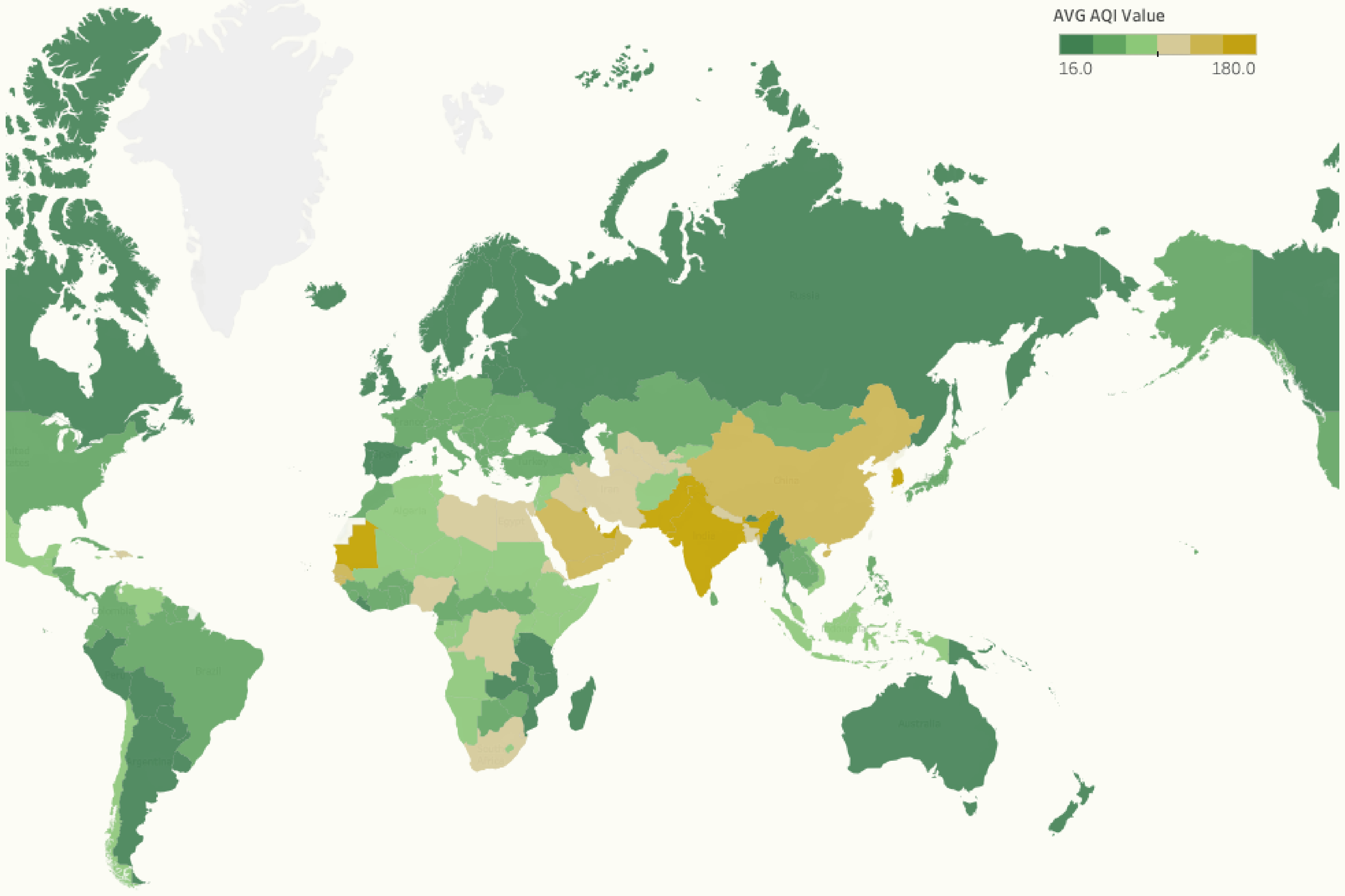
Mapping the Polluted Skies: Analyzing Air Quality Metrics

