On the Properties and Causes of Air Pollution

Mikautadze Sandro
De Muri Giovanni

1 Introduction

In the world, 7 million people a year die prematurely due to air pollution [1]. To put that in context, cancer kills 10 million people a year [2]. These are just some pieces of information that confirm a well-known fact: air pollution is a serious concern to global health. To further delve into this important topic, we have indeed chosen to analyze air pollution.

The structure of our research project is twofold: firstly, we will study some quantitative aspects of Italy's pollution. Namely, we will analyze different pollutants and understand if their distribution exceeds the guidelines set by the EU. Secondly, we will try to analyze which features can be used to explain the air quality in Europe by doing regressions. We motivate these choices by a simple fact: it is crucial to have a grasp of the condition in which we all live right now and understand where it might come from.

Given these initial remarks, we focused on the following research questions:

1. Is there enough statistical evidence to show that Italy's air pollution levels exceed the EU guidelines?
2. Is there enough statistical evidence to show that Italy's Northern, Central, and Southern air pollution levels exceed the EU guidelines?
3. Can we find statistically significant causal relations between demographic, socio-economical, and geographical factors and air pollution in Europe?

2 Data Processing

To address our research questions we used different data sets collected from different sources. Since the downloaded data were obviously built for different purposes than ours, we proceeded with a deep, long, and rigorous cleaning process for each research question. We decide not to illustrate the cleaning procedure here as it would require another report by itself. Therefore, for a detailed understanding, we ask the reader to check the cleaning code in the cleaning file on GitHub. As you will see, this cleaning process resulted in the creation of the data sets that we use in the analysis, which you can access in the respective folder of the GitHub repository.

The only thing that we mention here is that we decided to use as pollutants PM10, PM2.5, O_3, and NO_2, as we had more data and all observations refer to 2018. We saved the cleaned datasets for questions 1 and 2 under the name *Ita-PollutantName*, while *Europe-Causes* for question 3.

3 Analysis

3.1 Exploratory Analysis and Hypotheses

As far as the analysis of Italy goes, we can gain pretty interesting insights by doing an initial visual inspection of how the data is distributed across the whole country. However, we first need to highlight how the *AirPollutionLevel* column of the dataset was measured. Indeed, PM10's values correspond to the 36th highest daily mean concentration, while O_3's values are the 26th highest daily maximum of the running 8-hour mean. Instead, the values of PM2.5 and NO_2 represent the annual mean concentration. These metrics are chosen based on the way the EU sets the concentration limits, which we will later use in the analysis:

- PM10: not more than 35 days/year with a daily mean concentration exceeding 50 $\mu g/m^3$
- PM2.5: annual mean concentration must not exceed 25 $\mu g/m^3$
- O_3: not more than 25 days/year with maximum daily 8-hour mean concentration exceeding 120 $\mu g/m^3$
- NO_2: annual mean concentration must not exceed 40 $\mu g/m^3$

As we can immediately see in figure 1, Northern Italy tends to have higher concentrations of pollution, compared to the Central and Southern zones of Italy. Hence, we expect the North to be more likely to exceed the EU guidelines than other zones of Italy. Moreover, since all pollution levels seem on average pretty high, we would be tempted to believe that Italy's pollution levels, as a whole, are all beyond the EU references. However, as we can see by figure 2, only O_3 is distributed such that the median slightly exceeds the guideline. Therefore, even though the pollution levels are not low, we don't expect Italy's air pollution to be beyond the guidelines.

To also have hypotheses for the causes of air pollution, we can have a look at the correlation matrix. We are aware that the correlation is not a pure and direct measure of the relevance in a regression model, since there can be interaction effects that the correlation cannot capture and since the coefficient measures just the linear correlation between covariates, not other types of relationships. Nonetheless, it can still give us interesting insights into how the changes are related to each other.

One can indeed notice from figure 4 that just a few covariates have a correlation coefficient of over 0.5 (*GDP per capita*, *Agriculture VA* and *Primary energy consumption per capita*). Intuitively, we can expect at least one of them to be relevant in the regressions that we will run, but others covariates might be included too.

Before jumping into the statistical analysis, we remind the reader to check the analysis file and run it all along for a thorough understanding of the results that we will present in the next sections.

3.2 Questions 1 and 2: Testing

After the previous initial inspection, we now want to apply a statistical test to check whether we have enough evidence to conclude that the zones we consider exceed the EU guidelines.

We will apply a t-test. We have guarantees that it will reliable because that data is normally distributed (figure 2 and 3), and the variance is unknown. In particular, if we call μ_0 the EU reference guideline for the pollutant taken into consideration, we want to test

$$H_0 = \mu < \mu_0 \text{ vs } H_1 = \mu \geq \mu_0$$

at a statistical level of significance of $\alpha = 0.05$. Then, by analyzing the p-value of the test, we can understand whether there is sufficient statistical evidence to show that a specific pollutant exceeds the EU guidelines.

We have summarized the results obtained for question 1 - i.e. for Italy - in figure 6. As we can see, we retain H_0 for all pollutants. This means that there isn't enough statistical evidence to conclude that Italy exceeds the EU guideline for each pollutant. More specifically, the p-value is 1 for each pollutant except for O_3, where it is 0.59. By the previous visual analysis, these results are in line with our hypotheses.

