Air Quality Prediction using LUR Model: Parameter Reduction and Optimization

Vijay Kumar[1], **Vitt Patel**[2], Dr. Shantanu Sur[3], Dr. Suresh Dhaniyala[4], Dr. Supraja Gurajala[5], Dr. Sumona Mondal[1]

[1]Department of Mathematics, [2]David. D. Reh School of Business, [3]Department of Biology, [4]Department of Mechanical & Aeronautical Engineering, [5]Department of Computer Science

Introduction

- Air pollution is one of the most important public health risks, causing one in eight premature deaths globally[1].
- Pollutants such as PM, NO_x, SO_x, etc. are measured while monitoring air quality. PM2.5[2] are fine inhalable particles that lodge deeply into lungs causing various respiratory diseases.
- Conventional instruments for air quality measurements are expensive and difficult to maintain and hence these measurements are only available at low spatial resolution.

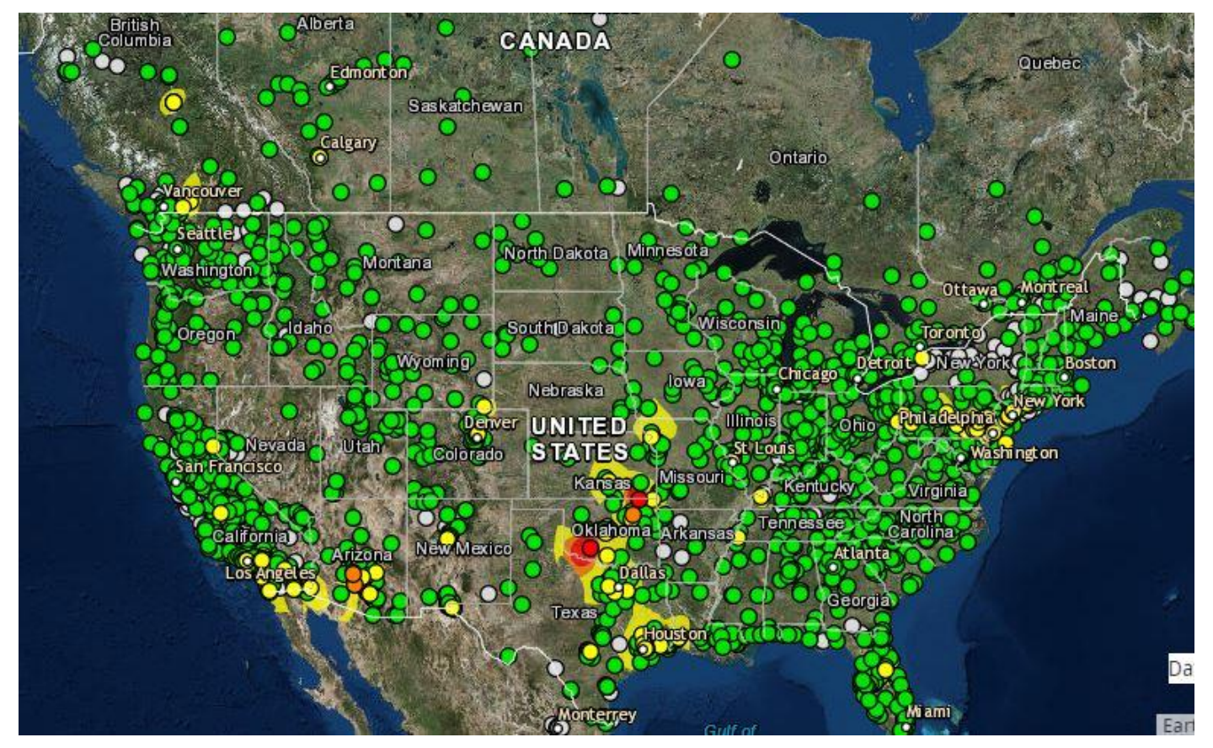

Figure 1 – Air Quality Monitoring Stations in the USA

- Land Use Regression (LUR) models have been built using a brute-force approach, employing all available parameters and are used to predict air quality at high resolution as an alternative to conventional instruments[3].
- This study will aid in building more robust and generic LUR models for prediction of PM2.5 where air quality is not monitored.
- By predicting and monitoring air pollution, the environment can be made more sustainable and air quality can be improved, which subsequently improves human health.

