Using Infrastructure Data to Predict Pollution Levels

Jaldhir Trivedi, Vincent Liu, Vimalesh Vasu

For 24-787: Machine Learning & AI.,
Carnegie Mellon University

Abstract

An attempt to build models that predict pollution levels of a US county based on infrastructure features has been done in this study. US EPA maintains the database of pollution levels of various pollutants out of which we have used 4 pollution data: PM2.5, SO_2, Ozone, NO_2. 15 infrastructural features in each county have been used to train various Supervised Learning algorithms. These features include Coal Plants, Airports, Public Schools, Hospitals, Fuel Stations, Land Area, Population etc. It was noted that the MSE errors for the regression models was very high and are not suitable for real life application. Furthermore, pollution which was originally in numerical form was converted to categorical data and classification model was carried out. This model gave very good results for PM2.5 (75%) accuracy, while returning <45% accuracy for SO_2, Ozone, NO_2. It is thus noted that the infrastructure data used by us is inadequate to predict levels of PM2.5, SO_2, Ozone, NO_2 within comfortable margin of error. Classification model may better serve our purpose but only for PM2.5, the accuracy for other pollutants is subpar. In conclusion, infrastructure data is not a good direct indicator of pollution but can be used to indirectly predict sources of pollution. We further discuss the limitations and possible future improvements to the model.

1. Introduction:

The first step towards solving our climate crisis is to identify how pervasive the problem really is around the world. To date, researchers have made exhaustive efforts towards documenting pollution levels across the world to better understand its causes. For example, the Environmental Protection Agency (EPA) has set up a system that tracks hundreds of types of pollution for each county on a continuous basis, which proves highly beneficial for climate-related research. However, many regions across the world do not have access such resources for tracking pollution, making it difficult for studying pollution levels and their sources. Such regions generally include low-income or remote areas.

According to EPA research, 39% of global energy-related carbon emissions are attributed to buildings, from heating, cooling, and lighting. Given such a large contribution towards pollution from buildings, this begs the question on if there is a correlation between infrastructure data and pollution levels. That is, is there a strong enough relationship between different infrastructure types, and their quantities, and pollution levels in a region for infrastructure to serve as a reliable indicator for pollution levels? Such a relationship would be advantageous because infrastructure data is well documented around the world, especially in areas where pollution data might not be available.

If such a relationship exists, this research aims to develop a model that uses existing pollution data mapped with infrastructure data in corresponding regions to predict to a certain degree the pollution levels in a region without the use of pollution data.

It should be noted that we do not aim to find a timeless relationship between infrastructure and pollution, as pollution levels change from year to year, but infrastructure data remain relatively the same over periods of many years. Rather, the key assumption that we are making in this data analysis is that for a given year, the correlation between existing infrastructure quantities and the current pollution levels is the same for all regions across the world. Thus, this model would be retrained on a yearly basis with updated EPA data. This assumption is disputed as a source of error in the conclusion (see conclusion).

Given the large set of data points to train our model with, and the potentially relatively endless number of features (infrastructure type) that could be integrated, such a model could prove to be useful for not only finding a general correlation, but also for attempting to explain specific contributors of high pollution levels. Namely, it could help indirectly track correlations not tracked by researchers (due to low funding, inaccessibility, etc.).

2. Related Work:

Although a direct, quantifiable relationship between infrastructure and pollution has not been published, the general effects of infrastructure has been well researched and documented in academia. This data and research is all published on the EPA website [1] for public viewing.

The EPA has determined that in general pollution is caused by four major sources: mobile sources, stationary sources, area sources, and natural sources. The mobile sources originate from the more than a billion active mobile vehicles that emit pollution from the burning of fossil fuels. The stationary sources originate from industrial sources like factories, and powerplants. Area and natural sources originate from a multitude of small pollutants that amalgamate into large sources in the large scale [2].