Then, we proceeded to solve the second question by doing the same tests for the same pollutants in the North, Center, and South. As we are doing multiple tests on a subset of the same sample used before, we run into the risk of multiple hypothesis testing, which is problematic since the likelihood of committing a type I error increases. Yet, for our scope, we are not going to consider any correction and we are still going to take a significance level of $\alpha = 0.05$ for the t-test.

The results are summarized in figures 7, 8, 9, 10. For PM10, PM2.5, NO_2, we have that all the p-values are 1 and so we don't have any statistical evidence to prove that each zone's air pollution level exceeds the EU guidelines. For O_3, we retain H_0 with the Center and the South, but we have enough statistical evidence to show that the North of Italy exceeds the EU guidelines, as the p-value is $3.7 \cdot 10^{-25}$.

Overall, even though we only rejected H_0 once, it is interesting to notice that Northern Italy always has a higher mean compared to the Center and South in all pollutants. Moreover, the estimates for the Center and South are always below the mean of the respective pollutant. This is in accordance with the preliminary analysis done in the visual exploration, where we saw that most of the measurements were below the EU limits and that the North was the most polluted zone. All the code can be found in the *TESTING - Questions 1 and 2* section of the code file.

3.3 Questions 3: Multiple Linear Regression

Let's focus on the causes of air pollution. To goal was to find a regression model that managed to explain the measured level of pollution for each air pollutant (PM10, PM2.5, NO_2, O_3), using as covariates *Total population*, *Population density*, *GDP*, *GDP per capita*, *Renewable energy consumption*, *Industry VA*, *Agriculture VA*, *Forest area*, *Primary energy consumption*, and *Primary energy consumption per capita*. Since our analysis is analogous for each pollutant, we are going to describe the general methodology first and then we are going to present the final results of each model.

After removing the rows containing missing values, we applied a multiple linear regression of the form:

$$Y = \beta_1 x_1 + \cdots + \beta_{10} x_{10} + e$$

where x_i are the causes, and Y represents the pollutant we are considering. To choose the best model for Y we ran three regressions using three different methods. The first was the model with all covariates; the second was obtained using the step-down method, at a significance level of 0.05; the third using the step-up method at the same significance level.

After having obtained three models, we wanted to do the model selection. We decided to use penalization criteria, and, in particular, the BIC. This is because here we are more interested in accurately selecting which covariates have an effect on air pollution, using a consistent model selection criterion. We avoided AIC model selection as it is more efficient in prediction, which we are not interested in.

Then we did a regression diagnostic on the optimal model, checking the normality of the residuals, with histograms and QQ-plots, and their homoscedasticity, through the plot of the residuals vs the fitted values and, in case of satisfied normality, applying the F-test. Notice that checking for homoscedasticity is very important since, if the assumptions are not satisfied, then the estimated regression coefficients are no longer valid.

Let's now see what the final optimal results are. For a deep view of the procedure of the different regressions, we ask the reader to run the code in the *Causes vs Air Pollution Levels - Question 3* section.

- Y is PM10. After removing the missing values, we remain with 36 observations. The optimal model, based on BIC, is given by the model obtained with the step-down method, which is actually the same as the step-up model. It is made of just 2 variables: *Renewable Energy*

Consumption and *Agriculture Value Added*. In this case, the p-value of the model is $7.43 \cdot 10^{-10}$, hence it can explain the data better than no model at all, and from the R^2 we see that 72% of the variance is explained. From the regression diagnostic, the mean of the residuals is very close to 0 and from figure 11 the normality assumption seems satisfied (excluding the effects of outliers). As for the homoscedasticity, the figure shows that there aren't big differences in the overall variance. This is further sustained by the F-test which tells us that the variance is indeed homogeneous.

- Y is PM2.5. With 33 observations, out of the three models, the best is given by the step-down method (same as step-up), containing the variables *Agriculture Value Added* and *Renewable energy consumption*. It has a p-value of $2.3 \cdot 10^{-5}$, so this model is statistically significant in explaining the data, and it manages to explain around 50% of the variance of the data. Figure 12 shows that the normality of the residuals is satisfied. As for homoscedasticity, the assumption seems to be satisfied too, but the F-test actually tells us that the variance is not homogenous. This is probably due to the effects of outliers, but further investigation would be required.

- Y is O_3. With 34 complete observations, we get that the best model is given by the step-down method (same as step-up), containing only the variable *Agriculture Value Added*. The p-value of the model is 0.002, so it manages to explain the data in a significant way, but the R^2 is only 0.25. Although there are some important deviations, from figure 13 we see that the residuals look almost normally distributed. However, homoscedasticity is not satisfied, neither graphically nor with the F-test. Therefore the results of the regression might not be very accurate.

- Y is NO_2. With 36 observations, the best model is, again, given by the ones of the step-down and step-up methods. The p-value of the model is 0.0029, so the model is better than no model at all, and the R^2 is 0.23. The only statistically significant covariate is *Renewable Energy Consumption*. The assumptions of normality and homoscedasticity can be easily verified by looking at figure 14.

4 Conclusions

For questions 1 and 2, we didn't find any statistically significant evidence to show that Italy's air pollution levels exceed the EU guidelines. The only exception goes for O_3 in the North, in which we could conclude that the pollution was beyond the given limits. Even though the data suggested that these were the most *obvious* conclusions, these results are still surprising for us, because it is well-known that Italy, especially the North, is one of the most polluted regions in Europe [3].

