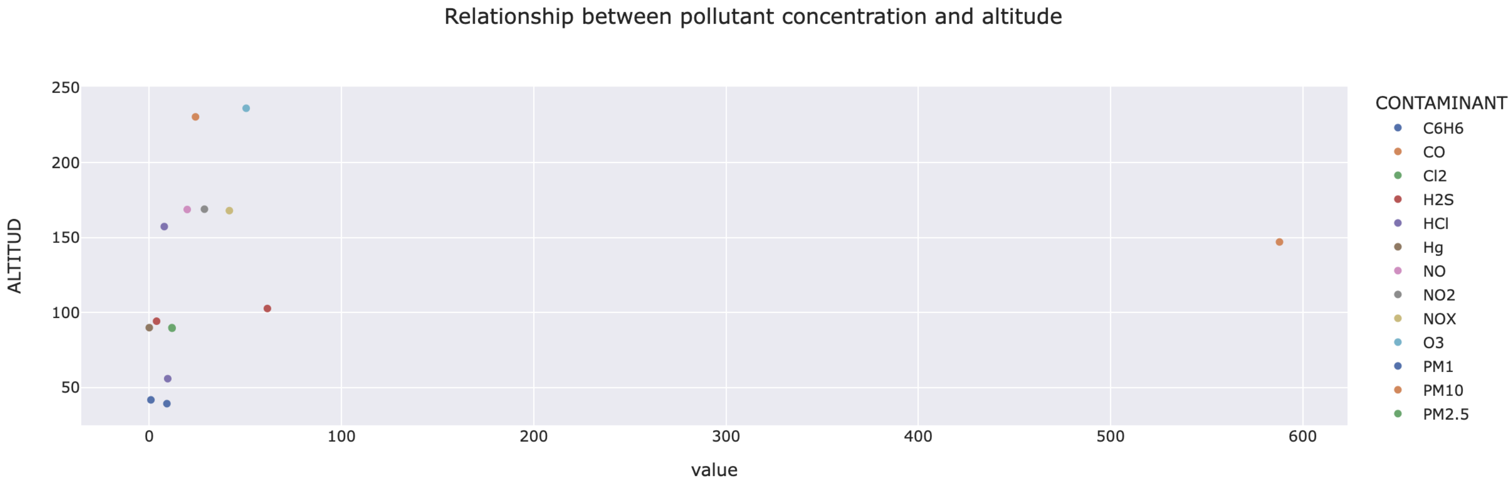
In terms of months, the best periods were identified as May through September with an average concentration of 65.28 µg/m3, while the worst months were November through March with an average concentration of 90.06 µg/m3. This is likely due to a combination of factors, including increased heating and energy consumption during the winter, as well as atmospheric conditions that trap pollutants close to the ground.

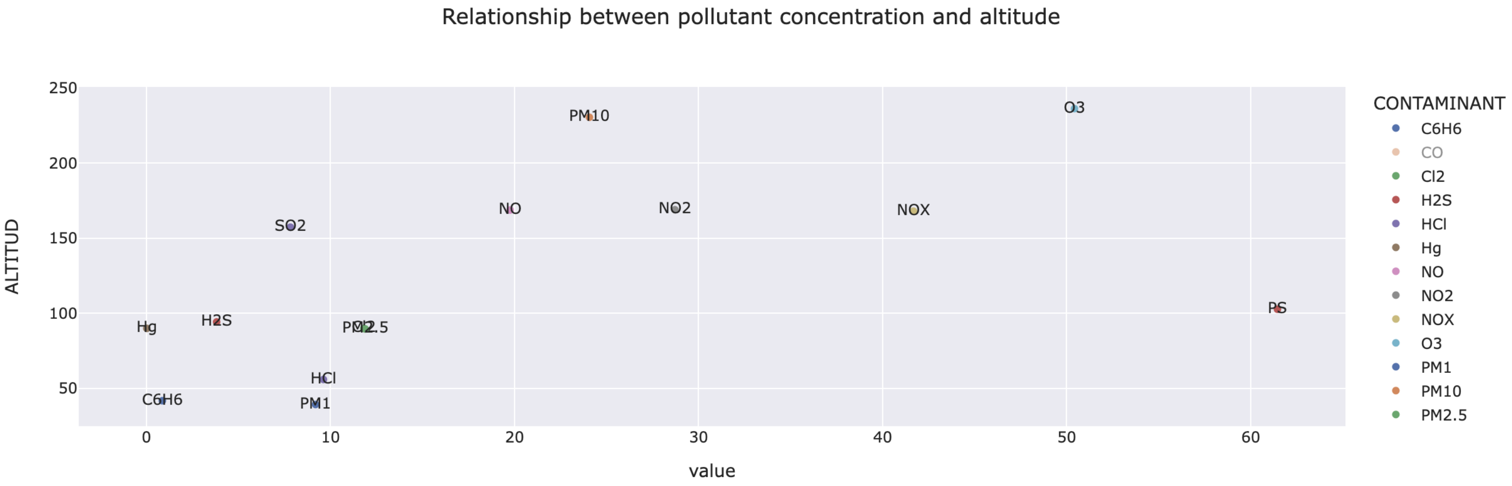
On the x-axis of each subplot, the hours of the day are represented, while the y-axis shows the average concentration of pollutant. Additionally, each subplot is labeled with the month of the year, providing a comprehensive view of the evolution of pollution levels throughout the year.



