








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@21e4b15](#) on November 26, 2024.

Published: November 18, 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 used for this project focused on modeling NO₂ concentrations in different locations of the different states of US using a Land-Use Regression (LUR) approach [1]. The data was collected from U.S Environmental Protection Agency (EPA) regulatory monitoring stations between 2000 and 2010. This NO₂ concentration measurement in different location is the dependent variable of the dataset.

The dataset contained different Land-Use geographic variables which are considered as the independent variables. These variables include impervious surface area (%coverage), population density, road lengths(major, residential, total), elevation of the location, distance to coast etc. These variables were evaluated over 22 buffers ranging from 100 m to 10 km to capture the local and regional land use pattern.

The dataset is a CSV file that contains 370 rows and 134 column which can be accessed from this link: <https://drive.google.com/file/d/1dCy4GJo4pk0tVJMhmtnC5ZF12hiWM91q/view> [Accessed 10/15/2024]

Proposal:

The main goal of this project will be to develop a predictive model which will predict the NO₂ concentration of any location of US from different land use variables given in the dataset. It would be very useful to predict the concentrations of NO₂ in any given location which will be helpful in identifying locations that immediate preventive measures and taking necessary actions.

Exploratory Data Analysis

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 NO₂ from 2000 to 2010. The dataset contains spatial and temporal concentration of NO₂ 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, NO₂ 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 NO₂ 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&oid=116140173519287299300

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 NO_2 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 NO_2 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 NO_2 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 NO_2 emission and also the absence of natural filtration effect with the absence of vegetation is another major source of NO_2 emission.

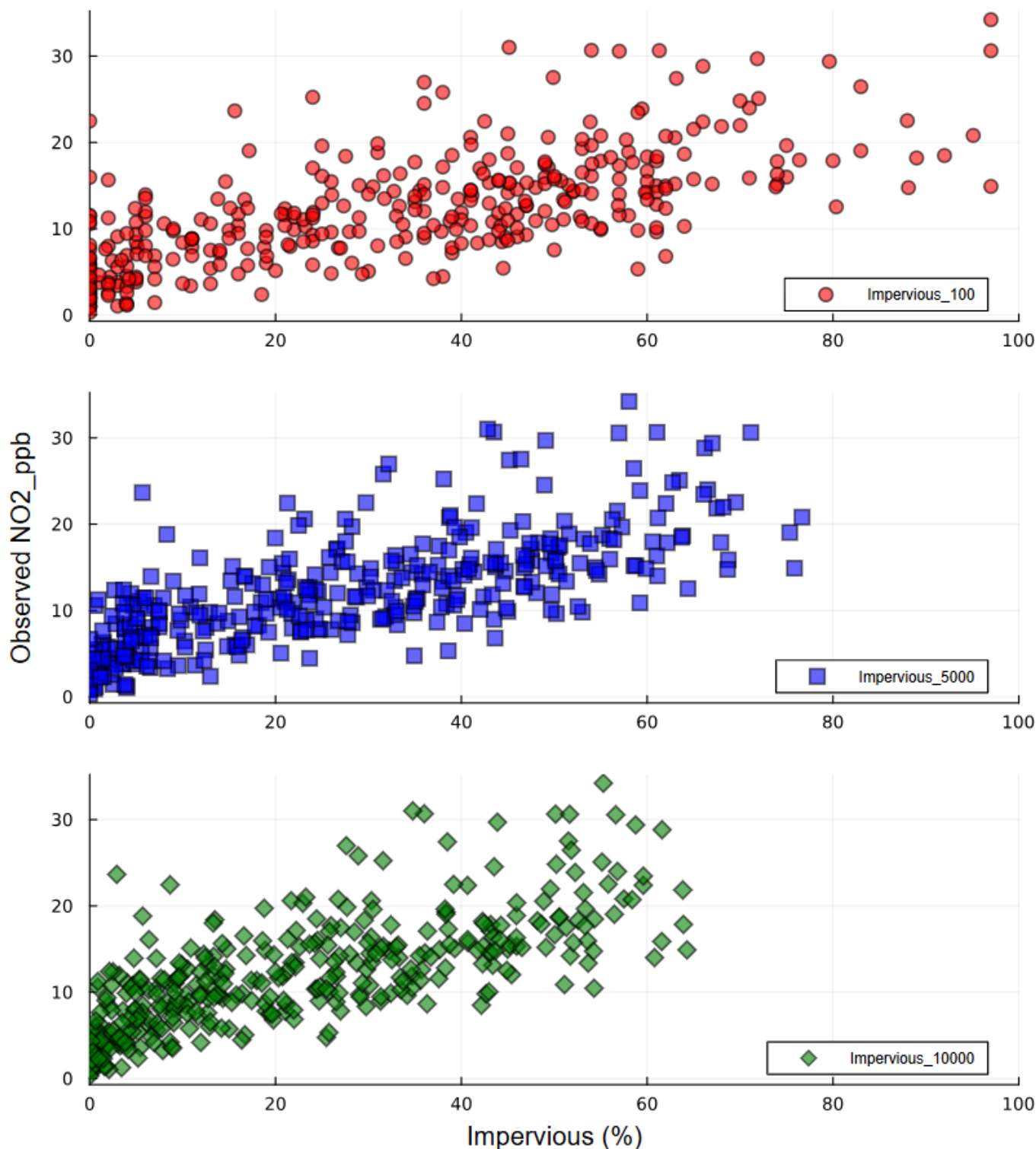


Figure 2: Variation of measured NO₂ concentration with the impervious surface at 100m, 5000m and 10000m radius around the monitor station.

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).

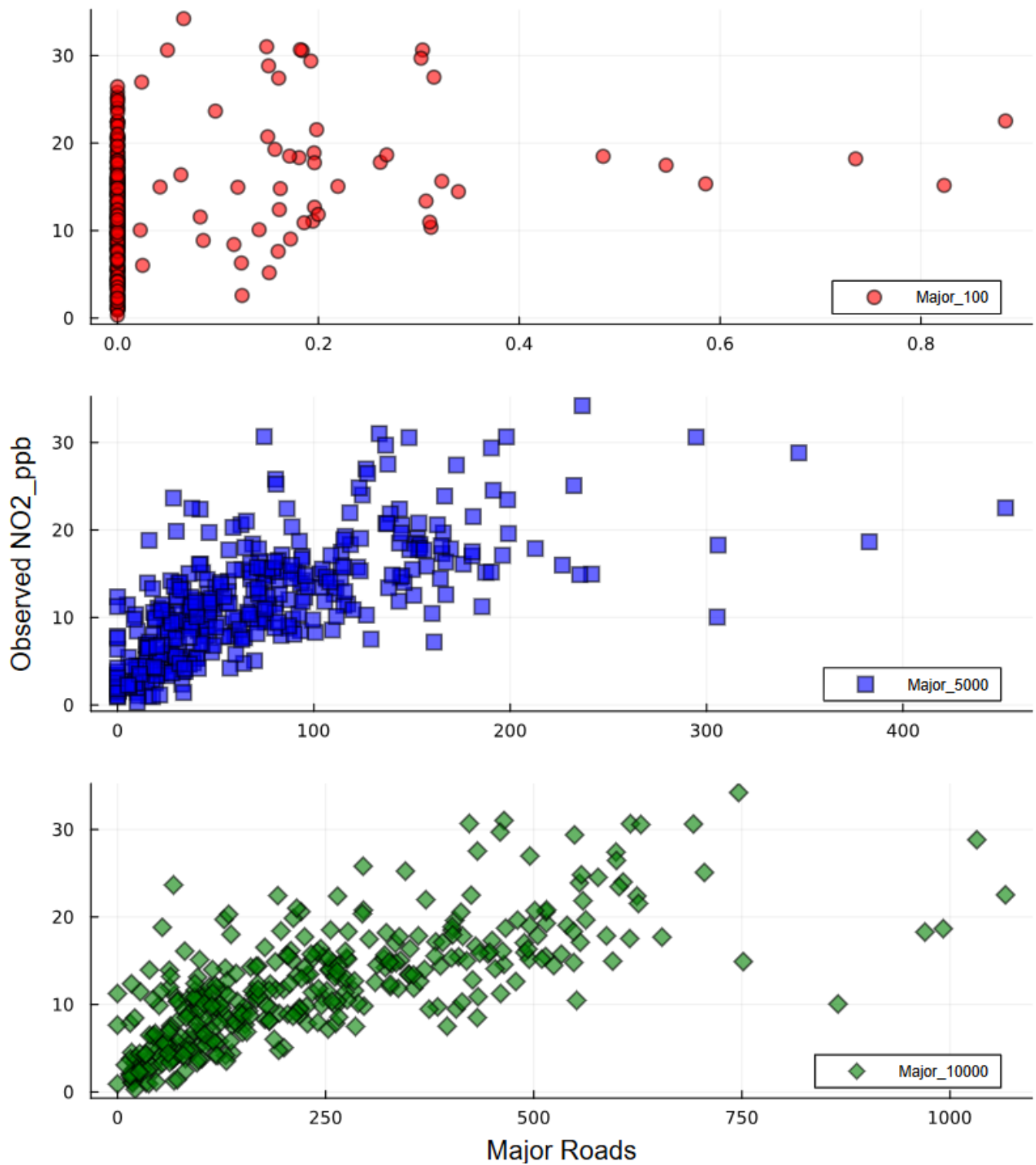


Figure 3: Variation of measured NO₂ concentration with the length of major roads at 100m, 5000m and 10000m radius around the monitor station.

