








Evaluation of NO₂ emission with different land-use pattern within different states in the US

This manuscript ([permalink](#)) was automatically generated from [uiceda/project-team492@ea6d0c3](#) on November 15, 2024.

Published: October 27, 2024

Authors

- **Siyoung Park** 
 -  [siyoung3](#)Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign
- **Tausif E Elahi** 
 -  [tausifeelahi](#)-Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign
- **Tabassum Nanzeeba** 
 -  [nanxee492](#)Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign
- **Rauf Momina** 
 -  [MominaRauf](#)Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign

✉ — Correspondence possible via [GitHub Issues](#) or email to Siyoung Park <siyoung3@illinois.edu>, Tausif E Elahi <tausife2@illinois.edu>, Tabassum Nanzeeba <nt23@illinois.edu>, Rauf Momina <momina2@illinois.edu>.

Project Description

Description of data set:

The dataset consists of

1. Air pollutants (O₃, CO, SO₂, NO₂, PM10) in different states (Illinois, California, Florida, North Dakota). of the US
- CSV file 1 Rows: 463,218 // Cols: 7

A data set shows air pollutant concentrations for 5 criteria i.e. O₃, CO, SO₂, NO₂, and PM10. The variable "pred_weight" shows concentrations of pollutants in µg/m³. The title lat and lon represent the latitude and longitude of the specific place where the data were measured and noted. O₃ and CO are measured as parts per million (ppm) whereas NO₂ and SO₂ are measured as parts per billion (ppb).

Threshold for pollutants:

O₃, CO, SO₂, NO₂ and PM10 concentration thresholds are based on standards set by WHO (World Health Organization) and EPA (Environmental Protection Agency).

- O₃: 0.070 ppm exposure for 8 hours.
- CO: 9 ppm (8 hours) and 35 ppm (1 hour). > Not to be exceeded more than once per year
- SO₂: 75 ppb (1 hour, 3 years average) and 0.5 ppm (3 hours, year)
- NO₂: 100 ppb (1 hour, 3 years average) and 53 ppb (year, annual mean)
- PM10: 150 µg/m³ (24 hours, 3 years average)

Dataset: <https://www.caces.us/data> [Accessed: 09/19/2024] In a CSV file

2. Temperature variation of the above states over years (using NOAA, National Centers for Environmental Information dataset from 1991 to 2020).

- CSV file 1 (Bismarck 2.4 NNW, ND, US) Rows: 13 // Cols: 313
- CSV file 2 (Springfield Capital AP, IL, US) Rows: 13 // Cols: 413
- CSV file 3 (Tallahassee AP, FL, US) Rows: 13 // Cols: 413
- CSV file 4 (Sacramento, CA, US) Rows: 13 // Cols: 385

A data set of 4 cities i.e. capital cities of four selected states of the US. The information of capital cities of selected states is given below

Regions	State	City
1	North Dakota (ND)	Bismarck
2	Mid Illinois (IL)	Springfield
3	South Florida (FL)	Tallahassee
4	California (CA)	Sacramento

The dataset consists of temperature variations in the abovementioned cities from 1991 to 2020. CSV file includes daily max, daily min, mean, standard deviation, cooling degree days, heating degree days, and mean number of days. The data are given in every month from 1991 to 2020.

Dataset: Palecki, Michael; Durre, Imke; Applequist, Scott; Arguez, Anothony; Lawrimore, Jay. 2021: U.S. Climate Normals 2020: U.S. Hourly Climate Normals (1991-2020). [indicate subset used]. NOAA National Centers for Environmental Information. <https://doi.org/>. Accessed [09/19/2024] In a CSV file

Proposal:

1. In this research, our team will track the air pollutants in different states of the US. O_3 , CO, SO_2 , NO_2 , and PM10 are the subjects of the investigation. The five states of the US, North Dakota (north), Illinois (mid), Florida (south), and California (west) are the regions for measuring air pollutants.
2. The historical trend (temperature change) of the air pollutants will also be investigated along with the US State data. From 1991 to 2020, the variation of climate (such as mean temperature by months) and air pollutants will be compared to evaluate their correlation.
3. The ultimate goal of this research will be to alert each investigated States of increasing air pollutant and to come up with ideas for mitigating them. Also, the monthly temperature and air pollutants will be compared to notice when the air pollutants are maximized with an understanding of their distribution.

Exploratory Data Analysis

Description and Characterization of Dataset

The dataset we are going to use is obtained from Bechle et al. [1] which was used for estimating air pollution in terms of NO_2 from 2000 to 2010. The dataset contains spatial and temporal concentration of NO_2 in ppb at different locations of the different states in US. It also contains Geographic Information System (GIS) data on land-use features such as impervious surfaces, population density, length of different types of roads-residential, major and total etc. These are commonly used as proxies for different pollution sources [2-4]. Based on the dataset, NO_2 concentration varies significantly from state to state depending on different land-use pattern and the value range between 0.31~34.21 ppb for different states. The distribution of the NO_2 pollutants across the US based on location are shown in Figure 1. Some of the explanations of the dataset are provided below:

Impervious_100: This represents the percentage of impervious surfaces such as roads, buildings etc. within a 100-meter buffer around the measuring station. Major_1000: It refers to the length of the major roads within a 1-kilometer radius around the measuring location. Resident_500: This indicates length of the roads within 500-meter radius of the monitoring station. Total_100: It represents the length of all the roads including major, minor and residential within a 100-meter buffer zone around the measuring station. Population_100: It denotes the population density within the 100-meter buffer area around the measuring station.

Dataset CSV file: https://docs.google.com/spreadsheets/d/1yo3cL23279-qwrjHSDbHc1e4t_8yl6h8CfVrMX-PIS4/edit?usp=drive_web&ouid=116140173519287299300

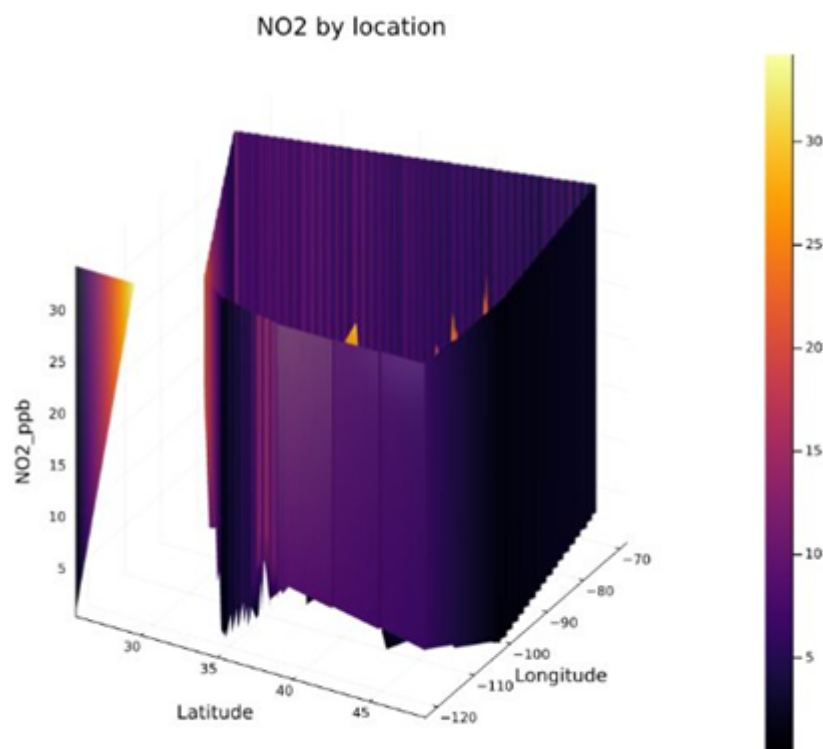


Figure 1: Distribution of NO_2 pollutants across the US based on latitude and longitude of monitoring stations.

Preliminary Analysis and Plots

From the given dataset, we did some preliminary analysis to visualize the dataset and the summary of the observations are described briefly: First, we tried to find out if there is any direct relationship between any of the land use characteristics and NO₂ concentrations measured at the monitor station. For this preliminary analysis, we considered the effect of this land use pattern within 100m, 5000m and 10000m radius of the station. The reason for selecting these three radii was to cover short, medium and long-distance land use behavior around the station. Figure 2-7 presents the effect of different land-use characteristic on the NO₂ concentration.

For impervious surfaces, for all three cases, we can clearly see there is a trend that with the increase of impervious surfaces around the station, the concentration of NO₂ increases gradually (Figure 2). As the impervious surface increases, it indicates there is increase in roads, sidewalks, parking lots, buildings, traffic and also there is decrease in vegetation areas and soil surface. Therefore, all these impervious surfaces are kind of indicator of high volume of vehicles, high population density which contributes to high NO₂ emission and also the absence of natural filtration effect with the absence of vegetation is another major source of NO₂ emission.