Relationship between altitude and concentration of particles in the air

After analyzing the relationship between altitude and pollutant concentration in the air, a small correlation was observed between the two variables. The analysis showed that, on average, the contaminants with higher concentration tended to be found at higher altitudes. However, this correlation was not significant for all the pollutants, and when removing CO, the correlation becomes more apparent. These results suggest that the relationship between altitude and pollutant concentration is not straightforward and is influenced by several factors.


Relationship between pollutant concentration and altitude


Relationship between pollutant concentration and altitude

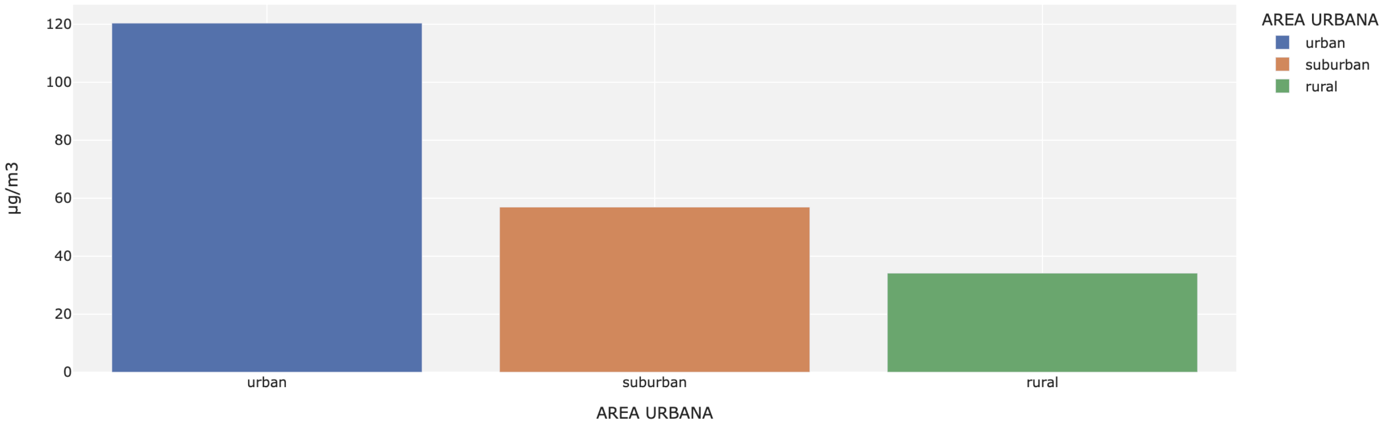
Concentration of pollutants in urban, suburban and rural areas

Based on the plots, it appears that there is a significant difference in the concentration of pollutants in urban, suburban, and rural areas. The urban areas have the highest concentration of pollutants, followed by suburban areas and then rural areas. This trend can be attributed to the difference in activities and human density in the different areas. Urban areas tend to have high levels of traffic and industrial activities, which lead to a higher concentration of pollutants in the air. In contrast, rural areas tend to have less traffic and industrial activities, resulting in lower levels of pollutants.

