CIV 551 Sensing and Machine Learning for Smart Cities Final Project

Analytical Data Methods for Predictions on PM 2.5

Mahsa Bargahi
Civil Engineering
Stony Brook University
Stony Brook, NY, USA
Mahsa.Bargahi@stonybrook.edu

Hojjat Barati
Civil Engineering
Stony Brook University
Stony Brook, NY, USA
Hojjat.Barati@stonybrook.edu

Daniel Ruiz
Civil Engineering
Stony Brook University
Stony Brook, NY, USA
Daniel.Ruiz@stonybrook.edu

ABSTRACT

PM 2.5 is an inhalable particulate matter which is used as a parameter for air pollution measurement. Therefore, studying PM 2.5 is important to prevent potential health and environmental concerns. The project consisted of developing models for short-term, long-term and interpolation data prediction of PM 2.5. Preliminary methods for solving each prediction task along with the analysis is presented. Moreover, different signal processing methods including moving-average and SVD is conducted for data signal processing. Subsequently, statistcal analysis for feature selection process is performed where different features were selected according to the task. Similarly, a different model was developed for each task. The three selected models are the following : cross-validation for short-term predictions, neural network for long-term predictions, and GPR for the interpolation task. The overall performance metric, NRMSE, for all tasks was 1.0528 including noise and 0.3688 without noise. The three selected models performed best with with zero noise level where each model yielded a NRMSE below 0.5. The highest NRMSE yielded was 2.8463 for the short-term predictions.

CSS CONCEPTS

• Evaluation Methods and Computing Design • Machine Learning • Human-Centered Computing

KEYWORDS

PM 2.5, Air Pollution Prediction, Forecasting, Analytical Data Methods, Regression Models

ACM Reference format:

Mahsa Bargahi, Hojjat Barati and Daniel Ruiz. 2021. CIV 551 Sensing and Machine Learning for Smart Cities Final Project: Analytical Data Methods for Predictions on PM 2.5. *Stony Brook University: Department of Civil Engineering. Stony Brook, NY, USA, 10 pages.*

1 Introduction

1.1 Background

One of the most significant environmental problems throughout the world is air pollution which threatens public health. There are numerous indices to investigate air pollution and PM 2.5 is one of the most frequently used parameters. PM 2.5 are inhalable particulate matter (PM). PM 2.5 particles have diameters that are predominantly 2.5 micrometers and smaller than human hair. To illustrate the perspective of the size of PM 2.5, the following figure demonstrates the size of PM 2.5.

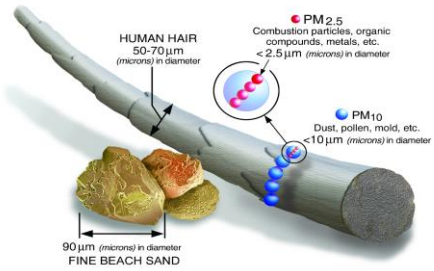

Figure 1: PM 2.5 Size Comparison[1]

Most PM 2.5 particles form in the atmosphere because of complex reactions of chemicals such as sulfur dioxide and nitrogen oxides, which are pollutants emitted from power plants, industries, and automobiles. A high indicator of PM 2.5 can impact human health and impact the environment. PM 2.5 has been researched and associated with premature death in people with heart or lung disease, heart attacks, and increased respiratory symptoms. Environmental impacts can consist of increased acidity and nutrient balance of oceans, lead to a decrease in soil nutrients, and impact on ecosystems.

1.2 Project Objectives

After discussing about the importance of PM 2.5, studying PM 2.5 data is important to prevent and alleviate the potential health and environmental effects. The goal of the project is to determine data analysis methods to provide measurements and predictions of PM 2.5. The development of the models will meet the objectives of the project. The objectives of the project can be divided into three goals. The first objective consists of short-term predictions. The goal will be to forecast hourly PM 2.5 over a period of three hours based on three days of data. The second objective consists of long-term predictions. The method developed will determine hourly PM 2.5 data for 24 hours at a specified location based on one-week data. The last objective of the project consists of interpolating missing PM 2.5 measurements from data of three days. The interpolation method will be able to determine 5-minute PM 2.5 data between 11:30 AM and 12:25 PM on the second day

of the data set. Lastly, the selected methods will each be tested for robustness against added zero-mean Gaussian noise with a standard deviation of 0.5σ and 1.0σ.