As for the regression model, a summary of the results is in figure 15. So, *Agriculture Value Added* and *Renewable Energy Consumption* were the only significant covariates that could explain the air pollution level in European countries. Even though it partially follows the hypotheses, it is still surprising to have only those two variables, since we expected a more relevant correlation between the other covariates as well. This might be explained by the fact that we are just checking the linear relationship between the features and the air pollutant, and not other types of relationships.

4.1 Limitations and Methods of Improvement

Overall, the investigation displayed a few weaknesses. As for the study in Italy, the main limitation was in the collection of the data. Indeed, for some countries we didn't have many observations and, for each station, we had access just to one measurement. Obviously, air pollution is a much broader problem and we would have needed more accurate data to precisely assess the air quality in a specific zone.

As for the regression part, a big limitation of these types of models was that we were not really finding the causes of air pollution, since correlation doesn't imply necessarily causation. In fact, we built the dataset choosing factors that *seemed like possible causes*. Hence, we decided *a priori* what "causes" to look at, and, by doing so, we have surely missed interesting insights that would've made our project more complete.

The regression analysis also contained another important limitation: the lack of data. Our sample was made of only 37 observations, one for each country. This meant that lots of local-level data were lost by taking the average air pollution as a measure of air quality and, hence, we could not properly capture true statistical tendencies with a single statistic per country.

Due to the lack of data and to the missing values that we had (see figure 5), we decided also not to use cross-validation, since removing further observations could have had a big impact on the accuracy of the model.

Still, we believe that there are various methods to improve the project. As we already said, more precise measurements and data at a local level are surely necessary in order to conduct a more rigorous study.

Moreover, we could have used more advanced and precise tools, for example by using the Kolmogorov-Smirnov test to accurately check the normality of the residuals, or, in the regressions diagnostic, by applying Box-Cox transformations to fix a possible non-normality of the residuals or to possibly find non-linear relations between the covariates and the independent variables.

Another way to improve the project is in the choice of regression models to consider. We used step-wise methods based on p-values but, even though the given models were better than the ones with all the variables, they were often too conservative and didn't include variables that would have explained part of the data. To have an intuitive understanding have a look at figures 16, 17, 18, 19. Analyzing the changes in AIC and R^2 at each step of the stepwise method, we see that mid-through the step-down procedure either R^2 dropped dramatically, removing an important covariate that could have explained the variance of the data, or the AIC increased, making it a worse model than before. So, fairly often, we find ourselves under-fitting the data. To test other models, it would be interesting to use more precise methods, like the step-wise method based on the AIC criterion. This might be more helpful in finding the best model and minimizing both bias (that is the cause of under-fitting) and variance (that is the cause of over-fitting).

For the first part, as a further analysis, we believe that it would be interesting to conduct a study of the changes in pollution over time and infer the trends. Moreover, it would be curious to study only Northern Italy more carefully, being a very peculiar zone pollution-wise as already mentioned.

For the second part, instead, it would be interesting to get more insights by studying other types of relationships between other covariates as well. Moreover, a further study might concern the possible effects that air pollution might have across Europe.

4.2 Final Comments

Regardless of the results and their optimality, the goal of our analysis was not that of finding significant and groundbreaking results. The objective was to explore a dataset and apply the main statistical tools that we learned in class. We have made a strong effort to carefully analyze the data and draw meaningful conclusions from the research questions. Yet, we do not claim that the described methodologies are the only ones that can be used to tackle the research questions.

Contents

References

[1] Max Roser. Data review: How many people die from air pollution?, 2021. https://ourworldindata.org/data-review-air-pollution-deaths.

[2] WHO. Cancer, 2021. https://www.who.int/news-room/fact-sheets/detail/cancer.

[3] Unknown Author. Pollution in the po valley: The most unhealthy air in europe. Shortened link: shorturl.at/dkmz1.

[4] EEA. Air pollutant concentrations 2018 (compared to eu values). Raw data for questions 1 and 2.

[5] Our World in Data. Primary energy consumption per capita. Raw data for questions 3.

[6] Our World in Data. Primary energy consumption. Raw data for questions 3.

[7] World Bank. Forest area. Raw data for questions 3.

[8] World Bank. Agriculture value added. Raw data for questions 3.

[9] World Bank. Industry value added. Raw data for questions 3.

[10] World Bank. Renewable energy consumption. Raw data for questions 3.

[11] World Bank. Gdp per capita. Raw data for questions 3.

[12] World Bank. Gdp. Raw data for questions 3.

[13] World Bank. Population density. Raw data for questions 3.

[14] World Bank. Total population. Raw data for questions 3.