Zitong LUO
Tsinghua University

Overview

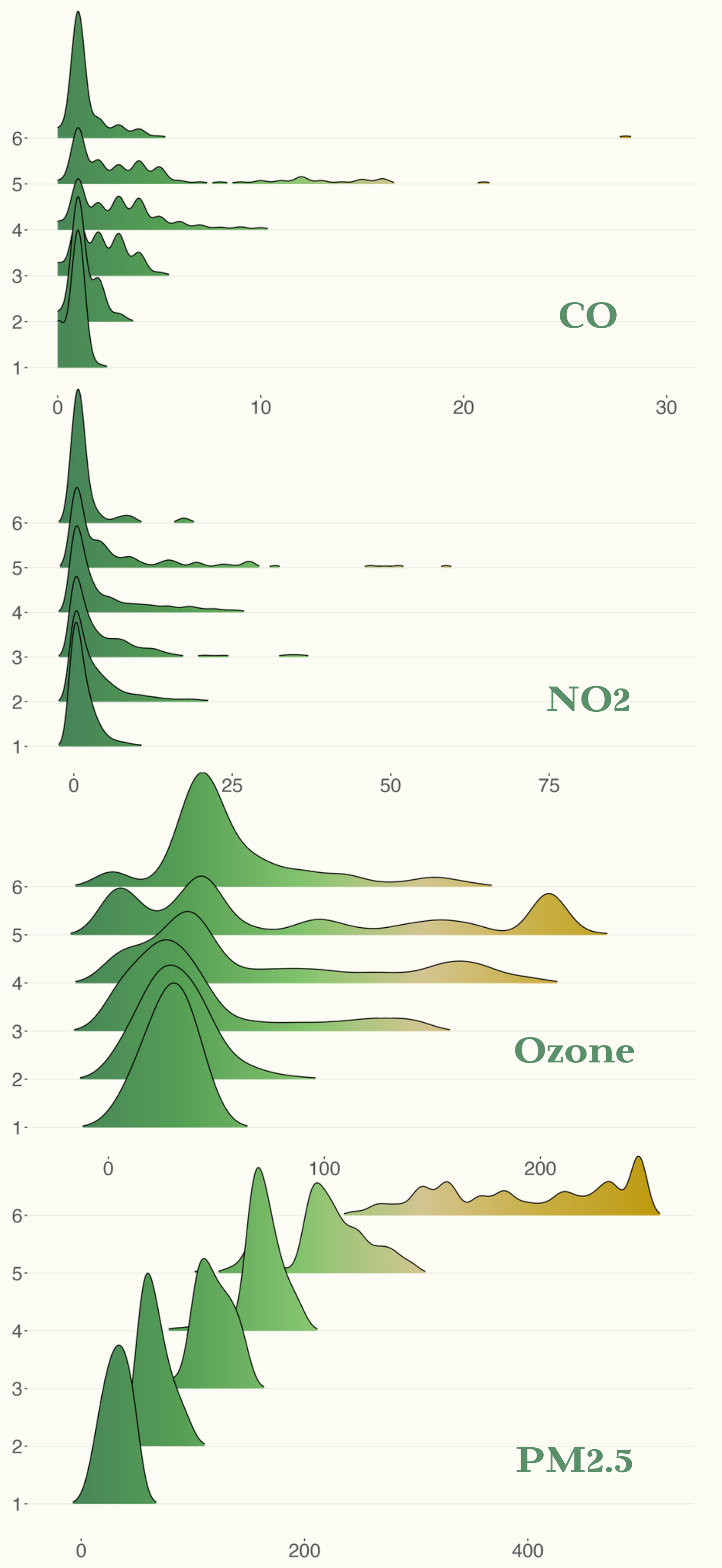


The **Average AQI Value World Map**, by providing a comprehensive overview of average AQI values, highlights areas that are highly polluted and those with relatively cleaner air. Notably, the map demonstrates that **Asia** and **Africa**, in particular, exhibit a concentration of highly polluted regions.

One striking observation is that **India** emerges as the 9th most polluted country. However, what sets India apart from other highly polluted countries is the largest standard deviation of its AQI, implying an unbalanced air quality conditions within India.

