The Silent Pandemic: Air Pollution in European Cities

Ivana Drabova, Ricard Punsola

December 2023

1 Introduction

In 2022, 98% of people in Europe lived in areas where air pollution exceeded the World Health Organization (WHO) recommendation of (5 $\mu g/m^3$) [DW 23]. According to a new study by the Barcelona Institute for Global Health, air pollution is the largest environmental cause of mortality in Europe [Eur23]. An estimated 300,000 to 500,000 people die prematurely in the EU each year as a result of air pollution [Nc23] [DW 23]. Air pollution is linked to heart disease, stroke, diabetes, and chronic pulmonary disease. Furthermore, air pollution decreases life quality through chronic diseases such as asthma or dementia [Nc23] [DW 23].

Air quality is measured by the particulate matter (PM), the solid and liquid particles in the air. PM 2.5 are particles with a diameter of less than 2.5 micrometers. Due to its small size, PM 2.5 can penetrate deep into the lungs and bloodstream and have a negative impact on the human body. The WHO, which tightened its air quality guidelines in 2021, warns that no level of air pollution can be considered safe but has set upper limits for certain pollutants at 5 $\mu g/m^3$ [DW 23]. Unfortunately, there is a lack of political will in the European Parliament, which voted in September 2023 to align the EU's air quality rules with the WHO's but in the end decided to delay doing so until 2035. Thus, air pollution is an invisible and largely ignored health hazard with no policy improvement in sight for the next decade in the EU.

The top sources of air pollution are traffic, coal and gas-related heating systems, and agriculture. However, it is worth noting that the results vary from city to city [Eur23]. Thus, we conduct our analysis on the city level.

We construct a novel data set with hourly pollution data for 227 European cities, spanning almost 3 years (Fall 2020 to Fall 2023). The data set marks the first contribution of our paper. The second contribution of our paper lies in creating infographics on the city level. We compare different cities in Europe as well as determine what is the best time during the day to spend time outdoors. Thirdly, we contribute to the body of research on the Environmental Kuznets Curve by analyzing the relationship between income and pollution on the city level. Fourthly, we analyze the relationship between pollution and city size. Lastly, in the case study of Barcelona, we aim to uncover and predict the relationship between local weather and pollution.

2 Literature review

The relationship between income and level of environmental degradation is widely researched in Environmental Economics. Environmental Kuznets Curve (EKC) is one of the leading hypotheses for describing the relationship between income and pollution. The EKC hypothesis suggests that as countries develop, they become more capable of investing in sustainability, and an initial increase in environmental damage is followed by a decrease in environmental damage. Therefore, as a country's income increases, there is expected to be an eventual decrease in environmental damage with the relationship resembling and inverted U-shaped curve. However, the validity of the EKC has been widely discussed: 'Truly, the relationship between economic growth and environmental quality has been an object of a long debate for many years' [Din04]. We extend the literature on Environmental Economics with our analysis of the relationship between GDP and air pollution on the city level. Furthermore, we contribute to the field of Urban Economics by studying the relationship between pollution and population.

3 Data

Our original data set on air pollution, GDP and population was constructed using data from Open Weather Map (using APIs) and Eurostat.

The data set contains 5,572,956 observations – the hourly measurements of air pollution (PM 2.5) [1] in 227 European cities. The cities are located in 28 countries and represent the total population of 239,955,890 citizens. Description of the data set is provided in Table 4. The correlation between the dataset variables can be found in Figure 8.

3.1 Eurostat Data

Eurostat provides data on population and GDP in European cities [Eura] [Eurb]. We used the GDP and population data for the year 2019, as it is the most recent data with relatively little missing data.
The main challenge with the Eurostat data was

1. City-level data and country-level data were mixed in the same data set.
2. City names were in the local language (e.g., *Praha* for Prague)
3. Latitude and longitude of the cities were not provided.

3.1.1 Challenges

The solutions to the aforementioned problems were as follows:

1. Countries were manually deleted from the dataset. Only cities remained.
2. Using Geocoding API, [Opeb] the city names were translated from local names to English. Special attention was paid to those cities, whose name is the same in different countries (e.g., Cordoba in Spain and Argentina). A manual check needed to be made for each city – of

[1] Fine particulate matter of concentration 2.5 micrograms per cubic meter (PM 2.5)

course, we used our coding skills to help us, such as detecting when in a sequence of cities from country X there was suddenly a city from country Y.

3. Having translated the city names, we obtained latitude and longitude of each city using the Geocoding API [Opeb]. Each city was then plotted on the map to uncover and delete cities of countries that are in the EU, but not in the European continent (e.g. some French territories and islands).

3.2 Open Weather Data

For each city identified by latitude and longitude, we used Open Weather API to download hourly data on weather and pollutants [Opea] [Opec]. The data on weather is available from Fall 2022 to Fall 2023. The data on pollution is available from November 2020 to September 2023.

4 Pollution in European cities

In Figure 1, we construct an interactive visualization of air pollution on a map. The safe limit set by the World Health Organisation is set at 5 micrograms per cubic meter. Many of the cities, however, surpass this limit. Cities in the Nordic countries and the south of Spain do relatively well and are close to the WHO limit of 5 micrograms per cubic meter of concentration of PM 2.5. Cities such as Paris and Milan are particularly affected.