The research most relevant to this project are those that have indirect relationships with infrastructure. For example, a major cause of pollution from infrastructure is from the emissions of heating, cooling, and lighting. Of the 39% of global pollution that buildings contribute to, 28% is due to heating, cooling, and lighting, and 11% is due to construction. Specifically, cooling systems like air conditioners release hydrofluorocarbons, which are 1000 times more potent than carbon dioxide, acting as a major source of global pollution [3].

The other major source of infrastructure pollution is from heating used for warmth, and more importantly, for cooking. According to EPA research, 58% of black carbon emissions in the world originate from cooking in residential areas. These cooking emissions originate from stove-top or fire cooking, which evidently is still a predominant method of cooking, despite the growth of electric-top cooking [4].

3. Data:

To train the model, we used the well-documented EPA databases for pollution levels in the United States and various infrastructure databases for corresponding infrastructure quantities. The regions were separated by FIPS county code, giving a dataset size of over 3,000 points to train and test our model.

3.1 EPA Data

EPA maintains database for a large number of counties and cities across the state in order to determine if the state/county complies with NAAQs [6]. Besides pollutants stated in the NAAQs, it also tabulates measurements of hundreds of Hazardous Air Pollutants (HAPs). The database that we use is an aggregate data based on certain criteria. For e.g., PM2.5 24-hour 2006 showcases the data at a certain location aggregated every 24 hours. Each county is identified with a unique FIPS code. Each county has many observation sites. EPA tabulates the arithmetic mean & various percentiles (99th, 90th, 75th, 50th, 10th Percentile) of the data.

3.2 Infrastructure Data

The infrastructure data was collected from a multitude of databases, due to the fact that there is no free centralized database of all infrastructure (Google geocoding services could have been used as a centralized source, if not for its large monetary price per query). A few data sources utilized were Kaggle, the US Department of Transportation, Data.gov, and other minor data repositories found around the internet.

The main challenge in collecting infrastructure data was mapping any data found into a common location unit, as the datasets represented data in many forms (longitude/latitude, address, zip code, etc.). To resolve this issue, the final accumulated dataset was decided to be organized by FIPS county code (Federal Information Processing Standards), as this seemed to be the easiest to convert to from the different location formats. Due to the tedium of this process, the data collection task proved to be the most time-consuming aspect of the project.

3.2.1 Longitude and Latitude to FIPS

For infrastructure locations given in longitude and latitude, we used a script written in R that inputted longitude and latitude and outputted the zip code of the location (initially, geocoding services like the Census Geocoding Services were attempted for this conversion, but they proved too inaccessible for our formatting needs). This list of zip codes was then run through a masking script that iterated through a two-column file that mapped zip codes to their corresponding FIPS code, to output a list of FIPS codes in the order following the original data file.

3.2.2 Address to FIPS

For infrastructure locations given in address form, the zip code was extracted either from column extraction or using comma delimiters. This list of zip codes was then run through a masking script that iterated through a two-column file that mapped zip codes to their corresponding FIPS code, to output a list of FIPS codes in the order following the original data file.

3.2.3 Final Feature Set

In the end, 15 infrastructure datasets were accumulated, and used as independent features for our model.

4. Methods

Labels. For Pollution Labels we collected 98th Percentile observation per observation unit. Each county has multiple of such units, the data was further aggregated by utilizing highest observation within each county. This was done for PM2.5 24hr 2012, SO2 24-hour 1971, Ozone 8-hour 2015, NO2 1-hour. Here, observations in each unit are aggregated based on aggregation time i.e., 24 hours for PM2.5, 1 hour for NO2 and so on. Then we added values to 15 features for all FIPS as available. We then divided the database into four individual databases, one for each pollutant. The null value rows were discarded, and the rest of data was ready for regression.