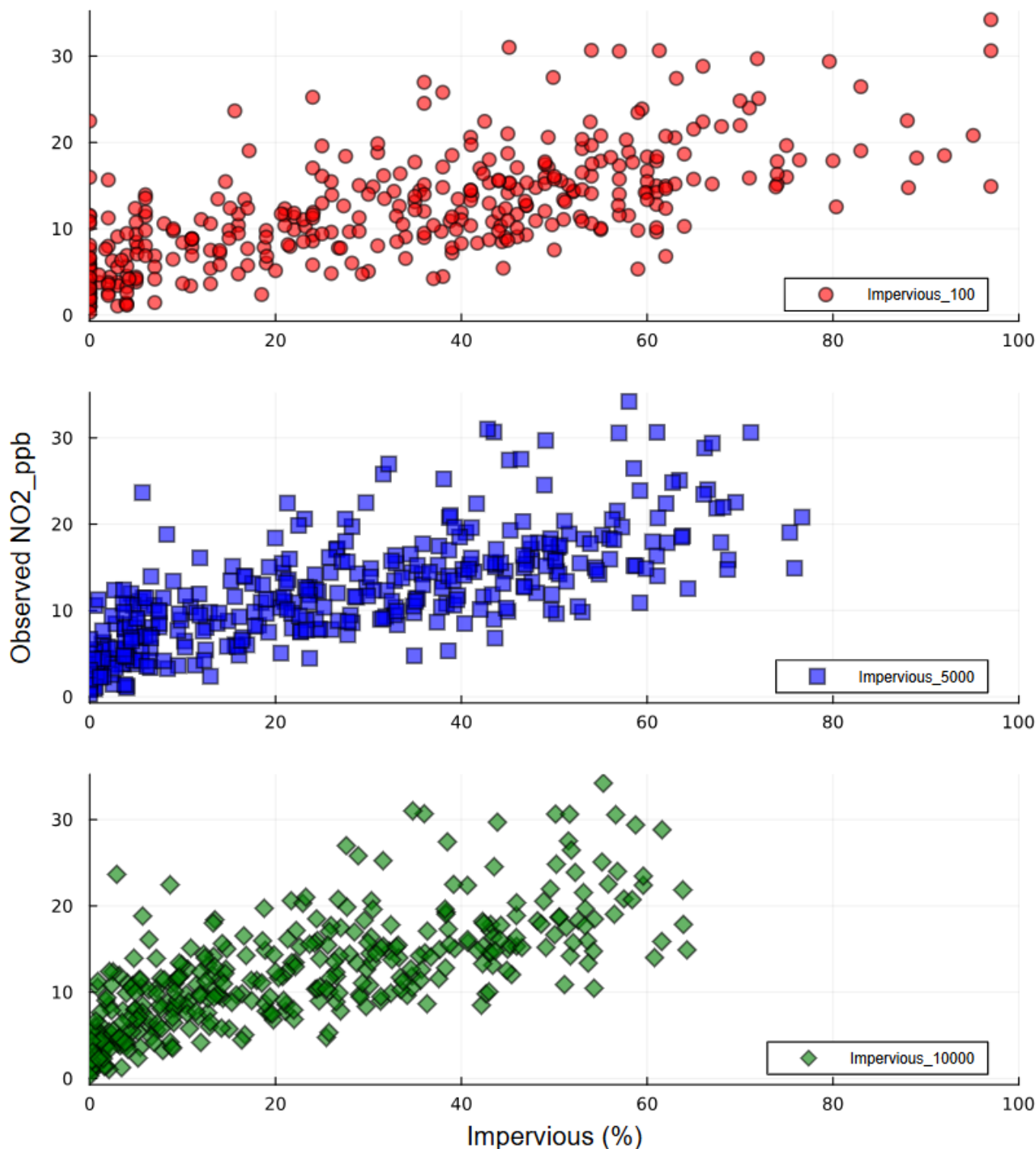


Figure 2: Variation of measured NO₂ concentration with the impervious surface at 100m, 5000m and 10000m radius around the monitor station.

In case of Major roads, we clearly see with the increasing length of major roads, there is clear increase in the concentration of NO₂ (Figure 3). Moreover, visually, it looks like there is a steep increase in the concentration of NO₂ initially with the increase of major roads, but the rate of increase slows down as the length of major roads increases further. The reason of such increase is understandable since the production of NO₂ is directly influenced by the volume of traffic and high traffic areas will release more NO₂ as the more diesel vehicles will be on the road contributing to high NO₂ emissions. Similar trend is observed for the relationship between NO₂ concentration and residential roads and total roads (Figure 4-5).

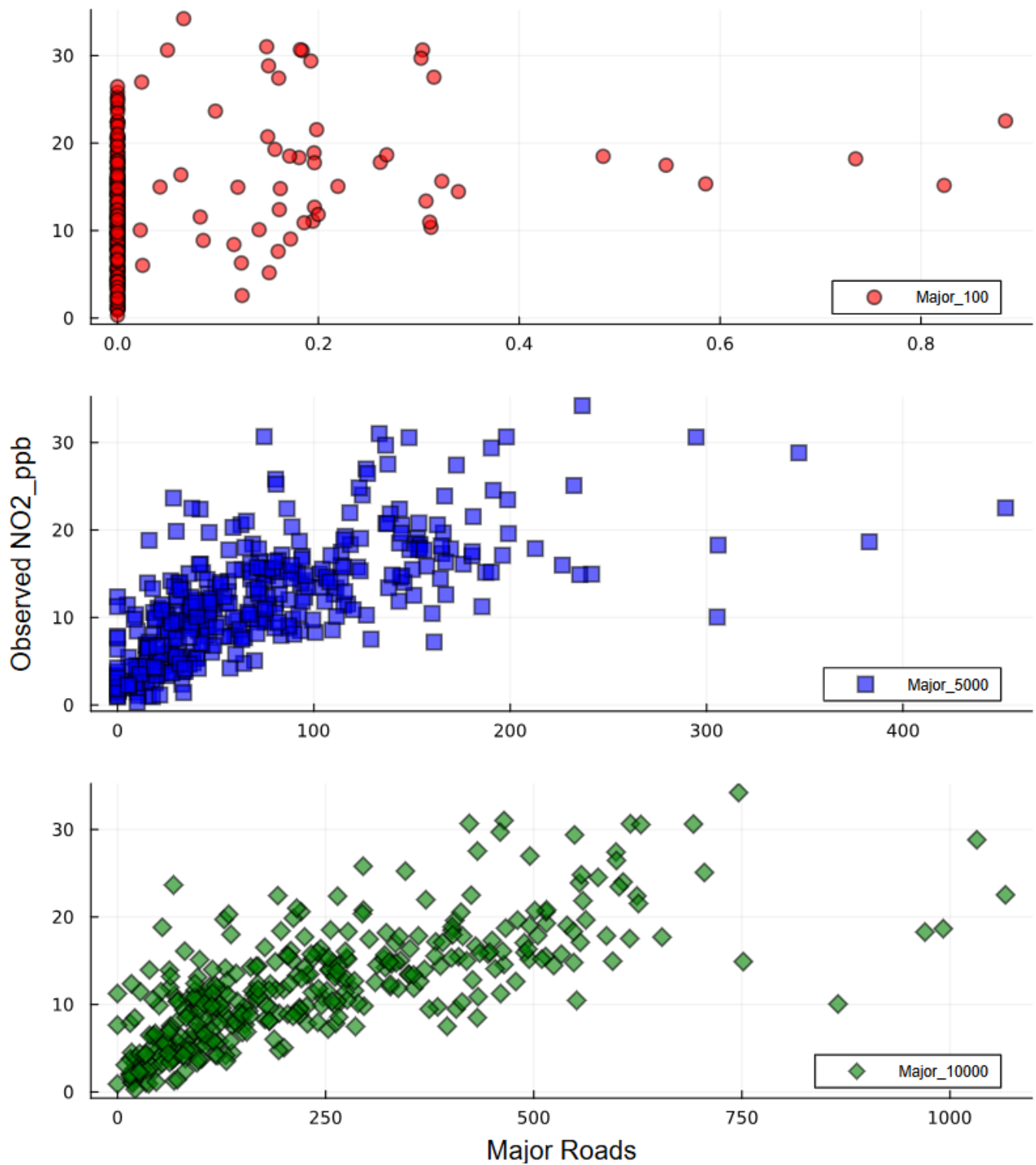


Figure 3: Variation of measured NO₂ concentration with the length of major roads at 100m, 5000m and 10000m radius around the monitor station.

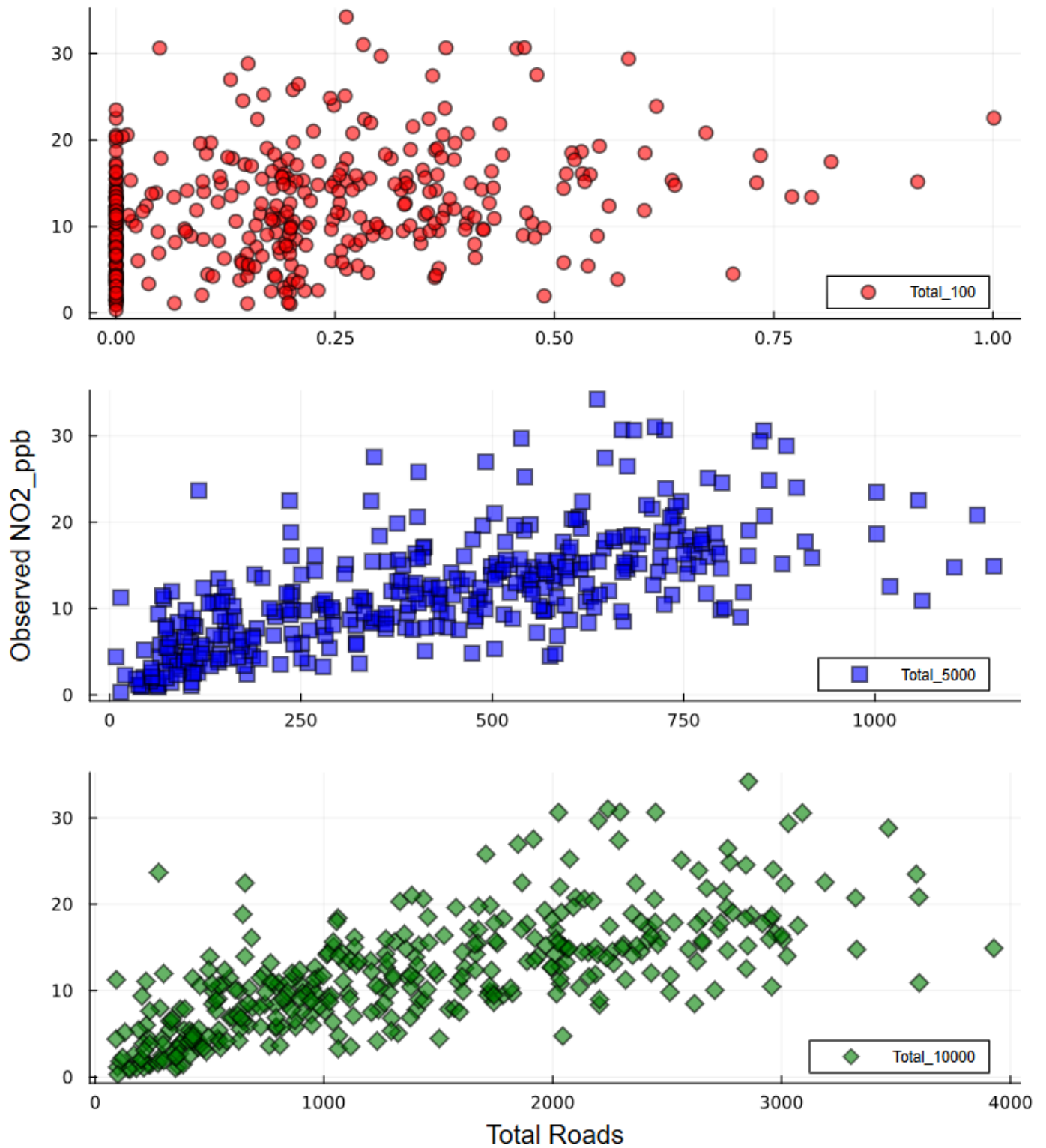


Figure 4: Variation of measured NO₂ concentration with the length of total roads at 100m, 5000m and 10000m radius around the monitor station.