Factor Insights

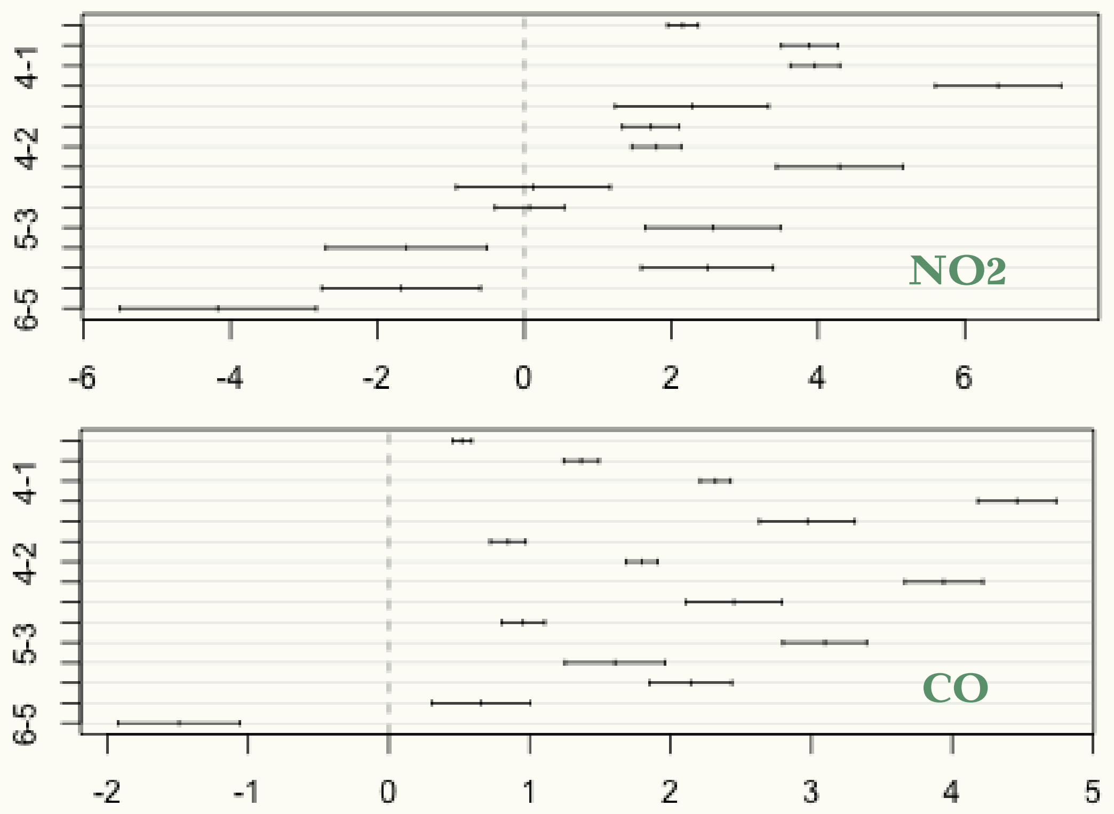
Pollutant AQI Values across Overall AQI Categories



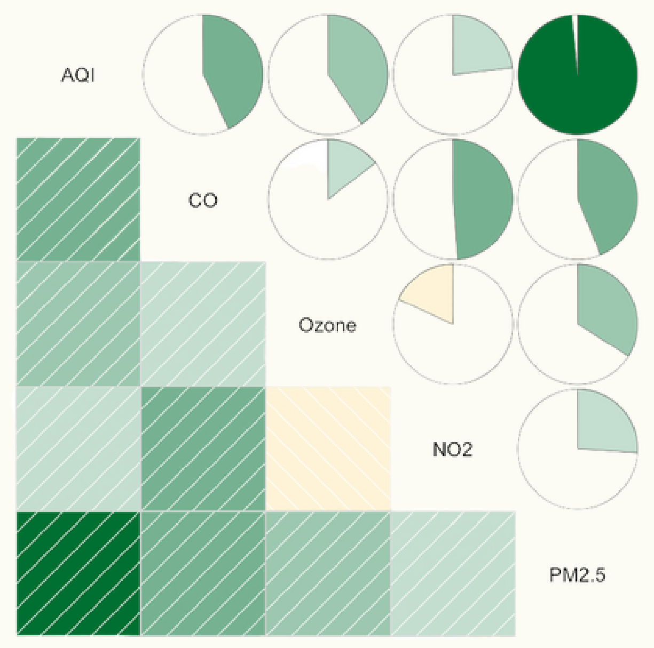
Labeling the AQI categories ('Good' to 'Hazardous') from 1 to 6, the **Ridgeline Charts** of the pollutant AQI values have shown a relatively significant difference in **Ozone** and **PM2.5** AQI values.