PCA. Initially, we thought of applying Principal Component Analysis (PCA) to our data since we believed many features were correlated, but found that it was unnecessary as we wanted to obtain information about every infrastructure feature. Our goal was to obtain the highest correlations and obtain useful information on how much each infrastructure feature factors into pollution. We aimed to add as much infrastructure data to achieve the highest accuracy.

Preprocessing. For preprocessing we used sklearn's StandardScaler() module which normalizes data from their raw measure into the following score:

$$z = \frac{x - \mu}{\sigma}$$

Where, z is the normalized score, x is the raw score, μ & σ are mean and standard deviation of samples. The samples were then split into test-train scores using train_test_split() module. The train/test sample ratio used was 70/30.

Univariate Visualization. For assessing individual feature correlation with pollution levels, we used a scatter plot and LinearRegression() module to plot the data and visualize the regression. We found that the individual correlation of features with pollution levels was very weak and the univariate distribution had a high kurtosis. Few of these plots are shown in Figure 1.

Feature Engineering. Instead of using raw feature value we tried to use area normalized (Feature/area) as our engineered feature but we noticed that the models are less accurate for such features using linearRegression() modules. Thus, we decided to use raw features to pass through StandardScaler().

Regression. Given that our labels were numerical values, we applied regression models. Initially, we started with applying linear and ridge regression using sklearn's LinearRegression() & Ridge(alpha=1) modules respectively. Furthermore, we also applied support vector regressor (with a rbf kernel), and random forest regressor using sklearn's SVR(C=1.0, epsilon=0.1), RandomForestRegressor() modules. Finally using Keras library we built Multi-Layer Perceptron (MLP) regressors. Hyperparameter tuning for MLP involved changing learning rate, number of hidden layers and neurons, and number of epochs. We used X_train & Y_train to fit the models. As our model metric, we initially used R2 scores to assess strength of the correlation. For the final results we used MSE using mean_squared_error() module. MSE as a percentage of average scores was our eventual metric for regression. For k-fold cross validation (k=5) we used train_test_split() inside a for loop to get the average MSE scores for every 5 folds.

Classification. Due to poor regression results, we decided to classify our labels and run classification models. We classified PM2.5 data into EPA designated Classification [5]:

PM2.5 (g/m3)	AQI Category	Classification
0-15.4	Good	1
15.5-40.4	Moderate	2
40.5-65.4	USG	3
65.5-150.4	Unhealthy	4

For the next 3 pollutants we designated ranges based on quantiles i.e.,1st quantile range: 75the percentile value to 100 percentile value and so on. We assumed that having a range of values for our features would allow for greater flexibility for our models to obtain reasonable predictions. After preprocessing our data, we used classifier modules such as LogisticRegression(), RandomForestClassifier(), svm.SVC(kernel='linear'), svm.SVC(kernel='rbf') and Keras built multi-layer perceptron (MLP) classifiers. Hyperparameter tuning for MLP involved changing learning rate, number of hidden layers and neurons, and number of epochs. We used X_train & Y_train to fit the models. For the Model metric we used sk learn's inbuilt score() function which returns mean accuracy of the classification model on our test samples. For k-fold cross validation (k=5) we used train_test_split() inside a for loop to get the average MSE scores for every 5 folds.