4.1 Relationship between pollution, population and income

Following the literature on the Environmental Kuznets Curve [CRB97], testing for an inverted U-shape requires fitting a polynomial model of degree 2.

$$pollution_c = \beta_0 + \beta_1 ln(GDP_c) + \beta_2 ln(GDP_c)^2 + e_i$$

Where $pollution_c$ is the city-level average value of air pollution measured by PM 2.5 micrograms per cubic meter in the years 2020 - 2023, and GDP_c is the gross domestic product of the city in the year 2019. Since we have data for PM 2.5 as well as other pollutants, such as CO, NO, PM 10 and SO2, we include these in our models. Our second model normalizes the GDP by population as it is an alternative approach in the literature on the Environmental Kuznets Curve. Our third model changes GDP for the population at the city level as we are interested in whether more populated cities are more polluted.

$$pollution_c = \beta_0 + \beta_1 ln(population_c) + \beta_2 ln(population_c)^2 + e_i$$

We do not find evidence for the Environmental Kuznets Curve (EKC). Figure 2 and Table 1 summarize the results. The interpretation of these models is that a one percent increase in the independent variable (population or income) is associated with an increase (or decrease) of the pollution level variable by (coefficient/100) units [2]. The squared term signifies the curvature, which means that the effect of population or income is allowed to change with its level. The marginal change denoted by the derivative of the equations is: $\beta_1 + 2\beta_2 ln(income)$. If the coefficient of the squared term is negative, the function is concave with a reverse U-shape such as we would want to see if EKC

[2] linear-log model

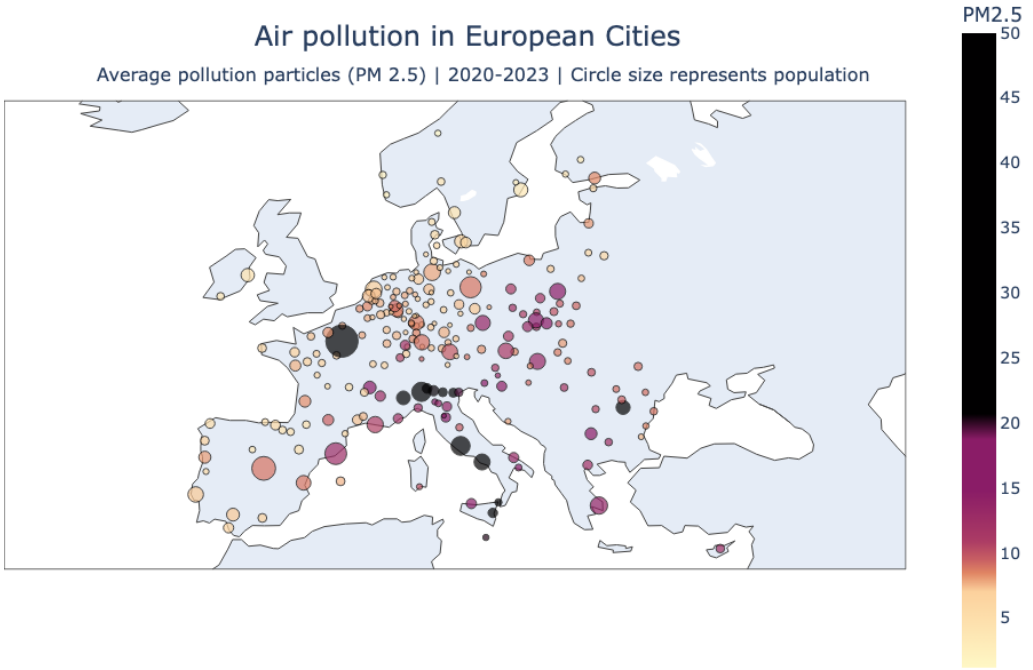


Figure 1: Interactive graph of air pollution (PM 2.5) in 227 European cities. The size of the circle indicates the population of the city. The air pollution variable is an average for all the years observed (2020-2023).

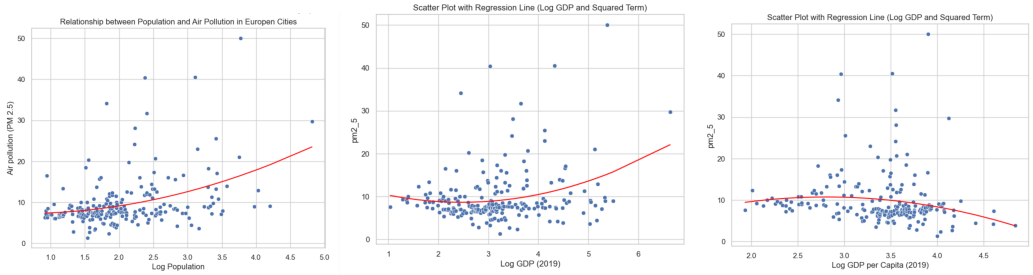
was true. If the coefficient on the squared term is positive, the function is convex with a regularly U-shaped curve.

In Table 1, for the dependent variable PM 2.5, neither the coefficient on income nor population is significant on the 95% confidence level. In Figure 2a and 2b, we see a slightly convex relationship between population and GDP, however, the linear coefficients are not significant. As we can see from the graph, there is no clear pattern. The reversed U shape, hypothesized by the EKC, can only be found in Figure 2c, but the coefficients are not significant, and thus there is no clear pattern in the data.