2 Preliminary Methods

2.1 Short-Term Predictions

Selection of the best representative model for large dataset seems to be a challenging task. To overcome this, miscellaneous models were optimized to maximize the accuracy of predictions for PM 2.5 values for short period of time. The set of preliminary methods used for short-term prediction includes: Linear regression, Polynomial regression, Ridge regression, Lasso regression, Neural Networks, and k-fold polynomial regression.

2.1.1 Linear Regression. As a very basic method of prediction, great results from linear regression were not anticipated. However, application of this method helped understand the importance of parameters in accuracy of prediction. To reach this merit, several combinations of parameters have been used and the accuracy of predictions have been examined. For example, in one model speed and humidity have been opted as the basic parameters, and in the other speed, temperature, and humidity have been used in the model. The final regression model based on linear regression is as following:

Eq. 1: Linear model short-term

$$PM2.5=121.93+0.291(Hmd)+21.33(Spd)+3.40(Temp)$$

$$R\text{-square}= 0.202$$

As we can see in the above equation, the precision of this model is not promising as demonstrated by the R-square value.

2.1.2 Polynomial Regression. The prediction task can be defined as multi-dimensional estimation. As it has been proved, using polynomial regression by itself for multi-dimensional regression would lead to multicollinearity problems. This deficiency is rooted in interdependency of parameters. Therefore, polynomial regression as it was expected, did not represent precise results.

2.1.3 Ridge and Lasso Regression. The most important feature of Ridge and Lasso regression is the coefficient of regression, which is mathematically defined as BETA but in Python libraries it has been indicated as Alfa. We used different values for Alfa (e.g., 0.01, 0.001, 0.0001). The best predictions were resulted by using Alfa = 0.001 for Ridge model and Alfa = 0.01 for Lasso model. The values of RMSE for Ridge and Lasso regression are 17.41 and 17.55 respectively. The mean value of the actual test dataset is 48.61.

Neural Networks and k-fold polynomial regression demonstrated better results. Therefore, these methods were selected to proceed in the final round to select the model for short-term predictions.

2.2 Long-Term Predictions

Polynomial fitting was considered during the preliminary phase to better understand the best methodologies. In the plot below, 3-degree polynomial fit, 3-degree polynomial with Ridge and 3-degree polynomial with Lasso are illustrated. As it can be seen, none of these models can capture the pattern of the data and the models do not yield accurate fits. Different degrees of polynomial fit were also tested (e.g., 5,7 and 20); however, these higher-order polynomial regression models had similar results as the 3-degree polynomial. Ridge Regression was considered as it regularizes the coefficients and reduces the complexity. As it can be observed, the model obtained from this approach is simple for the PM 2.5 data.

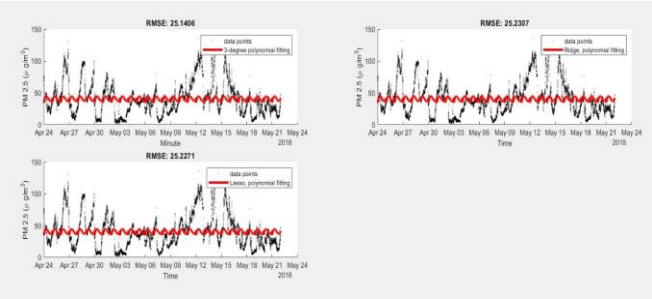

Figure 2: polynomial, ridge, lasso fitting

2.3 Interpolation Predictions

Three models were considered during the preliminary methods phase to understand which model would be the best approach to solve the interpolation objective of the project. The following sections will discuss ARIMA, Ridge Regression, and Gaussian Process Regression (GPR) that were developed. GPR was selected for the interpolation problem.

2.3.1 ARIMA. ARIMA was used as a preliminary method to predict PM 2.5. ARIMA was selected as a potential model due to its advantages. Some of the advantages include the integration of lag and shift during time-series data to predict future patterns. Furthermore, an ARIMA model provides a regression model with a moving average. To begin, an ARIMA model with only a nonseasonal AR polynomial lag of 10 was developed to predict PM 2.5. The outcome of the model is illustrated in the figure below.