5. Results

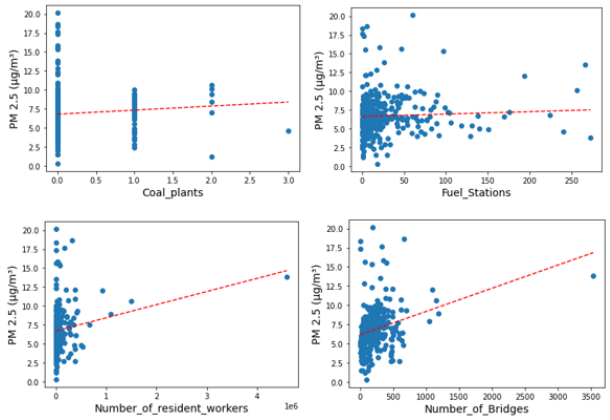

Figure 1: Univariate Visualization PM2.5 levels Vs Features

Univariate Analysis raised concerns that features that were collected may not have enough correlation for the model to work. This was surprising as features like coal plants and population did not show any correlation and univariate regression boundary had a very low gradient compared to the variance in pollution levels. This would precipitate in Model MSE being too large as compared to the average value pollution levels (as can be seen in Figure 1). We shall discuss more about the reasons for this later in this section. Univariate regression was prone to overfitting giving negative R2 scores for some test samples. This was mitigated in multi-variate regression and its results are discussed below. As the pollution levels are considered for across the spread of a county, it was logical to normalize the features with county area (county population density instead of county population). When we ran an initial test model for both area-normalized and unnormalized features with PM2.5 as label, the R2 score observed for the test samples was 0.13 & 0.3. This showed that raw feature would work better for the models rather than area-normalized feature. We also ran linear regression for raw features Vs normalized feature (through StandardScalar() module) and observed that normalized feature gave consistent result (a R2 score of 0.24 Vs 0.22 with LinearRegression() module). We also observed that the model run better with 98th Percentile Pollution data as compared to Arithmetic Mean which is why we used that for our regression and classification.

Regression. Table 1 tabulates the test mean squared error results after applying regression models on our data. Regression models have performed well for ozone but not for PM2.5, SO2, and NO2. Support vector regression performed the best for PM2.5, SO2, and NO2, while ridge regression performed the best for ozone. The final neural network architecture after hyperparameter tuning was a **4x50** layer network. We expected neural networks to perform the best, but even after a series of hyperparameter tuning our neural network model still performed less accurately than other models. One reason for this may be extreme outliers in our test set. For example, the Los Angeles County has a population of 10 million while the average county population is less than a million. Thus, neural networks may train on extreme outliers and skew our model to perform less accurately. Another reason may be due to data, in which the data that we obtained may only be just a small piece of the entire picture and that extensive data collection would be necessary to obtain the best results.

Table 1: Results for regression models

Pollutant (Average)	**PM2.5 (16.45)**	**SO2 (0.624)**	**Ozone (0.054)**	**NO2 (34.59)**
Linear Regression	7.97 (48%)	0.514 (82%)	0.00005 (0.1%)	93.08 (269%)
Ridge Regression	7.71 (47%)	0.472 (76%)	**0.000045 (0.1%)**	70.83 (205%)
SVR (rbf kernel)	**7.30 (44%)**	**0.437 (70%)**	0.000063 (0.1%)	**64.14 (185%)**
Multi-Layer Perceptron	11.82 (72%)	0.495 (79%)	0.000047 (0.1%)	65.97 (191%)
Random Forest	12.55 (76%)	0.538 (86%)	0.00005 (0.1%)	67.24 (194%)

Classification. Due to our regression not achieving expected results, we decided to pursue an alternative route in which we classified our pollutant values into EPA-defined brackets. We assumed having a range of values rather than a single number would allow more flexibility for our model to train and obtain better results. Below were our results in terms of test accuracy:

Table 2: Results for classification models

Pollutant (Average)	**PM2.5 (16.45)**	**SO2 (0.624)**	**Ozone (0.054)**	**NO2 (34.59)**
Logistic Regression	0.73	0.29	0.29	0.4

SVC (linear)	0.73	0.29	0.26	0.41
SVC (rbf Kernel)	**0.75**	0.27	0.23	0.38
Multi-Layer Perceptron	0.70	**0.34**	**0.30**	0.37
Random Forest	0.73	0.25	0.25	**0.42**

In table 2, classification results were good for PM2.5 but not good for SO2, Ozone, and NO2. Neural networks gave the best results for SO2 and Ozone while support vector classification performed the best for PM2.5 and random forest for NO2. Based on the results above, classifying our labels into ranges did not give high accuracy. Through these results, it became indicative that infrastructure data is not a good direct predictor of pollution. Rather, infrastructure data is better for indirectly predicting the big 3 pollution sources: automobiles, power/manufacturing plants, and buildings that produce heat. Another factor we must consider is that each county has vastly diverse land area, population, infrastructure, etc. for each region and thus not uniform for every county. There may be many other socio-economic factors that will influence pollution that we cannot ignore.

