Stats c173 Pollution in India

Yonatan Khalil

2/28/2021

```
## ------------------------------------------------------------
##  Analysis of Geostatistical Data
##  For an Introduction to geoR go to http://www.leg.ufpr.br/geoR
##  geoR version 1.8-1 (built on 2020-02-08) is now loaded
## ------------------------------------------------------------
```

Introduction

Across the globe, Pollution is one of the most serious issues contributing to the degradation of our environmental and health. Each industry has pooled in efforts to mitigate emissions, but one of the most important pieces for improving the collective issue is data. In this project we specifically look towards data from 152 monitoring sites in India collected in 2019. The dataset is sourced from Kaggle and is labeled "Air pollution dataset including PM2.5, PM10, OZONE, NO2, SO2, CO pollutant information" (https://www.kaggle.com/rabhar/air-pollution-dataset-india-2019). This data is geostatistical in nature as there are a select number of sites observing pollutants for the general area.

Of the six factors, our variable of interest is PM2.5 defined as fine inhalable particles with diameters that are generally 2.5 micrometers and smaller according to the Environmental Protection Agency. This variable was chosen due to the broad health concerns associated with such particles entering the lungs along with the importance for the general public to avoid PM2.5 exposure. The goal of this project is to find the best method of predicting such geostatistical data in order to truly see the broader picture of PM2.5 pollution in the region.

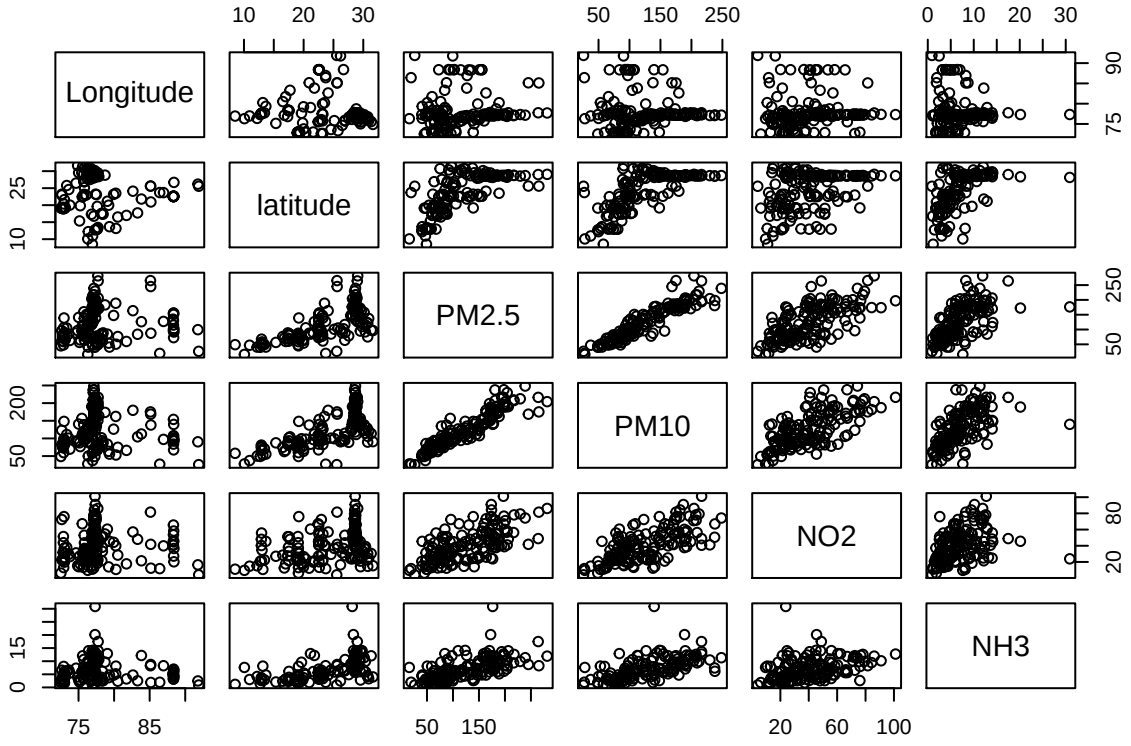
Data Cleaning

```
## [1] 152
```

```
##   Longitude latitude     PM2.5     PM10      NO2       NH3
## 1  75.80860 30.90280 137.44845 120.7457 24.28522  3.290378
## 2  76.87588 29.96694 117.50254 125.5901 33.04431  6.802246
## 3  76.41550 29.80060 109.90898 126.8460 24.69276 13.993138
## 4  86.41467 23.70791  19.39868  27.0815 11.73128  1.973568
## 5  77.23383 28.58028 174.14185 181.6426 84.26800  9.449165
## 6  88.34742 22.61197 155.27978 153.9334 53.48521  5.745091
```

The data cleaning process began with removing all incomplete observations with the complete.cases function. We then identified the uniques locations with the unique function and the id variable which defines the site values. There were 152 unique sites and in order to create a usable dataset, the mean of four predictors were chosen (PM2.5, PM10, NO2, and NH3). Finally, the columns of the dataset were named and the result was a clean dataset we could utilize.

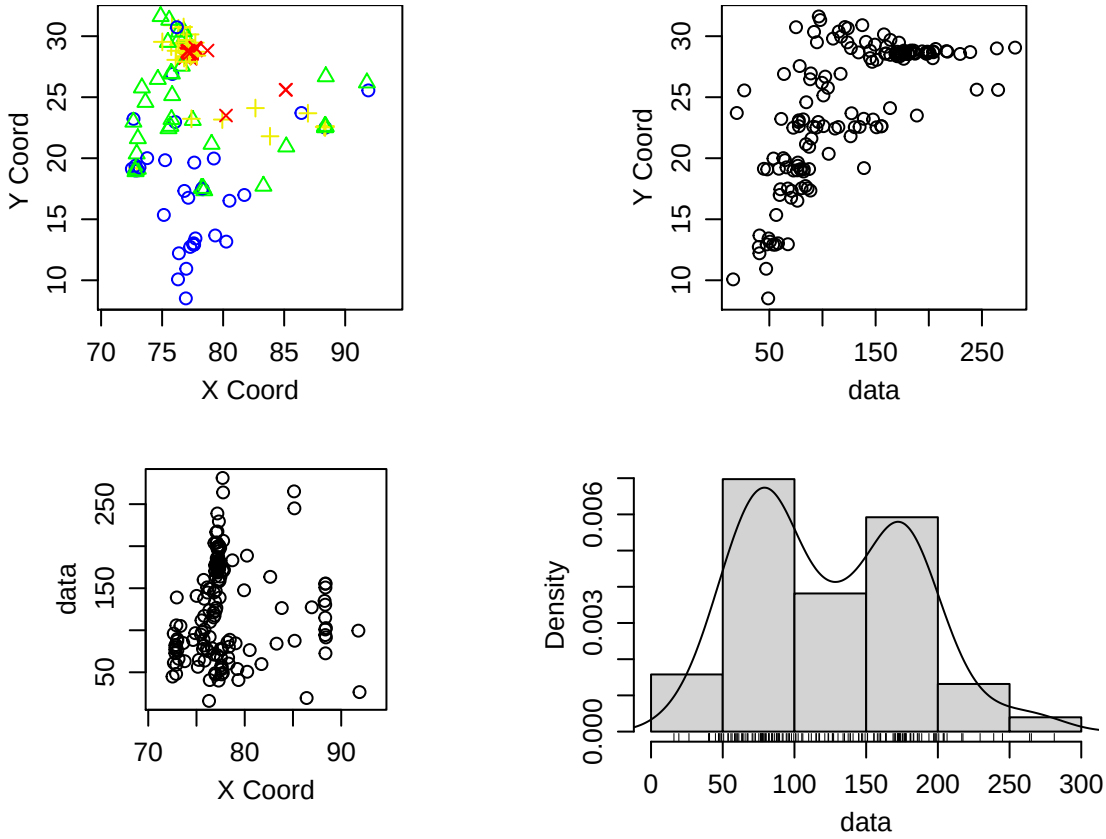
Non-Spacial Exploratory Analysis




```
##             Longitude    latitude      PM2.5       PM10        NO2         NH3
## Longitude  1.00000000 -0.02543218 0.05760045 -0.05906867 0.02565364 -0.07959489
## latitude  -0.02543218  1.00000000 0.71190852  0.71256348 0.37197592  0.57218701
## PM2.5      0.05760045  0.71190852 1.00000000  0.92375181 0.67015878  0.62588988
## PM10      -0.05906867  0.71256348 0.92375181  1.00000000 0.68498725  0.61156509
## NO2        0.02565364  0.37197592 0.67015878  0.68498725 1.00000000  0.37034008
## NH3       -0.07959489  0.57218701 0.62588988  0.61156509 0.37034008  1.00000000
```