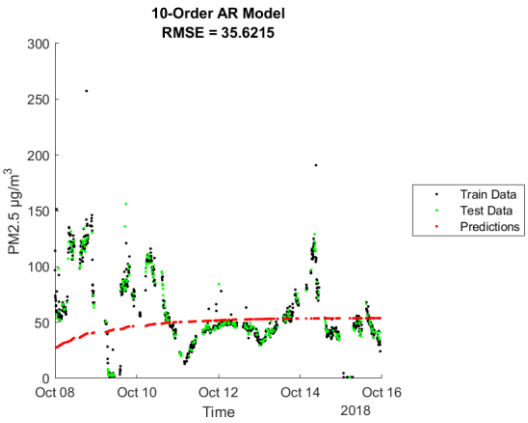

Figure 3: ARIMA (10,0,0) Model

To improve the model developed above, it was attempted to estimate PM 2.5 with a seven autoregressive polynomial degree with a moving average of 10th polynomial degree. The new model with the RMSE is indicated in the following plot.

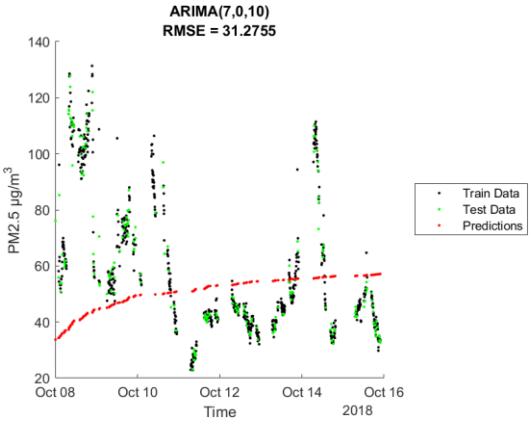

Figure 4: ARIMA (7,0,10) Model

It is observed from both ARIMA models that the method results in a significantly high RMSE. The models do not predict accurately PM 2.5. From both graphs, it can be observed that the predictions that are yielded are significantly different from the true (test) data even if the degree or MA of the model altered. Due to this limitation of accurately predicting the data, the ARIMA model was discarded as a possible method.

*2.3.**2** Ridge Regression.* The next model that was developed to predict PM 2.5 concentration was Ridge Regression. Ridge regression offers the ability for coefficient regularization and the reduced complexity of the model which can be important to prevent overfitting. The different parameters of Ridge regression are the linear regression polynomial degree and the regularization parameter, Lambda. Two different models were developed. The first model consisted of a degree of eight and lambda to be 0.1. The resulting model is illustrated below.

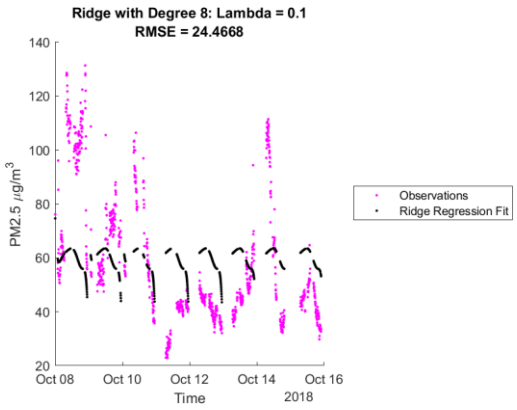

Figure 5: Ridge Regression Model (λ = 0.1)

From the plot, the Ridge regression fit is indicated in black. The original data is indicated in magenta. The Ridge regression fit does not perform well in fitting the data correctly. The Ridge regression performs better than the ARIMA model. However, the Ridge regression model still has a high RMSE value of 24.4668. The model was slightly improved by increasing the polynomial degree and decreasing the regularization parameter. This model is illustrated below in the following figure.

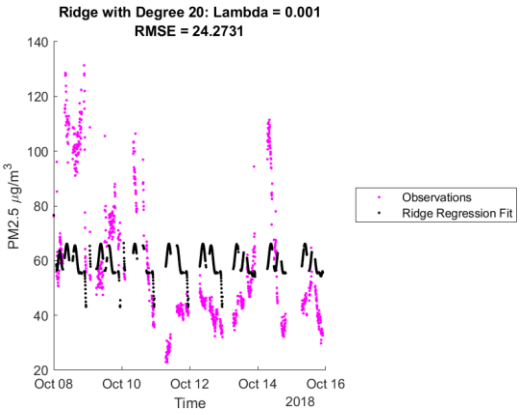

Figure 6: Ridge Regression Model (λ = 0.001)

The new model that was developed with Ridge regression performs slightly better compared to the previous Ridge model. However, the model will be discarded as well as a potential model for the project as it does not produce accurate predictions. It can be concluded that linear regression models yield models that are too simple and unable to fit the observations correctly.