For other pollutants in Table 1, namely CO and NO, we find significant relationships between CO and NO with both linear and quadratic term significant. However, we do not find an inverted U-shape as Environmental Kuznets Curve would suggest. On the contrary, the positive value on the quadratic term suggests a convex relationship, that is, the richer cities have increasingly more pollution of NO and CO (see Figure 3). For other pollutants (NO2, PM10, and SO2), we do not find a clear relationship (i.e. a relationship significantly different from zero). Therefore, our investigation yields no support for the EKC and in the case of some specific pollutants, it contradicts it by finding a regular U-shaped relationship instead of an inverted U-shaped relationship.

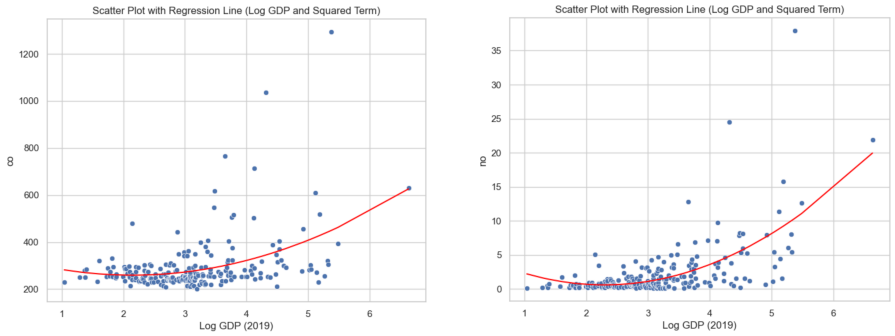
Table 1: Regression results of various pollutants on population and income variables, respectively. Stars denote the significance: * significance at p = 0.1, ** significance at p = 0.05, *** significance at 0.001

	PM 2.5	PM 2.5	PM 2.5	CO	NO	NO2	PM10	SO2
Const.	7.51*	13.36***	-2.57	343.53***	5.99**	8.13**	15***	4.5
log(GDP)		-3.82*		-78.32**	-4.71***	-1.85	-3.63	-1.14
$\log(\text{GDP})^2$		0.77		18.225***	1.03***	0.78**	0.79**	0.24
log(pcGDP)			9.46					
$\log(\text{pcGDP})^2$			-1.68					
log(pop)	-1.00							
$\log(\text{pop})^2$	0.90*							



(a) PM 2.5 and population (b) PM 2.5 and GDP (c) PM 2.5. and GDP per capita

Figure 2: Relationship between pollution (PM 2.5.) and city population, city GDP and city GDP per capita. There is no clear relationship in the data, as can see on the graph as well as from the Table 1, which shows statistically insignificant coefficients. Thus we find no evidence for the Environmental Kuznets Curve.



(a) Carbon monoxide (CO) and population (b) Nitrogen monoxide (NO) and GDP

Figure 3: Relationship between other types of pollution (NO and CO) and city GDP. We do not find the inverse U-shaped relationship that the Environmental Kuznets Curve suggests. On the contrary, the shape is convex. There is a clear relationship in the data with highly significant coefficients on the linear as well as squared term.

4.2 Prevention and mitigation: Good time for outdoor activities

Next, we turn our attention to prevention and mitigation. Given the fact that we have hourly data on pollution, we can analyze, what is a good time to spend time outdoors. While Figure 5 underscores widespread air quality issues across Europe, Figure 4 highlights the importance of awareness of peaking pollution and inspires mitigation strategies for spending time outdoors at times that minimize the exposure to harmful pollution.

In Figure 5, we plot the average value of PM 2.5 pollution in our sample of European cities. We observe a slight decrease in pollution between 10 AM and 3 PM, which is outside the commuting rush hours. Plotting this graph for individual cities, such as in Figure 4, we found that some cities (such as Paris or Milan) exhibit a much stronger pattern of peaks in the morning and evening and drop between 10 AM and 3 PM.

The takeaway from Figure 5, is that the average pollution throughout the day in Europe remains above the WHO recommended limit of 5 micrograms per cubic meter of PM 2.5. The takeaway from Figure 4 is that having access to pollution information is especially important in cities that experience high levels of pollution so that its citizens can decrease their exposure to harmful pollutants by planning their outdoor activities, such as running or taking children out for a walk.