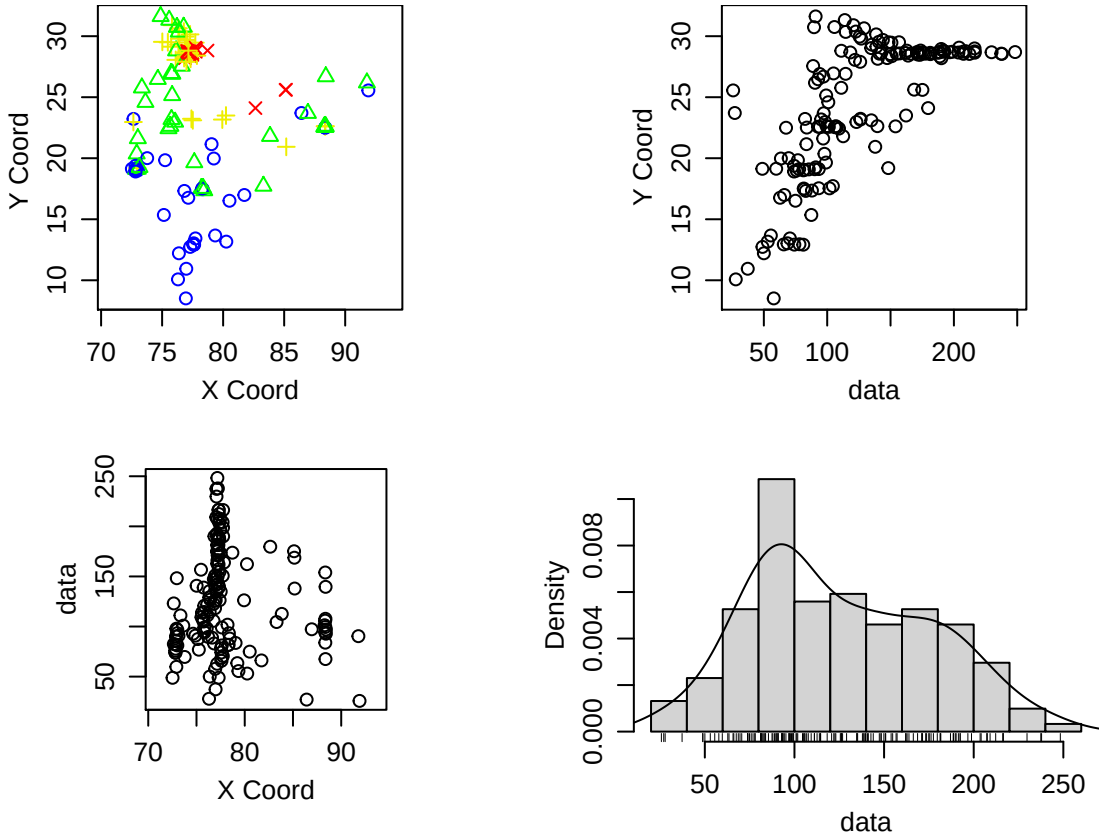
From the matrix plots and the correlation matrix we find that the four chosen pollutants have a very low correlation with longitude location values, but a moderately high correlation with the latitude location values. In addition, we should note that all pollutants are moderately correlated with each other and our variable of interest PM2.5 is significantly correlated with a related pollutant PM10 (.92375181).

PM2.5



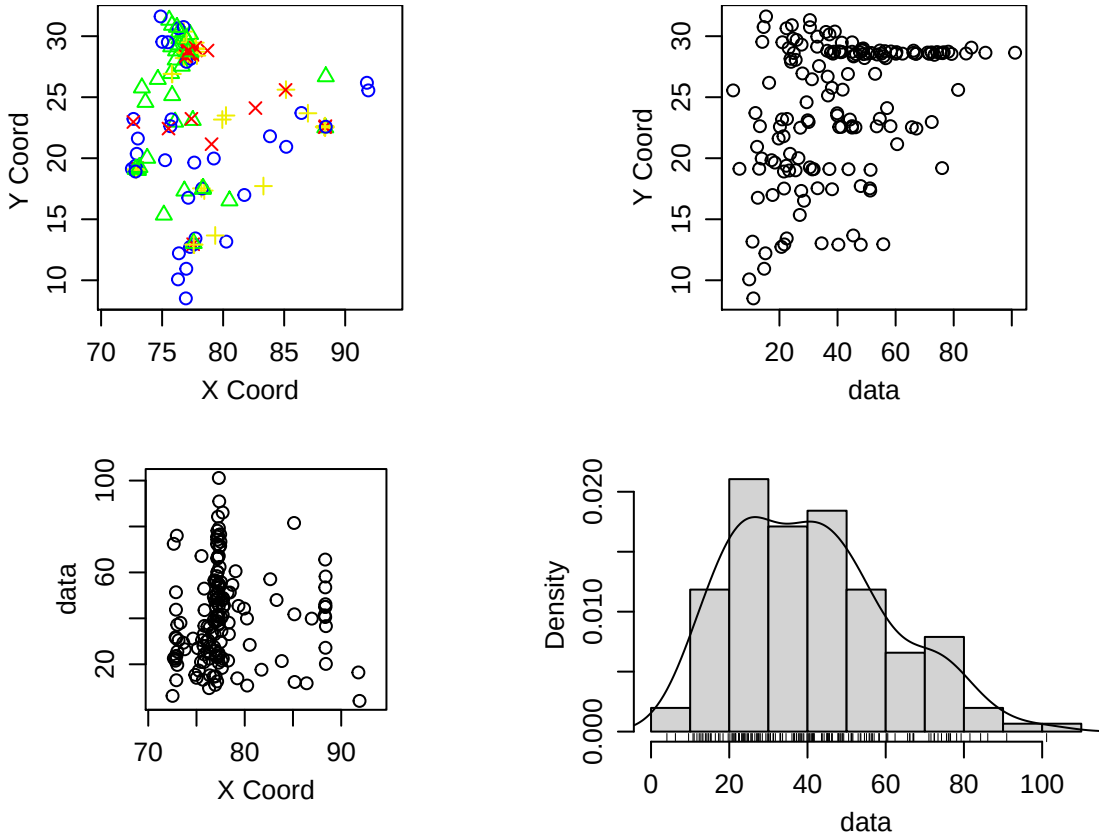
PM2.5 shows a nearly linear correlation to its Y coordinate or latitude values and the distribution of the variable is bimodal.