*2.3.**3** Gaussian Process Regression.* The next model that was analyzed was a Gaussian Process Regression (GPR). The GPR model has several advantages which include the specification of the fitting method, prediction method, Kernel function used in the model. Two different Kernel functions were used to develop a GPR model for interpolation. The results of the two models are illustrated in the following figures:

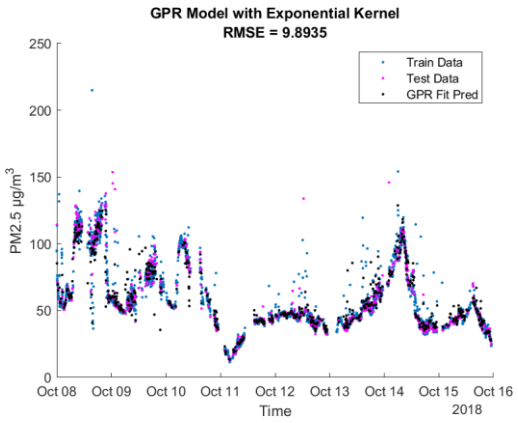

Figure 7: GPR with Exponential Kernel

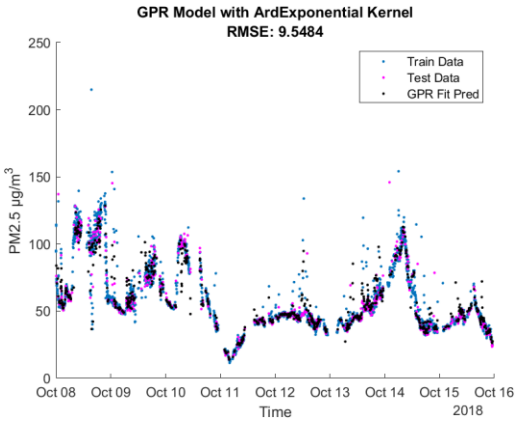

Figure 8: GPR with ArdExponential Kernel

The two plots illustrate the GPR model fitting against the testing data. The GPR predictions perform significantly better compared to the previous models that were analyzed such as ARIMA and Ridge Regression. It is observed that the RMSE values for both GPR models are significantly smaller. Although the RMSE values for GPR were relatively high, the model still performs well. From the RMSE values, the GPR model with an ArdExponential Kernel function performs slightly better than the GPR model with an Exponential Kernel. The fitting of GPR models can capture the trend of the observations which is important such that there is a balance between under and overfitting. Therefore, due to these results, a GPR model with a ArdExponential kernel function will be selected for the interpolation model.

3 Description of Measurement Data

Available data-frames for both short-term (Foshan) and long-term (Tianjin) prediction contain information every 3 seconds. The three types of problems are short-term prediction (1), long-term prediction (2), and interpolation (3). The output of the calibrated function will be three vectors (M-by-1) each for different problem types. To reach the desired vectors, a data-frame consisting of "data static" and "data mobile" cell arrays have been collected by static and mobile sensors, respectively. Furthermore, there are cells associated with each of the cell arrays. These cells constitute an N-by-7 Matrix. N is the number of samples and 7 indicates the different columns which are: time, humidity, temperature, vehicle speed, PM 2.5, latitude, and longitude. It is important to note that the same columns are manipulated for the train procedure. However, the matrix for train data will be P-by-7 which P<N. After developing the methods for short-term, long-term, and interpolation problems, the test procedure data-frame will be a M-by-6 matrix. This data-frame has been collected from a fixed location that is not the same as the location of sensors. The predictions results will be a “M x 1” matrix for problem type 1, 2, and 3 where “M” is 3, 24, and 12, respectively for the problem types.

4 Signal Processing

4.1 Short-Term Predictions

Signal Processing extracts meaningful information from signals. It does not create any new information but allows one to look at the same information in a different manner. The data collected from the sensors is a time-series data. A time domain signal is a record of what is happening to a system parameter as a function of time. In this context, a signal is a series of PM 2.5 values as function of time. Hence, visualization of the data over time basis will help us greatly to identify the trend of data. Therefore, the PM2.5 data has been plotted over time data. However, another method and technique can be utilized to understand the most repetitive values. By converting the time domain to frequency domain, the most important part of dataset can be analyzed in a different perspective. Fast Fourier Transform (FFT) is a technique which was employed to convert the time domain data to frequency domain data. The results of this conversion for one sensor by using FFT is demonstrated in the following figure:

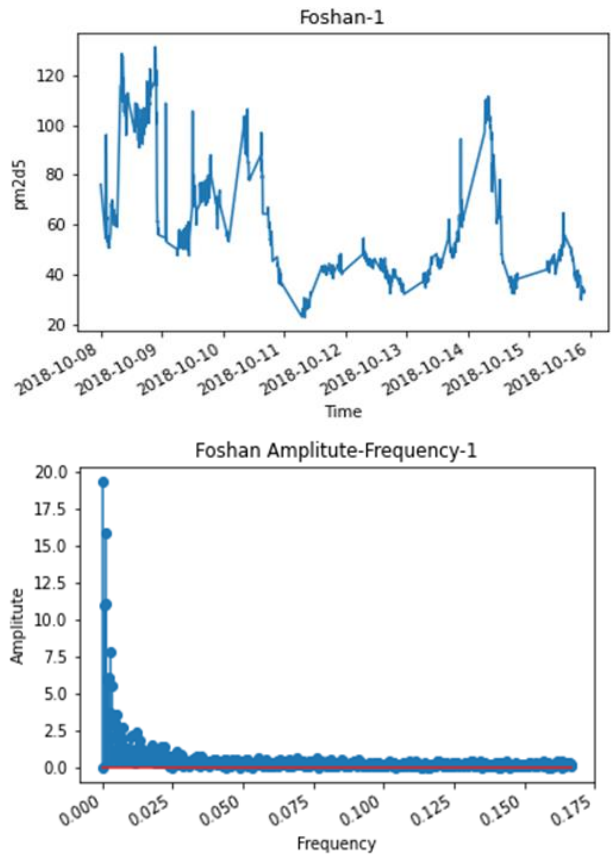

Figure 9: Time domain and frequency domain.

As illustrated in the figure, there is a high concentration of data closely to frequency of 0 and the amplitude varies between 0 to approximately 20. Therefore, this means that the repetition of PM 2.5 data has been occurred over an almost long period of time.
Noise detection and filtering of the data have been another issue during the analysis. Filters remove or modify some frequency components of a signal and pass others. Therefore, the resulted data after applying the filters would provide more smoothed data as an input which help to build more reliable models of prediction. The filtering process for short-term prediction has been applied by using moving average method. This method smooths the signal by taking average of a sliding window. Different values for sliding window can be selected. This method provides a smoother data by increasing the sliding window to a specific size. Moreover, one of the possible methods to find and remove noise is Singular Value Decomposition (SVD). As commonly known, SVD is widely used in many applications, such as the Principal Component Analysis (PCA). With application of SVD, the major patterns in the data were able to be discovered and understood how the patterns change over time. The results of the MA filtering and SVD are displayed in the following figures:

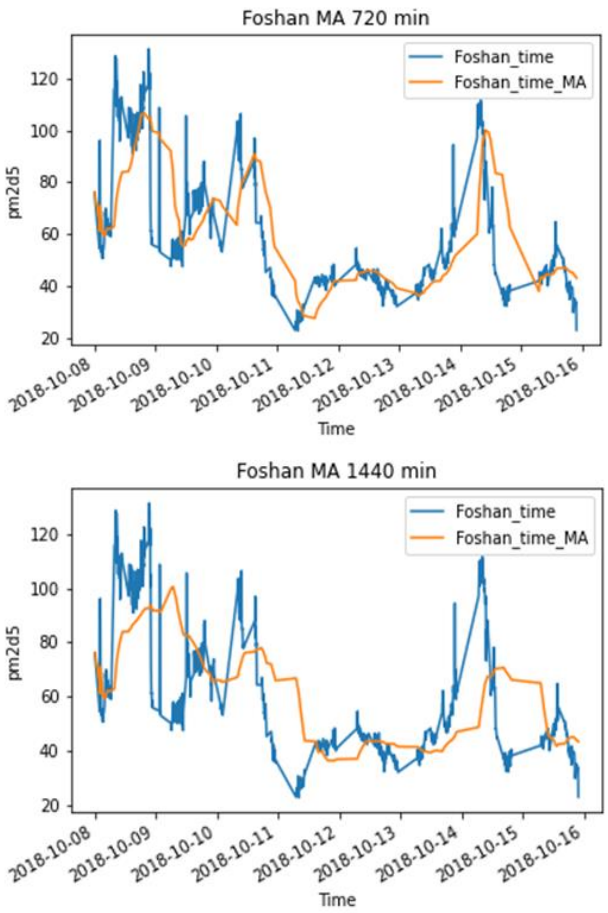

Figure 10: Applying moving average filter with different sliding window.

It can be found that using higher values for sliding window results in more smoothed data.