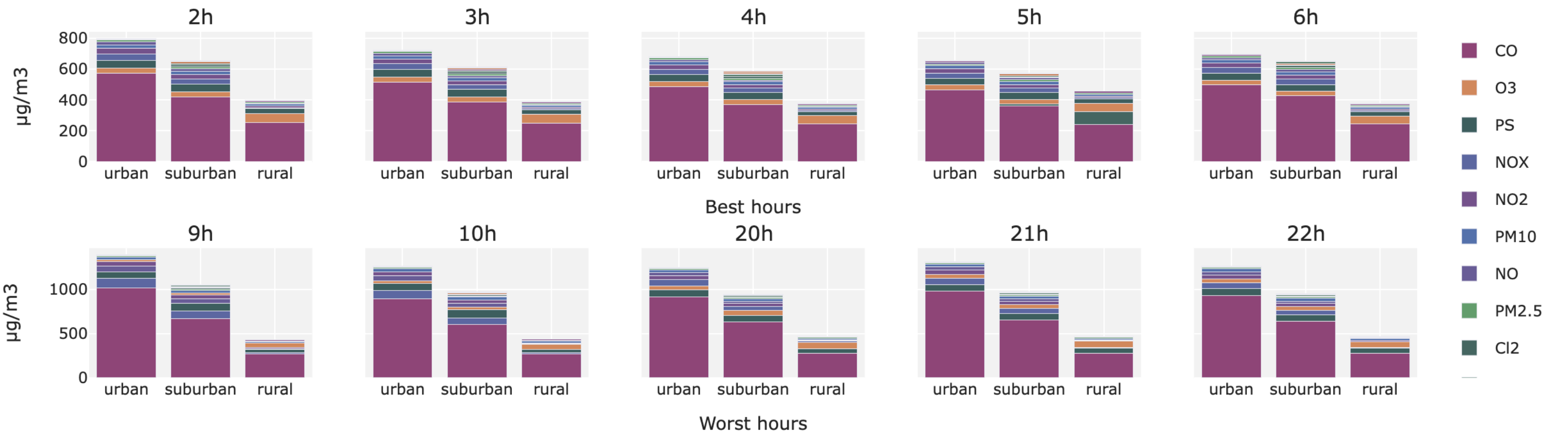
The graphs presented also show that CO and O3 are two of the most prominent pollutants in all three areas. The high levels of CO in all areas could be attributed to the high level of vehicular traffic, especially in urban areas, where there is a higher concentration of cars on the road.

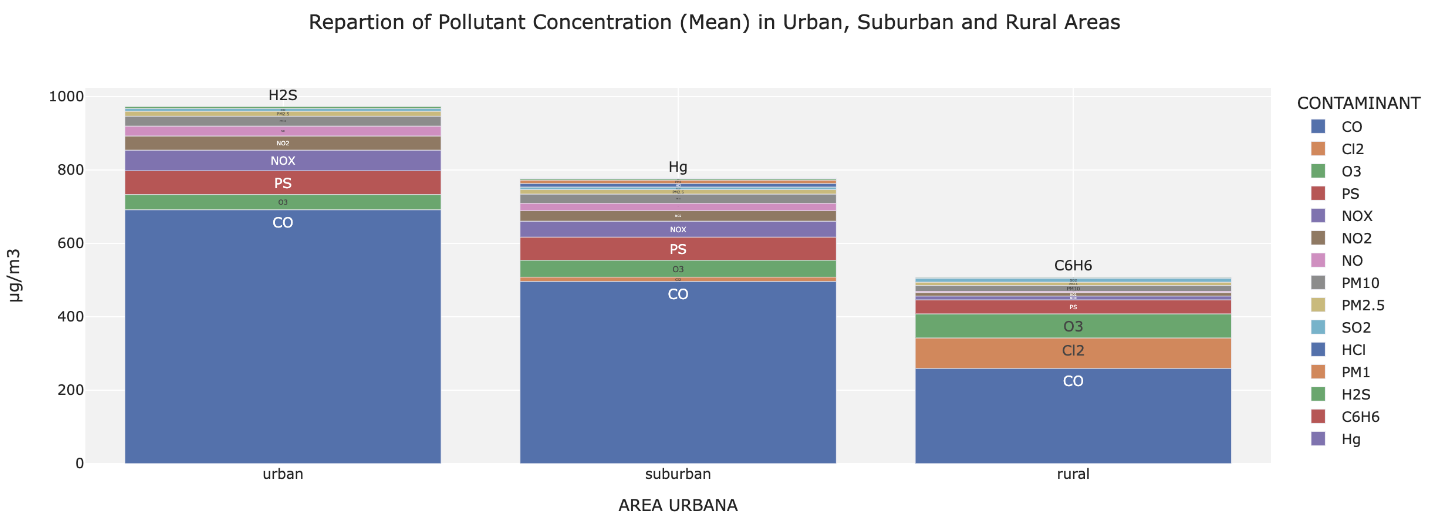
Overall, the plots suggests that urban areas have the highest concentration of pollutants, with traffic and industrial activities being the major contributors.