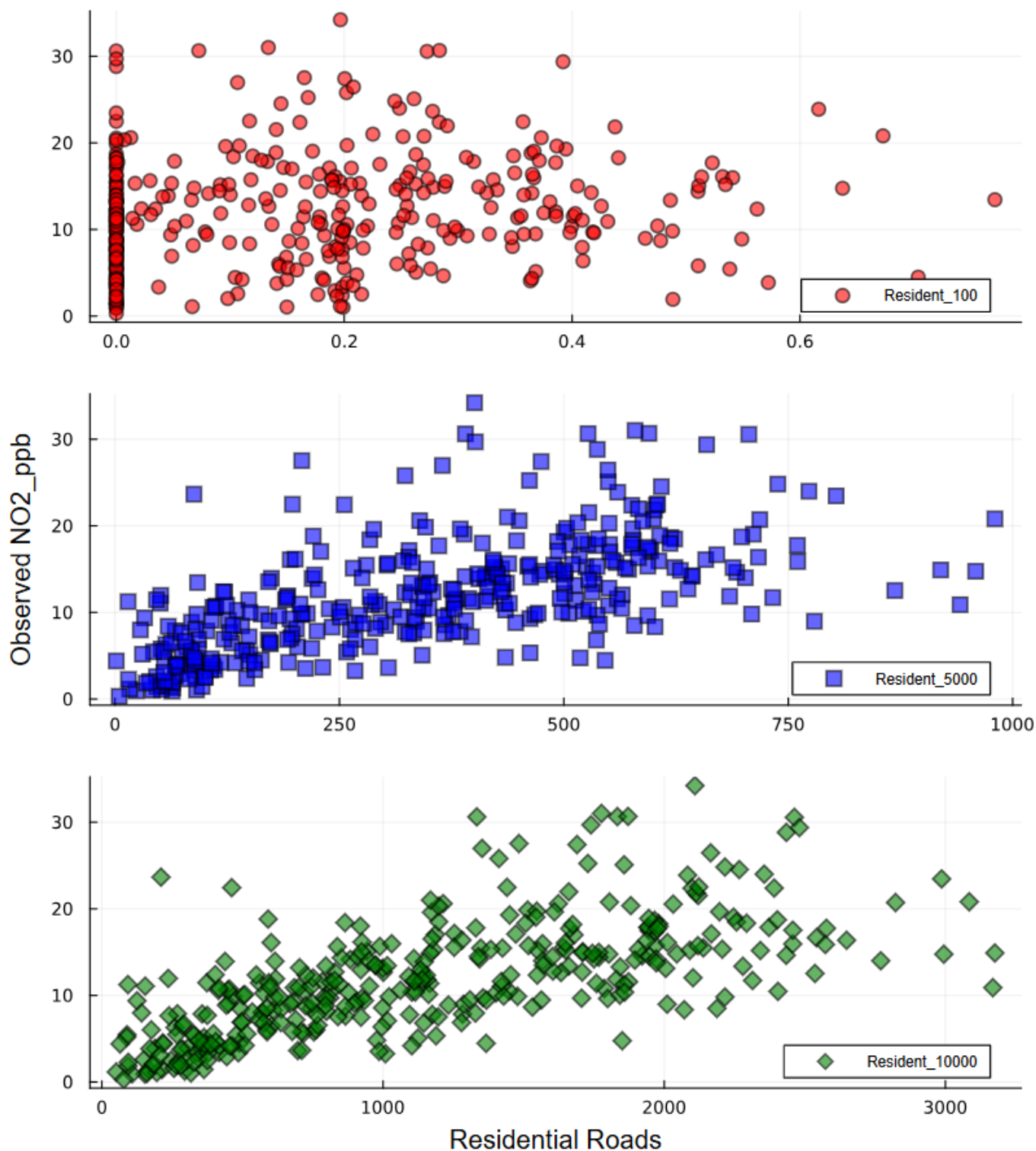


Figure 5: Variation of measured NO₂ concentration with the length of residential roads at 100m, 5000m and 10000m radius around the monitor station.

In case of population, it looked like an exponential curve which might describe the pattern very well where initially with the increase in population there is a drastic increase in NO₂ concentration which saturates at a certain point (Figure 6).

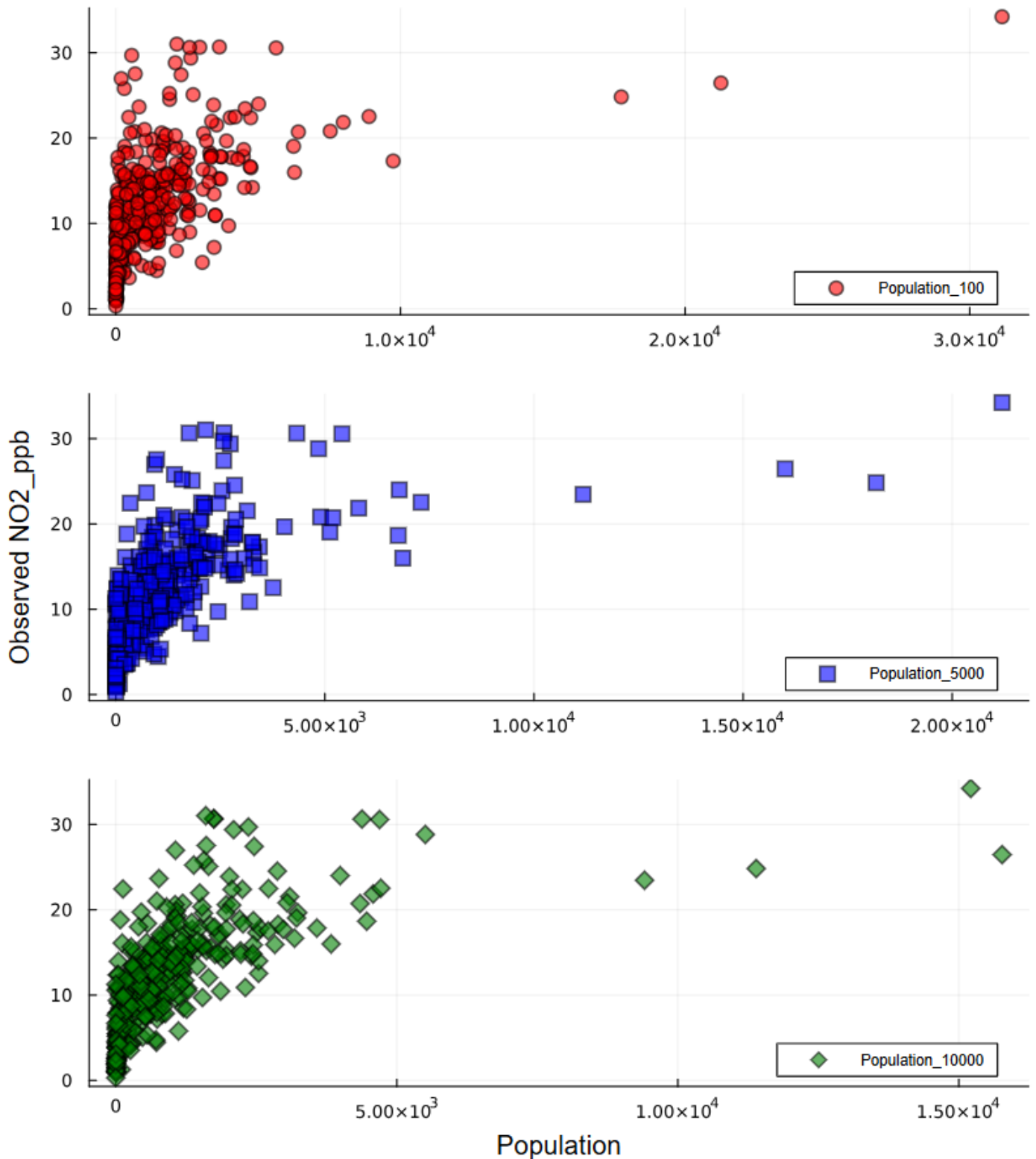


Figure 6: Variation of measured NO₂ concentration with the population at 100m, 5000m and 10000m radius around the monitor station.

Figure 7 shows the relationship between NO₂ concentration and the distance from the coast. In general, as the distance from the coast will be less, there should be lower concentration of NO₂ due to the ventilation from the winds. However, in the figure we can clearly see higher concentration of NO₂ in some of the places which are closest to coast. It indicates that although coastal distance have effect on NO₂ but it should be analyzed in combination with other land use pattern because even if the place is closer to coast but if there is high population density and roads, it will have higher NO₂. Overall, all of these land-use characteristics have their own effect on the NO₂ concentration and in some cases, there is strong relationship with NO₂.