Objectives

- Identification of optimal parameters for LUR models by Analysis of Variance (ANOVA).
- To build a model using optimal parameters to predict PM2.5 using polynomial regression.
- Validation of model using residual analysis.

Methods

- The data for this research was collected using Python Application Programming Interface (API) from the EPA sites for years 2015-16. Data was then stored in MySQL database.
- The outliers and noise in the data were cleaned. Post data collection and data cleaning, the data was statistically analyzed using RStudio v1.1.456 and SPSS v1.0.0.1174.
- Multi-collinearity check was run to determine the collinearity between the parameters.
- To identify the significant parameters that affect PM2.5, one-way ANOVA was implemented.
- To check higher-order interaction between parameters, two-way ANOVA was performed.
- The combination of optimal parameters and two-way interaction was used to build a model using the multiple polynomial regression technique that can predict PM2.5.

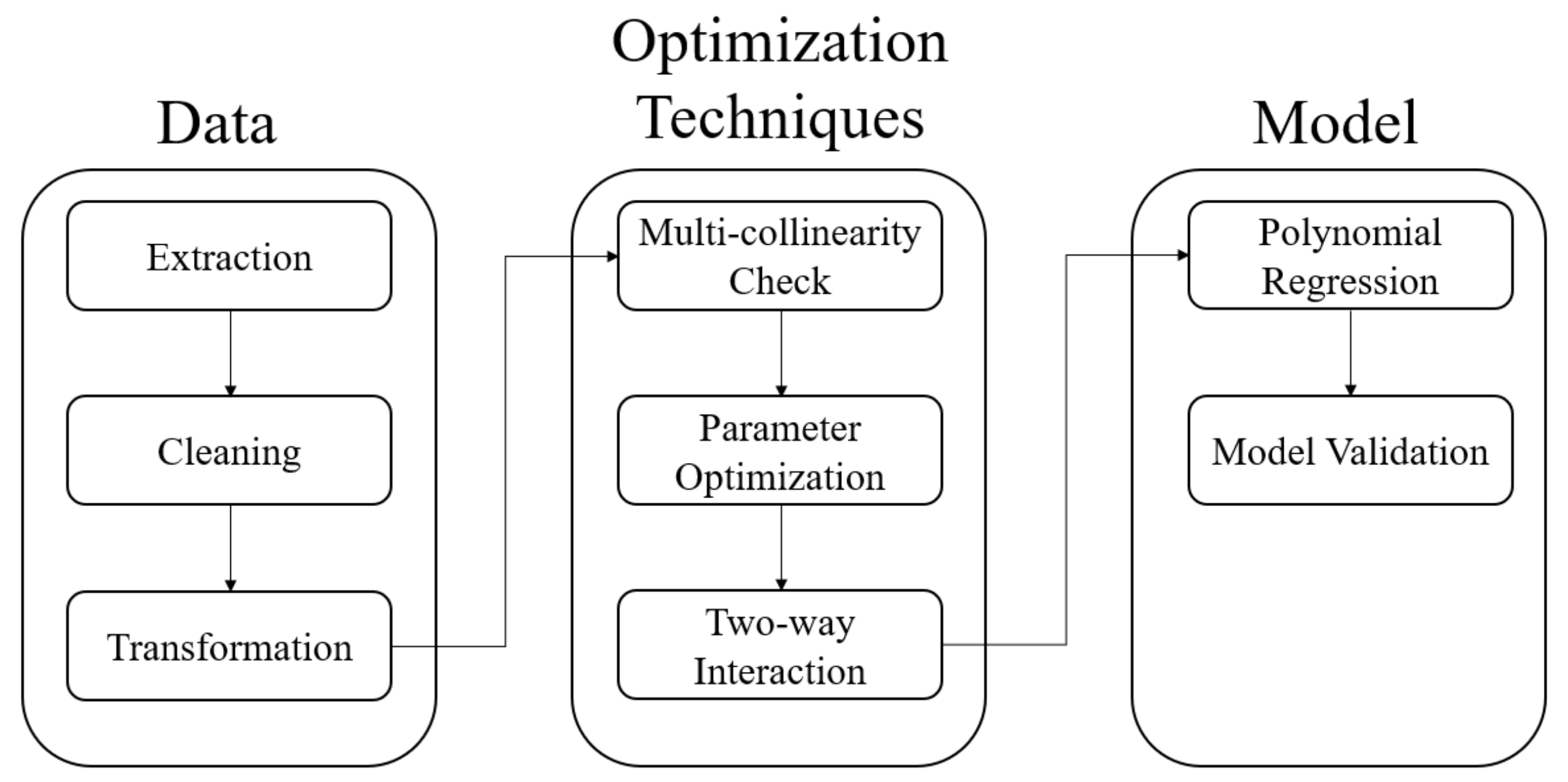

Figure 2 – Flowchart of the Methodology Used

Results & Discussion

Multi-collinearity Check:

Using correlation analysis, it is determined that all the parameters are independent of each other and the parameters have a very low correlation.

	Relative Humidity	Wind Speed	Outdoor Temperature	Barometric Pressure	Wind Direction	Black Carbon
Relative Humidity	1.00	-0.28	-0.22	-0.09	0.04	0.19
Wind Speed	-0.28	1.00	-0.05	-0.29	0.11	-0.33
Outdoor Temperature	-0.22	-0.05	1.00	-0.27	0.03	0.13
Barometric Pressure	-0.09	-0.29	-0.27	1.00	-0.21	0.08
Wind Direction	0.04	0.11	0.03	-0.21	1.00	-0.18
Black Carbon	0.19	-0.33	0.13	0.08	-0.18	1.00

Note: All p-values < 0.001

Table 1 – Results for Multi-collinearity Check

Parameter Optimization:

By implementing one-way ANOVA, barometric pressure, black carbon, wind direction, wind speed, outdoor temperature and relative humidity are the significant parameters for PM2.5.

Independent Variables	F-Statistics	p-value
Black Carbon	1832.001	$p < 0.001$
Wind Direction	223.581	$p < 0.001$
Outdoor Temperature	65.360	$p < 0.001$
Relative Humidity	63.423	$p < 0.001$
Wind Speed	32.959	$p < 0.001$
Barometric Pressure	17.034	$p < 0.001$

Table 2 – Results for One-way ANOVA

Two-way Interaction:

By performing two-way ANOVA to get significant interaction between two variables, it is found that 12 out of 15 interactions are significant.

Interaction between Parameters	F-Statistics	p-value
Wind Speed * Black Carbon	36.502	$p < 0.001$
Wind Direction * Black Carbon	31.831	$p < 0.001$
Temperature * Black Carbon	21.720	$p < 0.001$
Temperature * Wind Direction	15.035	$p < 0.001$
Relative Humidity * Black Carbon	14.528	$p < 0.001$
Relative Humidity * Wind Direction	13.241	$p < 0.001$
Barometric Pressure * Black Carbon	11.145	$p < 0.001$
Temperature * Relative Humidity	8.402	$p < 0.001$
Relative Humidity * Wind Speed	6.110	$p < 0.001$
Relative Humidity * Barometric Pressure	4.205	$p < 0.001$
Temperature * Wind Speed	3.571	0.003
Barometric Pressure * Wind Direction	2.708	0.004
Wind Speed * Wind Direction	1.798	0.095
Wind Speed * Barometric Pressure	1.617	0.152
Temperature * Barometric Pressure	1.29	0.258