To test the differences in means of the other two pollutants, CO and NO2, we opt to use the **Tukey's HSD test** for conducting the family-wise comparison.

95% Family-wise CI by Tukey's HSD



The results indicate that, with a 95% family-wise confidence interval, the AQI values within **CO** are significantly different. However, for two pairs(2-6, 3-4) in **NO2**, there is not much difference.



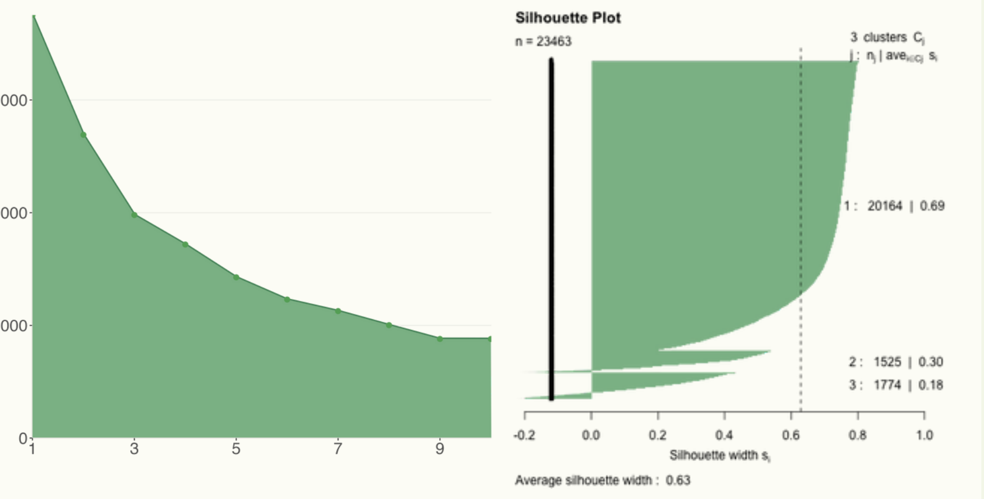
We can further look into the correlations among the AQI values. The **Correlogram** shows an overall positive correlations with one exception between Ozone and NO2.

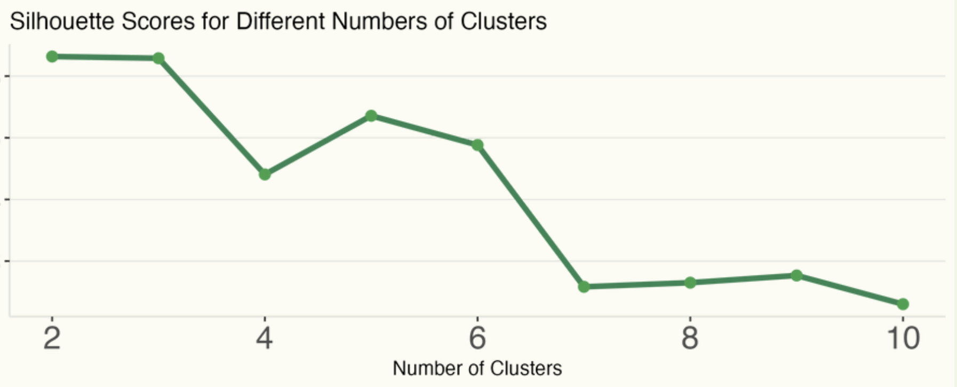
It is noticeable that the correlation between overall AQI values and PM2.5 AQI values has reached **0.98**. This implies that PM2.5 is supposed to be the main contributor.*