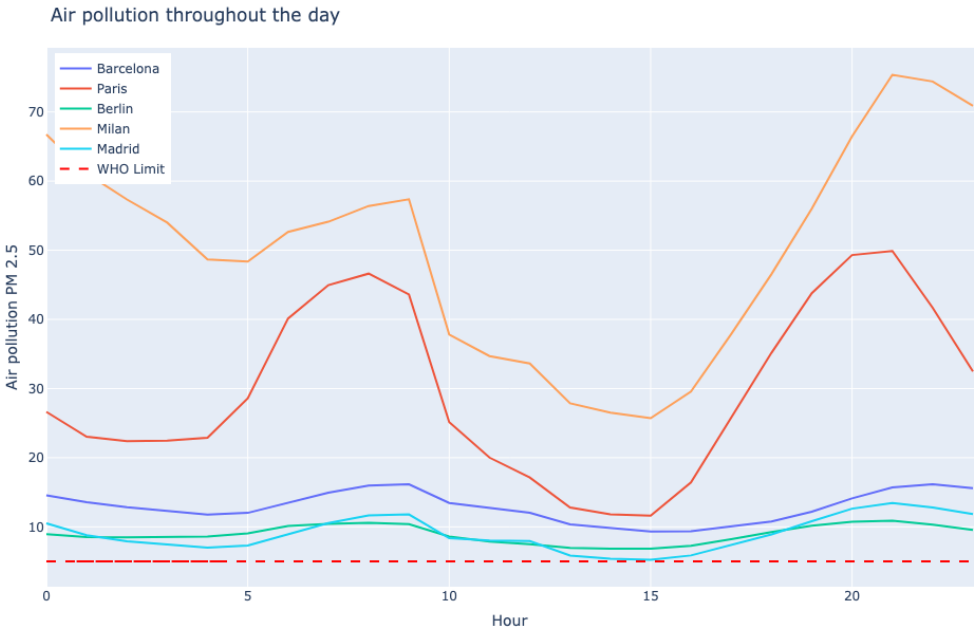


Figure 4: Average air quality, measured by PM 2.5. in selected European cities. The WHO limit of 5 micrograms per cubic meter of PM 2.5, highlighted in red color, is surpassed throughout the day. The cities achieve minimal pollution between 10 AM and 3 PM, although the difference between the peak and drop is more prominent in some cities, such as Paris and Milan, than in others, such as Berlin, Madrid and Barcelona.

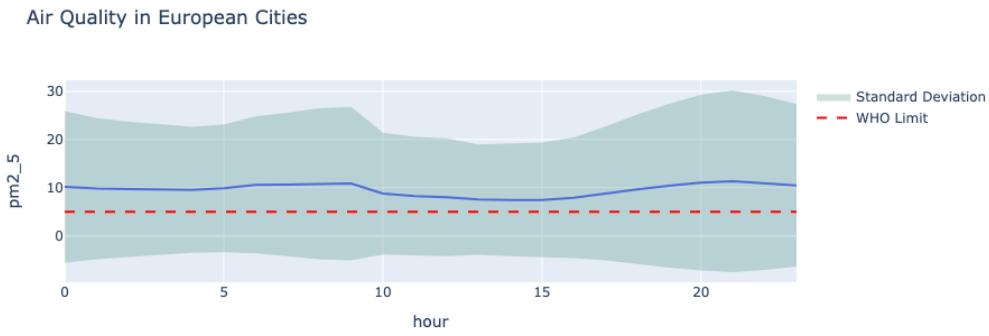


Figure 5: Average air quality, measured by PM 2.5, in European cities throughout the day.

5 Case study: Pollution prediction in Barcelona based on weather

In the case study of Barcelona, we attempt to explain as well as predict the pollution levels for the next hour based on some weather variables: Humidity (%), Atmospheric pressure on the sea level (hPa), Temperature (K), Wind speed (m/s), Cloudiness (%).
To explain the levels of PM 2.5 in Barcelona we checked whether the model satisfies the assumption of the classical linear model. Our main concern was a serial correlation since the data is time series. To try to solve this, we decomposed the pollution variable. We observed two cyclical components: daily and weekly. Unfortunately, the decomposition did not eliminate the serial correlation as we see in Figure 10 in the Appendix.

5.1 Weather-based prediciton

We fitted a linear model and we saw that all weather variables are significantly correlated with pollution. Our main hypothesis in this part is that the relationships between weather conditions and pollution are nonlinear, and we wanted to test this. To do so, we fitted this simple linear model as well as a linear model with interactions, and tree-based methods to account for non-linearities. We started with a regression tree, but since the main drawback of this model is its variance, we used random forest and boosting to get more stable predictions.

In Table 2 we can see that all the individual variables have some predictive power in the first model. However, in the complex model humidity loses its predictive power, but it is given to some interactions. This is a hint for some non-linearities in the "true model". We also observe that the adjusted R^2 is better in the second model. However, since we care about the predictive power we shouldn't focus on this metric. We should focus instead on the out-of-sample performance, as we will see in the next section.

	$PM2.5_{t+1}$	$PM2.5_{t+1}$
const	-591.0689^{***}	-1667.8109^{***}
main.humidity	0.1558^{***}	-0.0340
main.pressure	0.5935^{***}	1.6392^{***}
main.temp	-0.2766^{***}	37.4874^{***}
wind.speed	-0.8642^{***}	93.9613^{***}
clouds.all	-0.0828^{***}	7.0283^{***}
main.humidity_clouds.all		0.0031^{***}
main.humidity_wind.speed		-0.0004
main.pressure_clouds.all		-0.0072^{***}
main.pressure_main.humidity		0.0004
main.pressure_wind.speed		-0.0940^{***}
main.temp_clouds.all		-0.0033^{***}
main.temp_main.humidity		-0.0146^{***}
main.temp_main.pressure		-0.0362^{***}
main.temp_wind.speed		0.0282^{***}
wind.speed_clouds.all		0.0031
Adj. R^2	0.2324	0.2798

Table 2: Regression Results: Model with only variables and model with interactions

5.2 Results for Barcelona pollution prediction

Simple interactions introduced in the regression in the previous section may not sufficiently capture the underlying pattern of this data and try to capture these relationships with more complex models. In the following table, we observe the performance of each model in the testing sample, using the mean squared error. This loss function is a good choice in this case because we want to penalize more large errors.