Appendix

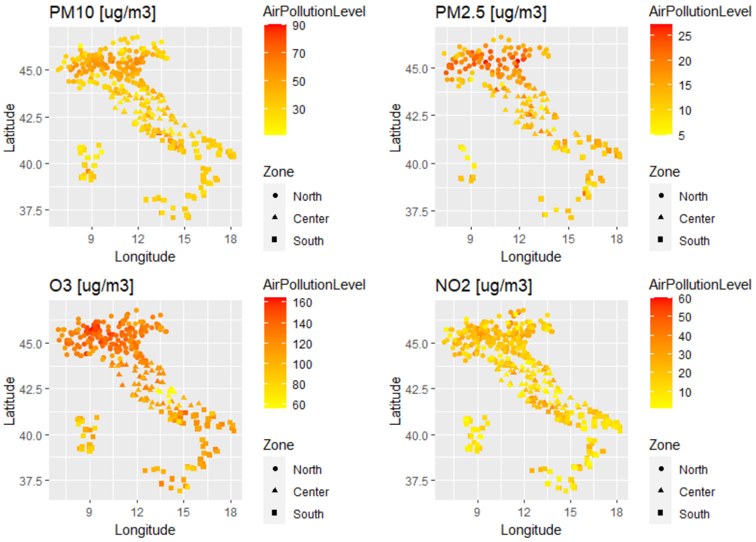


Figure 1: Air Pollution Levels of the different pollutants, grouped by zone

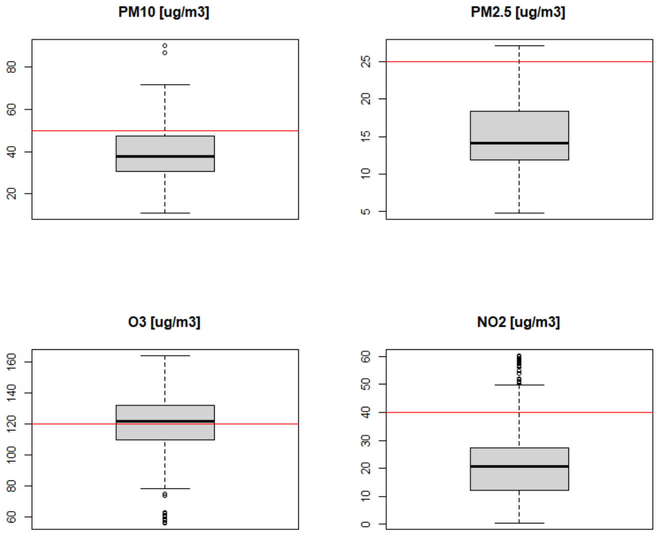


Figure 2: Boxplot of air pollution levels in Italy of the different pollutants. The red line represents the EU guideline

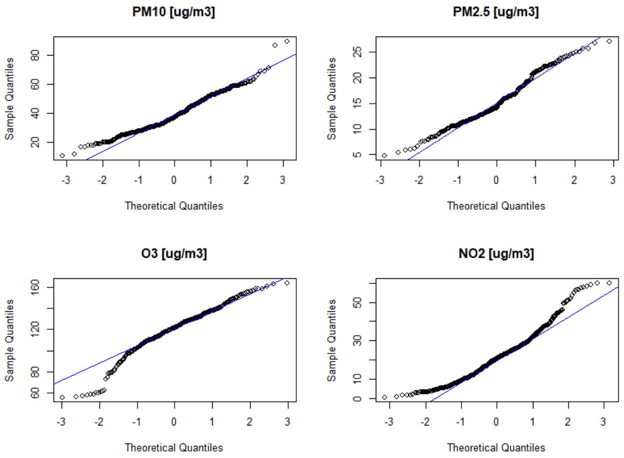
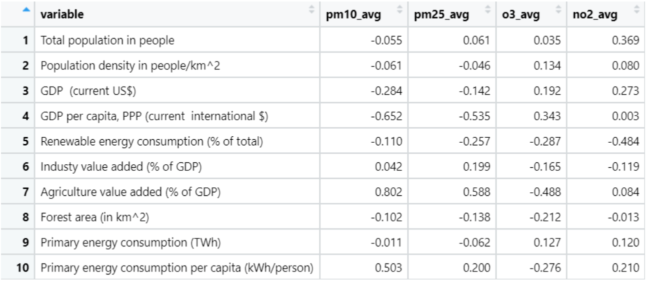


Figure 3: QQ plot to check the normality of pollutants in Italy



	variable	pm10_avg	pm25_avg	o3_avg	no2_avg
1	Total population in people	-0.055	0.061	0.035	0.369
2	Population density in people/km^2	-0.061	-0.046	0.134	0.080
3	GDP (current US$)	-0.284	-0.142	0.192	0.273
4	GDP per capita, PPP (current international $)	-0.652	-0.535	0.343	0.003
5	Renewable energy consumption (% of total)	-0.110	-0.257	-0.287	-0.484
6	Industy value added (% of GDP)	0.042	0.199	-0.165	-0.119
7	Agriculture value added (% of GDP)	0.802	0.588	-0.488	0.084
8	Forest area (in km^2)	-0.102	-0.138	-0.212	-0.013
9	Primary energy consumption (TWh)	-0.011	-0.062	0.127	0.120
10	Primary energy consumption per capita (kWh/person)	0.503	0.200	-0.276	0.210