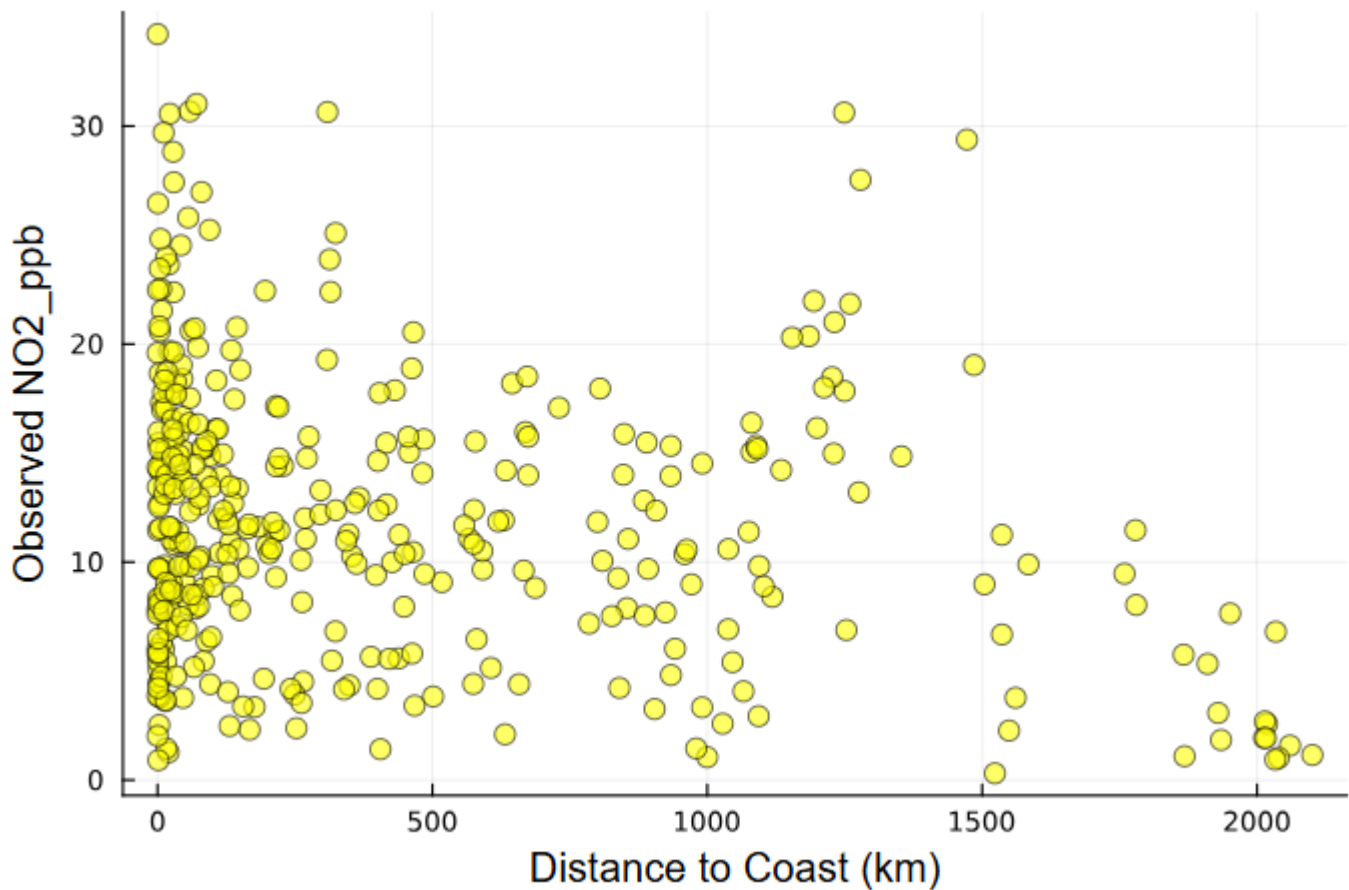


Figure 7: Variation of measured NO₂ concentration with the distance to coast from the monitor station.

As a next step, we also investigated the scenario of NO₂ concentration at individual state level and for the preliminary analysis we selected four states- IL, CA, FL and ND. The reason of selecting these four states was to capture the diverse representation of factors that might influence the NO₂ level which might be helpful for generalizing our analysis in future. CA is a highly urbanized and densely populated state with around 94 station available at the given dataset which is the reason we considered CA for our analysis. IL offers a perspective in the NO₂ pattern of Midwest's urban and suburban areas whereas FL is a coastal state which might help to understand the effect of breeze and humidity on NO₂. Lastly, ND is a low-population and rural environment with minimal urbanization which might help us to understand the effect of such characteristics on NO₂. The effect of the land use characteristics described earlier on NO₂ concentrations are summarized in Figure 8-15. Figure 16 presents the distribution of NO₂ pollutants of these states by monitoring station.

For CA state, visually we can clearly see there exists a correlation among impervious surface, population, length of the roads and NO₂ concentration (Figure 8-9). Interesting to see, although some of the places are very closer to the coast but it has significant concentration of NO₂. As discussed earlier, although the coastal distance is lower but other factors such as impervious area, population and length of the roads are so high that it affects the NO₂ significantly compared to the coastal distance from measuring station.

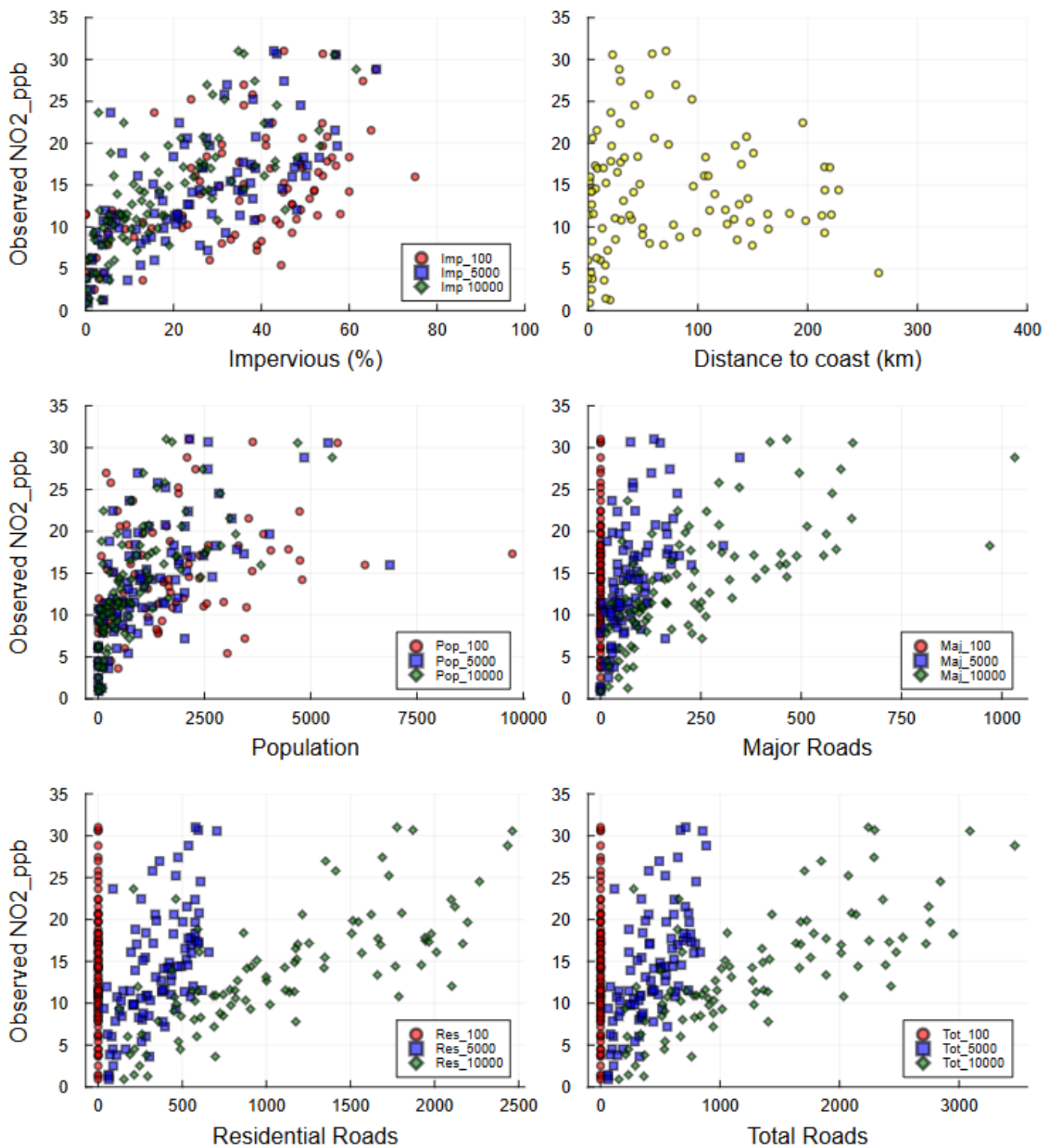


Figure 8: Effect of different land use characteristics on NO₂ concentration of CA state.

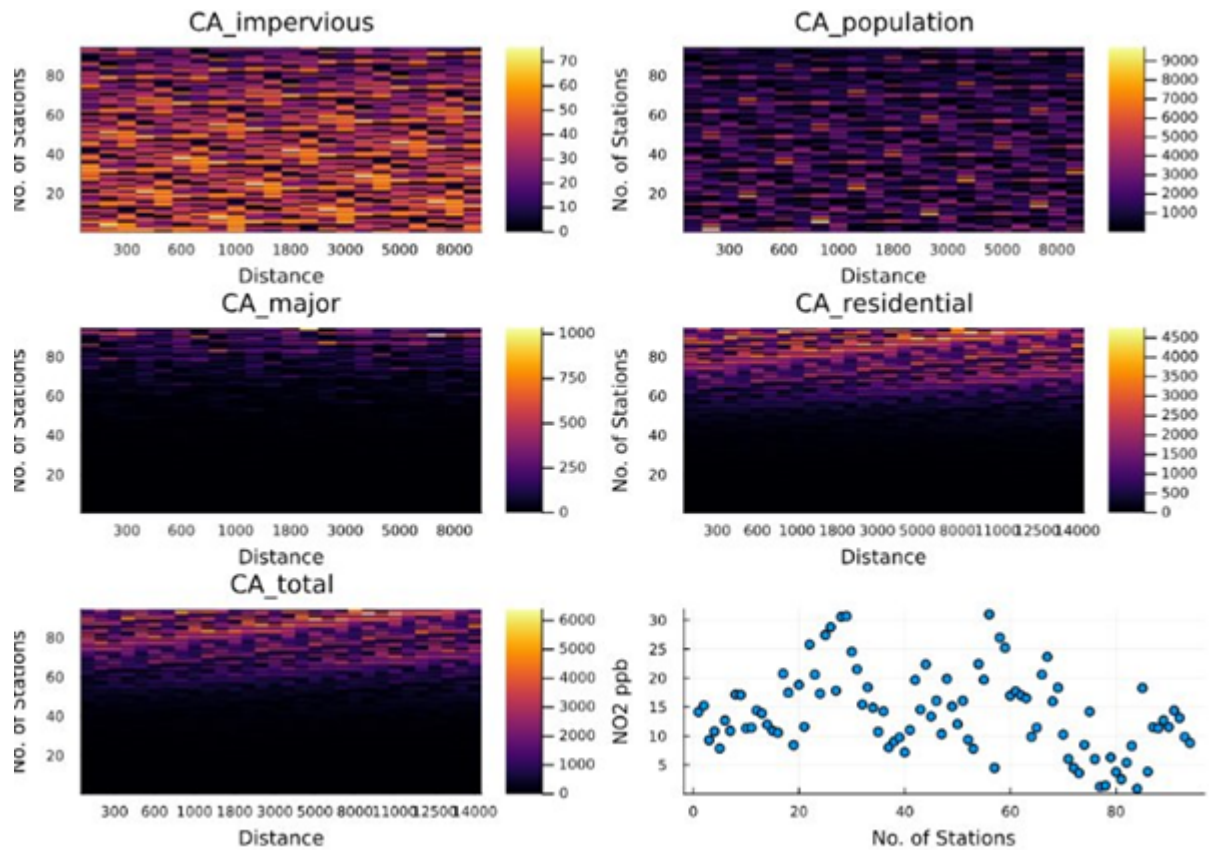


Figure 9: Heatmap to understand the effect of different land use characteristics on NO₂ concentration of CA state.