Model	MSE
Linear Model	47.981061
LM with interactions	50.211100
Regression Tree	166.441563
Random Forest	54.541369
Gradient Boosting	52.636345
Boosted Random Forest	46.688027

Table 3: Prediction results

The best-performing models are the Linear Model and Boosted Random Forest (see Figure 6 and 7).

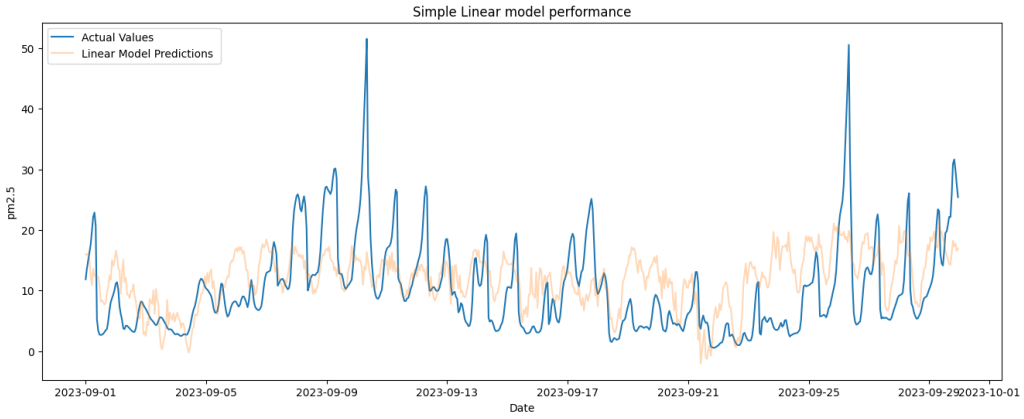


Figure 6: Performance of the simple linear model

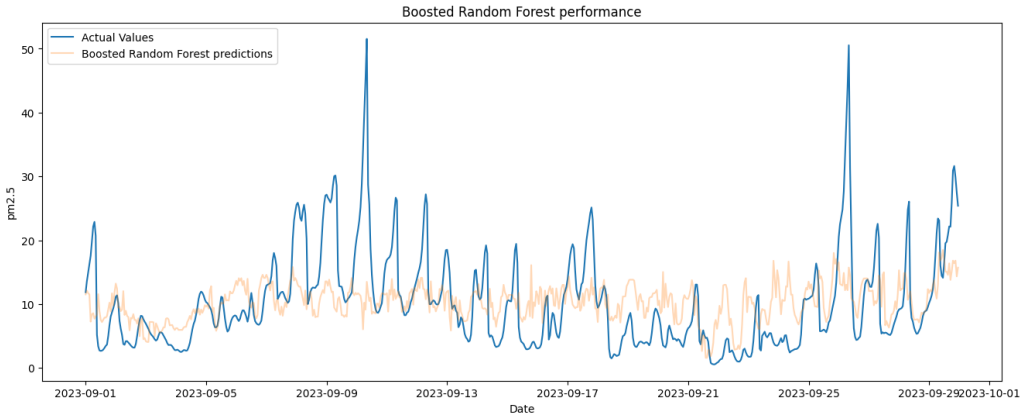


Figure 7: Performance of the Boosted Random Forest

Despite the more complex models such as boosted random forest slightly improving the MSE from the simple linear model, this improvement in out-of-sample performance is probably not worth the computational effort.

6 Conclusions

In this paper, we constructed a novel data set with pollution levels in 227 European cities. We found that neither population nor pollution alone explains the average pollution levels of PM 2.5. We did not find evidence for the inverted U-shaped curve between income and pollution that the Environmental Kuznets Curve (EKC) suggests. In other words, we cannot conclude that as cities get richer they, at first, become polluted to fund their growth but then become able to reduce the harmful pollution effects. If anything, we found the contrary, that for some pollutants, namely,

Carbon monoxide and Nitrogen monoxide (NO and CO), the relationship is a regular U-shaped curve, contradicting the EKC.
Future improvements of this investigation would be to exploit the spatial nature of our data, control for the country, and add other variables that could be used to explain pollution, such as the number of vehicles in a city or the quality of public transit.[3] Another limitation of our investigation is the time component. We only had a time series for pollution for 3 years and kept GDP constant due to missing values for many cities for later years. Perhaps the time frame in our study is too short to capture the investment and commitment to sustainability in European cities (e.g. through emission limits and investment in green mobility). In the ideal case, for the EKC investigation, we would have available a time series with income and pollution for each city for many years. Then we would proceed with time-based analysis on the city level (e.g. controlling for city in a regression model) because each city may have its own specific industry specialization and geographical setting (e.g. being located in a valley).
Our second conclusion is that the best time to spend time outdoors is between 10 AM and 3 PM. A future improvement to this analysis and Graph 5 would be to weigh the average by city size.
The third and last part of our analysis, the pollution prediction in the specific case of Barcelona, we conclude that a linear model is very good[4] despite the fact that there may exist some non-linear underlying mechanisms. It is also clear that the main source of prediction power of this model comes from seasonality. It would be interesting to analyze if a time-based predictor could perform better than our weather-based one, but unfortunately, our data doesn't have that level of detail.