Pollutant Concentration (Mean) in Urban, Suburban and Rural Areas

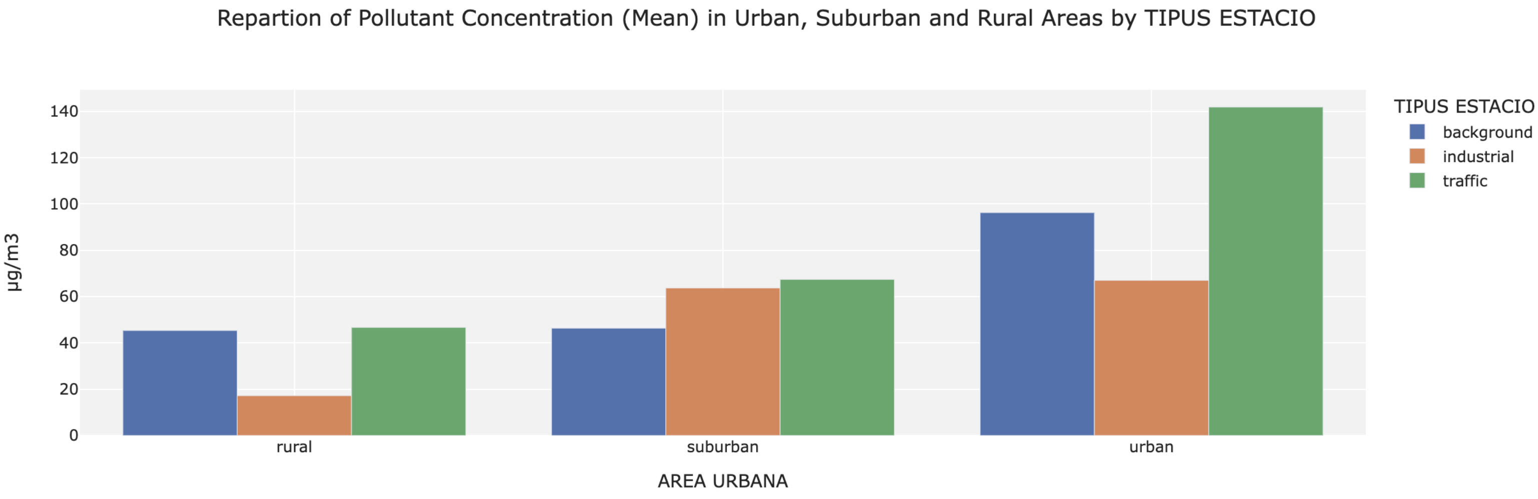


Comparative Analysis of Pollutant Concentration (Hours Of Interest) in Urban, Suburban and Rural Areas