PM10



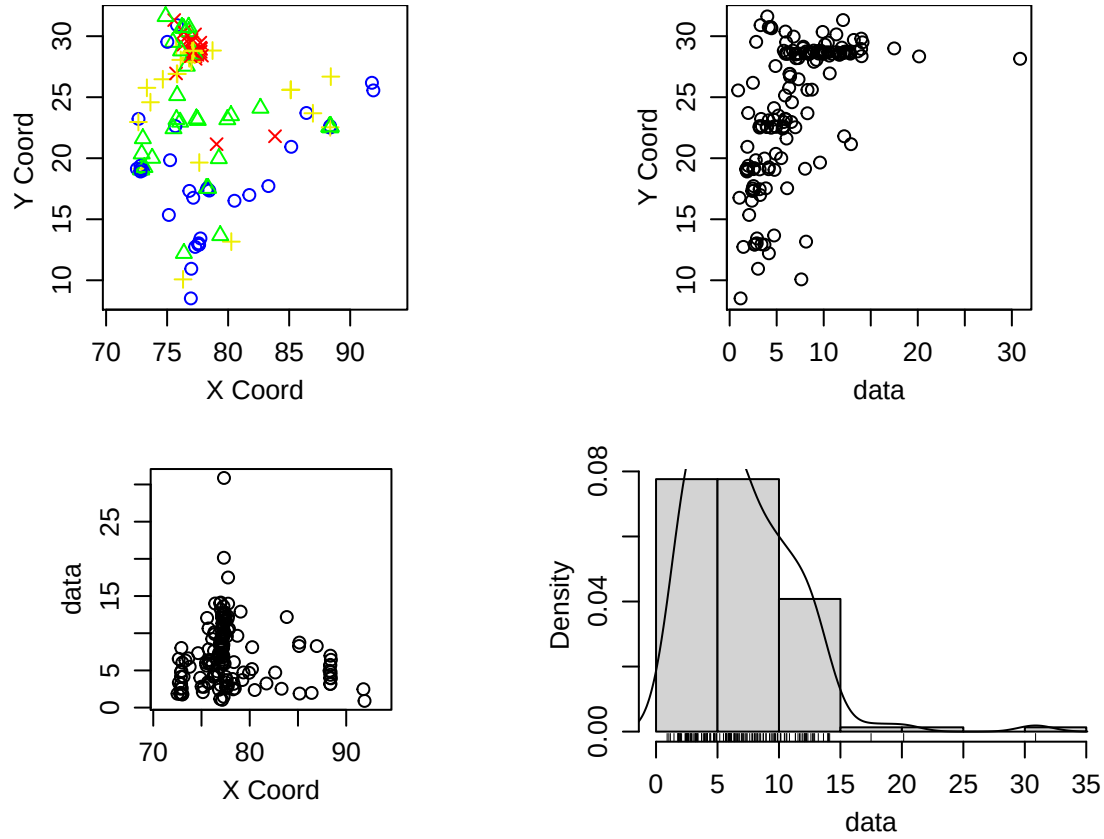
PM10 also shows a nearly linear correlation to its Y coordinate or latitude values and the distribution of the variable is Unimodal.

NO2

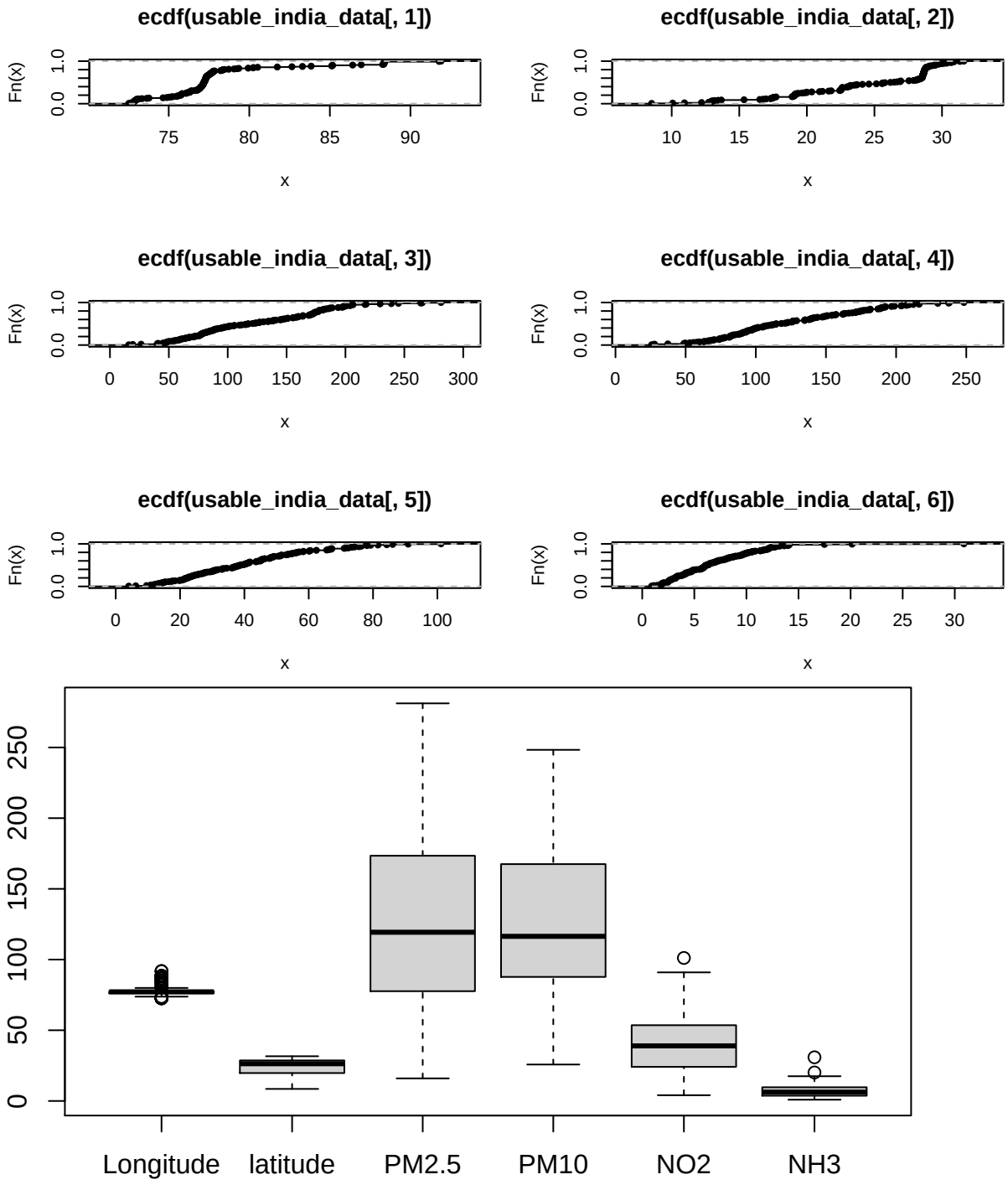


NO2 once again shows a very clear linear correlation to its Y coordinate or latitude values and the distribution of the variable is Unimodal.

NH3

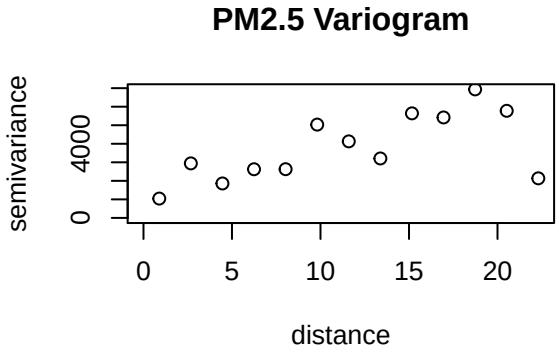


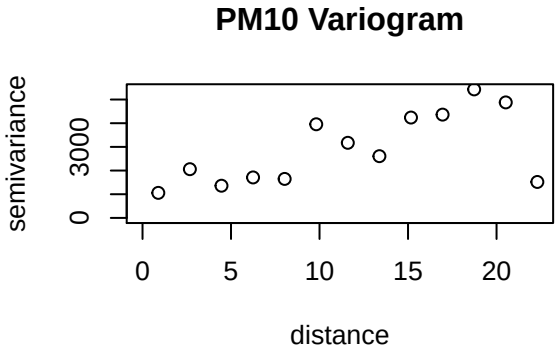
NH3 is the final predictor that shows a very clear linear correlation to its Y coordinate or latitude values and the distribution of the variable is Unimodal. However, there is a distinct skew in our data as most observations of NH3 fall between 0 and 15.