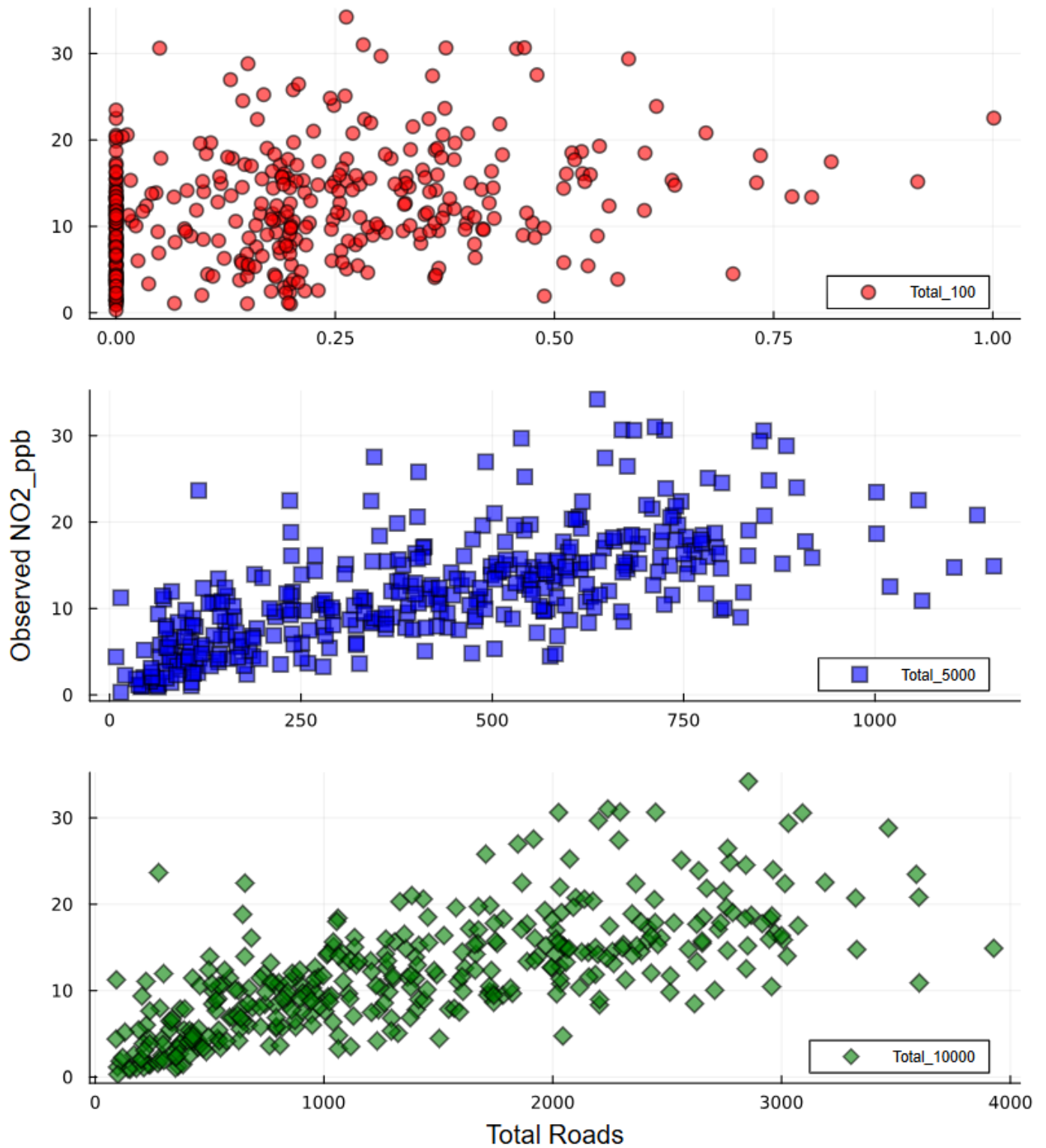


Figure 4: Variation of measured NO₂ concentration with the length of total roads at 100m, 5000m and 10000m radius around the monitor station.

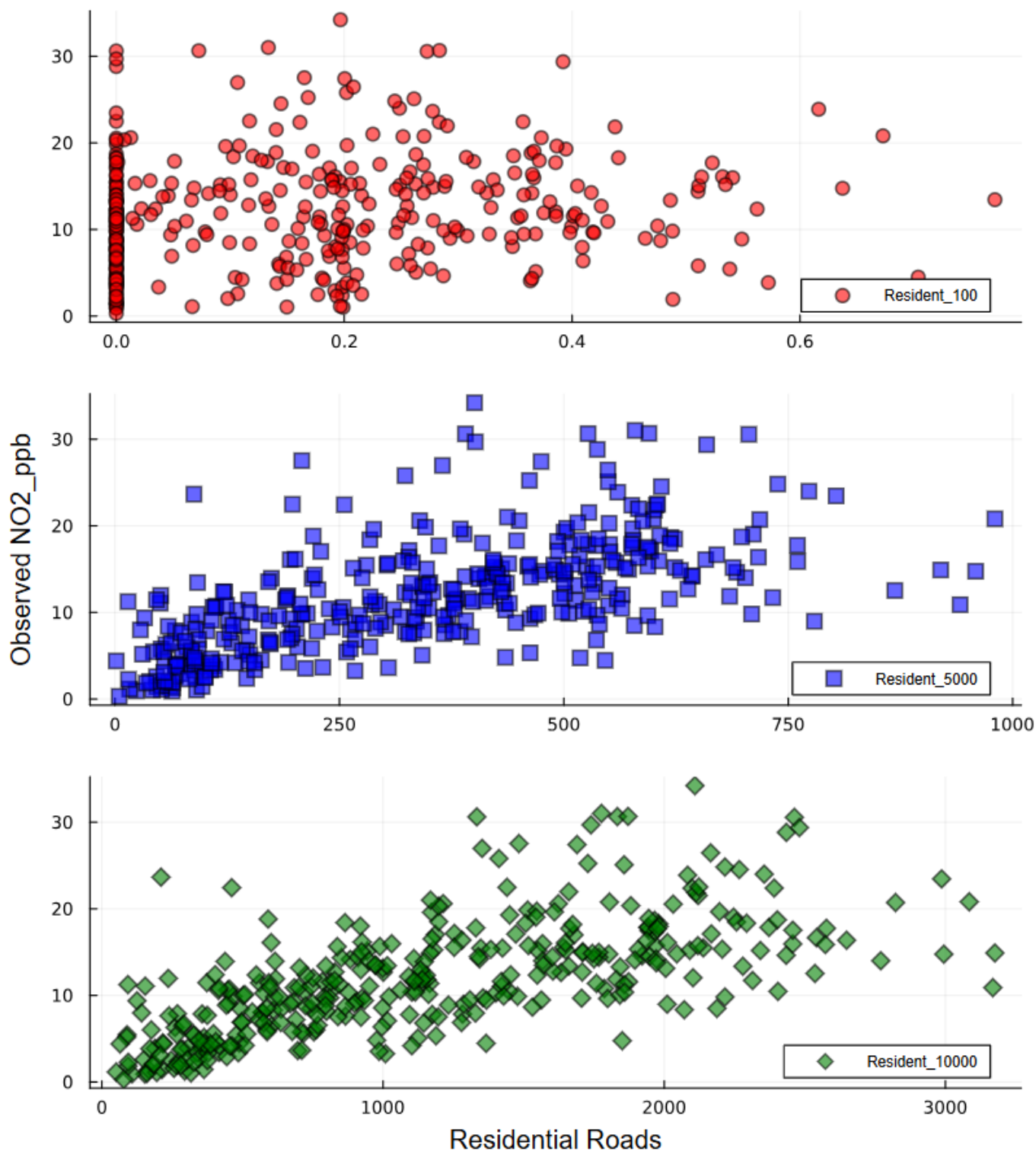


Figure 5: Variation of measured NO₂ concentration with the length of residential roads at 100m, 5000m and 10000m radius around the monitor station.

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).

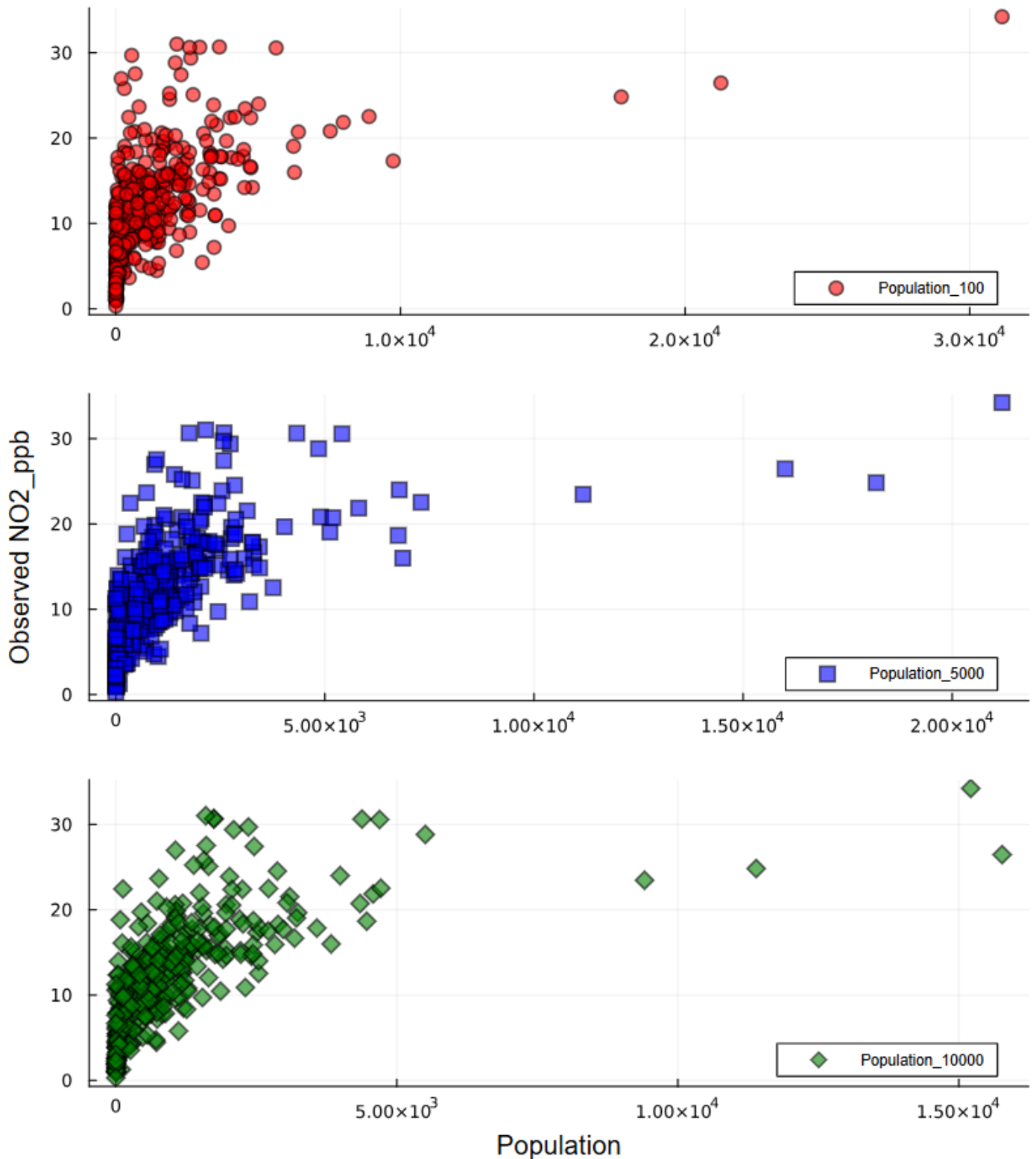


Figure 6: Variation of measured NO₂ concentration with the population at 100m, 5000m and 10000m radius around the monitor station.

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.