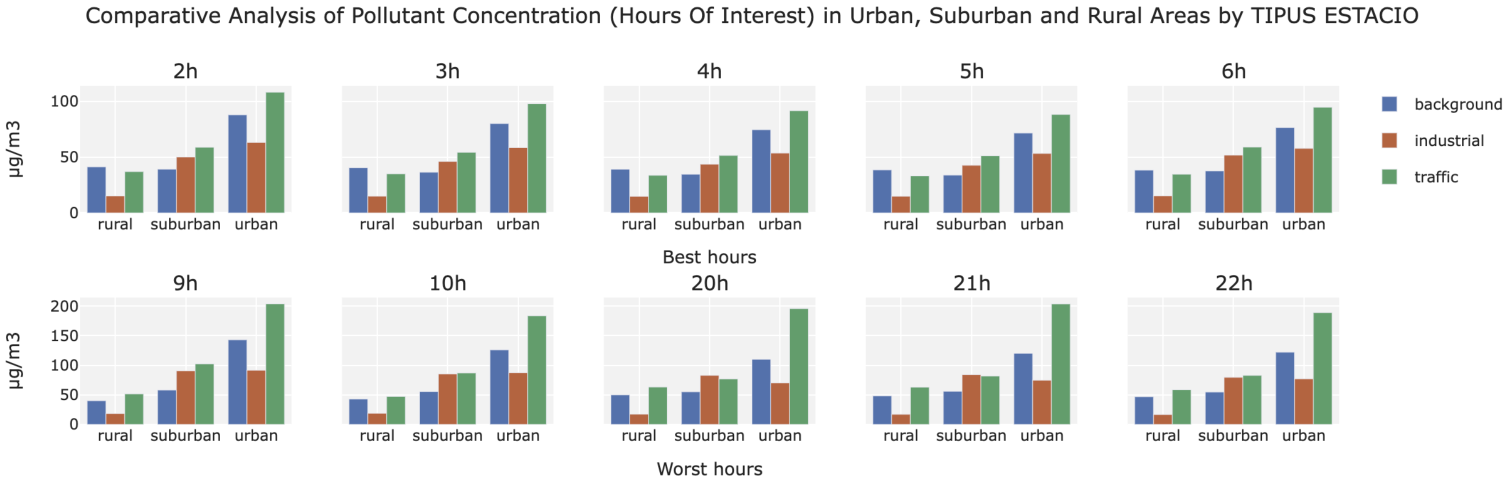




Repartion of Pollutant Concentration (Mean) in Urban, Suburban and Rural Areas



Repartion of Pollutant Concentration (Mean) in Urban, Suburban and Rural Areas by TIPUS ESTACIO



Comparative Analysis of Pollutant Concentration (Hours Of Interest) in Urban, Suburban and Rural Areas by TIPUS ESTACIO

Best/Worst cities in terms of pollution

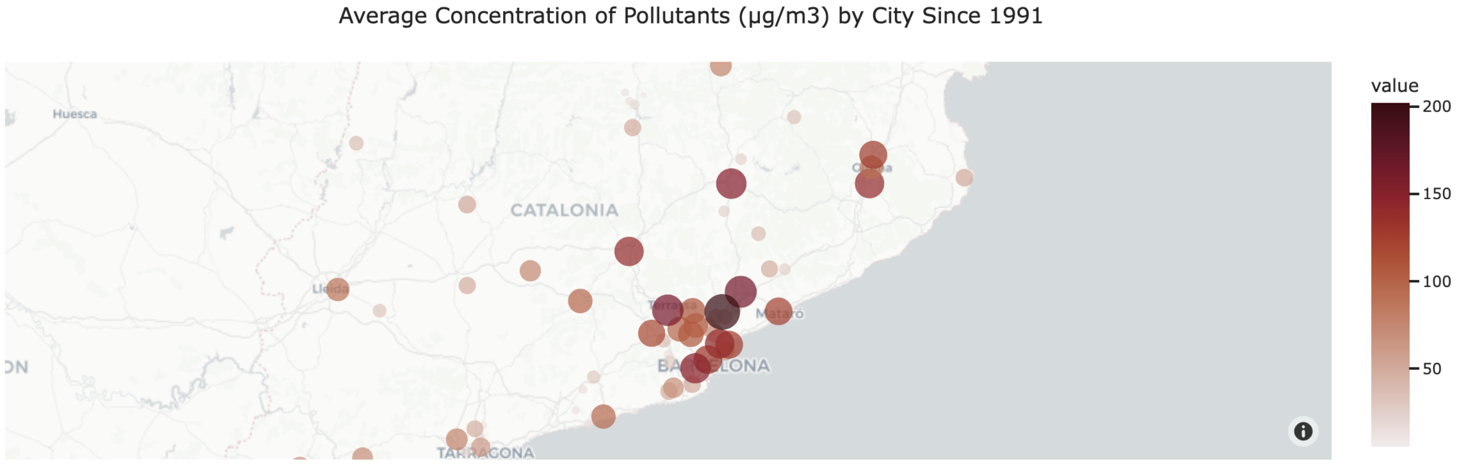
All cities

The best five cities, with the lowest average concentration of pollutants, are Flix, Nou de Berguedà, Alcanar, Sitges, and Perafort, with an average concentration of 7.579473. On the other hand, the worst five cities, with the highest average concentration of pollutants, are Hospitalet de Llobregat, Vic, Terrassa, Granollers, and Mollet del Vallès, with an average concentration of 163.817876.