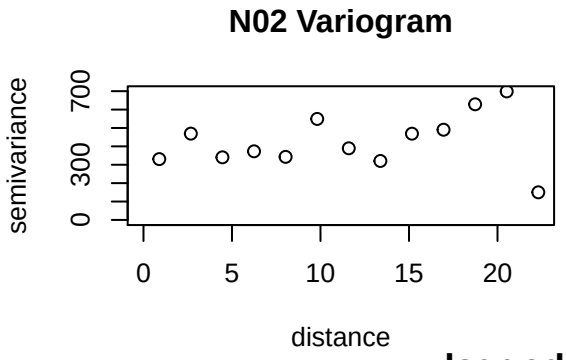


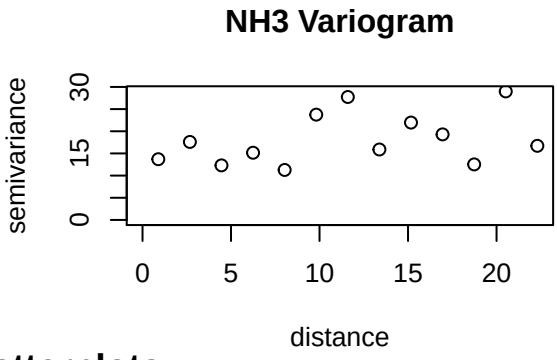
The boxplot above shows that the distribution of our target variable has a wider span and more clear distribution than that of other pollutants. Also, adding to what was found from the individual examinations of variables above we find that PM10 has a similar distribution to that of PM2.5.