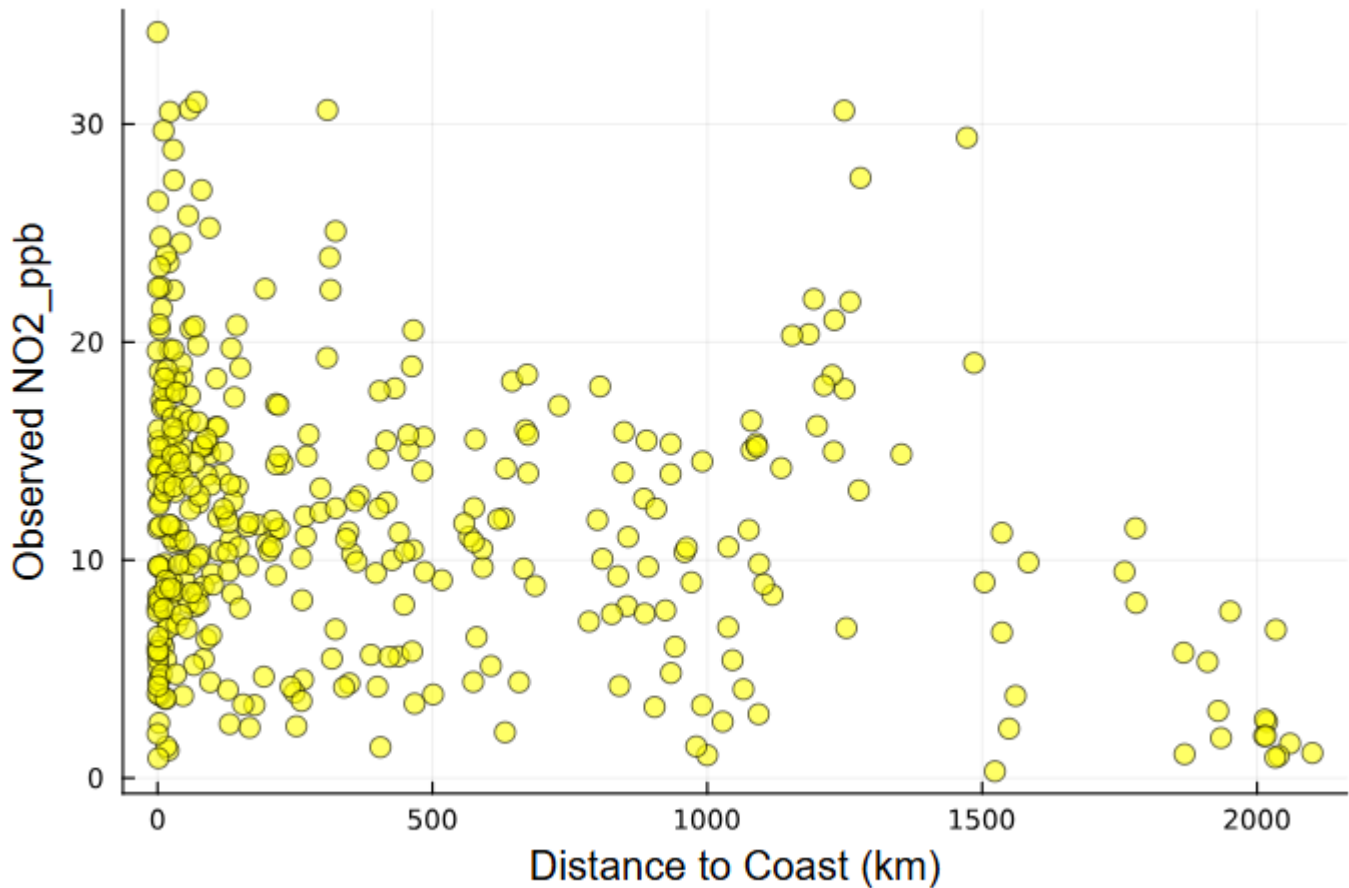


Figure 7: Variation of measured NO₂ concentration with the distance to coast from the monitor station.

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.

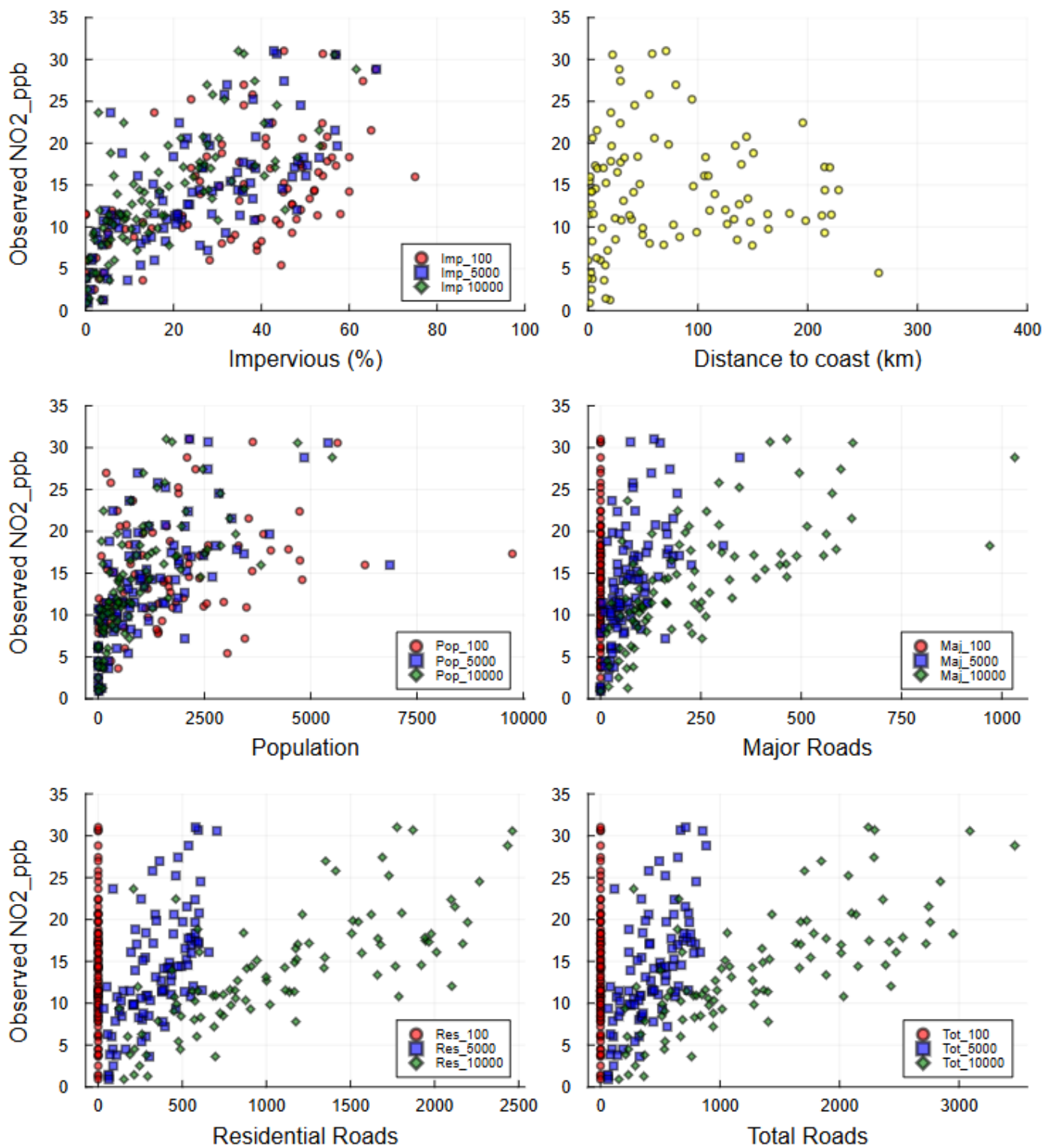


Figure 8: Effect of different land use characteristics on NO₂ concentration of CA state.

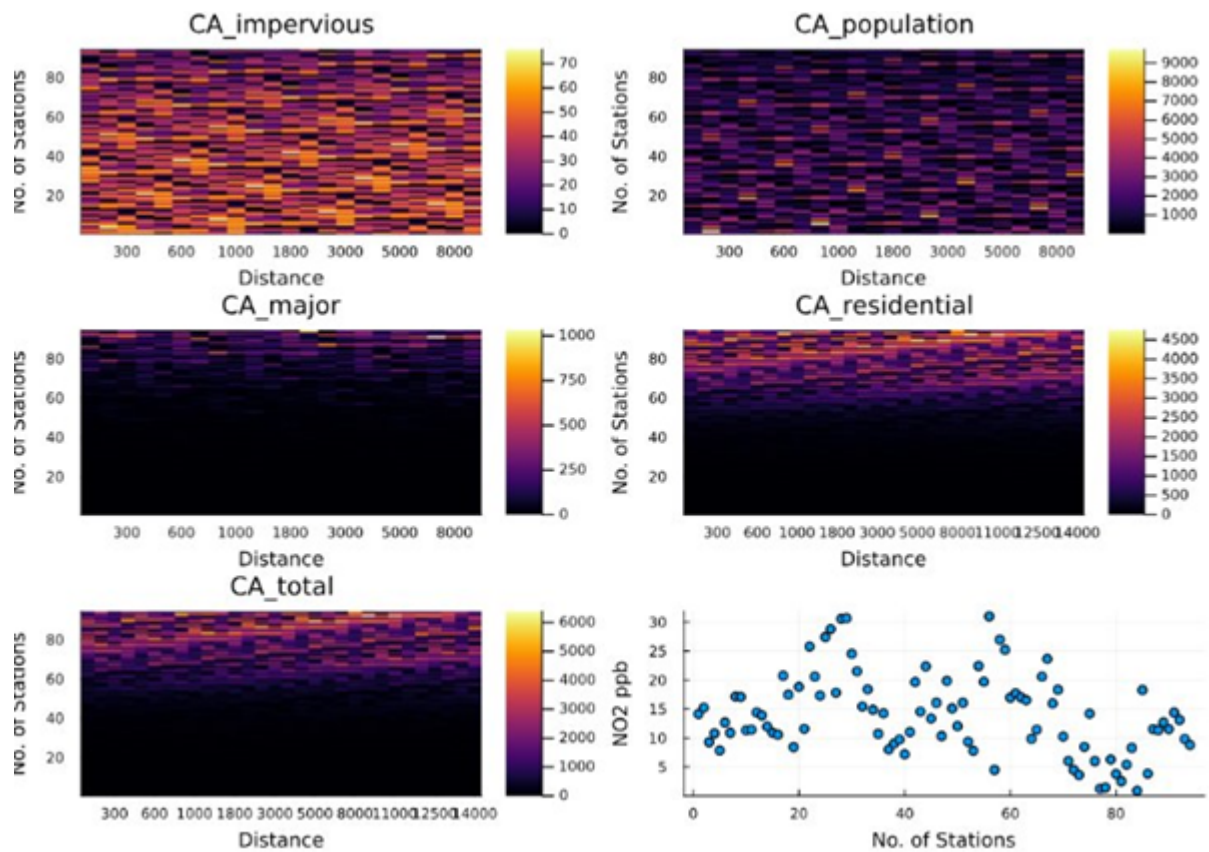


Figure 9: Heatmap to understand the effect of different land use characteristics on NO₂ concentration of CA state.

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.