References

[CRB97] M.A. Cole, A.J. Rayner, and J.M. Bates. "The environmental Kuznets curve: an empirical analysis". In: *Environment and Development Economics* 2.4 (Nov. 1997), pp. 401–416. ISSN: 1355-770X, 1469-4395. DOI: 10.1017/S1355770X97000211. URL: https://www.cambridge.org/core/product/identifier/S1355770X97000211/type/journal_article (visited on 12/04/2023).

[Din04] Soumyananda Dinda. "Environmental Kuznets Curve Hypothesis: A Survey". In: *Ecological Economics* 49.4 (Aug. 1, 2004), pp. 431–455. ISSN: 0921-8009. DOI: 10.1016/j.ecolecon.2004.02.011. URL: https://www.sciencedirect.com/science/article/pii/S0921800904001570 (visited on 12/04/2023).

[DW 23] DW News. *Nearly everyone in Europe is breathing polluted air – with deadly consequences — DW News*. Sept. 7, 2023. URL: https://www.youtube.com/watch?v=FkjtCKZx06E (visited on 11/28/2023).

[Eur23] Euronews. *Air pollution deaths: Exhaust fumes are the biggest killer in Europe*. euronews. June 30, 2023. URL: https://www.euronews.com/next/2023/06/30/air-pollution-related-deaths-exhaust-fumes-biggest-killer-in-europe-finds-report (visited on 11/28/2023).

[3] Perhaps using a principal component analysis
[4] Both in terms of efficiency and interpretablity

[Nc23] Ajit Niranjan and Ajit Niranjan Europe environment correspondent. "Toxic air killed more than 500,000 people in EU in 2021, data shows". In: *The Guardian* (Nov. 24, 2023). ISSN: 0261-3077. URL: https://www.theguardian.com/environment/2023/nov/24/toxic-air-pollution-eu-death (visited on 11/28/2023).

[Eura] Eurostat. *Average annual population to calculate regional GDP data by metropolitan regions*. Version met_10r_3pgdp. URL: https://ec.europa.eu/eurostat/databrowser/view/met_10r_3pgdp/default/table?lang=en (visited on 11/28/2023).

[Eurb] Eurostat. *Gross domestic product (GDP) at current market prices by metropolitan regions*. Version met_10r_3gdp. URL: https://ec.europa.eu/eurostat/databrowser/view/met_10r_3gdp/default/table?lang=en (visited on 11/28/2023).

[Opea] OpenWeatherMap. *Air Pollution API - OpenWeatherMap*. In collab. with Drabova, Ivana. URL: https://openweathermap.org/api/air-pollution (visited on 11/28/2023).

[Opeb] OpenWeatherMap. *Geocoding API - OpenWeatherMap*. In collab. with Drabova, Ivana. URL: https://openweathermap.org/api/geocoding-api (visited on 11/28/2023).

[Opec] OpenWeatherMap. *Historical weather API - OpenWeatherMap*. In collab. with Drabova, Ivana. URL: https://openweathermap.org/history (visited on 11/28/2023).

7 Appendix

Table 4: Description of the data set covering pollution, income, and population in 227 European cities. The cities are located in 28 countries and represent a total population of 239,955,890 citizens.

Country	Number of Cities in Sample	Total Population in Sample (thousands)	Average GDP per Capita (thousands)	Std. Dev. of GDP per Capita
Germany	63.0	48693.90	40.88	8.61
France	31.0	41178.99	33.13	7.29
Italy	22.0	28832.40	31.91	8.87
Poland	18.0	18055.21	15.11	5.53
Spain	15.0	26548.19	27.19	5.60
Netherlands	11.0	10550.00	45.91	8.67
Romania	9.0	7603.77	13.48	5.59
Belgium	5.0	3402.40	36.60	11.22
Hungary	5.0	4984.43	13.80	5.53
Austria	5.0	4985.37	49.51	4.69
Czech Republic	4.0	5672.18	21.97	7.36
Denmark	4.0	4035.00	49.52	10.94
Bulgaria	4.0	3224.30	10.00	4.87
Norway	4.0	2266.57	68.62	21.12
Sweden	4.0	5828.94	48.73	10.79
Portugal	3.0	5014.07	22.02	4.45
Finland	3.0	2675.71	45.86	9.49
Lithuania	2.0	1377.67	21.64	5.42
Greece	2.0	4663.92	18.96	6.34
Ireland	2.0	2831.07	104.31	30.87
Croatia	2.0	1614.27	15.57	6.27
Slovakia	2.0	1465.10	26.89	18.47
Slovenia	2.0	876.97	25.82	9.73
Latvia	1.0	999.64	21.42	0.00
Luxembourg	1.0	621.50	100.36	0.00
Malta	1.0	470.96	28.78	0.00
Cyprus	1.0	881.95	26.28	0.00
Estonia	1.0	601.41	28.56	0.00
Total	227.0	239955.89	992.83	217.72