Table 3 – Results for Two-way ANOVA

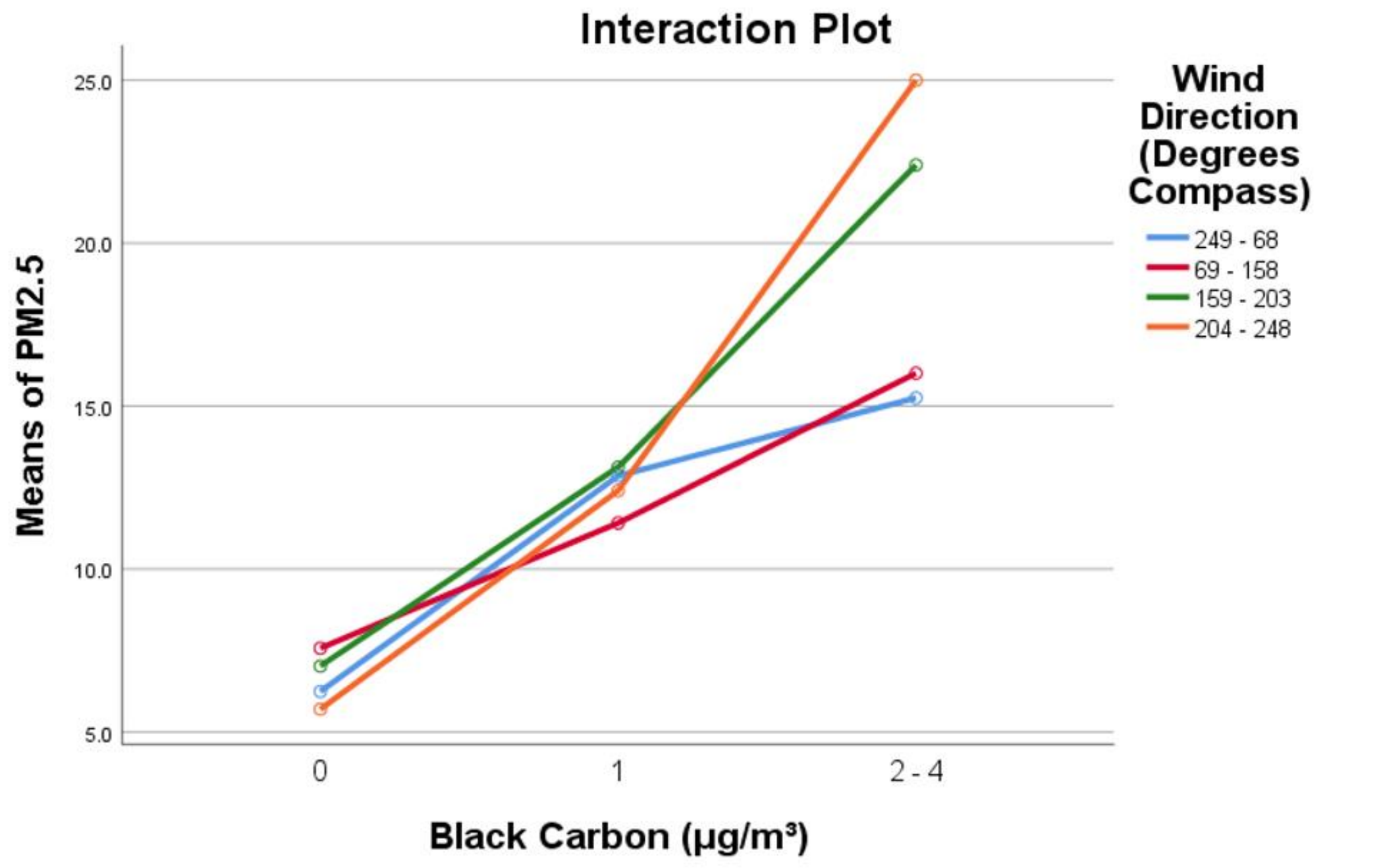

Figure 3 – Interaction plot for Black Carbon & Wind Direction

Polynomial Regression Model:

Models containing all significant parameters and two-way interactions were built.

$$PM2.5 = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2$$

Parameters used for Polynomial Regression Model	R² Obtained
Wind Speed, Black Carbon	0.213
Relative Humidity, Black Carbon	0.212
Wind Direction, Black Carbon	0.211
Temperature, Black Carbon	0.210
Barometric Pressure, Black Carbon	0.208
Wind Speed, Wind Direction	0.022
Relative Humidity, Wind Direction	0.021
Relative Humidity, Wind Speed	0.017
Temperature, Wind Speed	0.017
Temperature, Wind Direction	0.016
Wind Speed, Barometric Pressure	0.013
Barometric Pressure, Wind Direction	0.012
Temperature, Relative Humidity	0.010
Relative Humidity, Barometric Pressure	0.009
Temperature, Barometric Pressure	0.007