Figure 4: Correlation matrix of the causes of air pollution

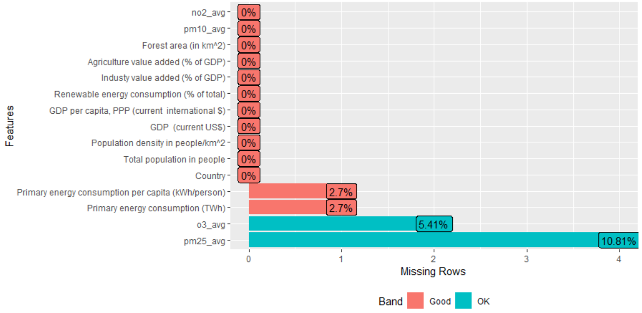
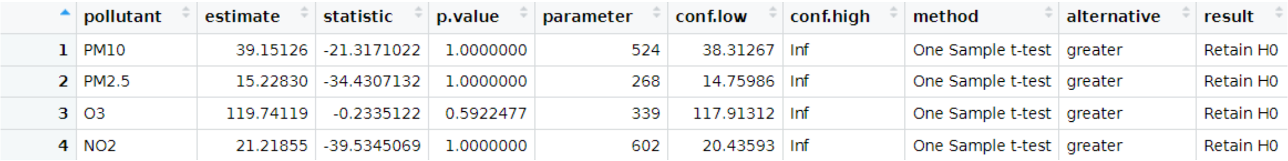
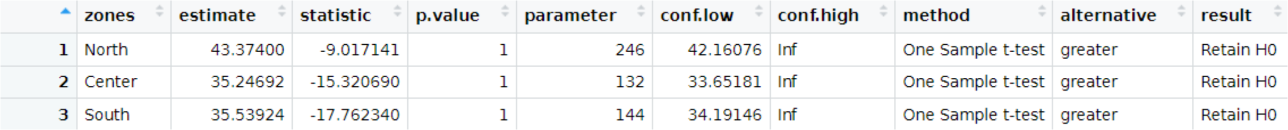


Figure 5: Visual analysis of the missing values in the causes regression model



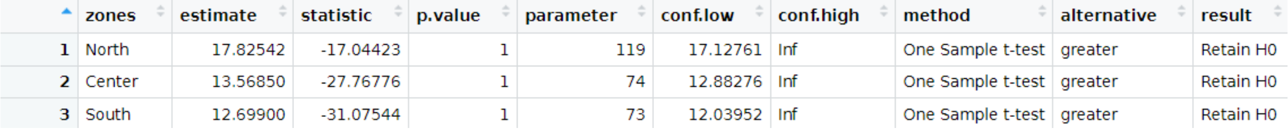
	pollutant	estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative	result
1	PM10	39.15126	-21.3171022	1.0000000	524	38.31267	Inf	One Sample t-test	greater	Retain H0
2	PM2.5	15.22830	-34.4307132	1.0000000	268	14.75986	Inf	One Sample t-test	greater	Retain H0
3	O3	119.74119	-0.2335122	0.5922477	339	117.91312	Inf	One Sample t-test	greater	Retain H0
4	NO2	21.21855	-39.5345069	1.0000000	602	20.43593	Inf	One Sample t-test	greater	Retain H0

Figure 6: Summary table of the results of each pollutant in Italy



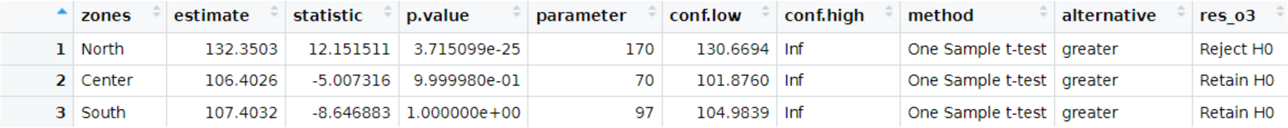
	zones	estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative	result
1	North	43.37400	-9.017141	1	246	42.16076	Inf	One Sample t-test	greater	Retain H0
2	Center	35.24692	-15.320690	1	132	33.65181	Inf	One Sample t-test	greater	Retain H0
3	South	35.53924	-17.762340	1	144	34.19146	Inf	One Sample t-test	greater	Retain H0

Figure 7: Summary of the results for PM10



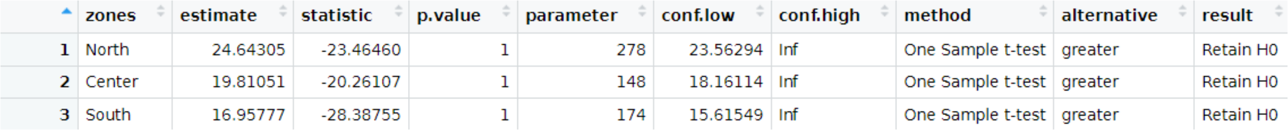
	zones	estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative	result
1	North	17.82542	-17.04423	1	119	17.12761	Inf	One Sample t-test	greater	Retain H0
2	Center	13.56850	-27.76776	1	74	12.88276	Inf	One Sample t-test	greater	Retain H0
3	South	12.69900	-31.07544	1	73	12.03952	Inf	One Sample t-test	greater	Retain H0

Figure 8: Summary of the results for PM2.5



	zones	estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative	res_o3
1	North	132.3503	12.151511	3.715099e-25	170	130.6694	Inf	One Sample t-test	greater	Reject H0
2	Center	106.4026	-5.007316	9.999980e-01	70	101.8760	Inf	One Sample t-test	greater	Retain H0
3	South	107.4032	-8.646883	1.000000e+00	97	104.9839	Inf	One Sample t-test	greater	Retain H0