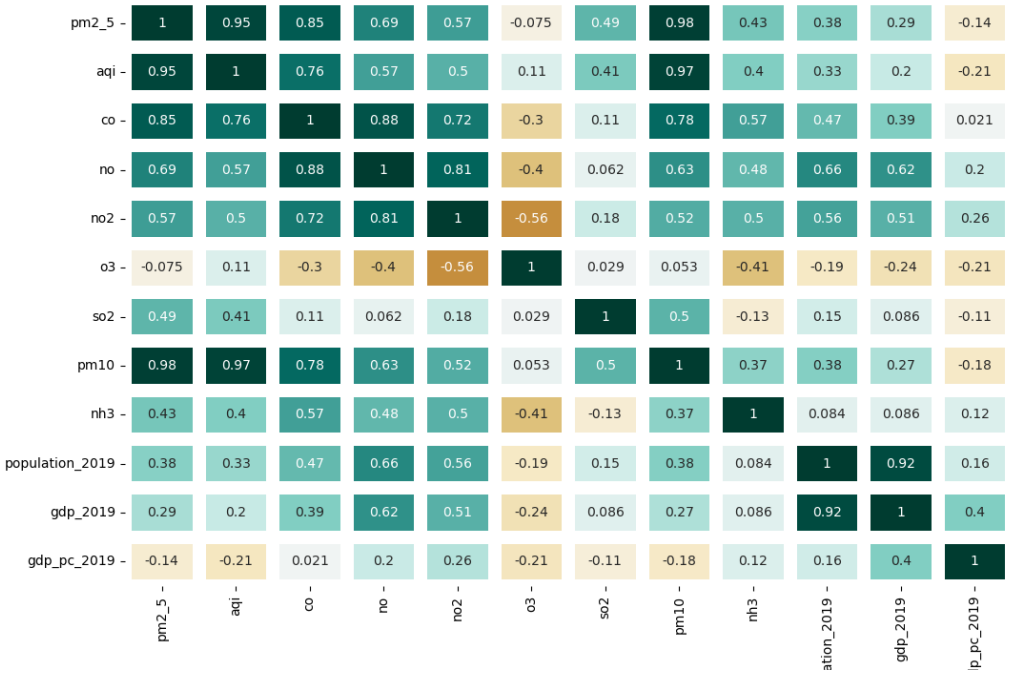


Figure 8: Correlation table between air pollution variables and population, income, and income per capita. Data is aggregated as city-level average. Air pollution variables include particulate matter 2.5 and 10 micrometers/ cubic meter (PM 2.5, PM10, respectively) and overall air quality index AQI, CO, NO, NO2, 03, SO2.

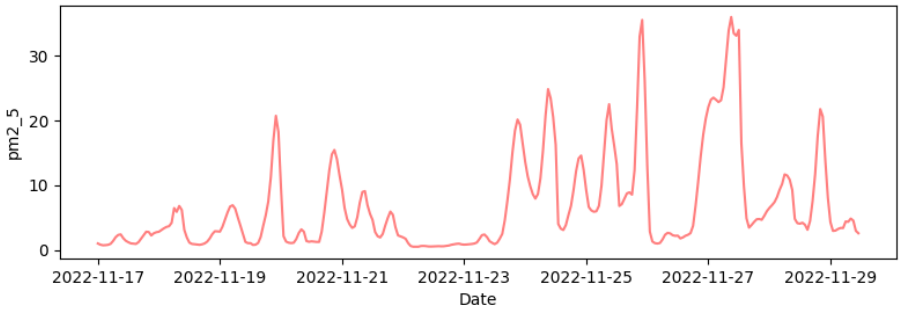


Figure 9: Closer look into pm2.5 time series

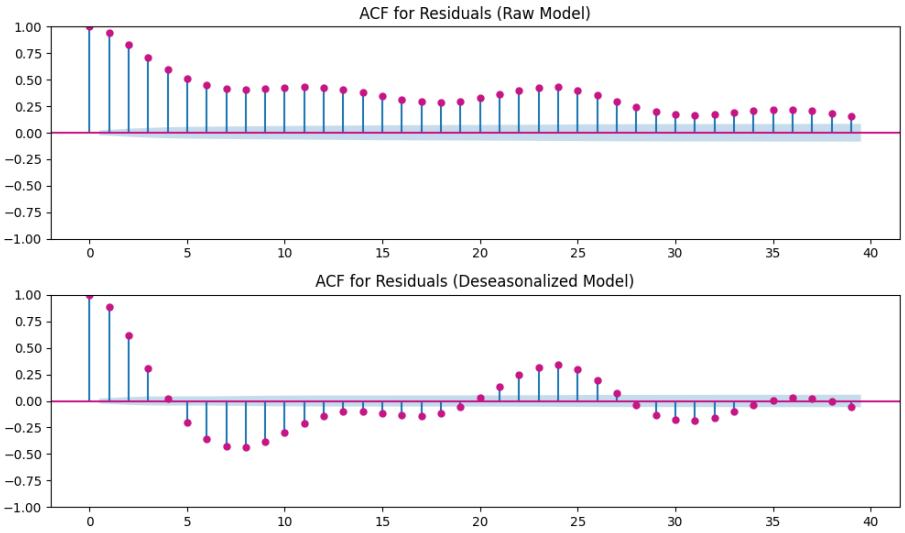


Figure 10: Auto correlation function (ACF) on the case study of Barcelona.