Random Forest. Interestingly, random forest gave a good look into how decision trees can be used to predict pollution. Below is the importance chart of the features that we used for our regression models:

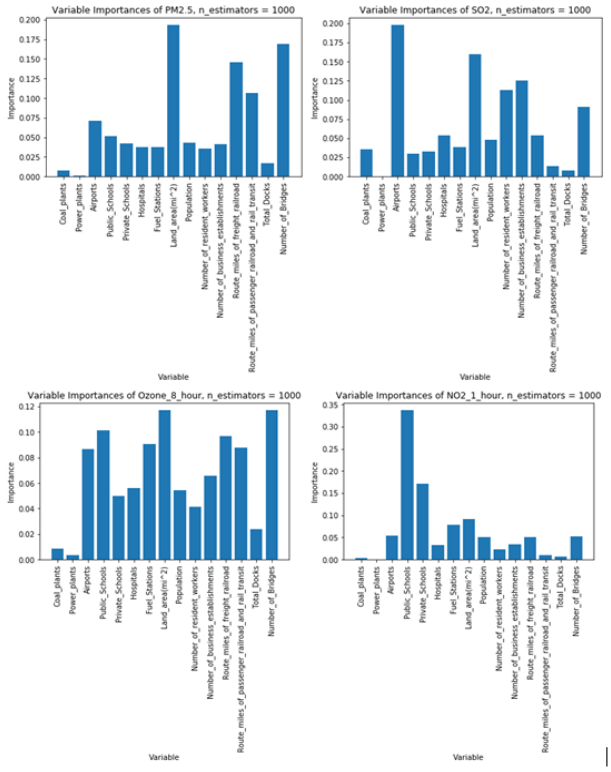

Figure 2: Importance charts from Random Forest

Above for PM2.5, we can see that land area and number of bridges had higher importance. One explanation for this is that having a larger land area would correlate with proportionally more population and infrastructure which results in higher pollution. SO2 on the other hand had the feature airports with the highest importance. This makes sense given that airplanes are one of the largest producers of pollution. For Ozone, we find that similarly land area had the largest importance which as explained earlier may be due to more infrastructure. Lastly, NO2 has highest important from schools which might be explained that more schools might be indicative of more automobiles in that county. There are many explanations to how each feature affects pollution but no clear correlation between each feature and pollutant.

Limitations. We did not find a good correlation between a county's infrastructure and its pollution levels due to the following reasons:

1. Features like coal plants, power plants, airports etc only accounts for the number of such plants or ports located in a county and do not say anything about the relative activity of those locations. For example, JFK airport, NY would have equal weight as GSP airport, SC, in our model, even though the former may be a substantially larger source of pollutants due to much higher levels of activity.

2. Our model works on the assumption that the pollution is a result of features within the county which is not entirely true. For e.g., a huge coal plant affects pollution levels of many counties around it.
3. Our model does not consider the past pollution activity. Features like power plant with huge pollution emission leave a lasting impact on the environment. Counties in and around it may experience high level of pollution years after the plant has closed. A county's history of pollution may prove to be of far-reaching consequence factor than existence of any feature at the instance.
4. We also have disregarded mitigating factors from our features. Factors like forest cover and local emission norms act as a counter measure to the pollution source thus driving the pollution down. Two different counties with similar pollution source may have different pollution levels if the mitigation of pollution is different for each of them.
5. Our motivation was to build a model which can accurately predict pollution levels based on a set of infrastructure features. This has limited many pertinent features to be used in our model which would otherwise prove to be give better ML models.