Urban cities

The best five urban cities, with the lowest average concentration of pollutants, are Cornellà de Llobregat, Vilafranca del Penedès, Gavà, Ripollet, and Rubí, with an average concentration of 47.194237. On the other hand, the worst five urban cities, with the highest average concentration of pollutants, are Hospitalet de Llobregat, Vic, Terrassa, Granollers, and Santa Coloma de Gramenet, with an average concentration of 160.121089.

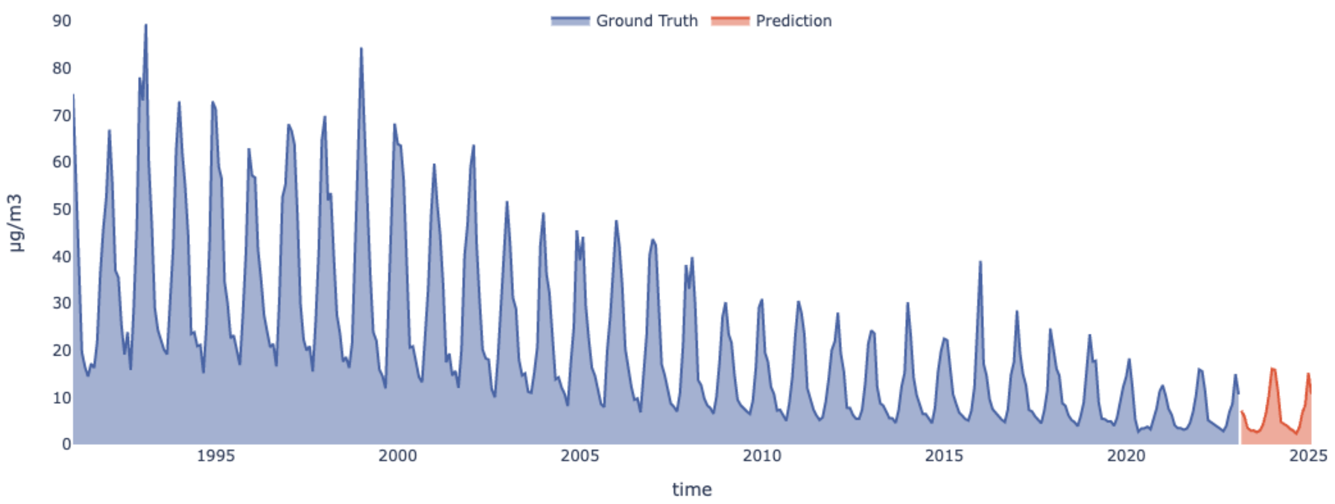
The ranking of cities based on pollution levels is influenced by various factors as we saw, such as the size of the city, its industrial activity, and the sources of pollution.