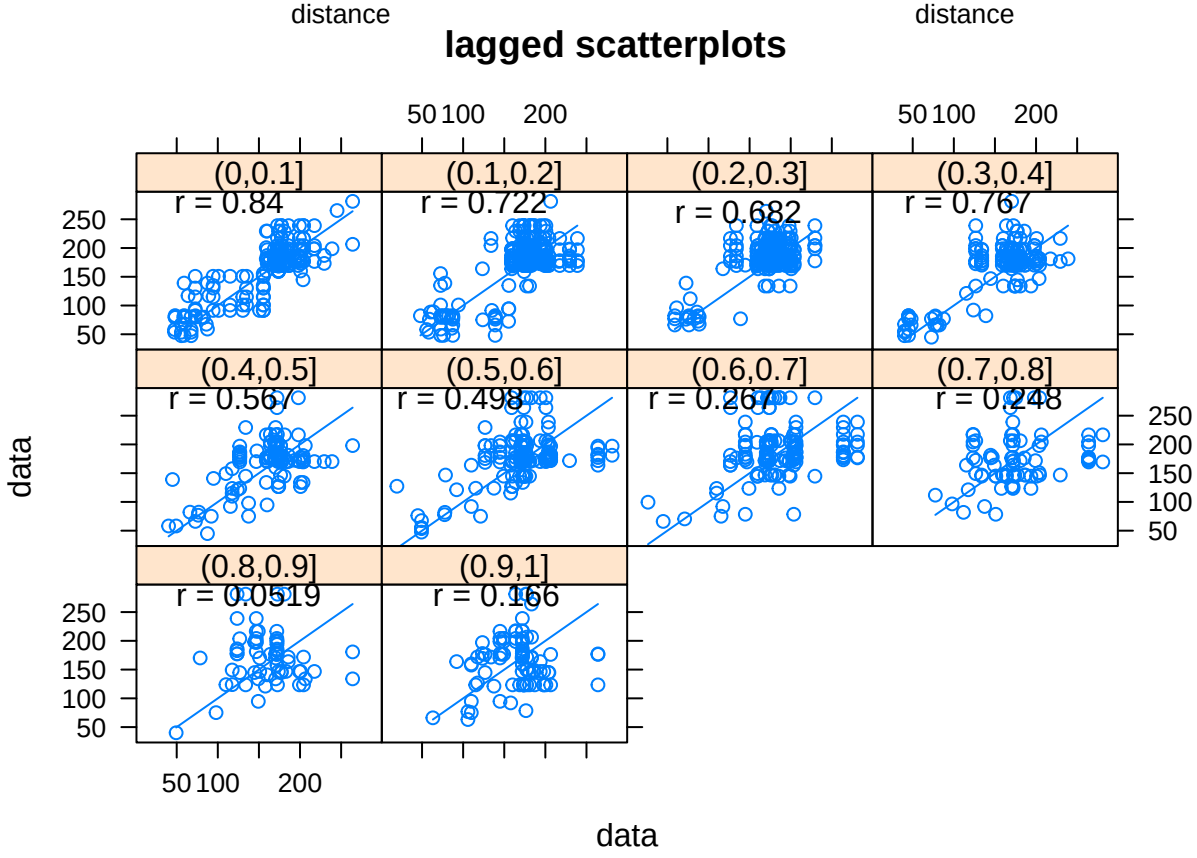
Spacial Analysis





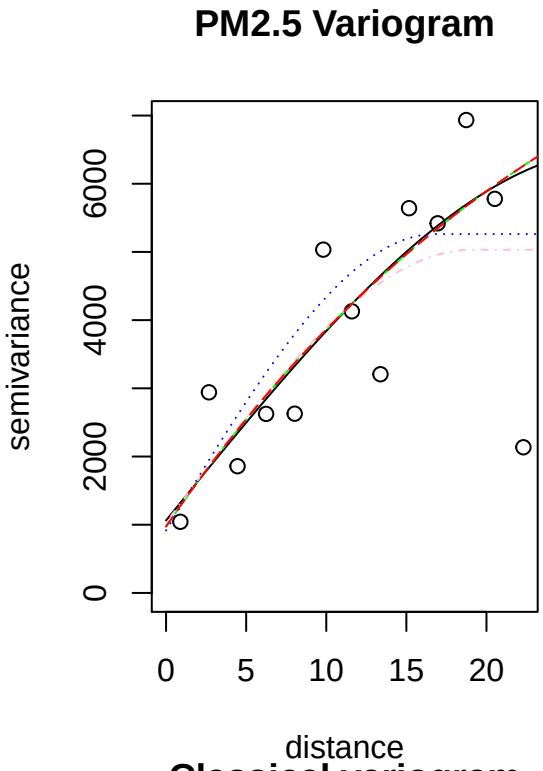


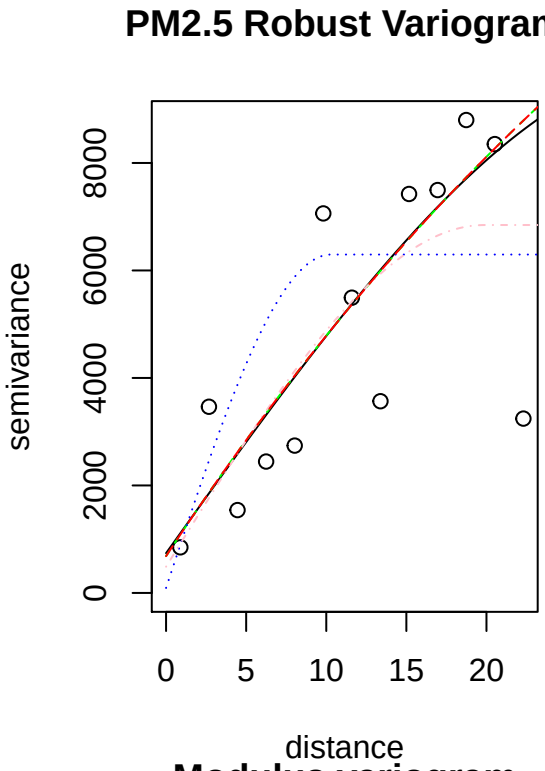


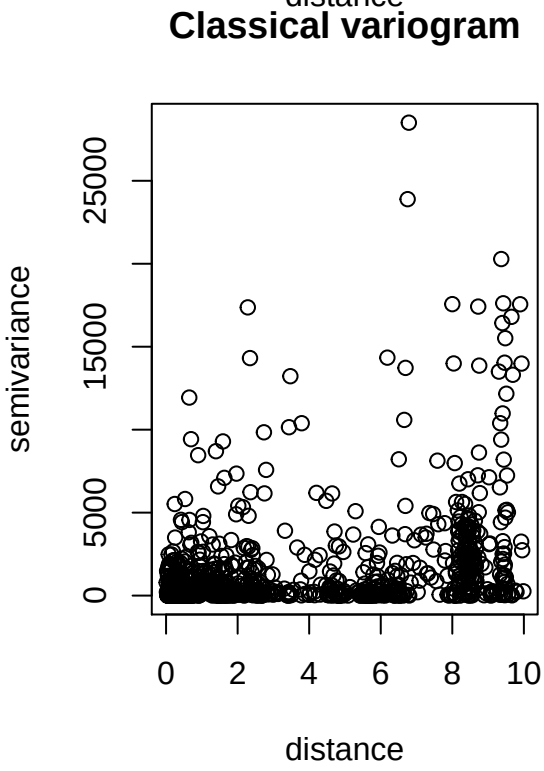


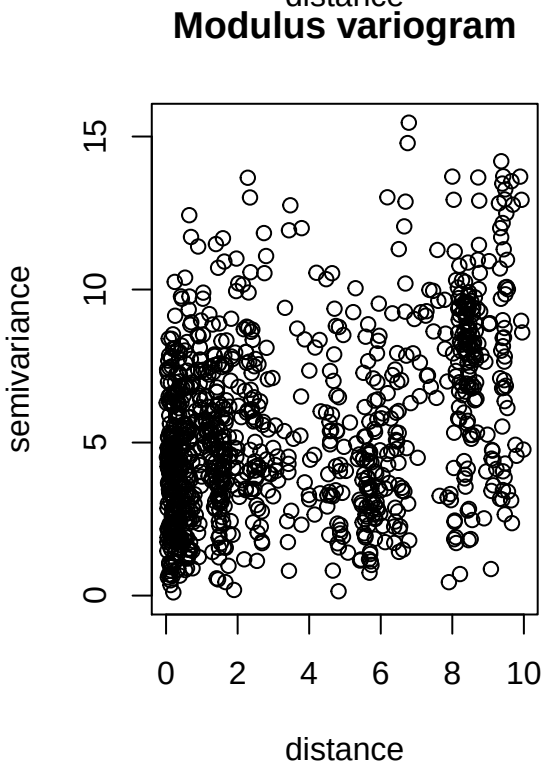
From the variograms above we notice that there is a spatial component to the variance of the data of each

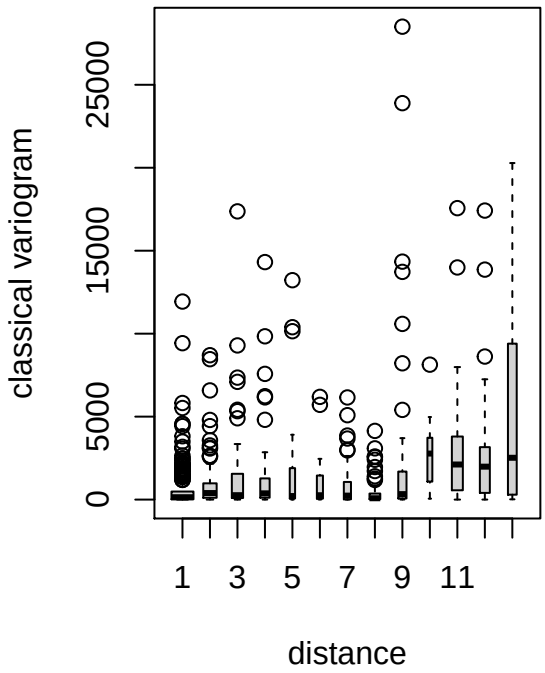
pollutant. Moving towards the h-scatterplot, we find relation between the correlation coefficient and h (the separation distance).