6. Conclusions:

From our results, we can conclude that infrastructure data is not a good direct indicator of pollution. Infrastructure trends are too random/complex to reliably be used as an indicator for pollution. Rather, infrastructure can help indirectly predict major sources of pollution such as automobiles, power/manufacturing plants, and inefficient buildings. It is however noteworthy that the classification model gave far better results for PM2.5 with accuracy of 75%, while returning <45% accuracy for SO_2, Ozone, NO_2. MSE error for all pollutants except Ozone are large for our models to be considered worthwhile. Nonetheless, we do assess from our study that pollution very scarcely depends on any indicators underlying our infrastructural features. Even though pollution is said to be caused by coal plants and vehicular activities and industries, it far more depends on other attributes which are not covered in our studies.

6.1 Limitations in Approach

6.1.1 Broadness in FIPS Counties as Data Points

A major limitation that was recognized in the data analyses is that FIPS counties are too large of a region to set as a data point. We learned that within a FIPS county, unaccounted for factors can cause major discrepancies in the pollution levels of counties that may have the same quantities of infrastructures. For example, a county that contains a city with a more densely populated population may have lower pollution levels than another county with similar population levels, but more spread out. This is due to the fact that city-living is a more efficient manner of living due to norms like living space sharing, and increased ride-sharing. Factors like this are not accounted for in our model, causing results to be inaccurate.

Unfortunately, however, there is no workaround for this issue because the smallest region of EPA pollution data collection is FIPS county code.

6.1.2 Economic Effects of Pollution

Another unobvious determinant of pollution level is the GDP of a region. GDP determines factors of energy efficiency and pollution sources that our model can't account for. For example, in the case of pollution caused by stove-top or fire-based cooking, it can be observed that countries with lower economic status emit larger proportions of black carbon. This is due to the fact that these lower economic regions have less access to efficient electric heating and cooking, and thus are more reliant on gas/fire-based heat sources, which emit black carbons. Therefore, given our model is accurate for predicting pollution in the US, it may predict inaccurate pollution levels for the same infrastructure in countries with varied GDP.

6.2 Future Extensions: Using Model to Indirectly Predict Major Sources of Pollution

From EPA research, we know that one definite predictor of pollution levels in a region are mobile sources. Thus, the better we can predict how many cars drive through a region on average, the better we can predict the pollution levels for that region to a significant extent. Although our data can't directly predict pollution levels, it can serve as a means for intuitively realizing high mobile sources. For example, from our Random Forest model, we found that two of the most significant features in our data were the quantities of public and private schools and the number of fuel stations. In the case of schools, it can be inferred that a county with a larger number of schools would have a larger active population, and thus a larger number of cars driving to and from the schools. In the case of the quantities of fuel stations, it can be inferred that a region has a higher number of fuel stations due to the increased traffic in that region. Based on the different quantities of fuel stations, a pseudo-scale can be generated for correlating the quantity of fuel stations to the density of cars in a region. These

indirect correlations can then be used to produce better models of pollution levels in a region.

Contributions:

Our group would like to thank Akanksh Shetty for providing guidance and support for our project. Each member of the group split up the work into three sections. Vimalesh worked on data collection, researched articles, and obtained background information on infrastructure. Jaldhir worked on coding up the linear regression, support vector models, and researching other algorithms. Vincent worked on data collection and coding the neural network and random forest models.