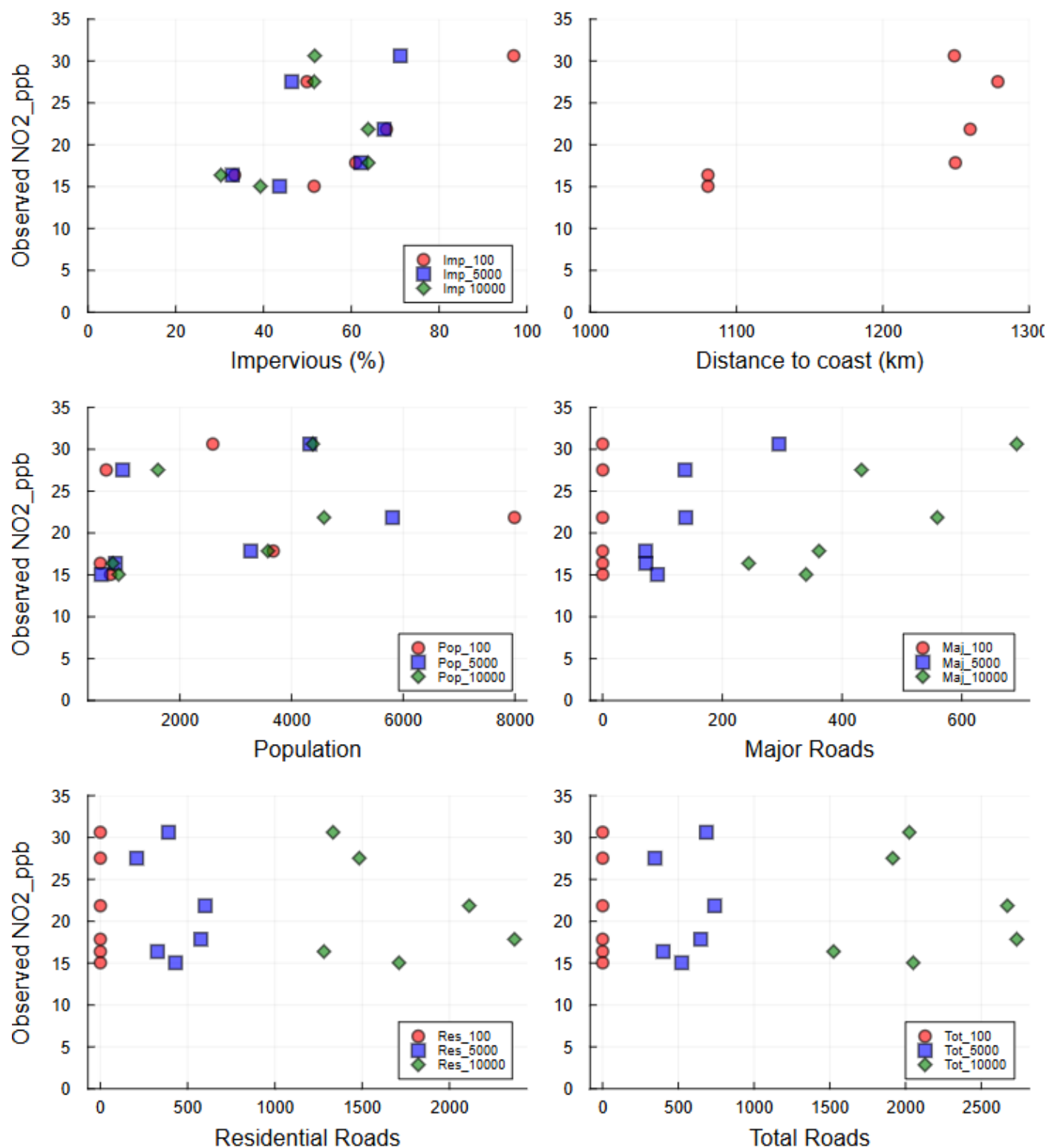


Figure 10: Effect of different land use characteristics on NO₂ concentration of IL state.

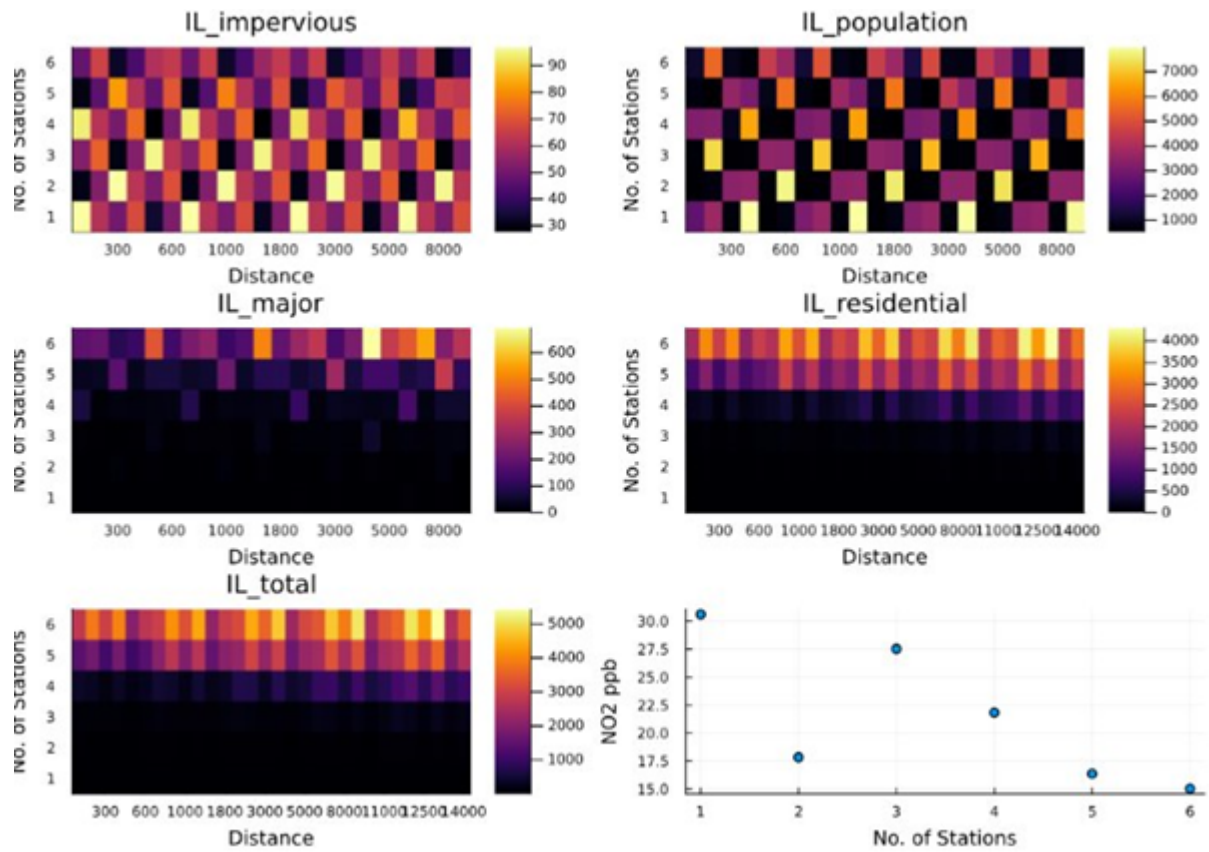


Figure 11: Heatmap to understand the effect of different land use characteristics on NO₂ concentration of IL state.

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.

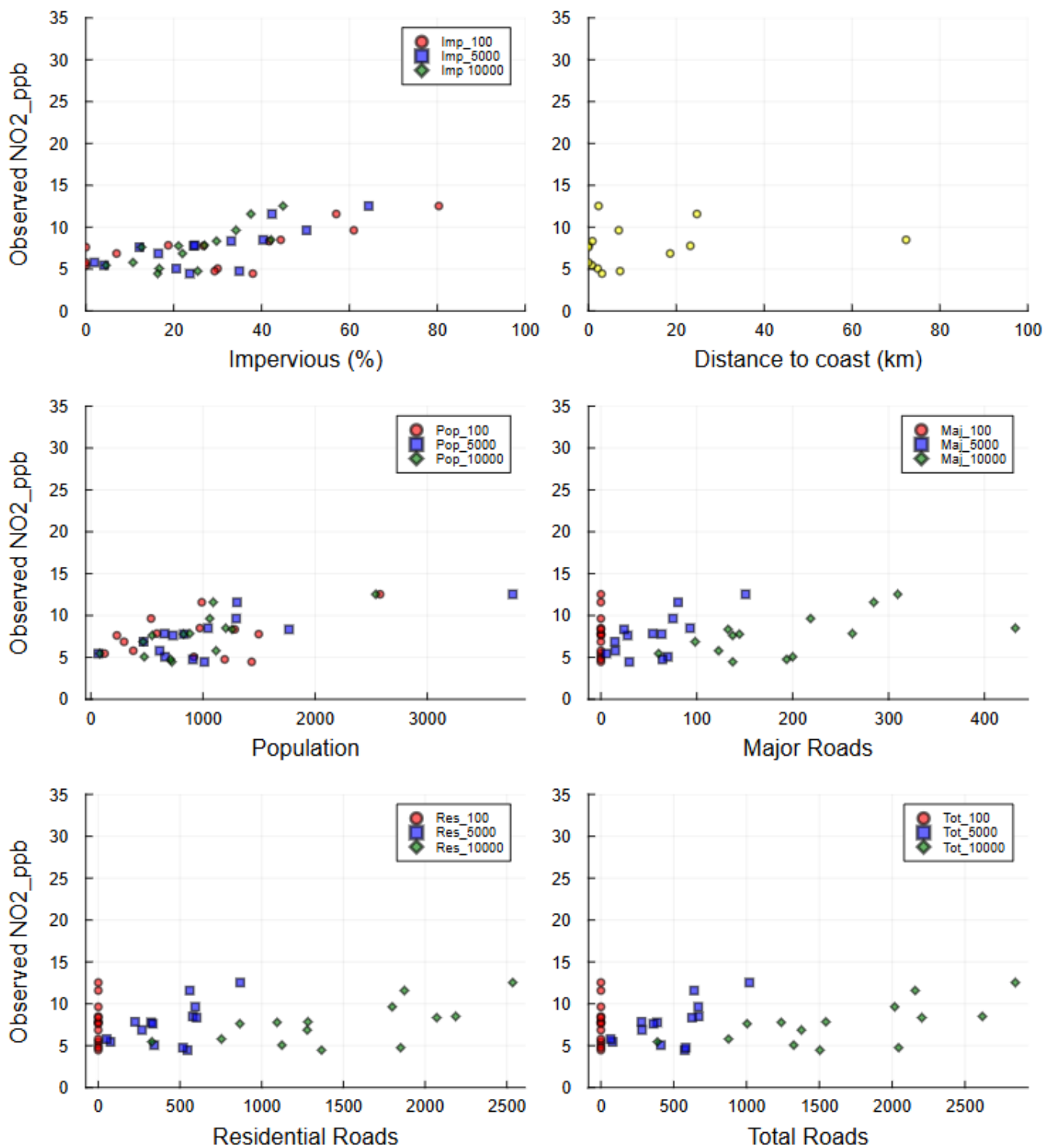


Figure 12: Effect of different land use characteristics on NO₂ concentration of FL state.