For IL state, there were only 6 stations, and the results suggests that there is a good relationship between impervious surface and NO₂ concentrations (Figure 10-11) . Also, since IL is far away from the coast it is clearly seen that NO₂ concentration has kind of linear relationship with coastal distance. Population and residential roads don't reveal any clear pattern but with the major roads, it is clearly visible that increase in the length of major roads correlates well with the increase in NO₂ concentrations.

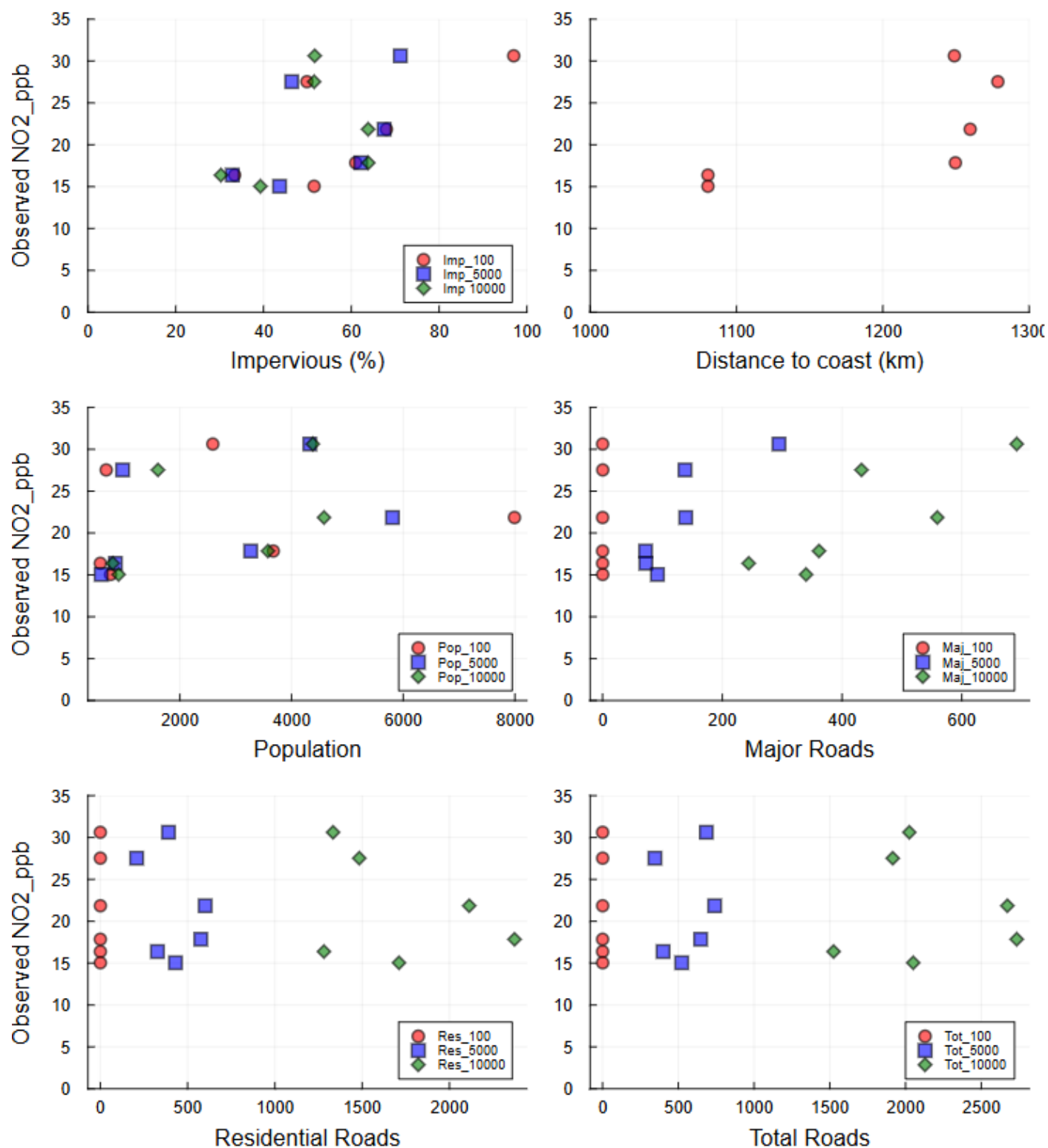


Figure 10: Effect of different land use characteristics on NO₂ concentration of IL state.

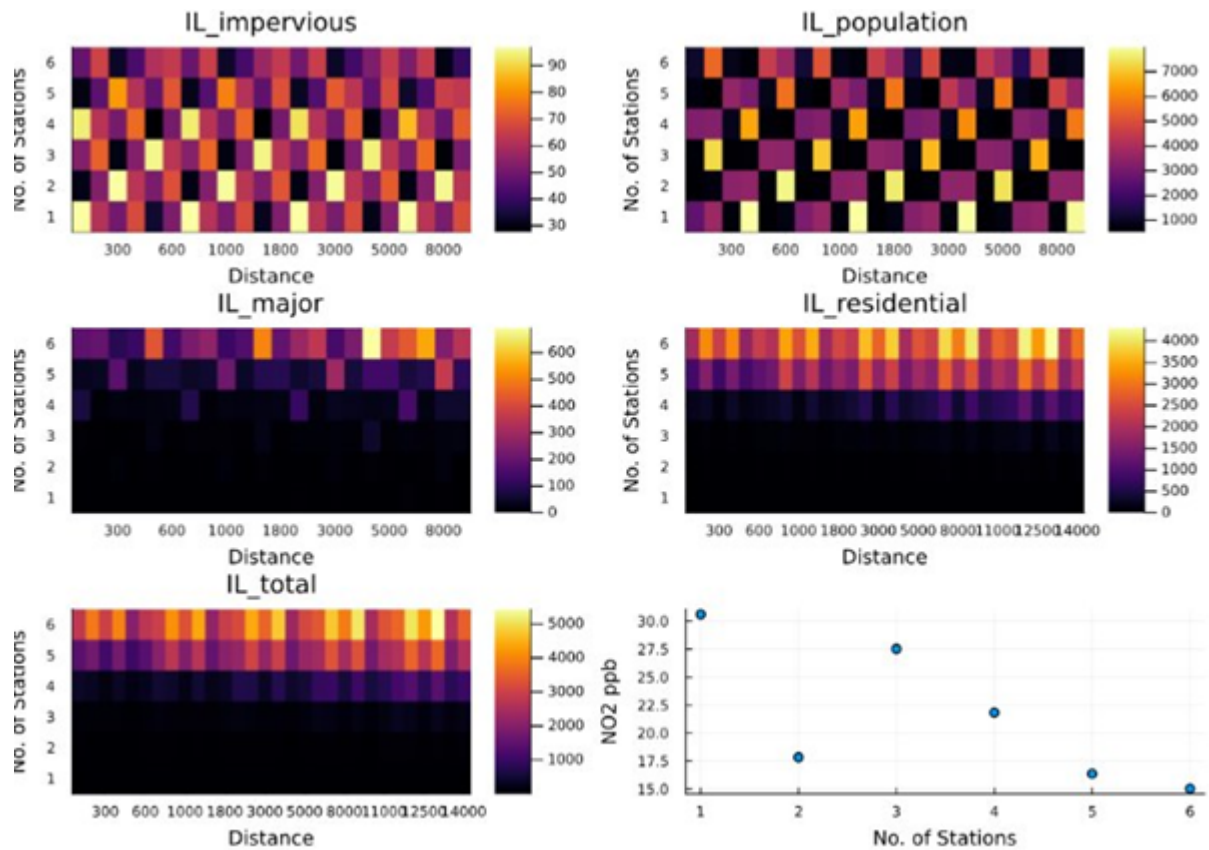


Figure 11: Heatmap to understand the effect of different land use characteristics on NO₂ concentration of IL state.

For FL state, impervious surface, population, roads all these parameters have kind of steady linear relationship with NO₂ concentration and with the increase in these parameters NO₂ increase is not that significant (Figure 12-13). For example, in IL state, some of the places with 70-80% impervious area has around 30 ppb NO₂ concentration whereas in FL, places with 60-80% impervious area has around 12 ppb NO₂. One of the major reason of this observed lower values could be due to the fact that all the stations in FL area are very close to the coast showing the noticeable effect of it on NO₂.