References:

[1] https://www.epa.gov/
[2] "Where Does Air Pollution Come From?" , January 2018 [Online]. Available:

https://www.nps.gov/subjects/air/sources.htm#:~:text=There%20are%20four%20main%20types,cities%2C%20and% 20wood%20burning%20fireplaces

[3] Kristina Costa "Reducing Carbon Emission Through Infrastructure" September 2019. Available:
https://www.americanprogress.org/issues/green/reports/2019/09/03/473980/reducing-carbon-pollution-infrastructure/#:~:text=Cutting%20transportation%20emissions%20through%20infrastructure,emissions%20in%20the%20United%20States.
[4] "Household Energy and Clean Cookstove Research". Available:

https://www.nps.gov/subjects/air/sources.htm#:~:text=There%20are%20four%20main%20types,cities%2C%20and%20wood%20burning%20fireplaces

[5] San Salvador, El Salvador, "Air Quality Index (AQI)," April 2012. [Online]. Available: https://www.epa.gov/sites/production/files/2014-05/documents/zell-aqi.pdf.
[6] EPA.gov, "Air Quality Statistics by County, 2019 (XLSX)," 8 June 2020. [Online]. Available: https://www.epa.gov/sites/production/files/2020-06/ctyfactbook2019.xlsx.

Supplementary Material

Random Forest hyper parameter training

n_estimators	Mean Absolute Error	Mean Squared Error
250	1.4127	3.4122
500	1.4158	3.43
1000	1.4181	3.4394
1500	1.4251	3.4597
2000	1.428	3.4752

Figure A1: Random Forest hyperparameter tuning, PM2.5

n_estimators	Mean Absolute Error	Mean Squared Error
250	0.379	0.1968
500	0.3687	0.1929
1000	0.3684	0.195
1500	0.3676	0.1972

Figure A2: Random Forest hyperparameter tuning, SO2

n_estimators	Mean Absolute Error	Mean Squared Error
250	0.0024	0.0001
500	0.0024	0.0001
1000	0.0024	0.0001
1500	0.0024	0.0001

Figure A3: Random Forest hyperparameter tuning, Ozone

n_estimators	Mean Absolute Error	Mean Squared Error
250	3.9003	22.2433
500	4.0033	23.2859
1000	4.0496	24.2085
1500	4.0635	24.2801

Figure A4: Random Forest hyperparameter tuning, NO2

Multi-layer Perceptron hyperparameter training

Hidden Layers	Mean Absolute Error	Mean Squared Error
(50, 50)	2.9914	12.2298
(50, 50, 50)	3.0021	12.7477
(50, 100, 50)	2.8933	15.1212
(50, 50, 50, 50)	3.1163	11.8240
(50, 100, 100, 50)	2.9166	12.5476

Figure A5: MLP hyperparameter tuning, PM2.5, epochs = 1000, Adam optimizer, loss = MSE, learning rate = 0.001

Hidden Layers	Mean Absolute Error	Mean Squared Error
(50, 50)	0.7444	0.9968
(50, 50, 50)	0.5655	0.5623
(50, 100, 50)	0.5496	0.4954
(50, 50, 50, 50)	0.5698	0.53802

Figure A6: MLP hyperparameter tuning, SO2, epochs = 1000, Adam optimizer, loss = MSE, learning rate = 0.001

Hidden Layers	Mean Absolute Error	Mean Squared Error
(50, 50)	0.0103	0.00024
(50, 50, 50)	0.0122	0.00036
(50, 100, 50)	0.0132	0.00074
(50, 50, 50, 50)	0.0049	0.00005

Figure A7: MLP hyperparameter tuning, Ozone, epochs = 1000, Adam optimizer, loss = MSE, learning rate = 0.001

Epochs	Hidden Layers	Mean Absolute Error	Mean Squared Error
1000	(50, 50)	13.8197	335.6640
1000	(50, 50, 50)	11.9656	99.8805
5000	(50, 50, 50)	10.091	65.9715
10000	(50, 50, 50)	9.8704	72.06467
1000	(50, 100, 50)	9.3301	73.3769
1000	(50, 50, 50, 50)	9.4213	80.07632
1000	(50, 100, 100, 50)	9.7859	67.24489

Figure A8: MLP hyperparameter tuning, NO2, Adam optimizer, loss = MSE, learning rate = 0.001