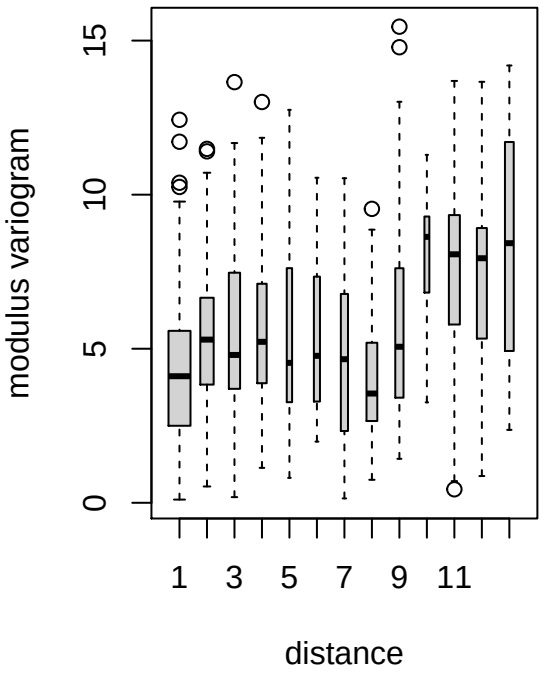






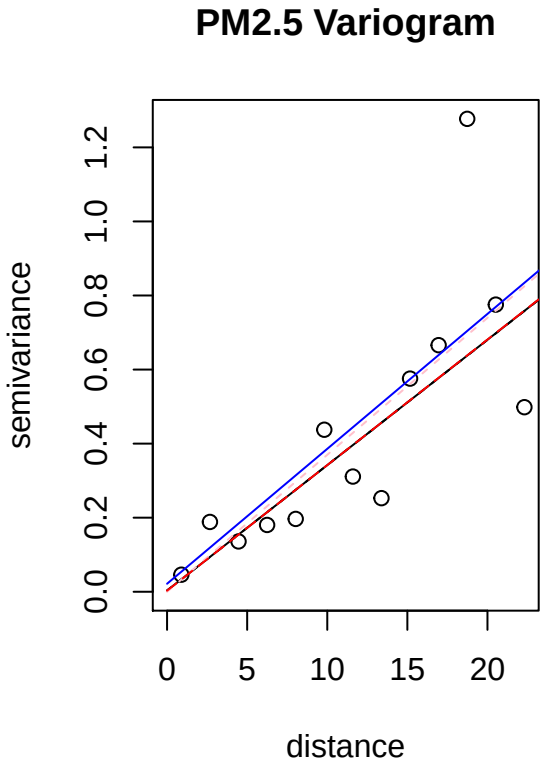


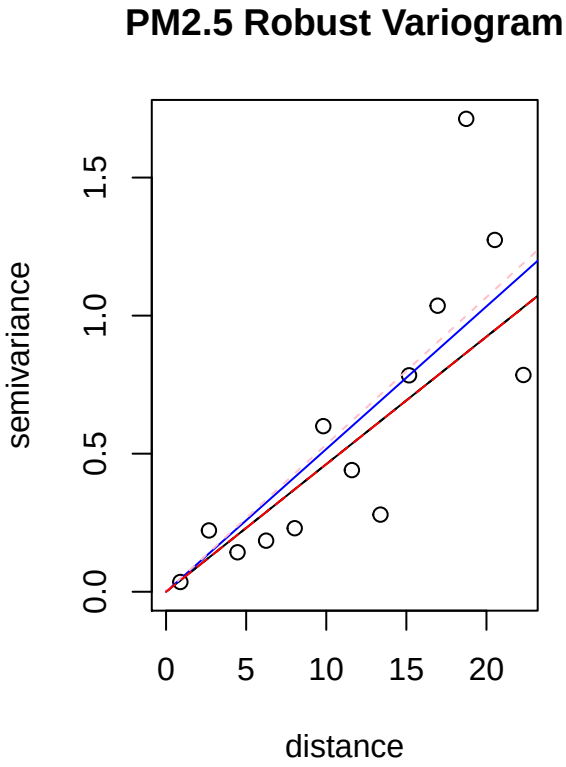




By plotting multiple variograms models on the plot we find that the best option from those plotted and the choice that is closest to the minimization model is a variogram model with partial sill 3000, range of 17.5, and nugget of 500.

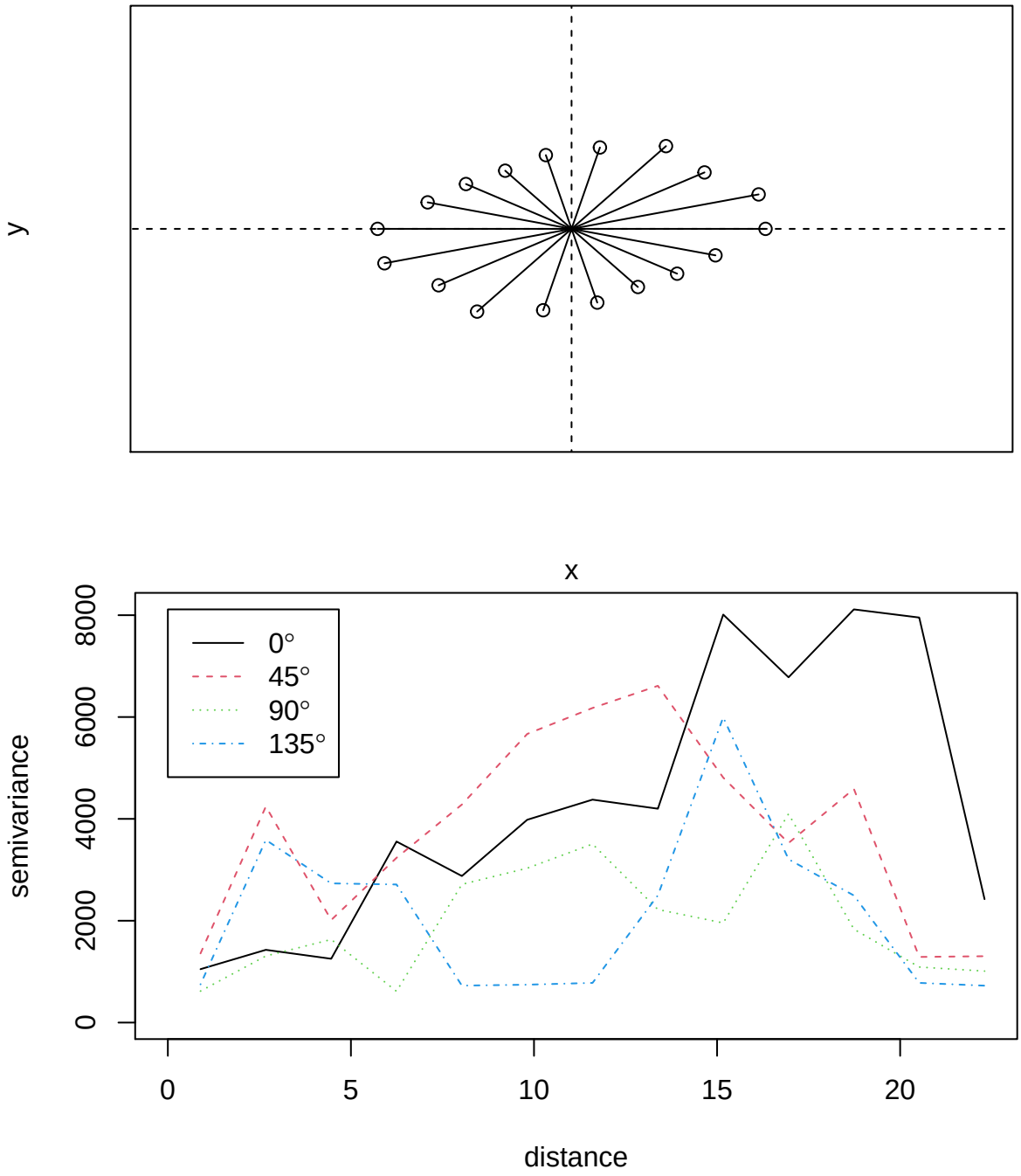
logarithmic variable of interest:





Plotting the variagram with identical metrics we find that the variogram model fits the plot and is very similar to the minimization variogram. When we include the original boxplots of PM2.5 values and its clear distribution, it seems that a log tranformation would disrupt the interpretability of our predictions rather

than improve accuracy and therefore we use the basic PM2.5 values instead of log(PM2.5)