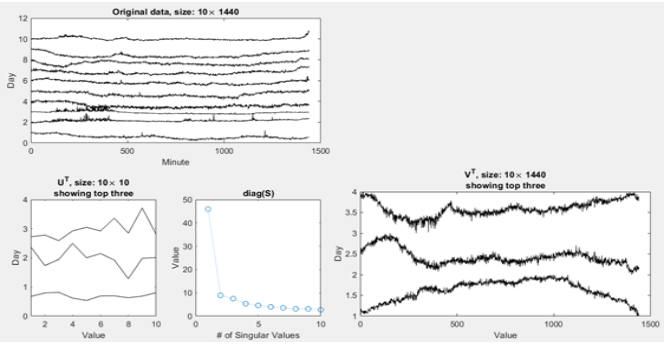

Figure 11: Singular Value Decomposition for short-term.

The above figure has been adopted by using the same code for the assignment of Smart Cities course.

4.2 Long-Term Predictions

Signal processing analysis was performed for the data set collected. The first figure is for the first static sensor and the second figure is for all sensors combined. In the second subplot the Fourier Transform of data from the plot of the time history in in the first subplot, is plotted. In the third and fifth subplot, data for PM 2.5 with 60-minute moving average and 12 hour moving average is shown. As the moving average window size increased, the temporal data becomes smoother and, in FFT plots, the magnitude of the high-frequency response is reduced in magnitude.

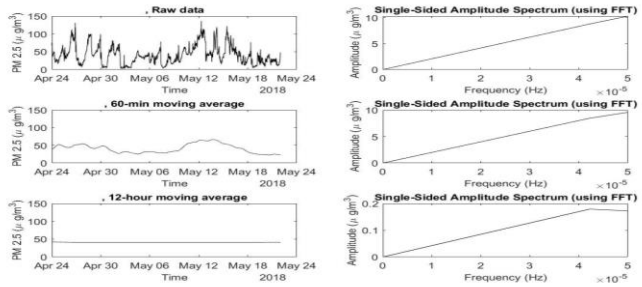

Figure 11: Signal processing and moving average for static sensor1

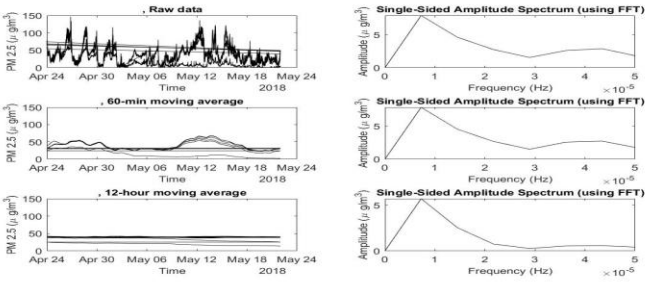

Figure 12: Signal processing and moving average for all sensors

4.3 Interpolation Predictions

Singular Value Decomposition was performed for signal processing on the Foshan dataset. SVD allows the ability for dimensionality reduction. With SVD, the possibility of balancing dimensionality and computing time was achieved. Having more features in the training dataset will impact the computational effort and efficiency of the model. SVD demonstrates the singular vector components that are important for the original signal. The significant singular value components are illustrated in the Explained Ratio figure below for the Foshan dataset.

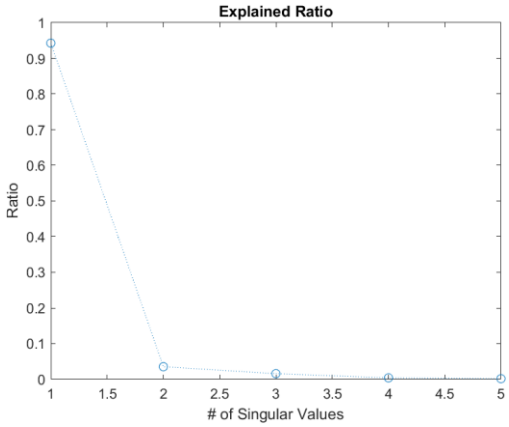

Figure 13: Explained Ratio for Foshan Data

The remaining unimportant singular vectors are interpreted as noise. Therefore**,** this implementation can clean the dataset. To construct an accurate estimation of the original signal, the number of principal components was increased until the explained ratio of each component reached the cumulative ratio of approximately 90%-95%. The following figure illustrates the original time series dataset normalized with the reconstructed normalized time series dataset.