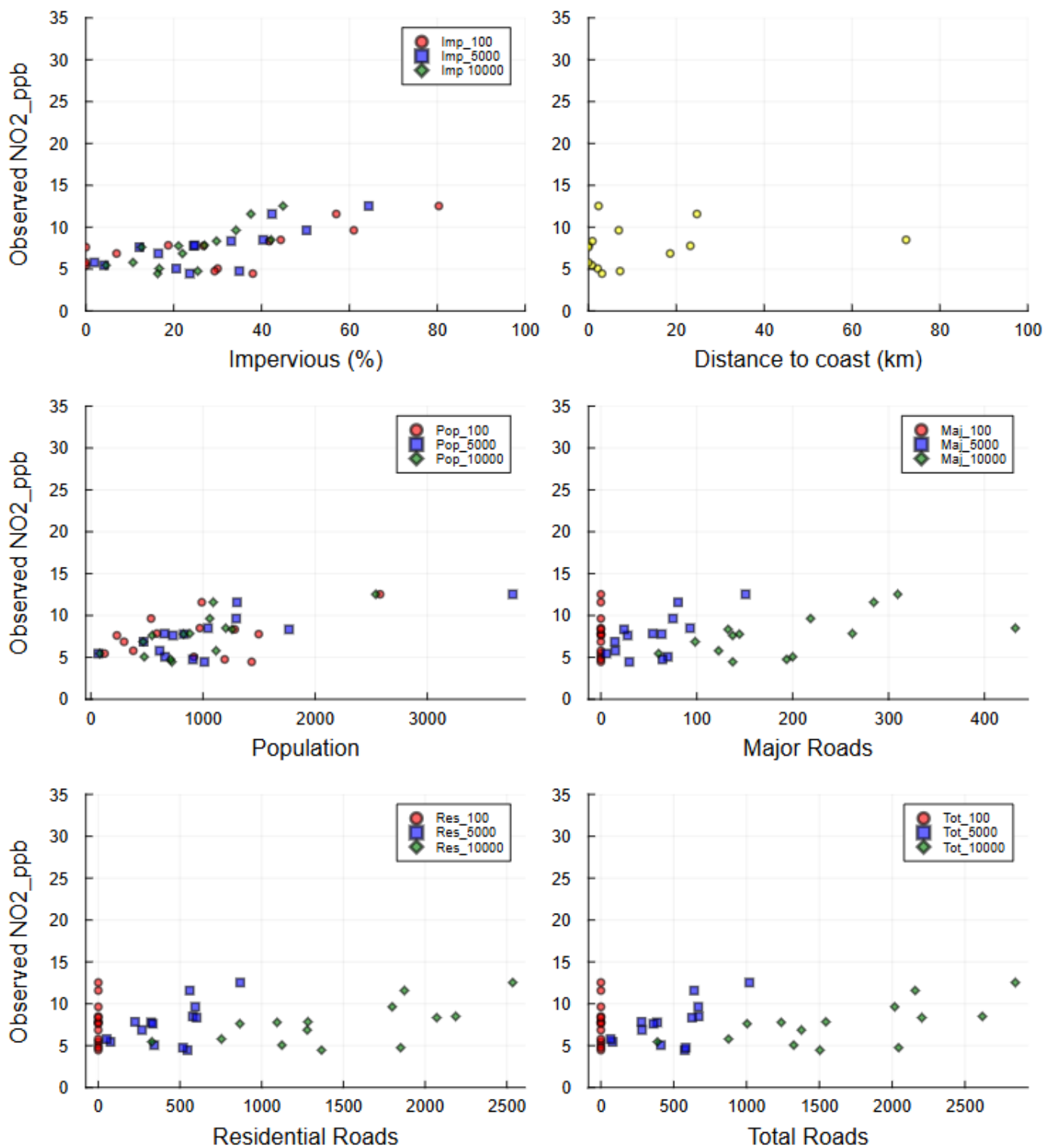


Figure 12: Effect of different land use characteristics on NO₂ concentration of FL state.

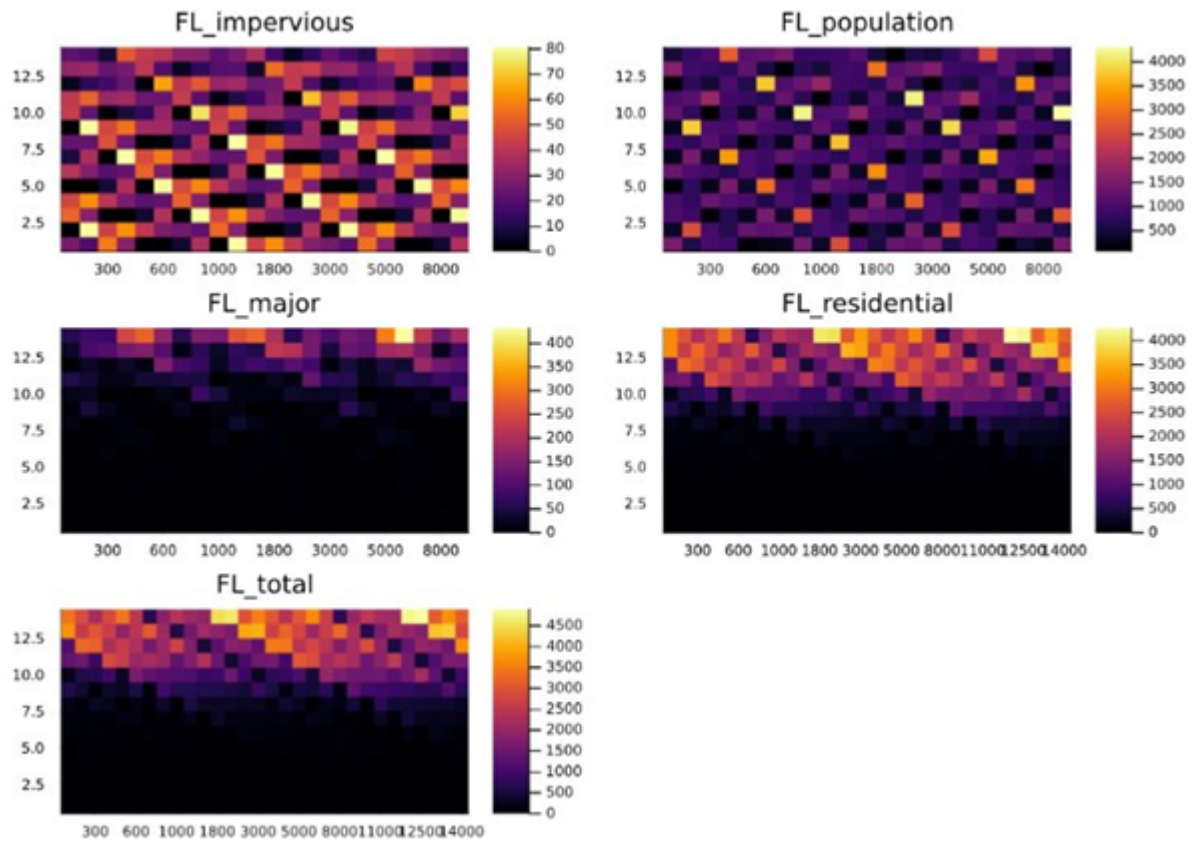


Figure 13: Heatmap to understand the effect of different land use characteristics on NO₂ concentration of FL state.

In case of ND state, although it is very far from the coast but still it has very low concentration of NO₂ (highest being ~6%) (Figure 14-15). It could be attributed to the fact that all the stations had very low population and the length of the roads are the lowest among all the four states considered in the preliminary analysis. Overall, it is seen that out of all the land use characteristic considered, all of the factors do not have similar effect on NO₂ concentration, and the effect varies from state to state.

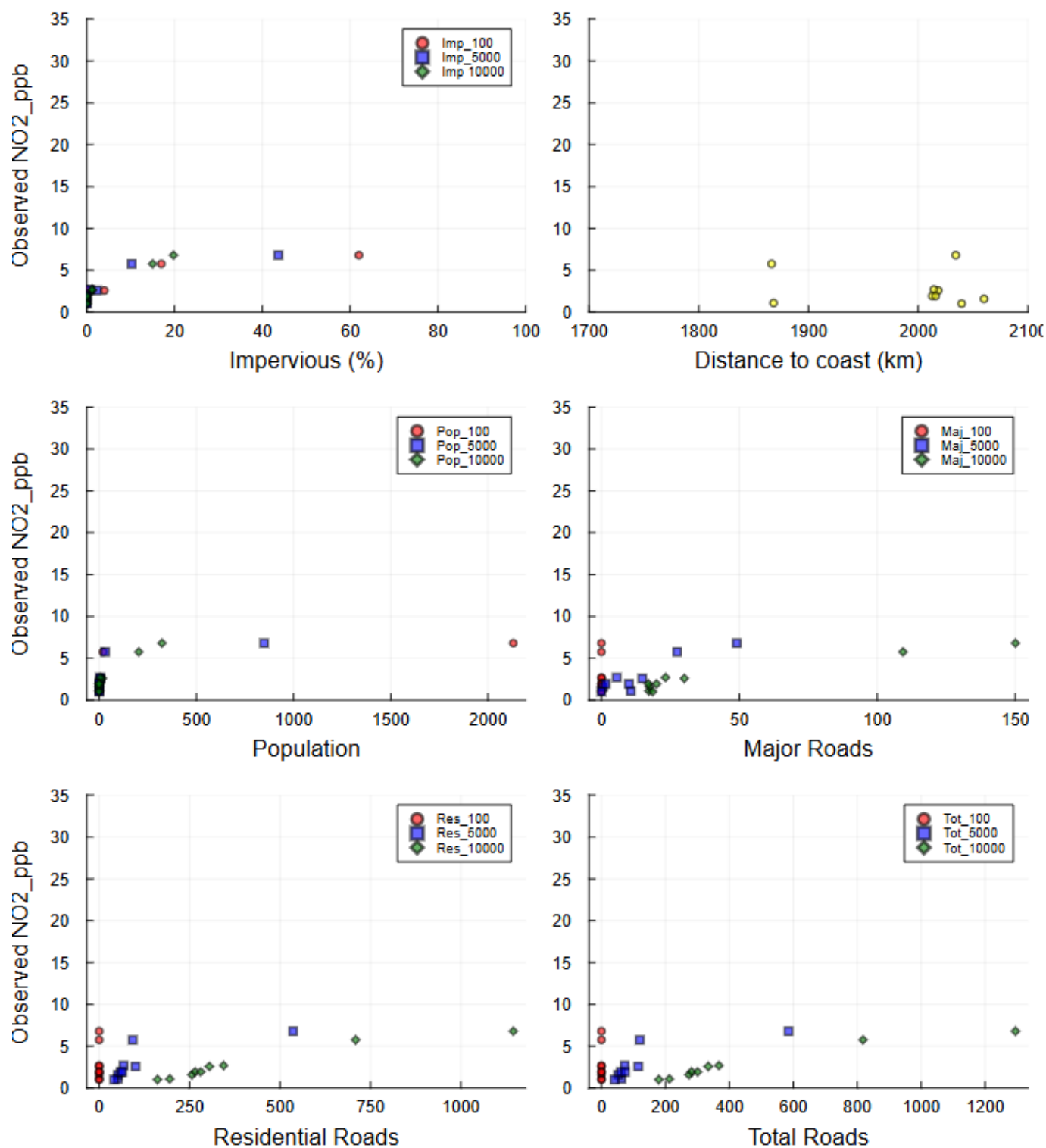


Figure 14: Effect of different land use characteristics on NO₂ concentration of ND state.