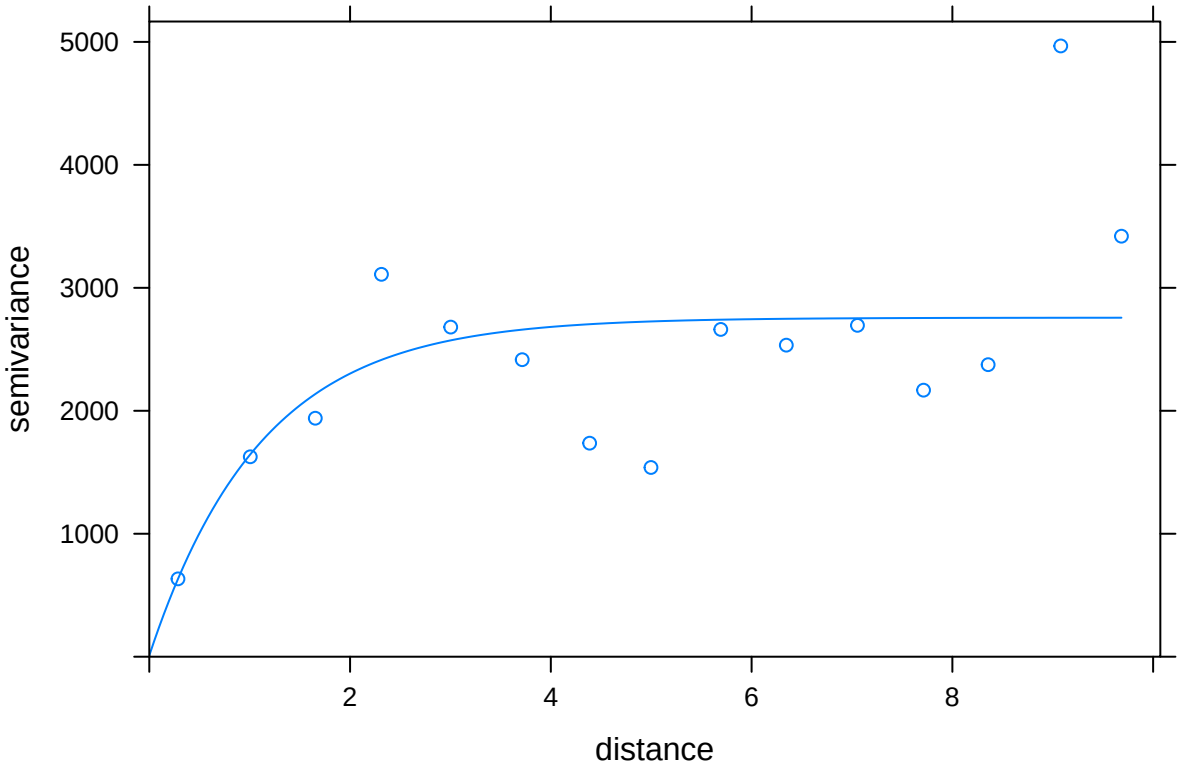


The final aspect of our spacial analysis involves checking for a trend based on the direction of our variogram. At first the rose diagram was believed to indicate that there was a trend in our data, but when examining directional variograms for 4 directions we find that the given variagrams are not easily distinguishable nor is there a parabolic shape indicating a clear trend. Based on the four directional variogram the trend of our data was found to be negligible.

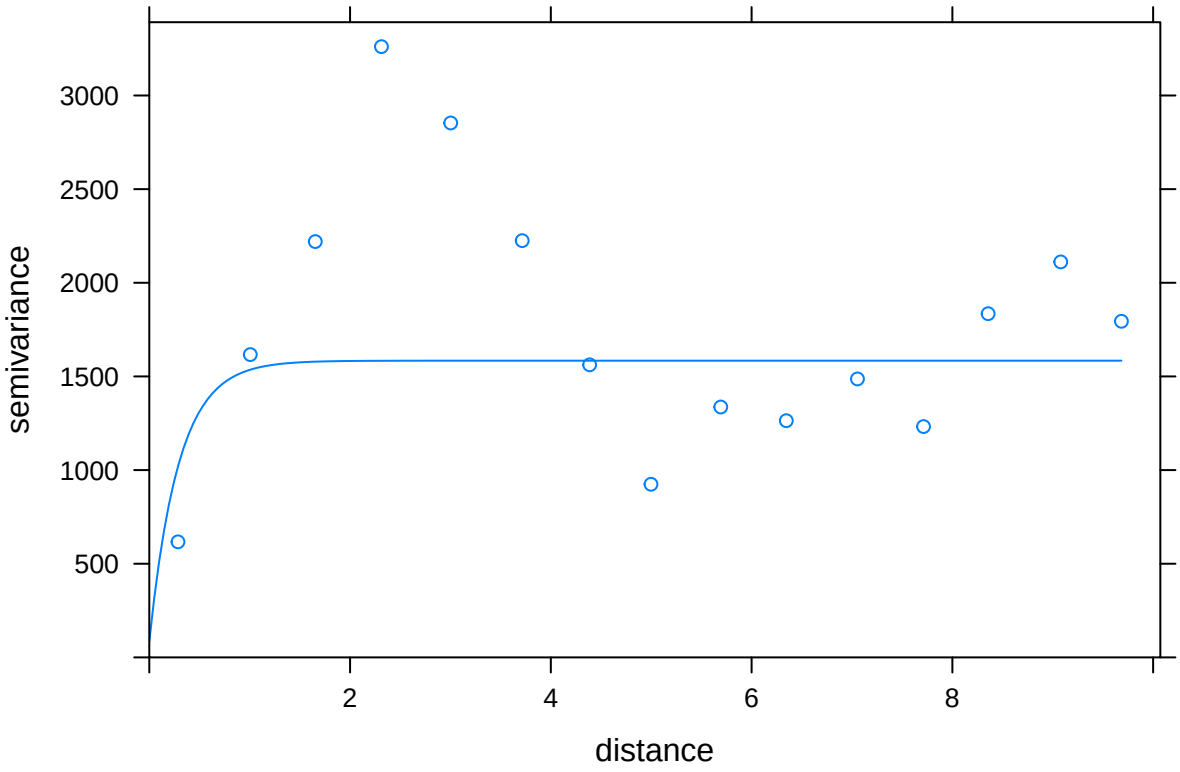
Spacial Predictions

Using the variogram model found above we use Kriging, a method of interpolation, to predict the values on a grid. However, in order to select the best type of kriging we first do cross validation on a subset of points.