Table 4 – R-squared Values Obtained for Models

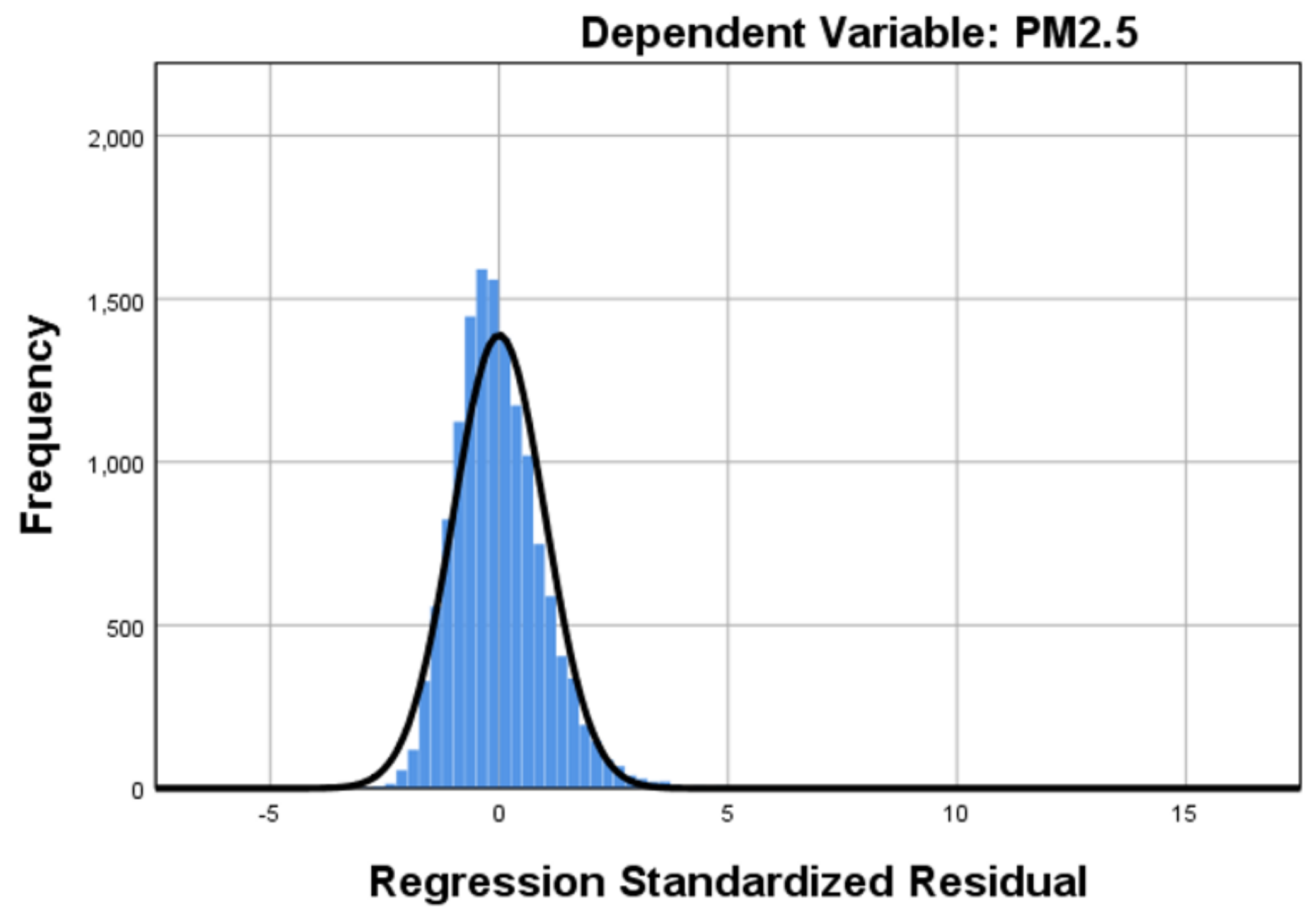

Figure 4 – Histogram for Regression Standardized Residual

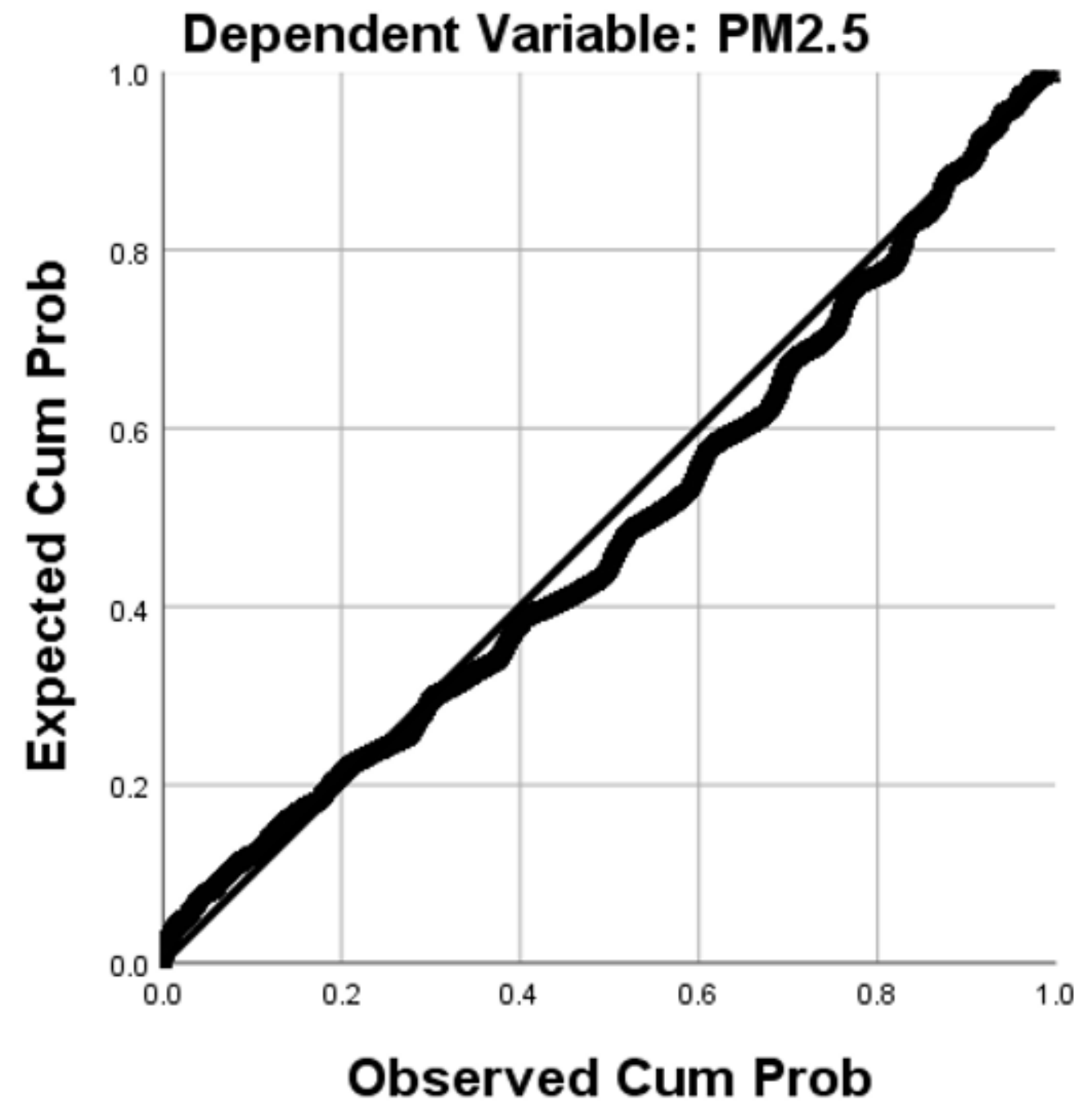

Figure 5 – Normal P-P Plot of Regression Standardized Residual

Conclusion & Future Work

Though all parameters and their two-way interactions are statistically significant, Black Carbon with Temperature, Relative Humidity, Wind Speed and Wind Direction explains utmost variability which all parameters together can exhibit. Hence, keeping Black Carbon with those parameters is important to build any LUR model for predicting PM2.5.

In this study, it was found that all six parameters and interaction between them is significant. In the future, high-resolution data from traffic, population, and land-use variables can also be incorporated, optimized and analyzed with the help of dimension reduction techniques and ANOVA. Subsequently, optimal parameters will be obtained and included to make more robust LUR models using machine learning techniques such as DSA. Such models will allow for accurate prediction of PM2.5 at high resolution necessary for health studies.

Acknowledgements/References

We would like to thank Srishtee Dalal for assisting us throughout our research.

[1] World Health Organization ambient air pollution: A global assessment of exposure and burden of disease 2016.

[2] Particulate Matter (PM2.5) Trends *Retrieved from* https://www.epa.gov/air-trends/particulate-matter-pm25-trends

[3]Philip K. Hopke, Hourly Land-use Regression Models Based on Low-cost PM Monitor, Environmental Research 167 7-14 2018.