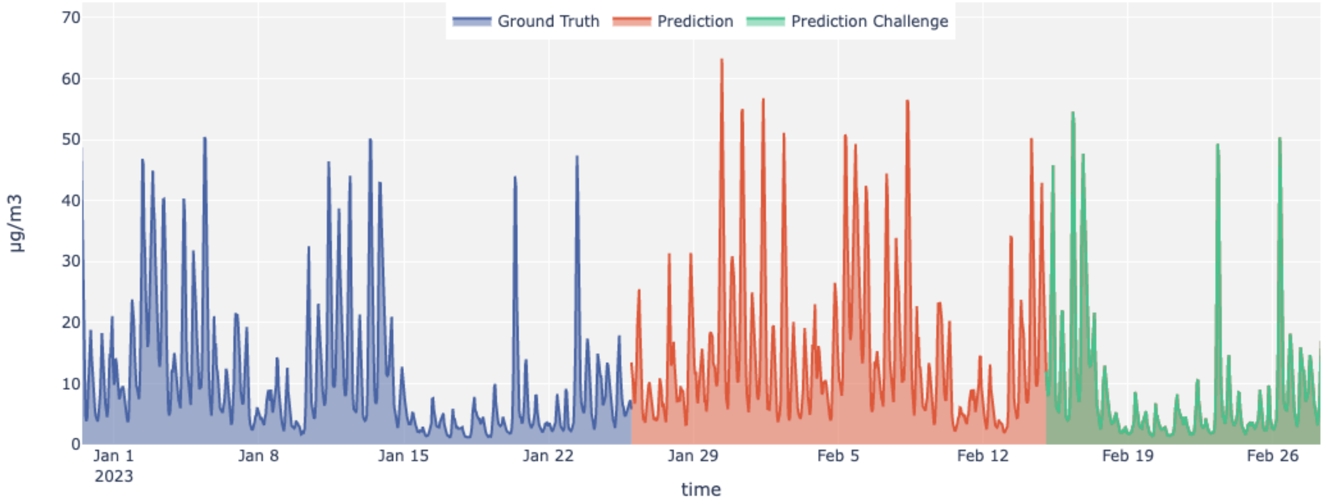
Predictions results

Below are the predictions for the pollutant NO for the next 24 months, as well as the average hourly predictions for all stations between 15-02-2023 and 28-02-2023. The trend appears to be consistent.

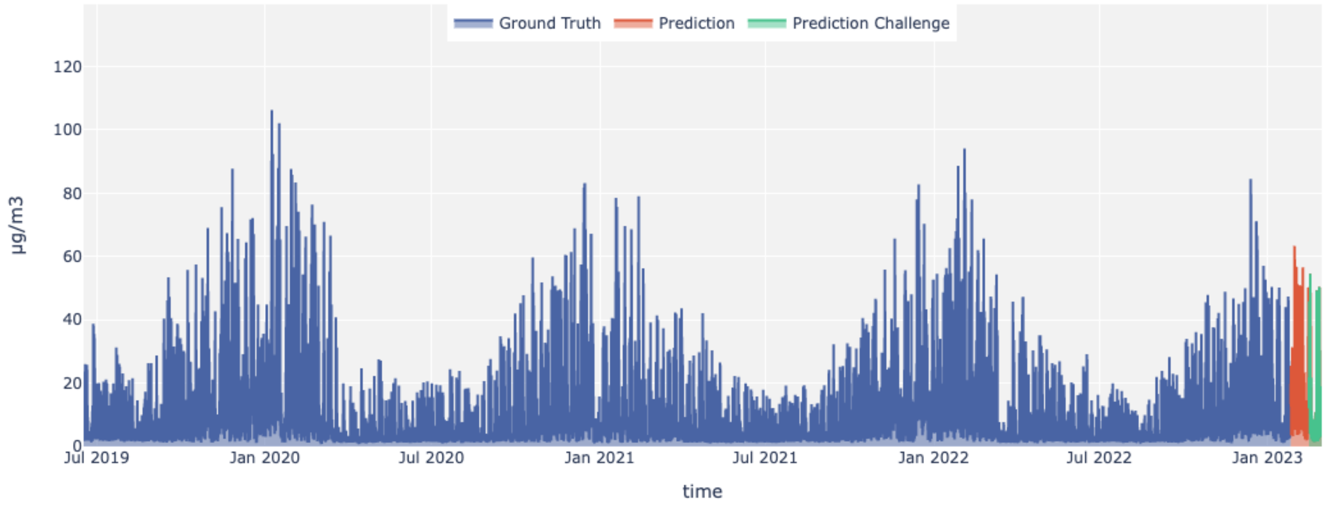
Prediction of the evolution of the polluant NO for the next 24 months