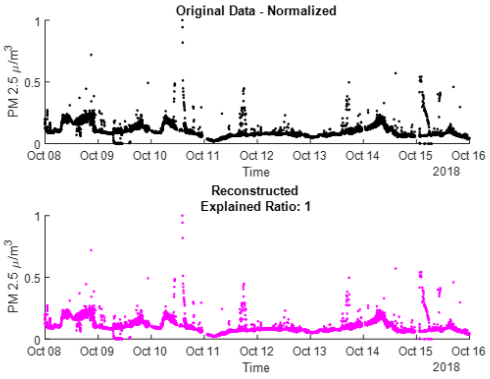

Figure 14: Reconstruction of Time-Series Data with SVD

5 Statistical Models

5.1 Short-Term Predictions

Feature selection is the corner stone of the creation of the models. Correlation matrix is a tool to apply to our dataset to select the most appropriate explanatory variables. The best representative parameters are those that have the least correlation with each other and most correlation with the PM2.5 parameter. Obtained results from linear regression and data visualization also assisted in this selection. The selected parameters are Temperature, Humidity, Speed, and PM2.5 as the dependent variable. Speed parameter did not show significant importance in model calibration by itself.

However, the simultaneous presence of Speed helped the improvement of the model. Thus, this parameter was also used for short-term prediction. The correlation matrix can be seen in the following figure.

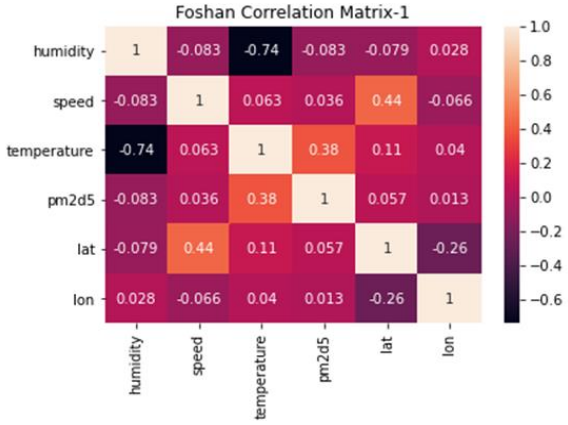

Figure 15: Correlation Matrix short-term

5.2 Long-Term Predictions

To select the features for long-term predictions, statistical modeling was performed similarly as for the short-term task. It is critical as the feature selection is one of the most important parts of the project. Two statistical models were used. The first model was Variable importance, and the second model was a correlation matrix. From these two methods, it is observed that latitude and longitude are highly correlated, and they cannot be both present in the model. After an initial analysis, latitude was excluded from the long-term prediction model. Subsequently, the remaining best features were selected which were humidity, temperature, and speed. There were some doubts about whether speed should be included as a factor. Therefore, a model with and without speed was developed. The results with the inclusion of speed were better. Thus, speed was selected as a feature for the final model.

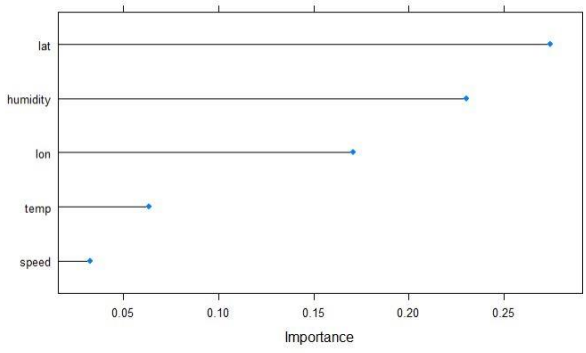

Figure 16: Variable Importance

Figure 17: Correlation Matrix long-term

5.3 Interpolation Predictions

Statistical modeling was important for the project to identify important relationships between the different variables. Statistical modeling enabled the ability to select important features for appropriate model training to make accurate predictions. An F-test was performed in order to determine the most important predictor variables. Each F-test tests the hypothesis that the response values grouped by predictor variable values are drawn from populations with the same mean against the alternative hypothesis that the population means are not all the same. A small p-value of the test statistic indicates that the corresponding predictor is important. The bar graph below demonstrates the F-Test results on the PM 2.5 predictor variables. The output scores are –log(p). Therefore, a large score value indicates that the corresponding predictor is important.