Figure 9: Summary of the results for O_3



	zones	estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative	result
1	North	24.64305	-23.46460	1	278	23.56294	Inf	One Sample t-test	greater	Retain H0
2	Center	19.81051	-20.26107	1	148	18.16114	Inf	One Sample t-test	greater	Retain H0
3	South	16.95777	-28.38755	1	174	15.61549	Inf	One Sample t-test	greater	Retain H0

Figure 10: Summary of the results for NO_2

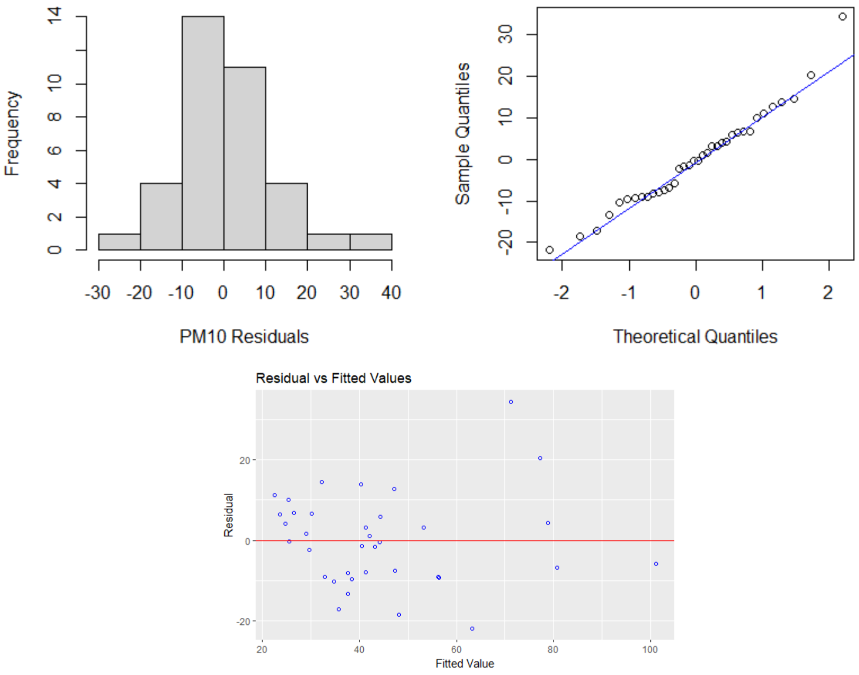


Figure 11: PM10 Regression diagnostic

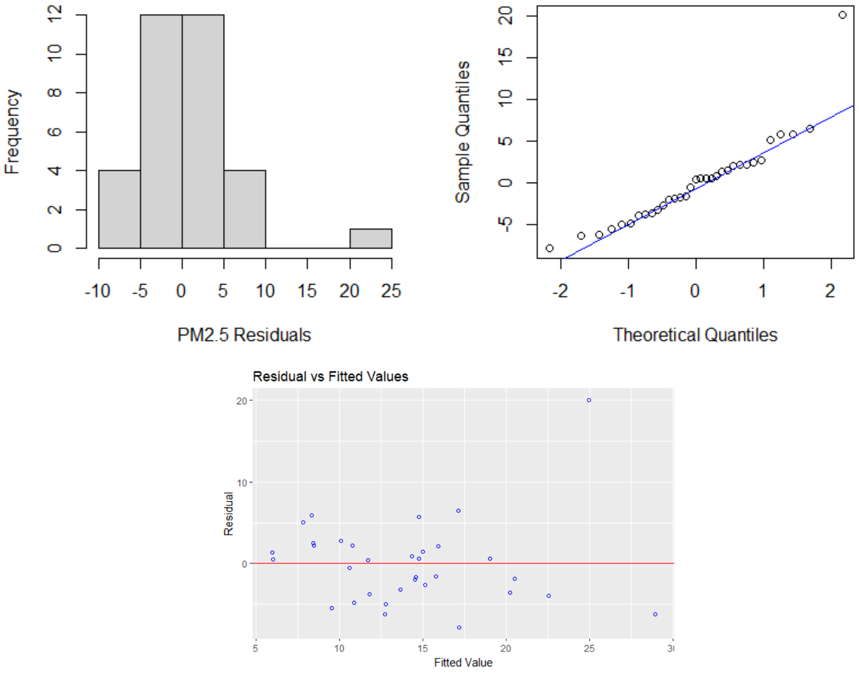


Figure 12: PM2.5 Regression diagnostic

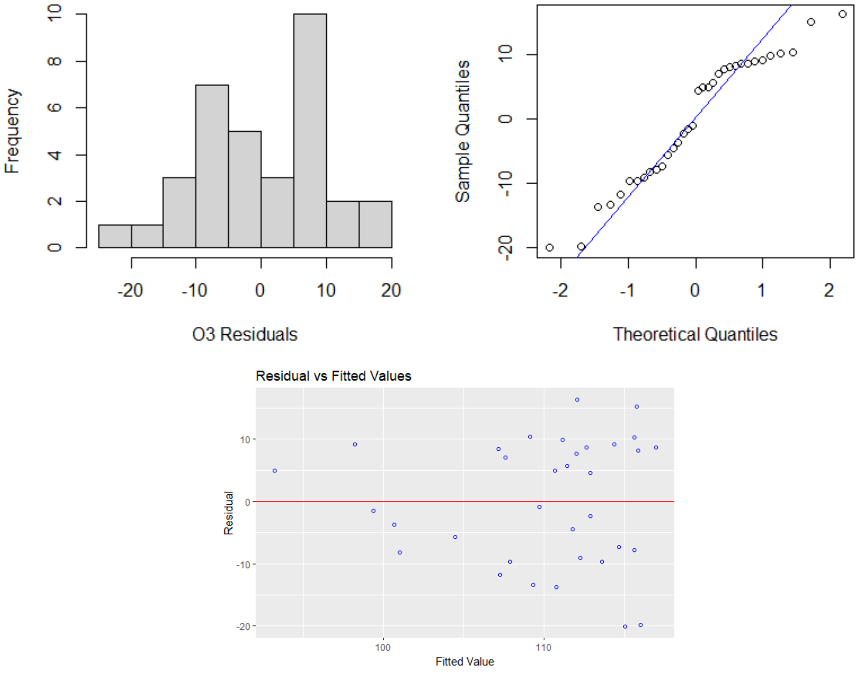


Figure 13: O_3 Regression diagnostic

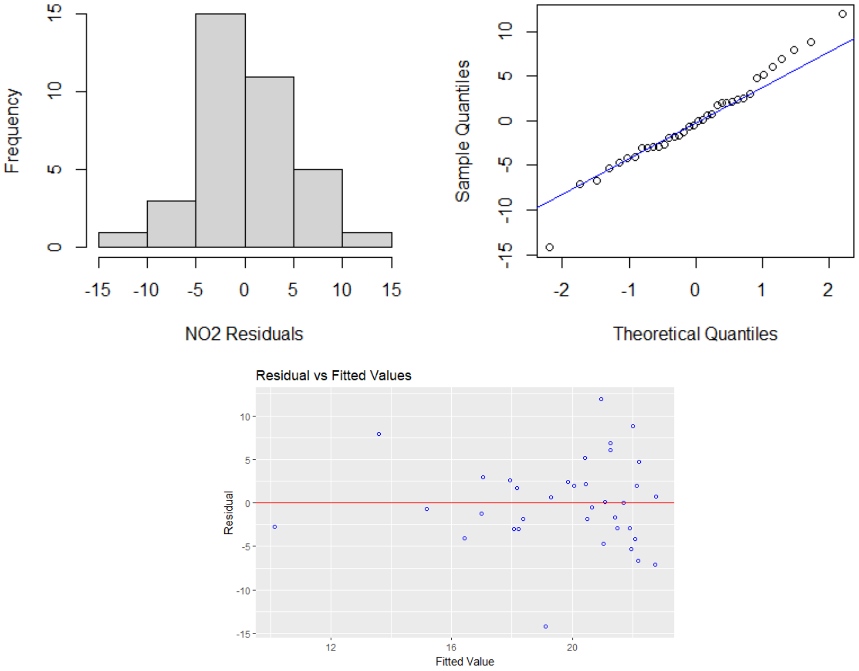
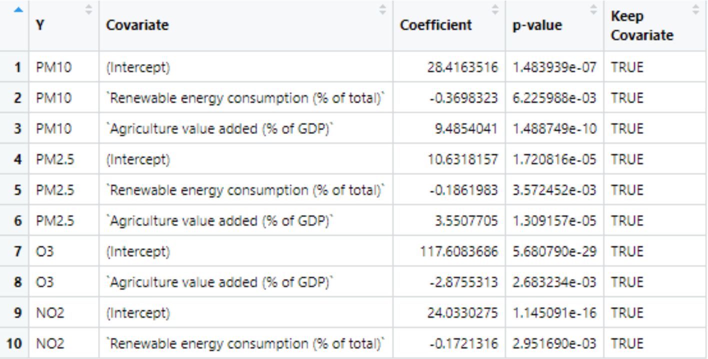


Figure 14: NO_2 Regression diagnostic



	Y	Covariate	Coefficient	p-value	Keep Covariate
1	PM10	(Intercept)	28.4163516	1.483939e-07	TRUE
2	PM10	`Renewable energy consumption (% of total)`	-0.3698323	6.225988e-03	TRUE
3	PM10	`Agriculture value added (% of GDP)`	9.4854041	1.488749e-10	TRUE
4	PM2.5	(Intercept)	10.6318157	1.720816e-05	TRUE
5	PM2.5	`Renewable energy consumption (% of total)`	-0.1861983	3.572452e-03	TRUE
6	PM2.5	`Agriculture value added (% of GDP)`	3.5507705	1.309157e-05	TRUE
7	O3	(Intercept)	117.6083686	5.680790e-29	TRUE
8	O3	`Agriculture value added (% of GDP)`	-2.8755313	2.683234e-03	TRUE
9	NO2	(Intercept)	24.0330275	1.145091e-16	TRUE
10	NO2	`Renewable energy consumption (% of total)`	-0.1721316	2.951690e-03	TRUE