Prediction of the evolution of the polluant NO between 15-02-2023 and 28-02-2023



Prediction of the evolution of the polluant NO between 15-02-2023 and 28-02-2023



Observations and Conclusions

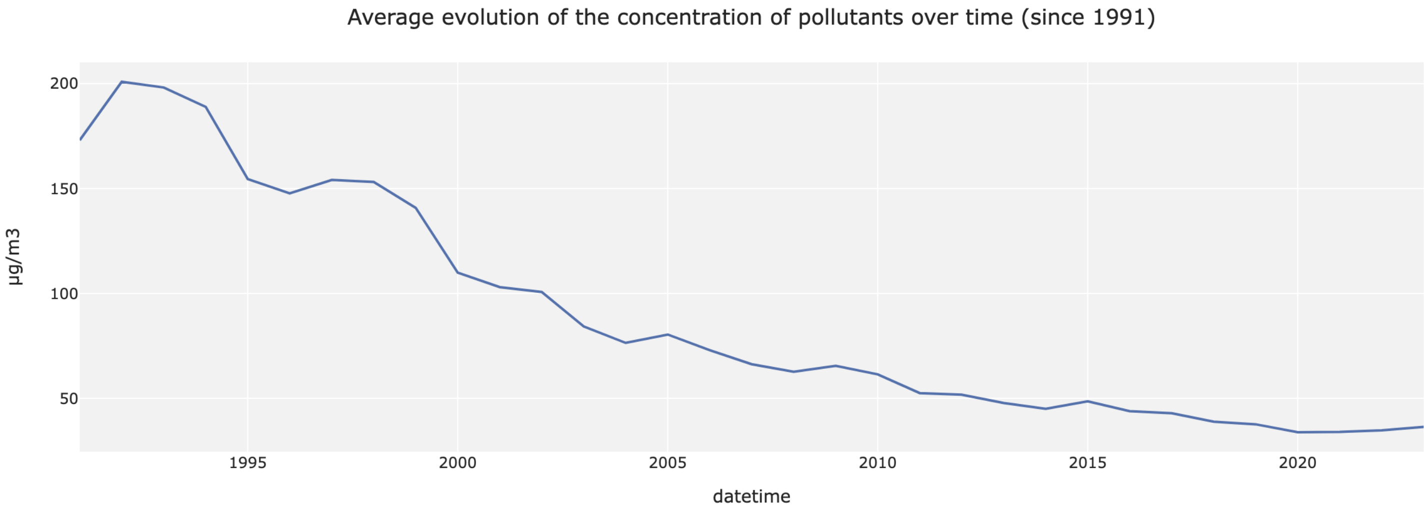
The analysis of air pollution in Catalonia over the past three decades revealed that the concentration of pollutants in the air varies depending on the time of day, month, altitude, and location. The worst hours for all pollutants were found to be between 9am and 10pm, likely due to increased traffic and activity during typical working hours. The worst months were from November through March, likely due to increased heating and energy consumption during the winter, as well as atmospheric conditions that trap pollutants close to the ground.

The analysis also showed a small correlation between altitude and pollutant concentration, with contaminants with higher concentrations tending to be found at higher altitudes. The relationship between altitude and pollutant concentration is influenced by several factors and is not straightforward.

Urban areas were found to have the highest concentration of pollutants, followed by suburban areas and then rural areas, which can be attributed to the difference in activities and human density in these areas. CO and O3 were identified as two of the most prominent pollutants in all three areas, with traffic and industrial activities being the major contributors. The ranking of cities based on pollution levels was influenced by various factors such as the size of the city, its industrial activity, and the sources of pollution.

These findings can inform policies to mitigate the effects of air pollution in Catalonia and beyond. Strategies could be implemented to reduce traffic and industrial activities in urban areas and to promote the use of clean energy in winter.

And if we take a step back, we can still see that we have reduced air pollution since 1991 and that a lot of efforts have been made as shown in the graph below.



Limitations and Recommendations

Challenge

The pollution monitoring challenge we recently tackled was both engaging and timely, given the ongoing need to assess and manage air quality. The task presented interesting analytical challenges.

One notable limitation was the lack of molecular mass data, which made it impossible to convert parts per million (ppm) measurements into micrograms per cubic meter (µg/m3). Without molecular mass data, we were unable to accurately compare the pollutant levels across different stations, which made it difficult to identify overall trends or forecast future levels.

Ocean Protocol

This was my first experience using the Ocean Market platform, and I found it to be a promising platform, especially for data sharing. The platform's user interface is straightforward and easy to use.

In the context of the challenge, one feature that could be interesting to develop is the possibility to have logs when using C2D. This feature would allow users to track the progress of their jobs and identify any issues that may arise during the execution of the code.

Another feature that could be helpful is the ability to cancel jobs that are running. This feature would be useful in situations where the user has mistakenly started a job, or the job is taking longer than expected.

Although not a priority, it would be helpful to have the option to add a complete folder rather than having to upload just one file. This feature would save users time when working with large datasets and complex directory structures (cleaning file, feature preparation file...).

Overall, I think that the Ocean Market platform has a lot of potential, and I look forward to using it again in the future.