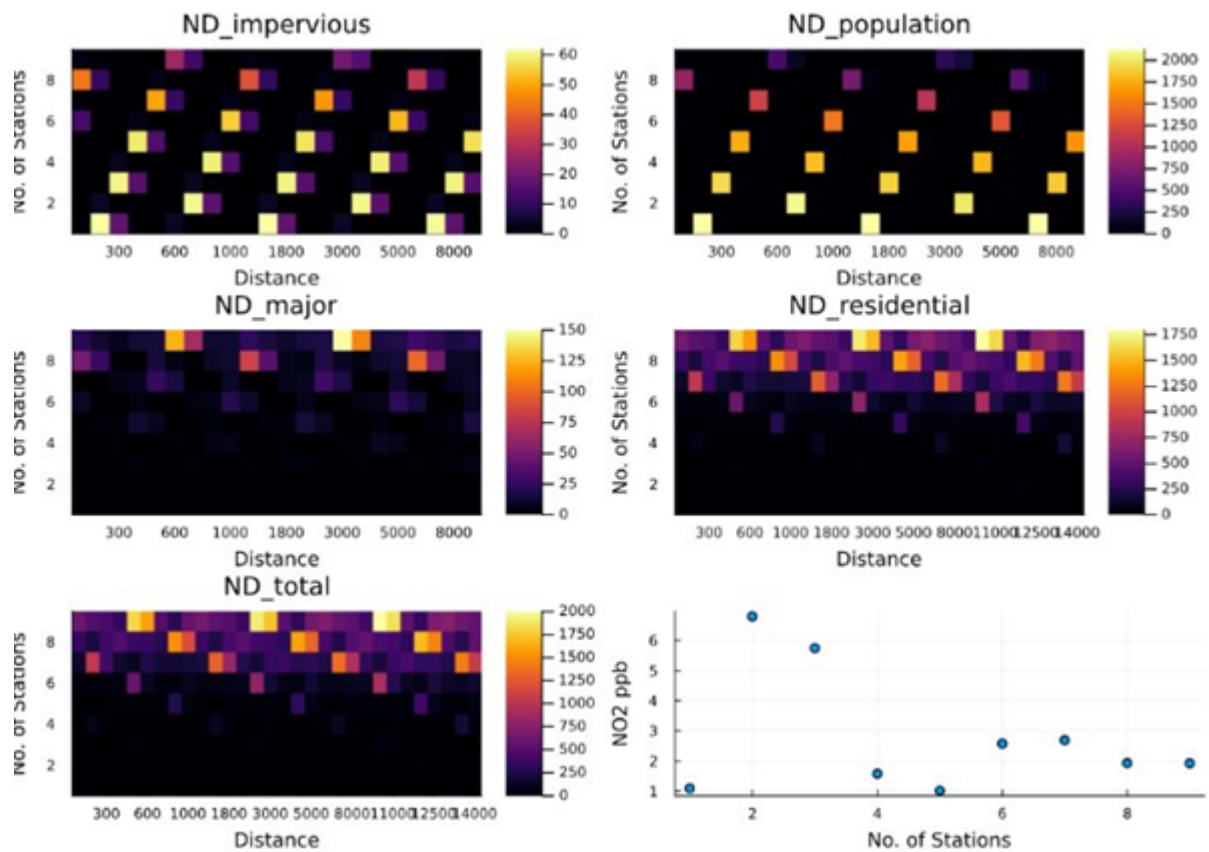


Figure 15:Heatmap to understand the effect of different land use characteristics on NO₂ concentration of ND state.

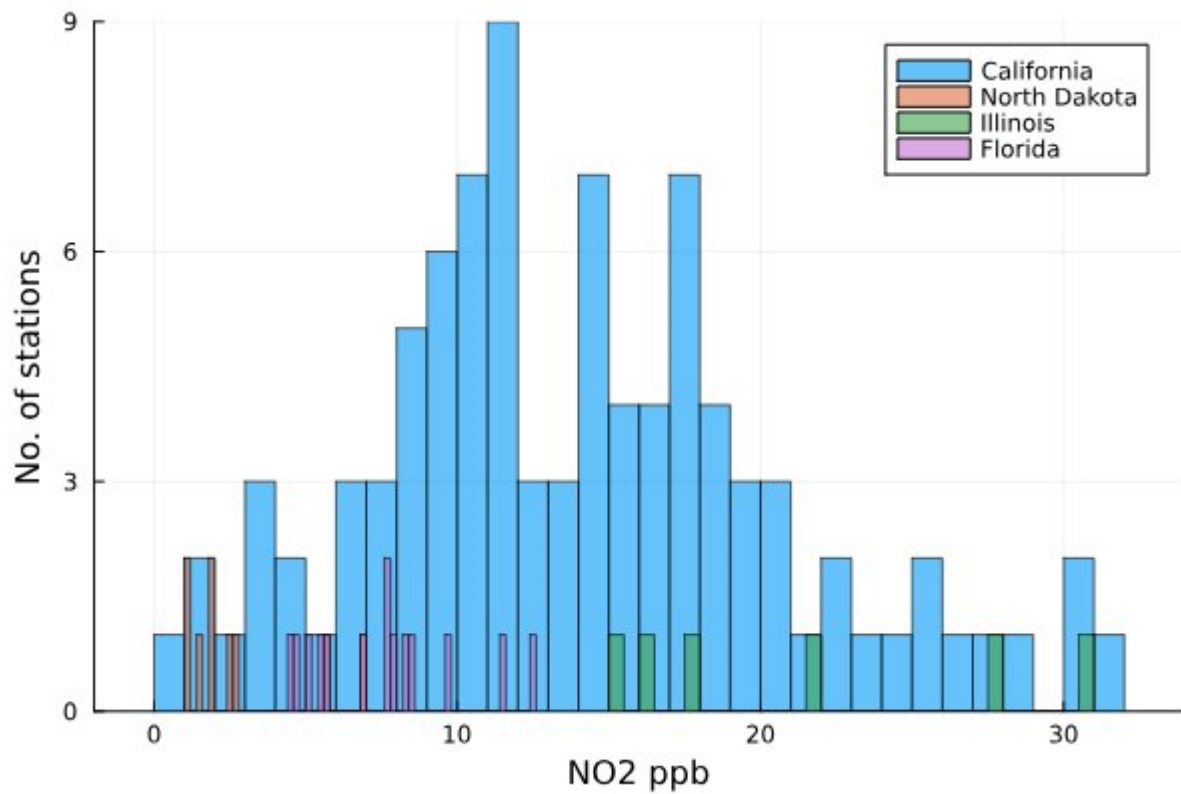


Figure 16:Distribution of NO₂ concentration across CA, ND, IL, FL based on the station.

Predictive Modeling

Based on the preliminary analysis of our dataset, it is evident that all these land-use parameters, such as impervious surfaces, road density, population distribution, etc. play a crucial role in shaping NO₂ concentrations. These parameters exhibit a strong correlation with NO₂ levels, underscoring their significance as predictors. Our next objective is to quantify the specific effects of each land-use parameter on NO₂ concentration and identify which factors most significantly influence these levels. By pinpointing the primary contributors, we aim to refine our understanding of pollution sources and dispersion.

Building on these findings, we will develop a predictive regression model capable of estimating NO₂ concentrations across different states in the U.S., factoring in the varying land-use patterns with high accuracy. For developing this model, we will consider interaction terms such as population density x road, impervious surface x major roads etc to capture the combined effects of multiple variables. We will do further correlation analysis to identify strong predictors of NO₂ emissions. Using PCA, we will try to reduce dimensionality if there are many correlated features simplifying the model without losing the predictive capacity. Then, we will explore different machine learning algorithms to find out which one works better for our purpose, followed by model training and cross-validation. We will consider different performance metrics like R-square, Root mean squared error values etc. to evaluate the model accuracy. Eventually, we will apply the model to predict NO₂ levels in regions that do not have measured data on NO₂ but have land-use information.

Such a model has the potential to be instrumental for multiple applications. By assessing long-term health impacts associated with chronic exposure to pollutants, it can provide insights into the risk of respiratory and cardiovascular conditions associated with NO₂. This type of analysis is invaluable to public health agencies tasked with identifying regions and populations at greater risk of pollutant-related diseases [5-6]. Moreover, predictive modeling of NO₂ concentrations can guide policymakers and city planners in designing urban environments with better air quality. By predicting pollutant dispersion, decision-makers can strategically zone residential areas, schools, and recreational spaces away from high pollution zones, thus enhancing community health and safety. Finally, our model will help highlight areas where pollution levels are worsening, providing actionable insights for immediate interventions. This capability will empower environmental agencies to prioritize regions for pollution control efforts, thus contributing to a healthier and more sustainable living environment for all residents.

Predictive Modeling

Description and Characterization of Dataset

In the predictive modeling, machine learning was used to make a predictive model of the dataset. The sequence of this project was as follows: 1. Data selection and clearing 2. Machine learning - 2.1 Normalization and Regularization, 2.2 Perform analysis 2.3 Check the validity. In conclusion, all machine learning techniques were compared in terms of accuracy, speed, and simplicity.

Data selection and coordinate transform

First, the data was sorted by using the correlation plot. Since there were numerous variables in the dataset, selective dependent variables introduced in the exploratory data analysis section were also used for the machine learning. This contains distance to coast, Impervious 100, Impervious 5000, Impervious 10000, Population 100, Population 5000, Population 10000, Major 100, Major 5000, Major 10000, Residential 100, Residential 5000, Residential 10000, Total 100, Total 5000, Total 10000.

The correlation comparison was segmented into five different groups at first, since to visually inspect, it was impossible to compare all. Also, depending on the naming (i.e. impervious), this process was expected to classify necessary data. Figure 1 shows the results of each correlation plot.