Figure 15: Summary of the optimal regression models

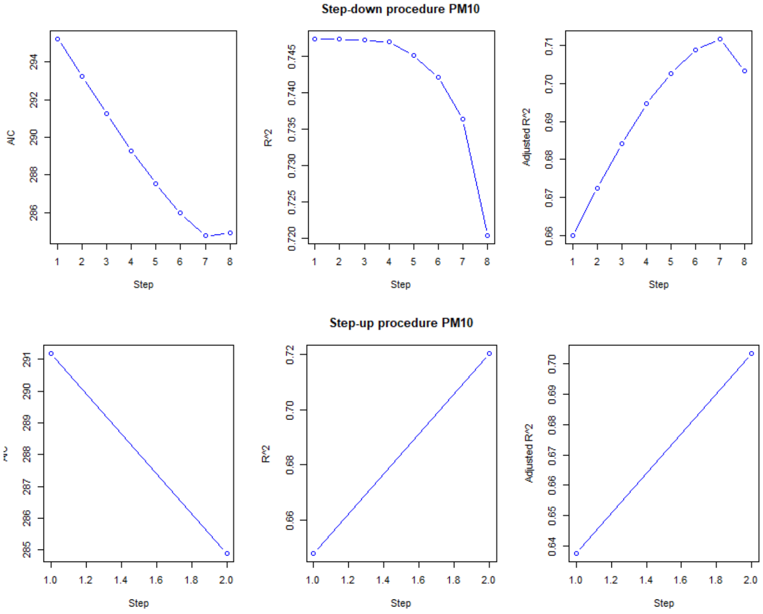


Figure 16: AIC, R^2 and adjusted R^2 results at each step from step-down and step-up for PM10

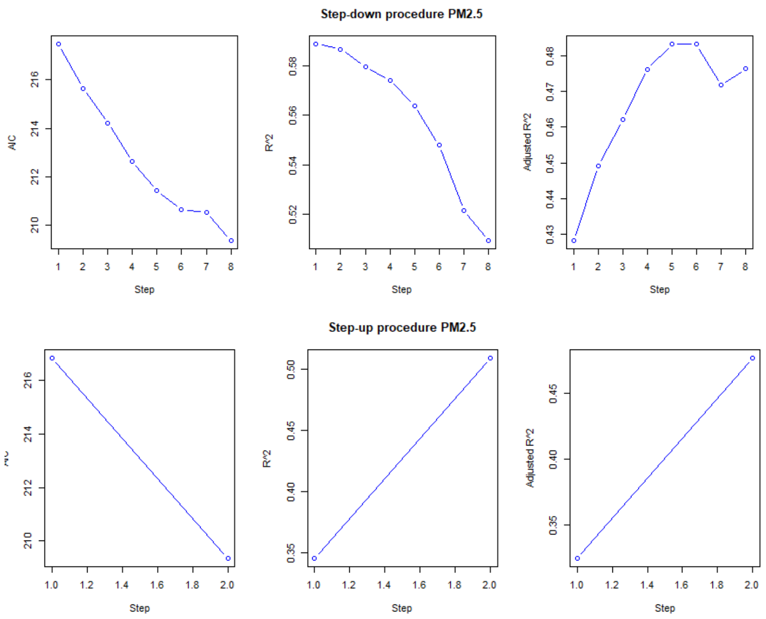


Figure 17: AIC, R^2 and adjusted R^2 results at each step from step-down and step-up for PM2.5

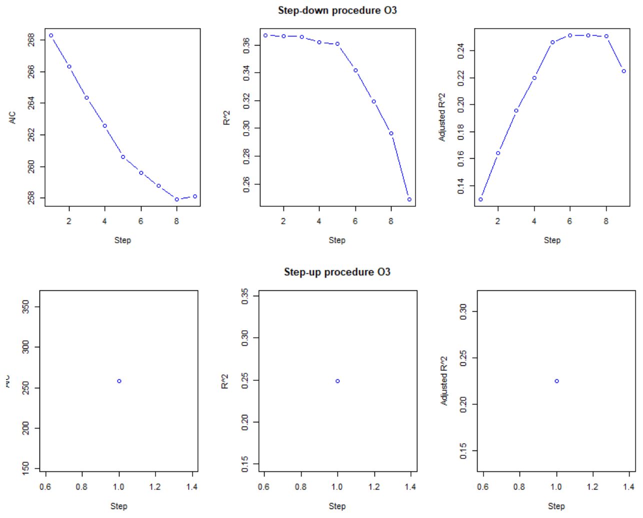


Figure 18: AIC, R^2 and adjusted R^2 results at each step from step-down and step-up for O_3

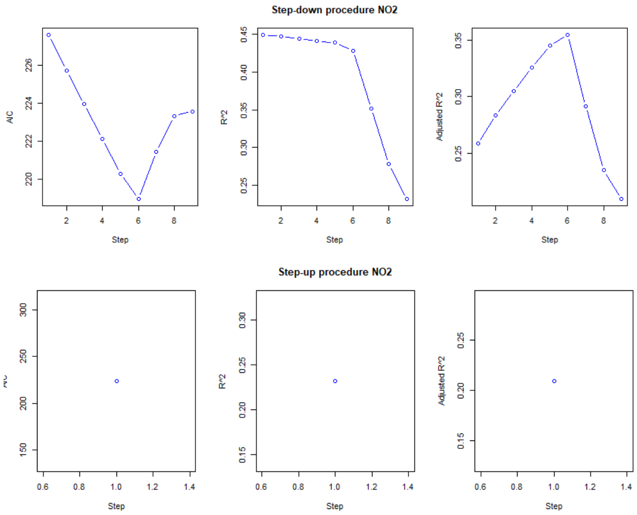


Figure 19: AIC, R^2 and adjusted R^2 results at each step from step-down and step-up for NO_2