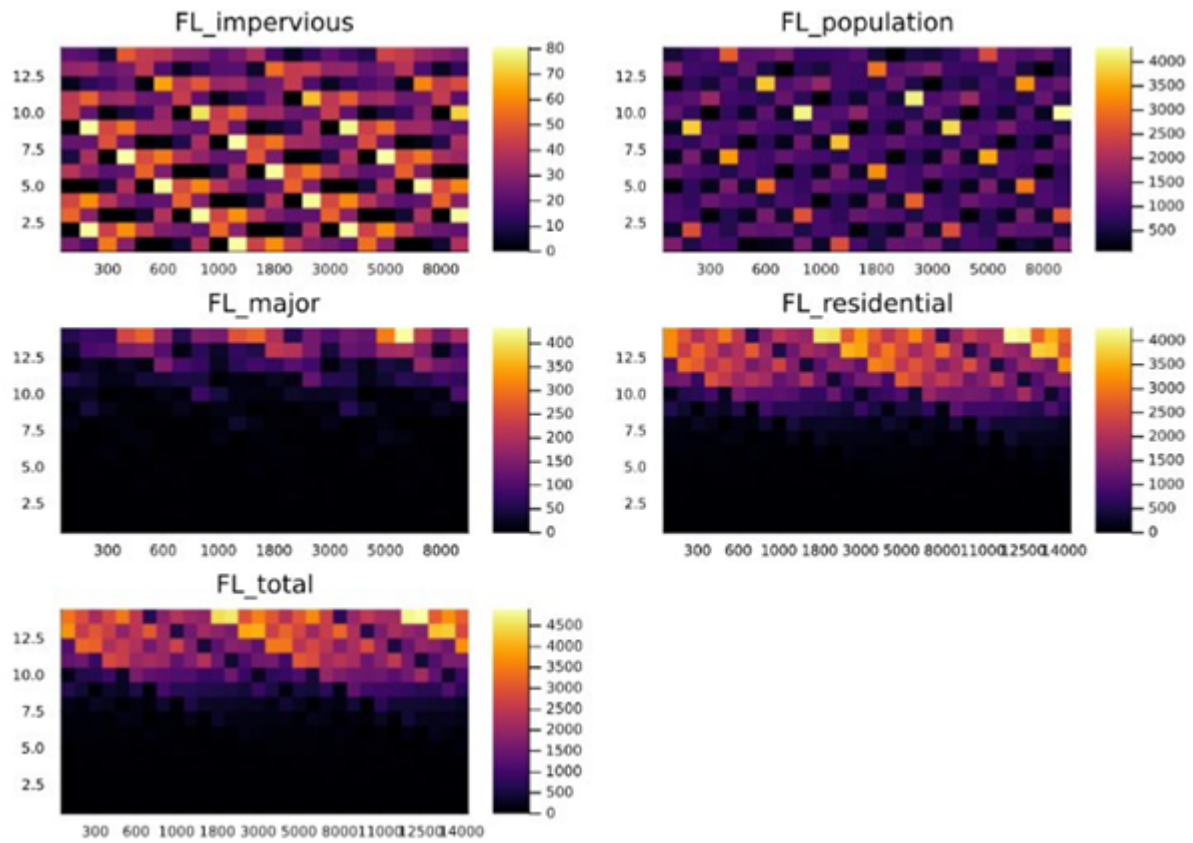


Figure 13: Heatmap to understand the effect of different land use characteristics on NO₂ concentration of FL state.

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.

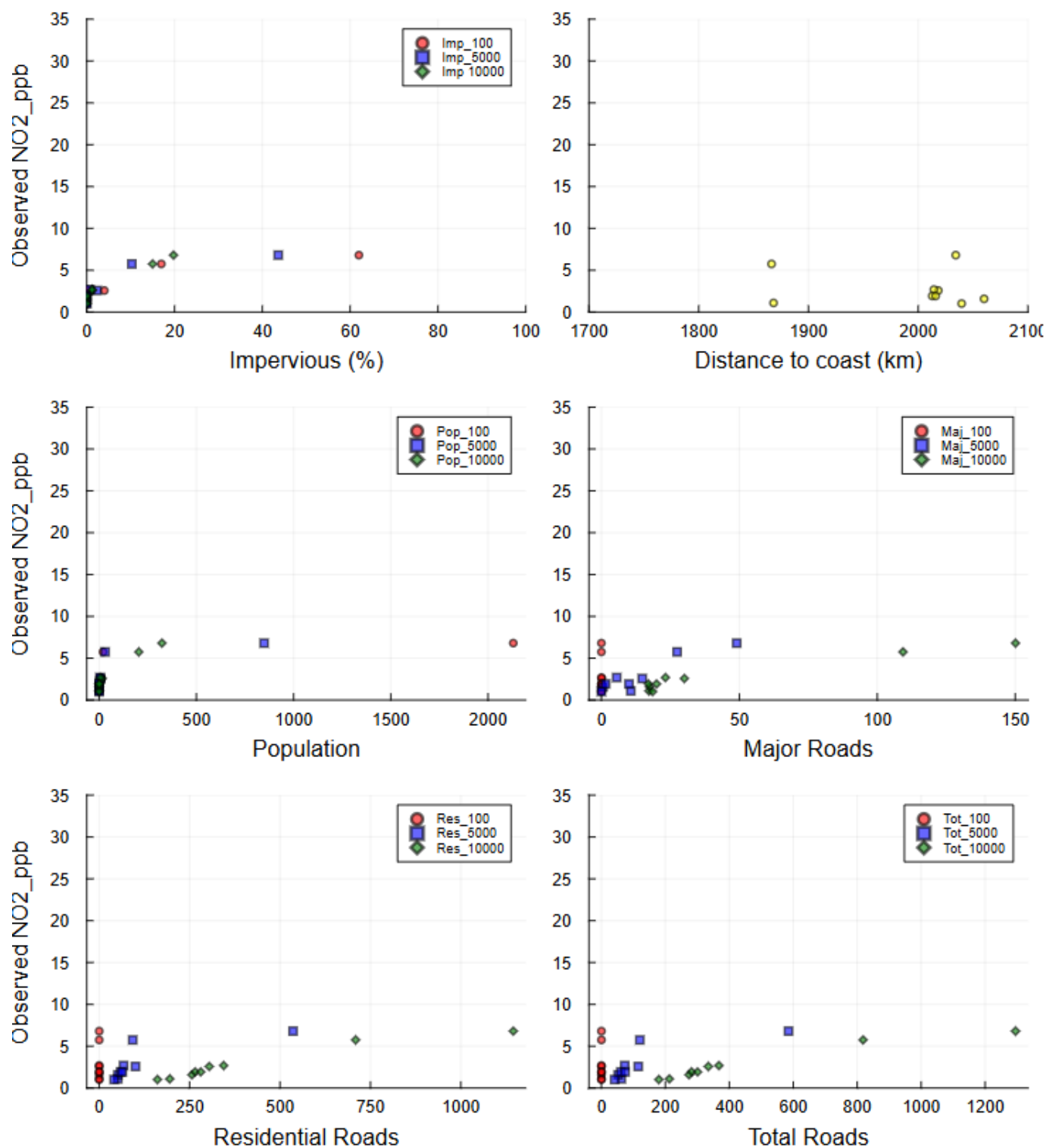


Figure 14: Effect of different land use characteristics on NO₂ concentration of ND state.

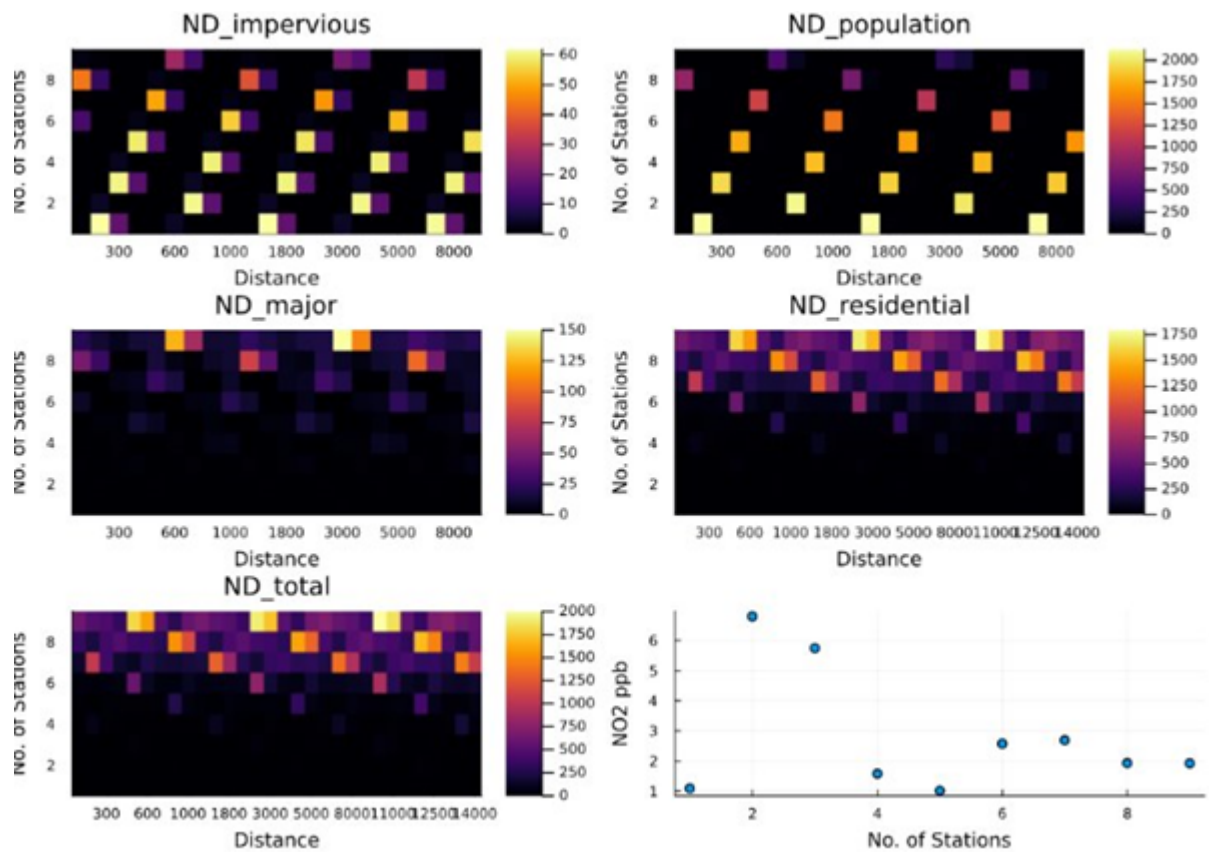


Figure 15:Heatmap to understand the effect of different land use characteristics on NO₂ concentration of ND state.