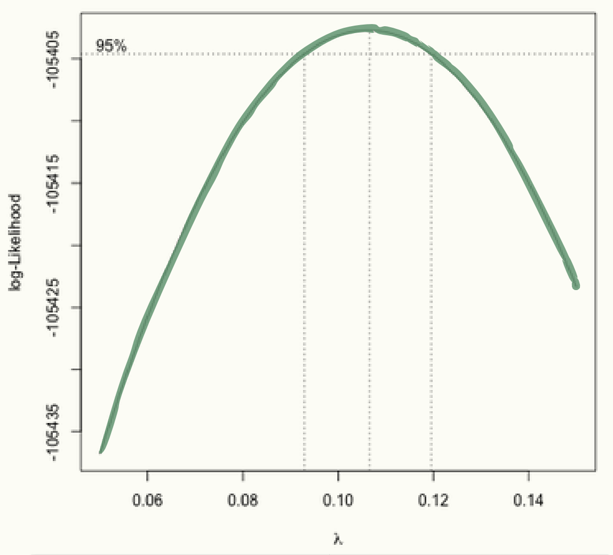
**Overall AQI = Max{categorical AQI}, where PM10 and SO2 are not included in the dataset*

Model Building & Diagnostics

Firstly, we apply **K-Means Clustering** to see if the data can be well grouped based on 4 categorical AQIs. By plotting Within cluster Sum of Squares against the number of clusters, it seems that there is no significant **elbow** point. Taking k = 3 as an example, the **Silhouette Plot** in the right panel shows an average silhouette width of 0.63. But the **Silhouette Score Plot** has shown that the data may not be able to well clustered with k > 2.





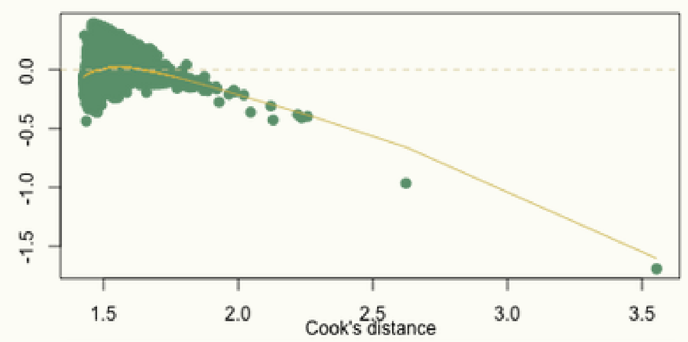


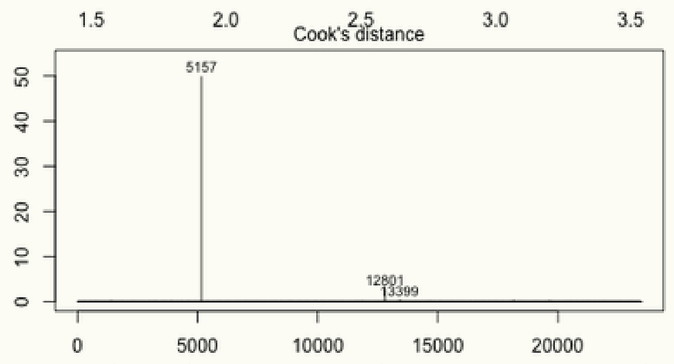
	Estimate	t value	Pr(>|t|)
(Intercept)	1.4209163	1381.79809	0
Ozone	0.0011742	54.91086	0
NO2	0.0047635	36.74853	0
CO	0.0141433	38.27783	0

Considering that PM2.5 is the primary contributor to the AQI values, we propose building a **Linear Regression Model** to estimate PM2.5 based on CO, Ozone, and NO2. As PM2.5 is right-skewed, we derive a Box-cox Plot to determine the appropriate transformation. Here we choose the lambda of 0.1.

After fitting the model, we get the estimates all significantly differing from 0. By plotting the residuals against fitted value, we observe **heteroskedasticity**, **non-linear relationship** and **possible outliers**.

To address the issue of **unequal variance**, we attempted to use a **Weighted Least Squares (WLS) model**. However, the results were not satisfactory. Additionally, we examined possible outliers using the **Cook's Distance Plot**, revealing that the 5157th and 12801st observations are considered influential due to their exceptionally high CO and Ozone AQI values, respectively.