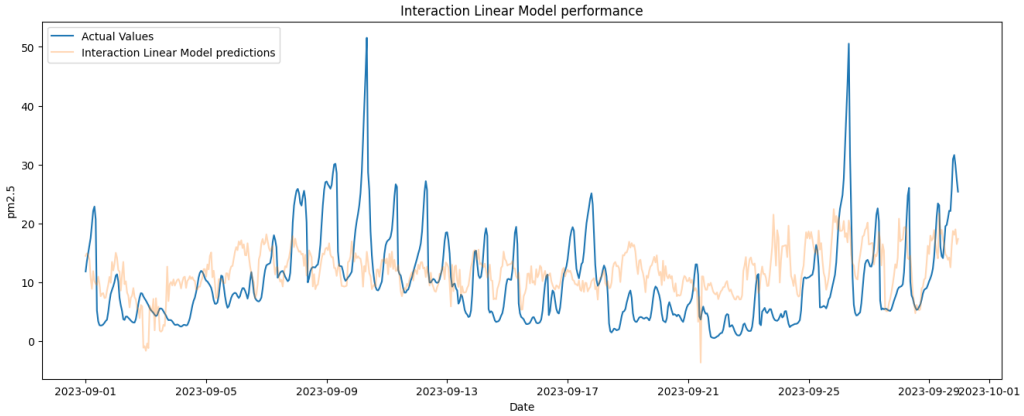


Figure 11: Performance of the interaction linear regression

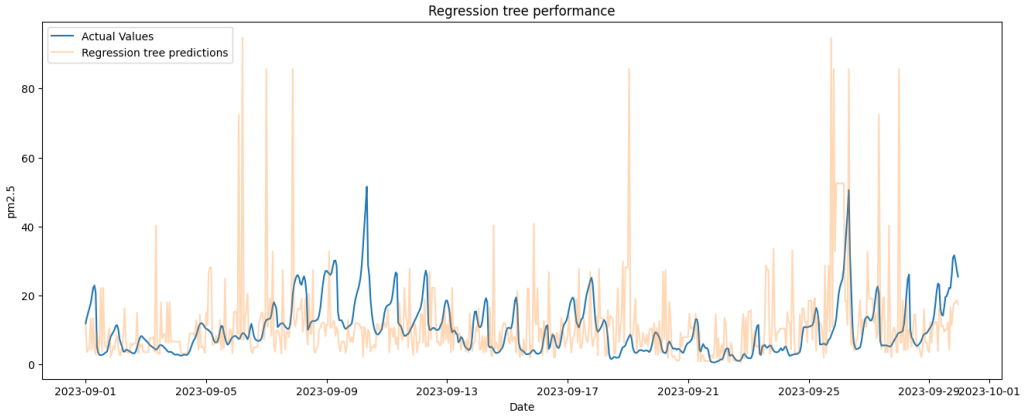


Figure 12: Performance of the Regression Tree

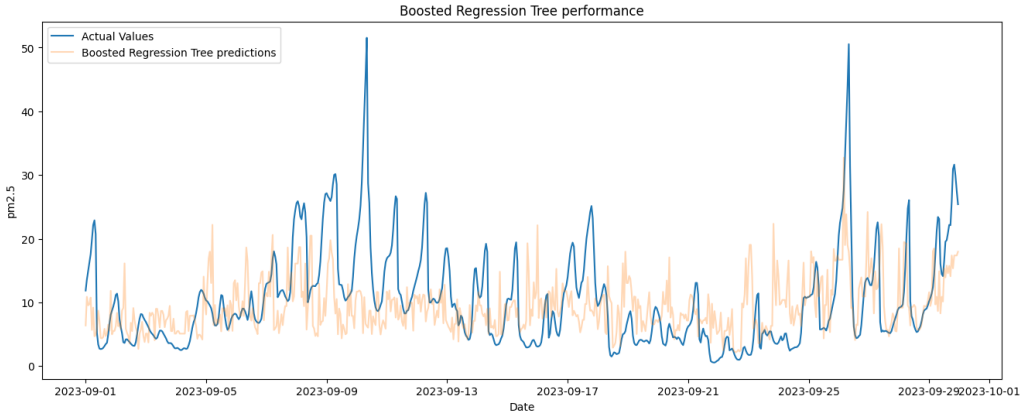


Figure 13: Performance of the Boosted Regression Tree

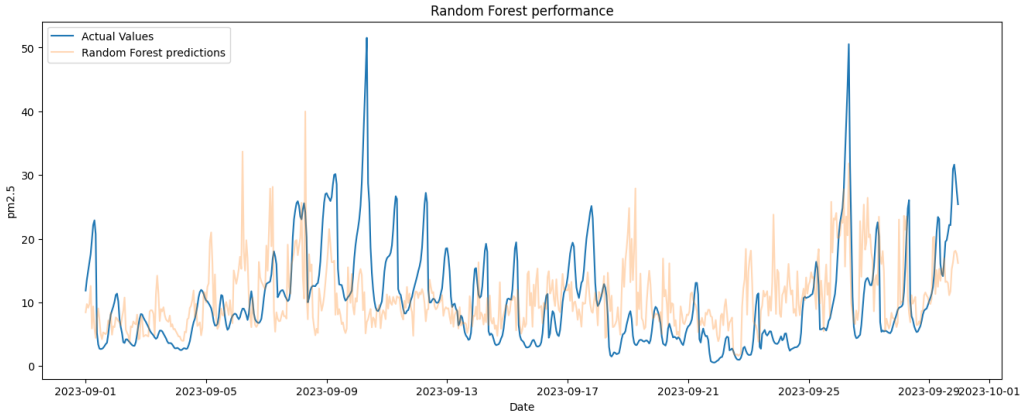


Figure 14: Performance of the Random Forest