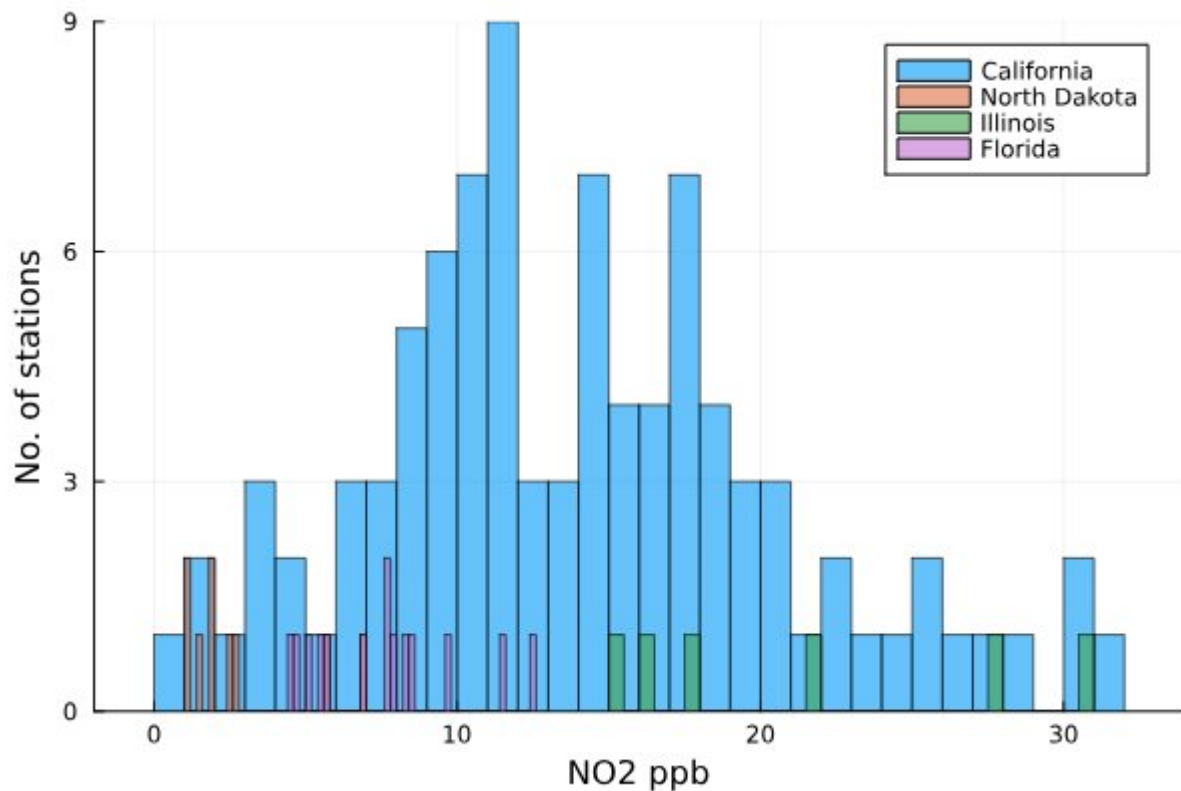


Figure 16:Distribution of NO₂ concentration across CA, ND, IL, FL based on the station.

Plan for 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.

The main objective of this project is to develop a predictive model to predict the NO₂ concentration accurately based on the land use pattern of a particular location. The dataset contained around 370 observations and 128 land use pattern variables such as Impervious_100, major_100, Population_100, Resident_100, total_100, distance_to_coast etc. Therefore, first important step of developing a predictive model was to select the important features required for the model where correlation plot and Lasso regularization techniques were utilized. Once the feature selection was done, different machine learning models were trained and used for predicting the NO₂ concentrations in test data. The model accuracy was evaluated in terms of three metrics-Mean Squared Error (MSE), Root Means Squared Error (RMSE) and R² value. Finally, a comparative analysis has been done to find out the accuracy of different models used for prediction in this study.

Data selection and Feature Engineering

To find out the correlation between different independent variables, the data was sorted by using the correlation plot. Since there were numerous variables in the dataset, initially selective dependent variables introduced in the exploratory data analysis were considered which are_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 analysis was initially divided into five distinct groups to facilitate comparison, as visually inspecting all features simultaneously was impractical. This segmentation also allowed for classification of the data based on different features, aiding in the identification of relevant variables. Figure 1 presents the correlation plots for each group, illustrating the relationships between features.

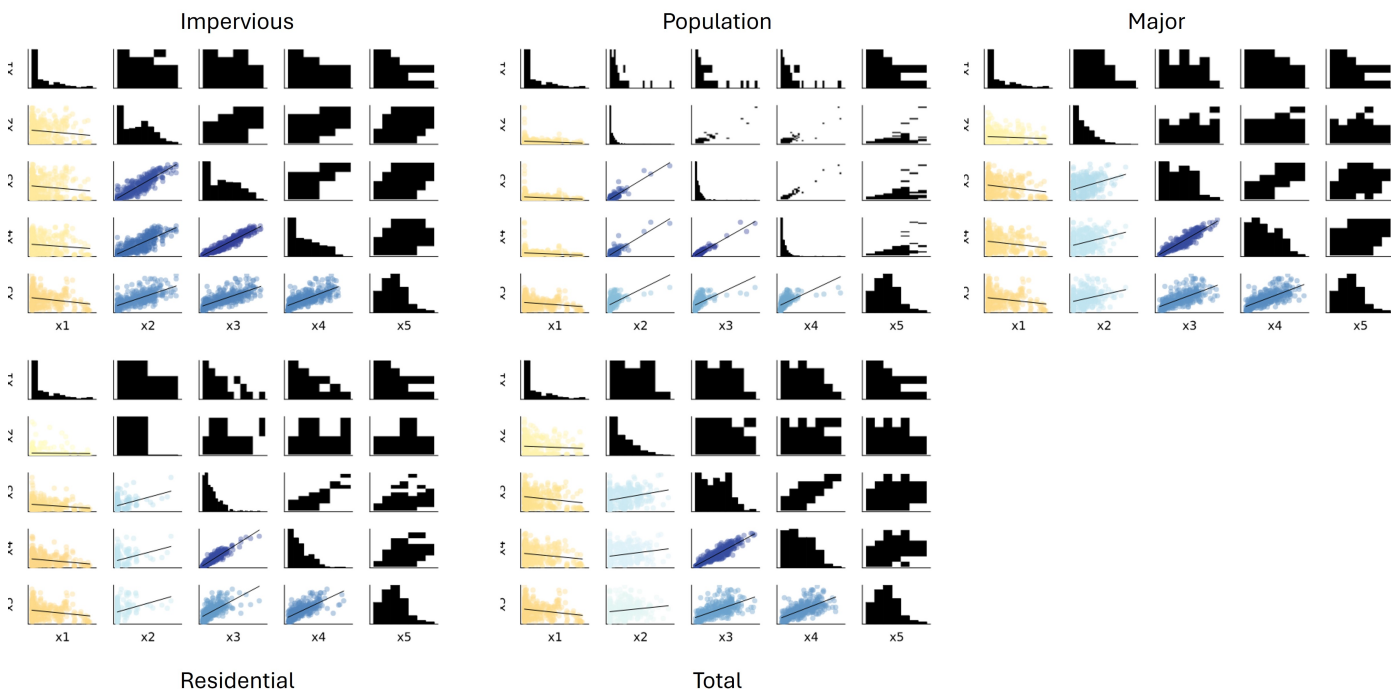


Figure 1: Correlations of each dependent variables

Based on the correlation plots, it was observed that as the color of the curves became darker, the degree of correlation appeared to decrease. By comparing all plots, it was concluded that features such as distance to coast, impervious 100, major 100, major 5000, resident 100, resident 5000, total 100, and total 5000 exhibited lower inter correlations among themselves. To ensure greater accuracy, these findings were reanalyzed through detailed correlation plotting, as illustrated in Figure 2.

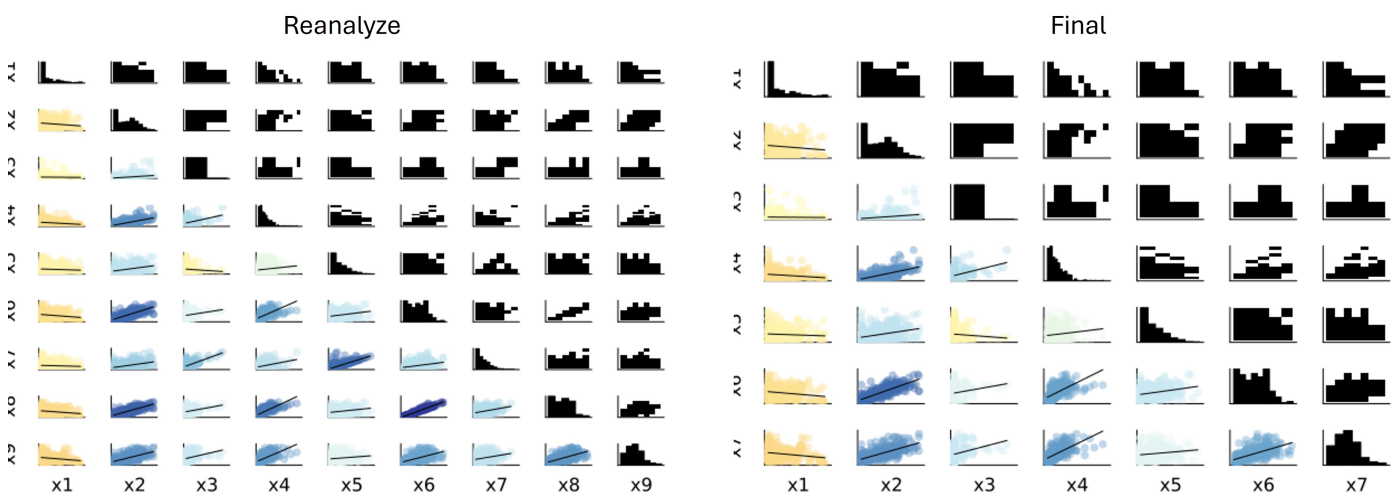


Figure 2: Reanalyze and final correlation checking

To aid this feature selection process with correlation plot, Lasso Regression (Least Absolute Shrinkage and Selection Operator) was applied to the dataset for selecting the most relevant features to predict observed NO₂ concentrations. The main advantage of Lasso regression is that it reduces complexity

by shrinking less important feature coefficient to zero. The reason of selecting Lasso was as it performs both feature selection and regularization, simplifying the model by irrelevant and redundant features.

```

• begin
• XX = Matrix(data[:, Not([:Observed_NO2_ppb, :State, :Monitor_ID, :Latitude,
• :Longitude, :WRF_DOMINO])]) # Replace 'NO2' with your target variable name
• yy = Vector(data[:, :Observed_NO2_ppb])
•
• Random.seed!(42)
•
• # Standardize the features to have mean 0 and standard deviation 1
• X_standardized = (XX .- mean(XX, dims=1)) ./ std(XX, dims=1)
•
• # Fit a Lasso model with cross-validation to find the best lambda
• fit = glmnetcv(X_standardized, yy, alpha=1.0) # alpha=1.0 for Lasso
•
• # Find the lambda that corresponds to the minimum cross-validation error
• cv_errors = fit.meanloss # Cross-validation errors for each regularization
parameter
• best_param_index = argmin(cv_errors) # Index of the best regularization
parameter
• best_regularization_param = fit.lambda[best_param_index]
• println("Best regularization parameter (lambda): ", best_regularization_param)
•
• # Refit Lasso using the best regularization parameter to get feature coefficients
• lasso_model = glmnet(X_standardized, yy, alpha=1.0,
lambda=[best_regularization_param]) # Note: lambda as array
•
• # Extract the coefficients of the selected features
• coefficients = lasso_model.betas[:, 1] # 'betas' contains coefficients,
selecting the first column for best lambda
•
• # Find indices of features with non-zero coefficients
• selected_features_indices = findall(coefficients .!= 0)
•
• # Get the feature names corresponding to the selected features
• feature_names = names(data, Not([:Observed_NO2_ppb, :State, :Monitor_ID,
:Latitude, :Longitude, :WRF_DOMINO]))
• selected_feature_names = feature_names[selected_features_indices]
•
• # Display the selected features with their coefficients
• println("Selected features and coefficients:")
• for i in 1:length(selected_features_indices)
•     println("Feature: ", selected_feature_names[i], " - Coefficient: ",
coefficients[selected_features_indices[i]])
• end
• end

```

Figure 3: Lasso Regression code used for selecting the input features for predictive modeling.