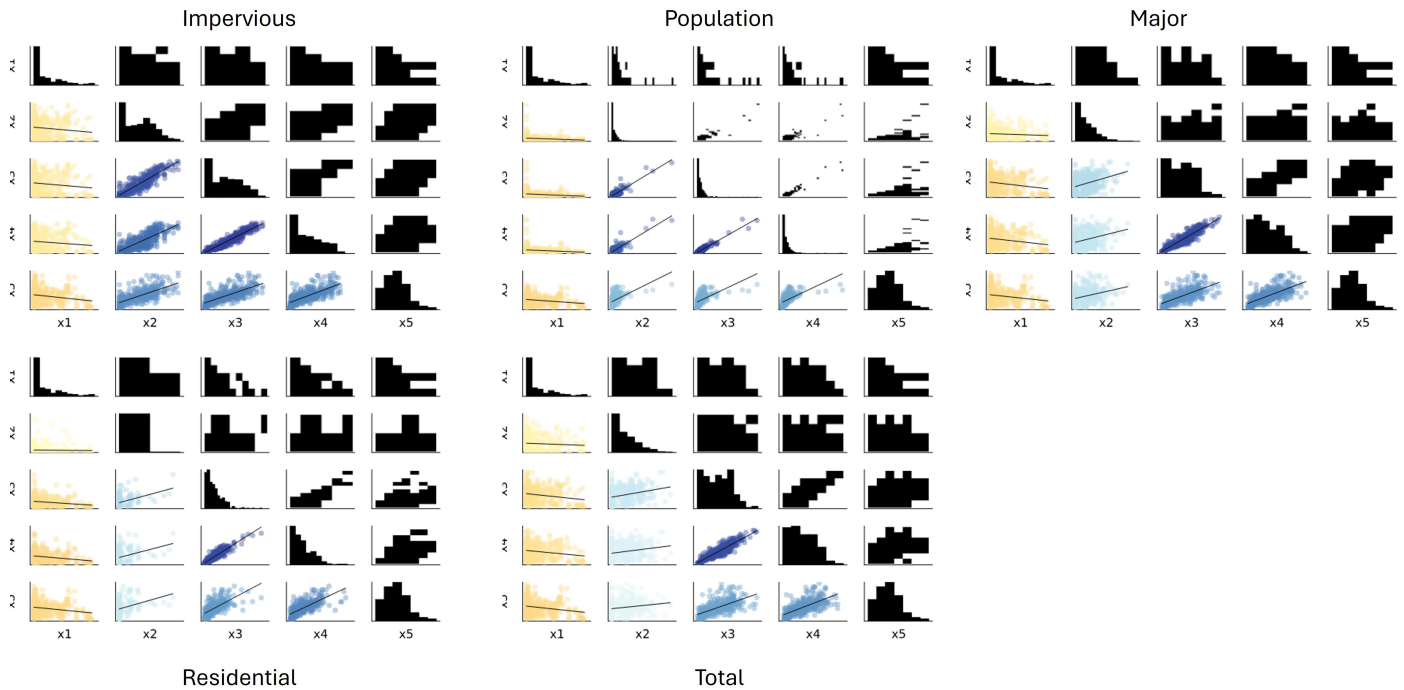


Figure 1: Correlations of each dependent variables

As per the correlation plots, it was assumed that as the color of the curves get darker. By comparing all, it was concluded that distance to coast, impervious 100, major 100, major 5000, resident 100, resident 5000, total 100, and total 5000 have less correlation. This was reanalyzed through correlation plotting as shown in figure 2 to be more accurate.

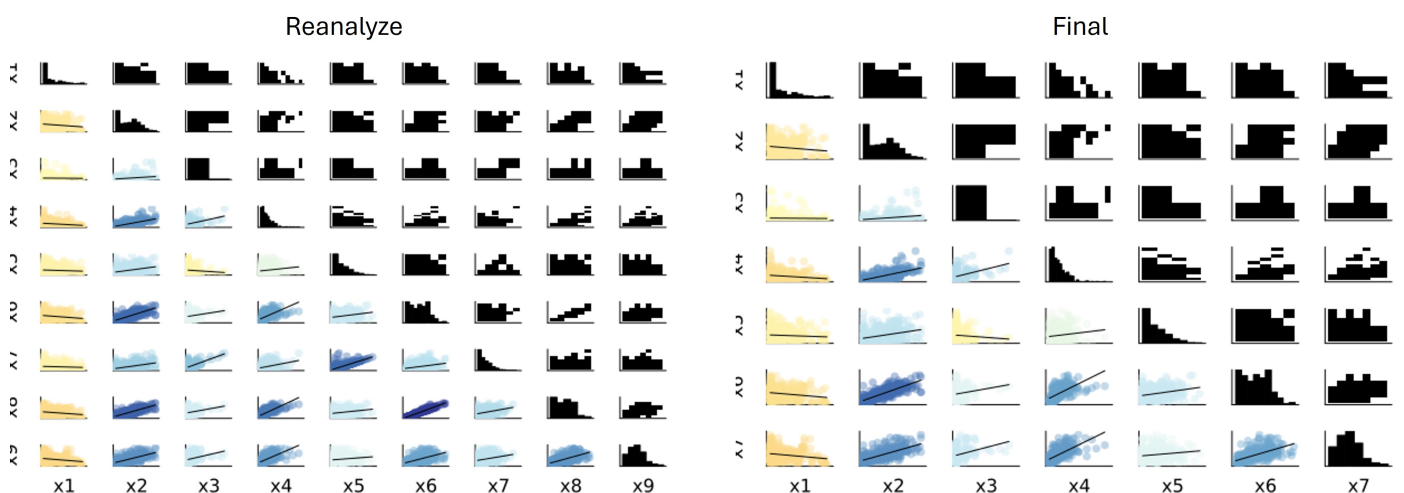


Figure 2: Reanalyze and final correlation checking

Finally, Distance to coast, Impervious 100, Major 100, Major 5000, Resident 100, and Resident 5000 were selected as dependent variables. The dataframe consists of 369 rows and 8 columns (including state information and independent variable - Observed NO2 ppb).

df2 =

	State	Distance_to_coast_km	Impervious_100	Major_100	Major_5000	Resident_100
1	"AZ"	313.0	59.4431	0.0	166.584	0.61637
2	"AZ"	323.8	72.0	0.0	232.425	0.26126
3	"AZ"	308.4	53.0	0.15677	115.958	0.3946
4	"AZ"	309.0	61.3099	0.30378	198.04	0.07244
5	"AZ"	269.5	12.0	0.19457	20.6286	0.0
6	"AZ"	314.7	66.0	0.0	143.582	0.28342
7	"AZ"	275.1	57.0	0.0	122.248	0.33082
8	"AZ"	271.6	48.0	0.0	76.5419	0.0
9	"AZ"	112.4	16.0	0.0	35.9344	0.47504
10	"AR"	629.7	24.0641	0.0	46.9212	0.37524
⋮ more						
369	"WY"	1536.0	0.0	0.0	20.0424	0.0

Figure 3: Training data

df2 =

	State	Distance_to_coast_km	Impervious_100	Major_100	Major_5000	Resident_100
1	"AZ"	313.0	59.4431	0.0	166.584	0.61637
2	"AZ"	323.8	72.0	0.0	232.425	0.26126
3	"AZ"	308.4	53.0	0.15677	115.958	0.3946
4	"AZ"	309.0	61.3099	0.30378	198.04	0.07244
5	"AZ"	269.5	12.0	0.19457	20.6286	0.0
6	"AZ"	314.7	66.0	0.0	143.582	0.28342
7	"AZ"	275.1	57.0	0.0	122.248	0.33082
8	"AZ"	271.6	48.0	0.0	76.5419	0.0
9	"AZ"	112.4	16.0	0.0	35.9344	0.47504
10	"AR"	629.7	24.0641	0.0	46.9212	0.37524
⋮ more						
369	"WY"	1536.0	0.0	0.0	20.0424	0.0

Figure 4: Training data

Figure 3 is training dataset and figure 4 is testing dataset. As more data are used in analysis, the training gets more accurate. Thus, all state information with pre-selected variables were used for training. After machine learning, each results were compared to the four states information and/or all states information.

Machine learning

Machine learning was performed in five different methods: linear regression, decision tree, k-means clustering, neural network, [].

References

- [1] M.J. Bechle, D.B. Millet, J.D. Marshall, National Spatiotemporal Exposure Surface for NO₂: Monthly Scaling of a Satellite-Derived Land-Use Regression, 2000–2010, *Environ Sci & Technol.* 49 (2015) 12297–12305. doi:10.1021/acs.est.5b02882.
- [2] L. Smith, S. Mukerjee, K. Kovalcik, E. Sams, C. Stallings, E. Hudgens, J. Scott, T. Krantz, L. Neas, Near-road measurements for nitrogen dioxide and its association with traffic exposure zones, *Atmos Pollut Res.* 6 (2015) 1082–1086. doi:<https://doi.org/10.1016/j.apr.2015.06.005>.
- [3] G. Hoek, R. Beelen, K. de Hoogh, D. Vienneau, J. Gulliver, P. Fischer, D. Briggs, A review of land-use regression models to assess spatial variation of outdoor air pollution, *Atmos Environ.* 42 (2008) 7561–7578. doi:<https://doi.org/10.1016/j.atmosenv.2008.05.057>.
- [4] E. V Novotny, M.J. Bechle, D.B. Millet, J.D. Marshall, National Satellite-Based Land-Use Regression: NO₂ in the United States, *Environ Sci & Technol.* 45 (2011) 4407–4414. doi:10.1021/es103578x.
- [5] H. Saki, G. Goudarzi, S. Jalali, G. Barzegar, M. Farhadi, I. Parseh, S. Geravandi, S. Salmanzadeh, F. Yousefi, M.J. Mohammadi, Study of relationship between nitrogen dioxide and chronic obstructive pulmonary disease in Bushehr, Iran, *Clin Epidemiol Glob Heal.* 8 (2020) 446–449. doi:<https://doi.org/10.1016/j.cegh.2019.10.006>.
- [6] J.E. Hart, J.D. Yanosky, R.C. Puett, L. Ryan, D.W. Dockery, T.J. Smith, E. Garshick, F. Laden, Spatial Modeling of PM₁₀ and NO₂ in the Continental United States, 1985–2000, *Environ Health Perspect.* 117 (2009) 1690–1696. doi:10.1289/ehp.0900840.