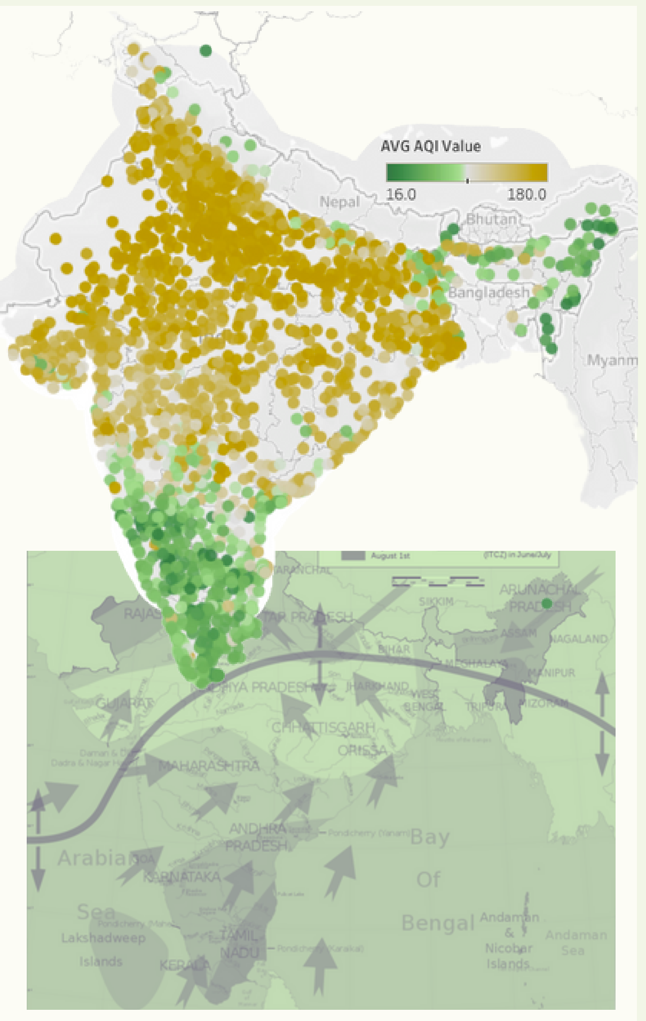




In an effort to capture any **non-linear relationships**, we considered including quadratic&interaction terms in the model. We employed **Mallow's Cp** as the criteria. However, the problem remained unresolved and the **interpretability** was significantly reduced.

The above diagnosis revealed unsatisfactory results, which could be attributed to the **limitations of the existing data** fitting and **significant variations in air quality patterns** across different regions worldwide.

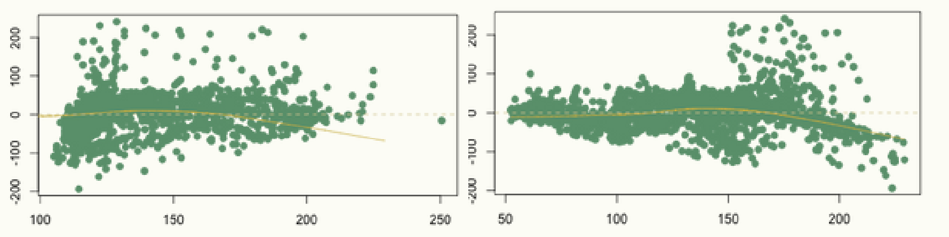
Therefore, we proceeded with an analysis focused on a single country, as mentioned in the beginning, **India**, which exhibits interesting phenomena in its air quality patterns. The AQI gradually increases from the **south** and **east** towards the **northwest**, a phenomenon that is well reflected in regression fitting as well. In the regression results of this single country, there is relatively good satisfaction in terms of **linearity**, **homoscedasticity**, **normality**, and other assumptions.



One possible explanation for this pattern is the influence of the **monsoon** season.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	113.508646	16.7830275	6.763300	0
Latitude	6.604978	0.1851698	35.669852	0
Longitude	-1.488516	0.2149950	-6.923492	0

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	97.1893493	2.0336427	47.7907692	0.0000000
Ozone	0.5758403	0.0410150	14.0397378	0.0000000
NO2	3.4522725	0.5898734	5.8525650	0.0000000
CO	-0.1780252	1.6543328	-0.1076115	0.9143153



From Pollution to Protection

The importance of transitioning from a reactive approach to a proactive stance in safeguarding air quality has always been paramount. Through further analysis using more detailed data, we can enhance our understanding of the issue and pave the way for comprehensive protection measures to be implemented.