Lasso Regression along with cross-validation was applied that performs k-fold cross-validation (which is by default 10) to find the optimal regularization parameter, lambda. Alpha value of 1.0 specifies exactly Lasso regularization rather than Ridge or elastic net. Thereafter, using the best lambda value, a new Lasso model was fit which enable the model to learn the optimal coefficient for selected features. Finally, coefficient of features was evaluated and non-zero coefficients were calculated by Lasso for using the selected features for further modelling.

Sl. No	Selected Features	Coefficients
1	Distance_to_coast_km	1.035150
2	Impervious_100	1.396751
3	Impervious_1800	0.919251
4	Impervious_10000	0.510092
5	Elevation_truncated_km	1.123214

Sl. No	Selected Features	Coefficients
6	Major_100	0.298449
7	Major_5000	0.324849
8	Resident_100	1.082713
9	Resident_5000	1.235198
10	total_100	0.112781
11	total_5000	0.175802
12	Population_10000	0.076178

For training and test data distribution, 75/25 ratio was followed where 75% of data points were considered training data and rest of the data points were test data. Out of 369 observations in the dataset, first 277 points were training data and rest of the data points were testing data. Model was first trained on training data and then the model was applied on testing data to find the efficiency of the predictive model.

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: Dataset used in this project: first 277 rows were considered as training data and rest of the points were testing data.

Predictive Modeling

Machine learning methods

Predictive modeling using machine learning was performed using four methods: linear regression, decision tree, random forest regression, and neural network. The dataset, which has 369 data in eight variables, was used for training. The validity of the model was confirmed visually by plotting it linearly and checking the RMSE and R2 values. RMSE is a root mean square error, in which the best value is 0 and the worst value is near infinite [7-8]. As the range of the criteria is infinite, it is difficult to say at which point the data has a good prediction. However, the data is well-trained when the value is near 0. R2 is the proportion of the variance in the dependent variable that is predictable from the independent variables [7-8]. R2 is best when close to 1 and worst in - infinite. Each machine learning method was compared with these criteria to show their effectiveness.

1. Linear regression

The first method used for prediction was linear regression model. This method is quite simple but it would give us a sense of machine learning and the complexity needed for training data. Mean squared error was used to minimize the error and in the linear model and independent variable and gradient descent parameter was used in the model structure. The training set was standardized and normalized to enhance the accuracy of prediction. Figure 5 shows the main flow of the coding.

After training, the dataset was trained as shown in Figure 6. By performing root mean square error, the scoring was 4.3. Still, it is not as low as expected. For future modeling, more dataset could be used for higher accuracy, polymodel rather than linear model can be tried.

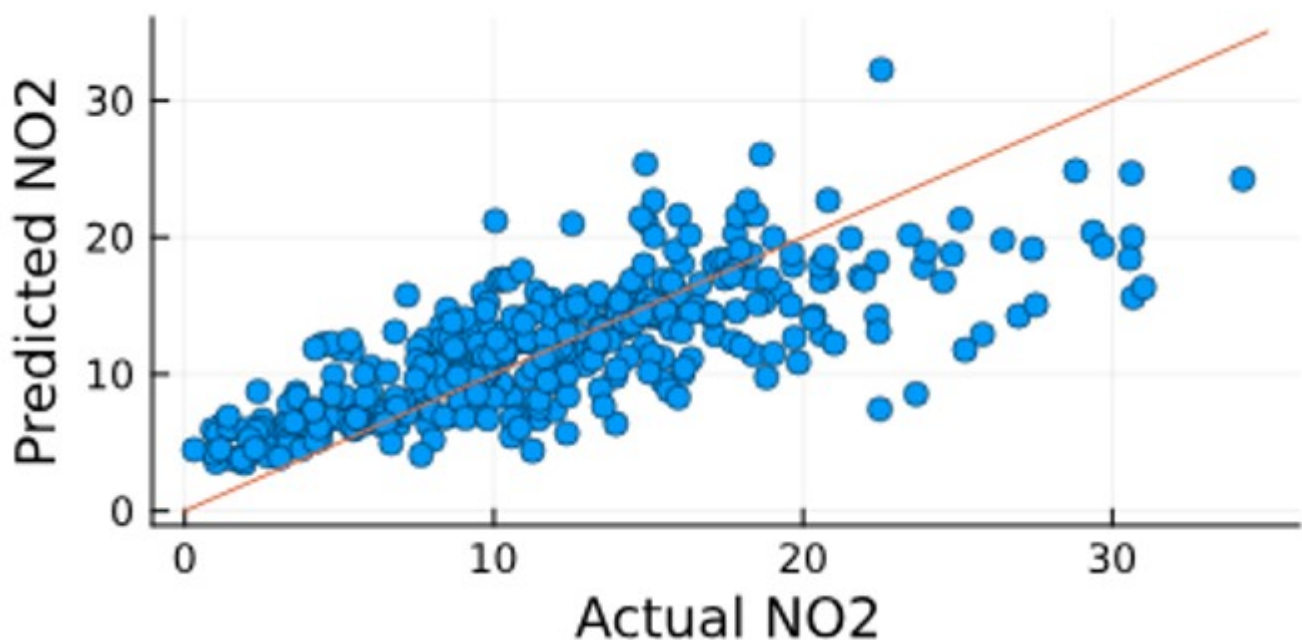


Figure 6: omparison between the NO2 values predicted from Linear Regression model and the observed NO2 values for all the data.

2. Decision tree

Decision tree was also used for predicting the dataset. Six variables that were assigned in the dataset selection, each were gini plotted for setting up the baselines. Figure 7 shows the code flow and Figure 8 shows the gini plot with each variable. Dependende on the ppb value, <10 was decided to be low, >20

was decided to be high, and between that was decided to be medium. These values are the dominant range of ppb level in all states.

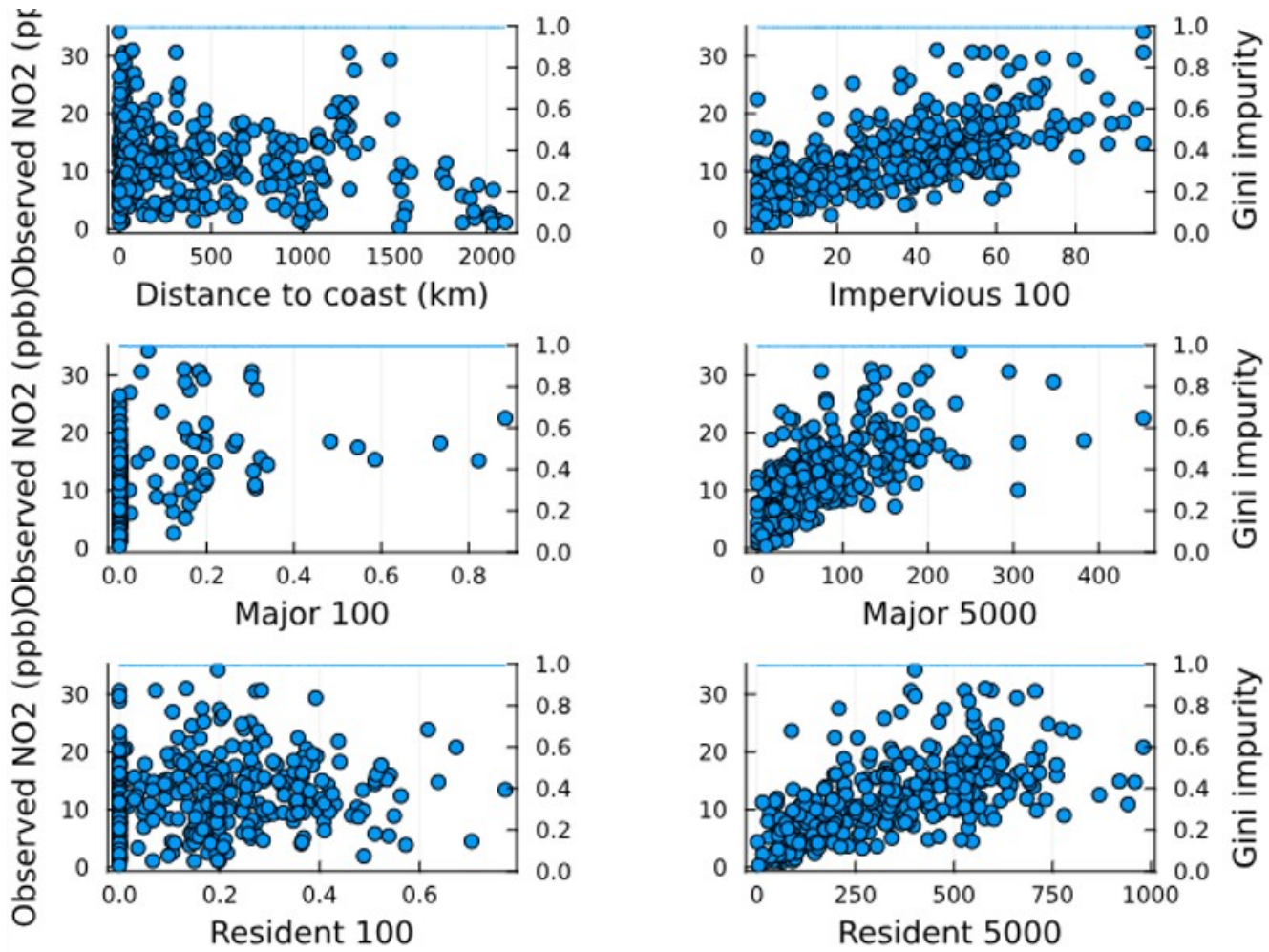


Figure 8: Gini plot shown for selected six variables

With comparing all plots, the cutting points were decided as: 1) Distance to coast (km) > 1500 : low, Distance to coast (km) < 100 : high, Impervious_100 < 20 : low, Impervious 100 > 60 : high, Major 5000 > 200 : high, Major 5000 < 50 : low, Resident 5000 < 200 : low, else : medium (Figure 7).

The model's precision was calculated by measuring recall divided by the average of the prevision and recall, which got value of 0.078. Decision tree method was proven to be quite accurate is classifying NO2 pollutant value in high, medium, and low range.

3. Random forest Regressor

Random forest is an ensemble learning method that aggregates predictions from multiple decision trees to reduce overfitting and improve generalization. Some key parameters of this algorithm are:

- n_trees:** It specifies the number of trees in the forest. A large number improves the stability and reduces variance but increases computational cost. Considering all these factors, the value of n_trees were considered 200 in this project.
- max_depth:** It limits the minimum number of samples required to split and internal node. A depth of 20 is considered in this study which is suitable for capturing complex pattern of land use pattern and correlation with NO2 data without overfitting.

- c. `min_samples_split`: It is the value of minimum number of samples that are required to split an internal node. The large values prevent overfitting by avoiding overly specific splits. A value of 15 was considered in the model.
- d. `min_samples_leaf`: The minimum number of samples required to be at a leaf node which ensures that leaf nodes represent significant data, improving model generalization.