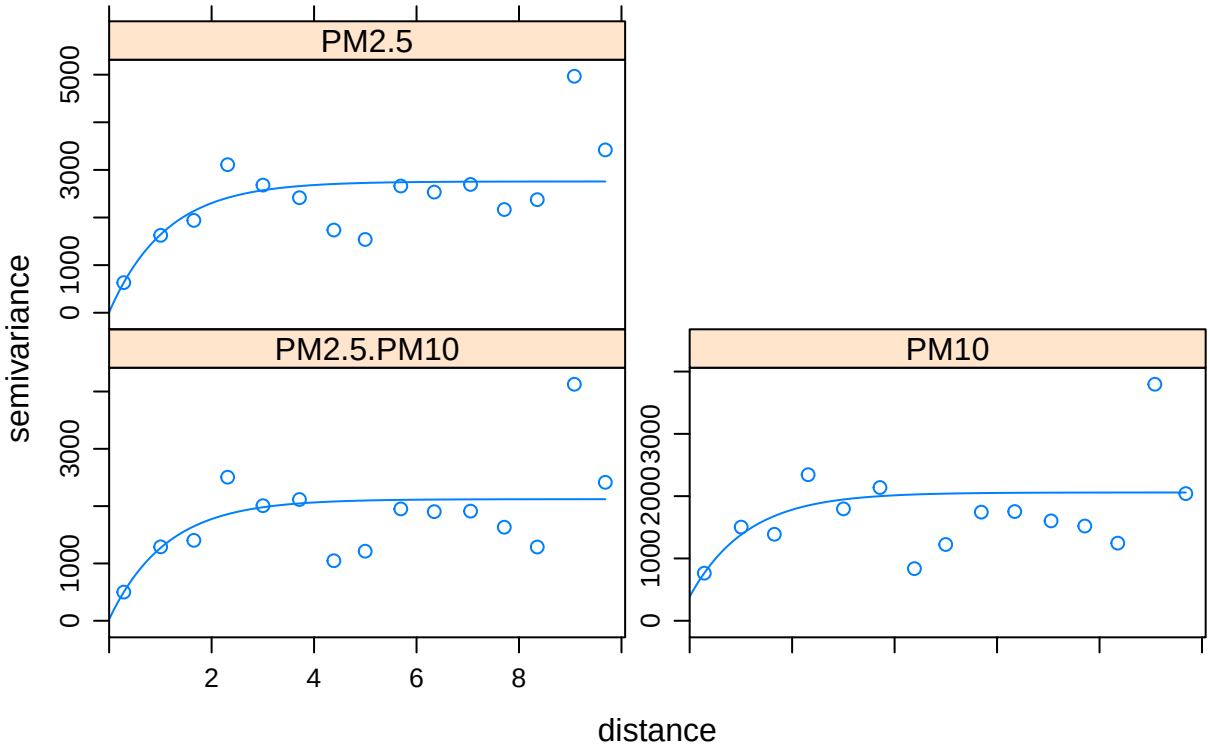
Ordinary Kriging



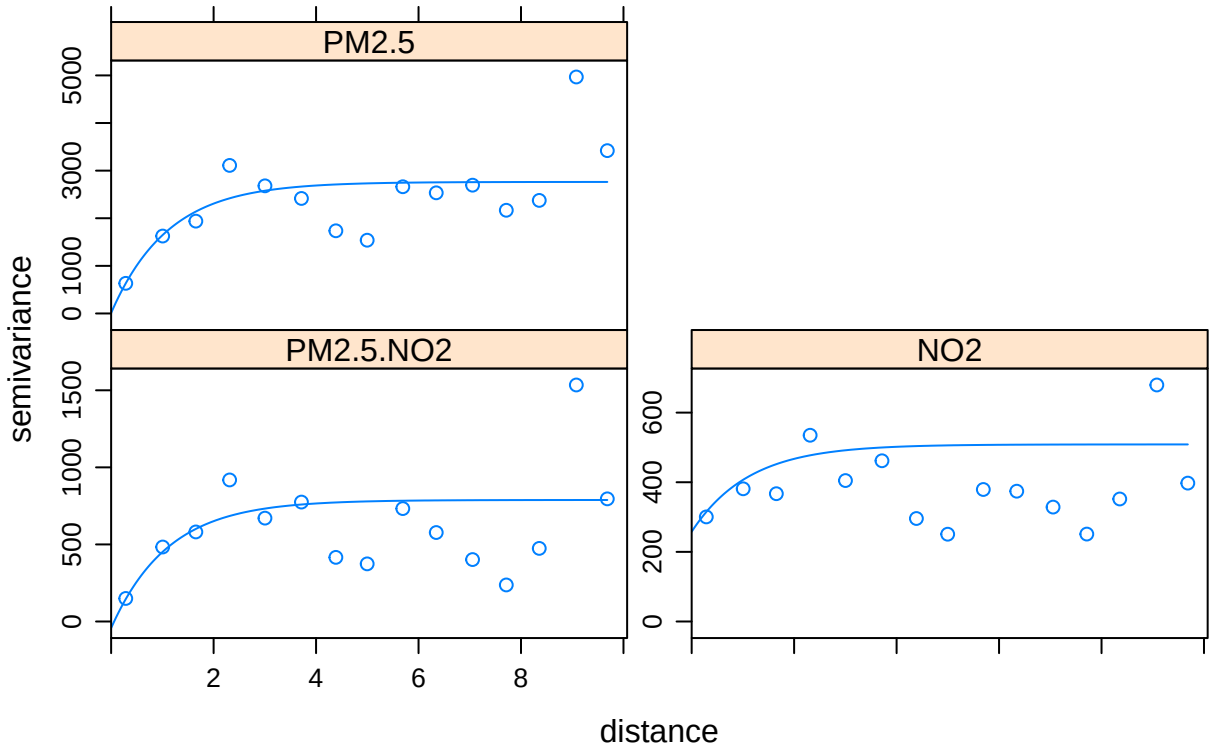
Universal Kriging



Co-Kriging with PM10



Co-Kriging with NO2



Comparing the different sum of squares:

```
## [1] 913.9521
```

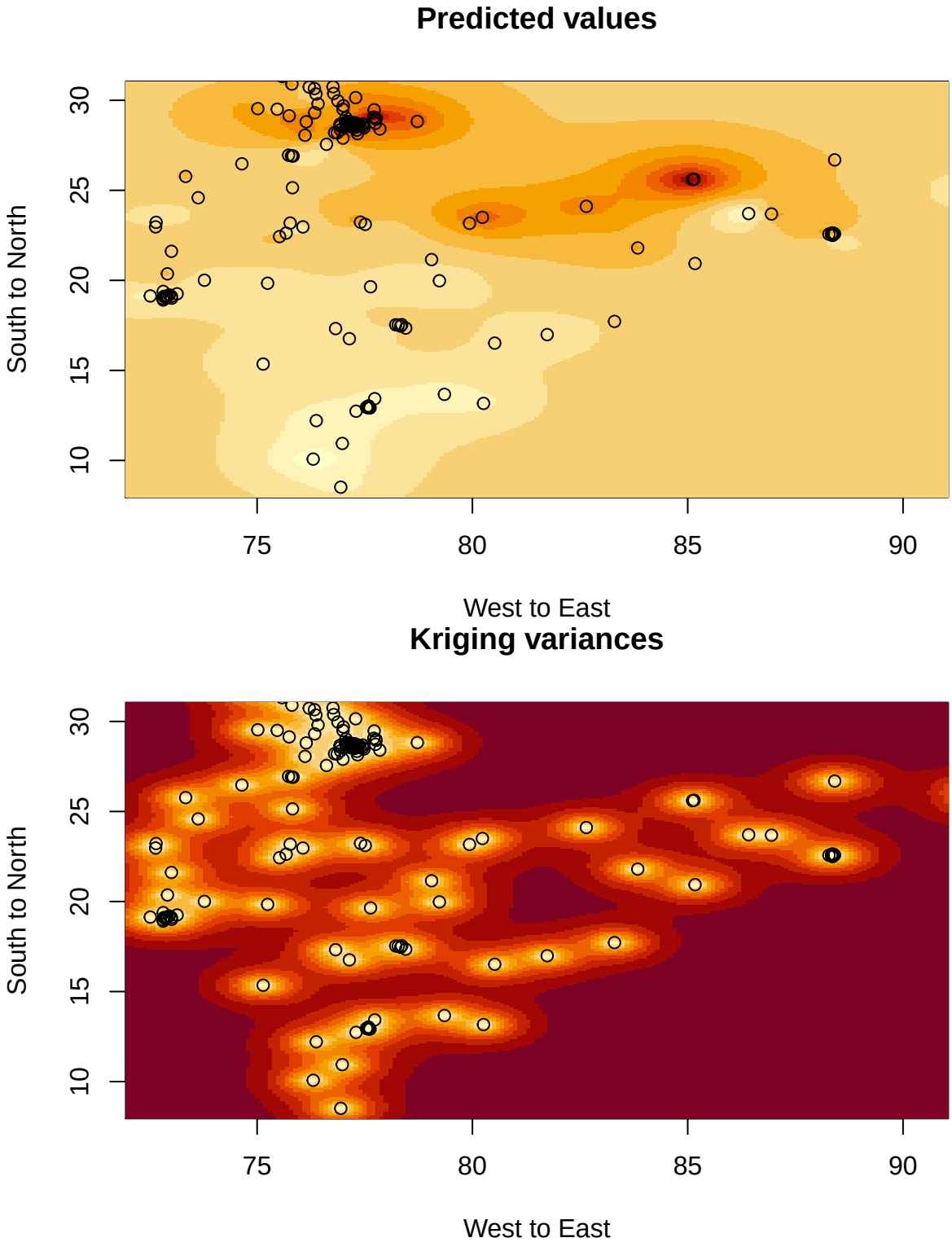
```
## [1] 130083.9
```

```
## [1] 466.2751
```

```
## [1] 807.7858
```

From the cross validation of Ordinary Kriging, Universal Kriging, Co-Kriging PM2.5, and Co-Kriging PM10 given above in the same order the method with the lowest sum of squared residuals is Co-kriging with PM10.

CoKriging with PM10 Raster Map



Conclusions

Overall the raster map above helps us get a clearer picture of our PM2.5 situation in the plotted region of India. It should not surprise us that Co-Kriging with PM10 was the most accurate prediction method as the predictor is highly correlated with PM2.5. Meanwhile, one of the most important aspects an observer should note from the raster map is the localization of heavy pollution specifically PM2.5. One can speculate that perhaps a high number of sites in one region indicates a city where more pollution would be prevelant or maybe warmer climates could be linked with PM2.5 concentration. In order to further understand the cause of PM2.5 prevelance geographically, research within this field should be continued.