Firstly, the features were standardized by subtracting the mean and dividing by the standard deviation for each column. This ensures that all features are on the same scale, preventing dominance of features with large numerical ranges during model training. For the analysis, DecisionTree package of Julia was used. A typical 80/20 split for training and testing is used to evaluate the model's generalization on unseen data. As an evaluation metric, means squared error (MSE), root mean squared error (RMSE) and R2 score was used for assessing the accuracy of the predictive modelling.

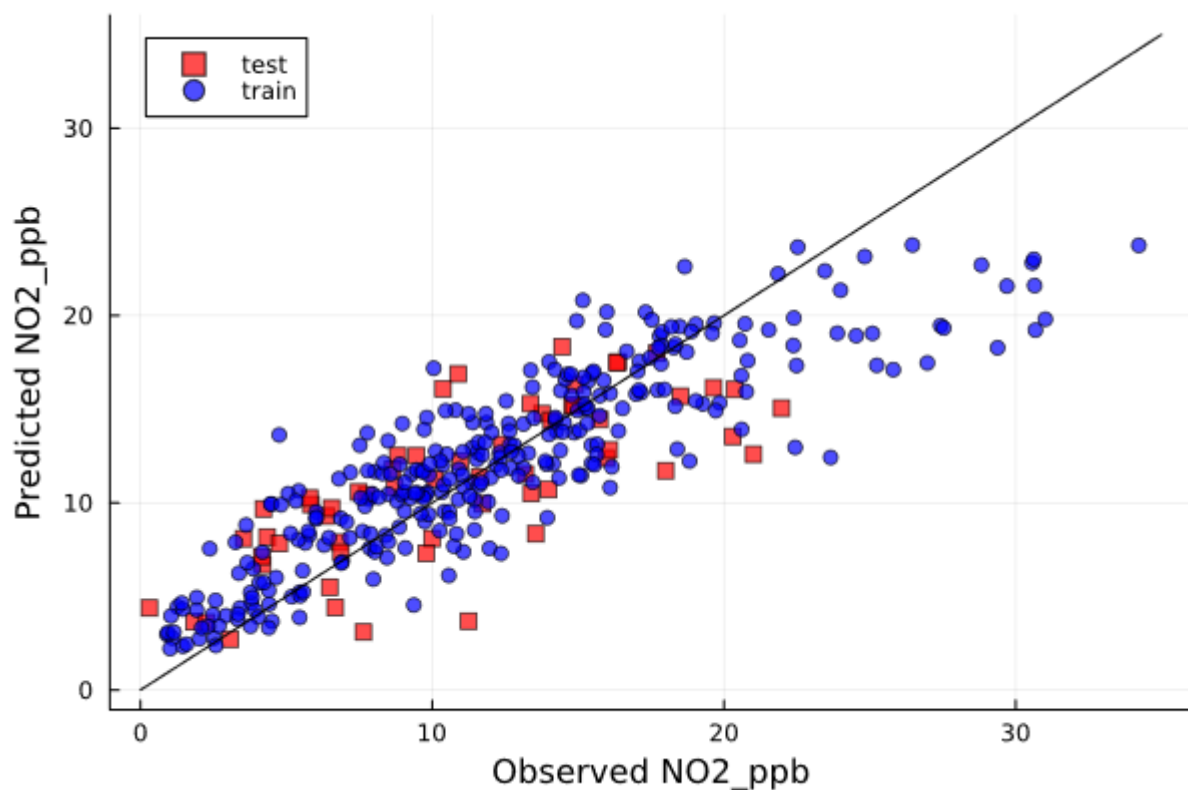


Figure 10: Comparison between the NO₂ values predicted from Random Forest Regressor model and the observed NO₂ values for all the training and test data.

For this predictive model, MSE value was 12.82, RMSE value was 3.58 and R² value of 0.54 was obtained. While R² value of ~0.54 suggests moderate predictive ability, the remaining 40% of the unexplained variance indicates the room for improvement. Potential area for optimizing and improving the models are optimizing the Random Forest hyperparameters, using alternative and strong algorithms like Gradient Boosting etc. Moreover, Since there were quite a large number of independent variables, feature selection also needs to be analyzed more cautiously. However, given the complexity of the given environmental data, R² value of 0.54 can be considered a moderate one.

4. Neural Network Algorithm

A neural network is a computational model which is inspired by the human brain, consists of layers of interconnected neurons. It is commonly used to learn patterns and relationships in a dataset. This model has been applied to explore its predictive capability compared to other techniques used.

The main neural network architecture has been defined using the chain structure which takes 12 features as input and passes them to first hidden layer. Three hidden layers have been added with 128, 64 and 32 neurons along with ReLU activation. Dropout of 0.5 has been kept to drop 50% of the neuron randomly to prevent overfitting. A single neuron outputs the predicted NO₂ value. To minimize the loss function, MSE and L1 regularization has been incorporated in the code. L1 regularization penalty adds a penalty proportional to the absolute values of the model's weights.

The training process is conducted updating the model's weights to minimize the loss function where descent optimizer was used to update the model's weights using gradient descent with a fixed learning rate of 0.001. The model is set for 500 epochs and Flux.train! has been used to compute the loss function with respect to model weights. The Flux package in Julia has been utilized for this modelling.

For this predictive model, MSE value was 10.61, RMSE value was 3.12 and R² value of 0.61 was obtained. Although the R² value increased compared to other predictive models, it is still very low compared to the predicted accuracy for neural network model. Potential improvement scope would be to tune the number of layers, neurons and learning rates. Adding additional L2 penalty might be helpful to optimize the model. Evaluating the model's generalizability with k-fold cross validation might be another option to fine tune the model. But most importantly, selecting the proper input feature might be the most important one which would determine the better prediction accuracy. While revising and optimizing the predictive model, we will try to use feature importance, PCA and other methods to select the proper input features for modelling.

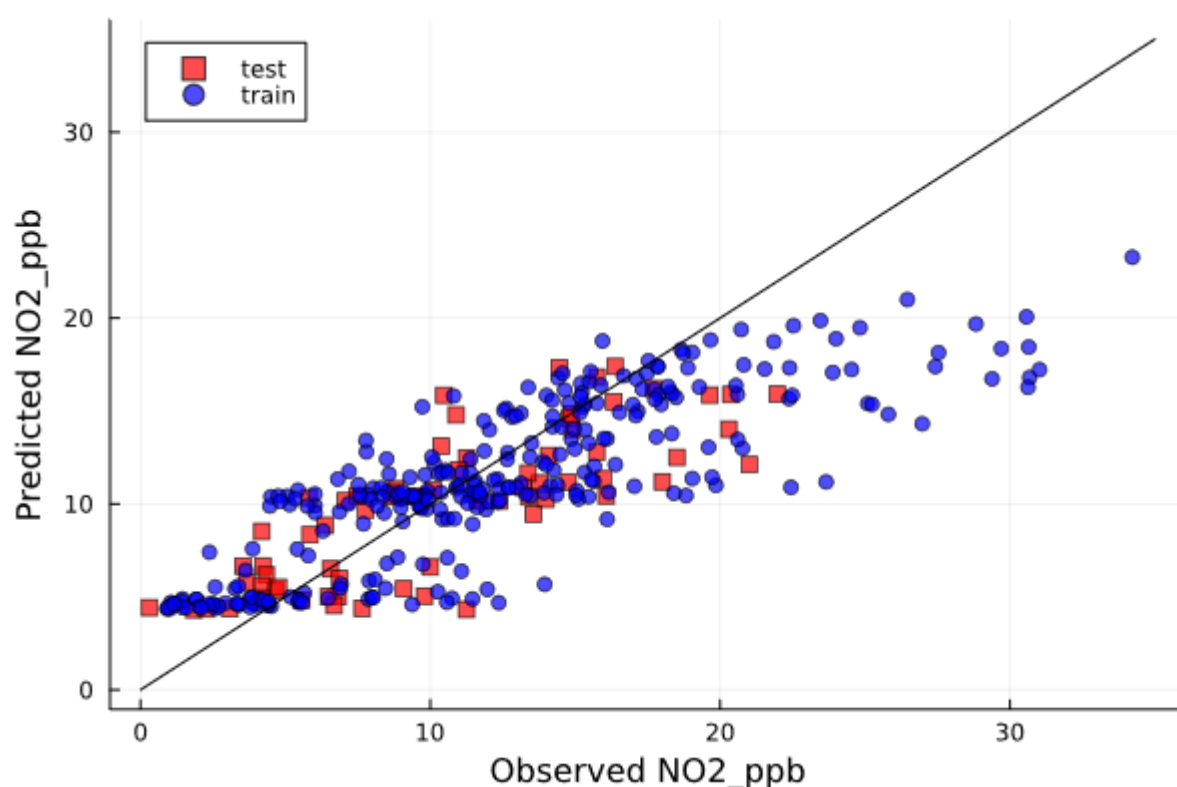


Figure 12: Comparison between the NO₂ values predicted from Neural Network model and the observed NO₂ values for all the training and test data.

Comparison Among Different Predictive Models

The analysis utilized four predictive modeling techniques to predict NO₂ concentrations using a dataset of various land use pattern in different locations across US. The evaluation metrics of all the predictive models used in the analysis are summarized in the table below. The accuracy of the predictive model is as follows: Neural Network model > Random Forest Regression > Linear

Regression. A decision tree classifying technique has been also explored which had a precision of 0.08 indicating the effectiveness of this technique in classifying the NO₂ concentrations into three categories-high, medium and low. It is expected that linear regression will be poor fit for this dataset because of the complexity of the dataset and non-linear correlation between the independent variables and NO₂ concentrations. Random Forest model with R² value of 0.54 suggested a moderate predictive capability whereas R² value of Neural Network model indicates that it explained 61% of the variance in NO₂ concentration. These results highlight the importance of accounting for the non-linear interactions and feature complexities when modeling NO₂ concentrations, with Neural Network demonstrating their strength in capturing such patterns. In the current analysis, Neural network model did not enhance the predictive capability that much and the reason could be due to highly non-linear relationships between the features which reduced the advantage of applying neural network model. Moreover, the number of observation ~370 might not be enough data to effectively learn complex patterns and even with L1 penalties, overfitting might be an issue. However, further improvements could be made in optimizing the predictive model specially in terms of feature selections and explore other advanced ensemble methods to enhance predictive accuracy and generalizability.

Sl. No	Technique of the Predictive Model	MSE/Precision	RMSE	R ²
1	Linear Regression	18.49	4.30	0.41
2	Decision Tree	0.08	–	–
3	Random Forest Regression	12.82	3.58	0.54
4	Neural Network	10.61	3.12	0.61

References

- [1] M.J. Bechle, D.B. Millet, and 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, and 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, and 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, and 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, and 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, and 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.
- [7] D. Chicco, M.J. Warrens, and G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, *PeerJ computer science.* 7 (2021) e623. doi:10.7717/peerj-cs.623
- [8] U. Agbulut, A.E. Gurel, Y. Bicen, Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison, *Renewable and Sustainable Energy Reviews.* 135 (2021) 110114. https://doi.org/10.1016/j.rser.2020.110114