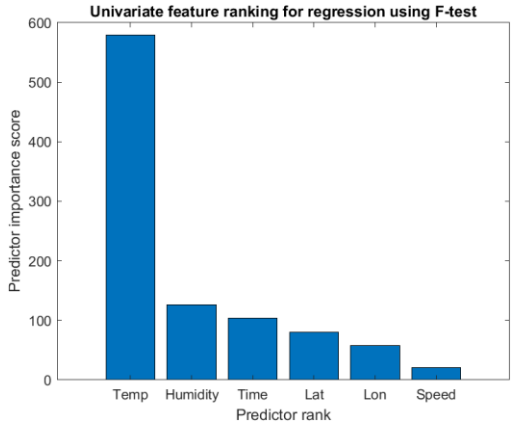

Figure 18: **F-Test Results**

The results indicate that temperature is the most important predictor. The next two valuable predictors are humidity and time. The last three variables do not demonstrate a significant response variable to the PM 2.5 predictions. These three variables are latitude, longitude, speed. Therefore**,** the features selected for the PM 2.5 interpolation model were temperature, humidity, and time. To appropriately train the model, the data was partitioned into train and test data. This procedure was performed to understand

the performance of the model against data not used in the training process. This also allowed to prevent overfitting for the model which is important. This will allow the model to predict accurately against a new dataset. The metric used to measure the performance of the model was the RMSE value. The RMSE value of the test data predictions was interpreted to understand the performance of the model. The model performed well if the RMSE values were minimized; this is an indication that the test predictions were approximately close to the true values.

6 Results

6.1 Short-Term Predictions

After initial inspection of models, Neural Networks and k-fold polynomial regression have been developed for the final models. In the structure of Neural Networks, different activation layers are defined, such as “relu”. It is noteworthy that different values for learning rate, epochs, batch-size, and layer numbers have been tested. The final and best representative results of PM2.5 prediction by using NN is summarized in the following table:

Table 1: PM2.5 prediction values using neural network.

Optimizer	RMSE	yactual_test	Ratio
Adam	40.221	81.738	0.524
SGD	32.258	81.738	0.398

Eventually, 5-fold and 10-fold polynomial regressions have been implemented for final prediction. As there was not significant improvement by using 5-fold or 10-fold model, a 5-fold regression was selected. Also, different degrees have been examined for polynomial regression. Moreover, the accuracy of prediction for robustness has been examined by adding 0.5 times standard deviation and 1 times standard deviation of the PM2.5 data. The summary of the RMSE values and their comparison with actual test data can be found in following table:

Table 2: PM2.5 prediction values using 5-fold cross validation

Type	Ideal Polynomial Degree	Avg RMSE	Avg Yactual test	Ratio	Avg R^2
Raw Data	6	13.840	48.746	0.284	0.512
0.5(Std) noise	6	19.779	48.808	0.406	0.399
1(Std) noise	6	24.05	48.979	0.491	0.249

Finally, the best calibrated model has been used to predict the test dataset values. Therefore, the results of using 5-fold cross validation regression and comparing the results with the ground truth values are represented in following table and figure:

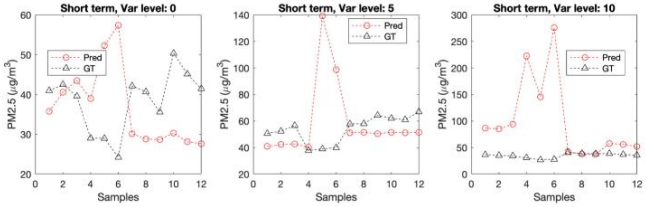

Figure 19: PM2.5 test dataset using 5-fold cross validation.

Table 3: PM2.5 test dataset NRMSE using 5-fold cross validation.

Test Type	Zero-Noise Variance	0.5σ Noise Variance	1.0σ Noise Variance
NRMSE	0.3046	0.4758	2.2474

The results of the regression illustrate that the accuracy of our model for predicting the short-term problem has been in high levels of validity and accuracy.

6.2 Long-Term Predictions

Neural Network was the model selected for the long-term predictions. For training the model, 80% of the data was used as training dataset, 10% as the validation dataset and 10% as the test dataset. Furthermore, different activations were added to the model. The optimal model had 150 epochs. The next process consisted of epoch optimization. Although increasing epochs may

indicate accuracy increase, after an epoch threshold the accuracy does not change or even decrease, and the model is only consuming time. Also, both optimizers “Adam” and “SGD” were tested. It was concluded that SGD was a better optimizer for this model prediction and dataset. To insert the train and test data set provided for submission, the data was down sampled to hourly rate. The prediction vs actual data plot for the initial training dataset is shown below. Also, a table the compares different neural network models and their performance is added.

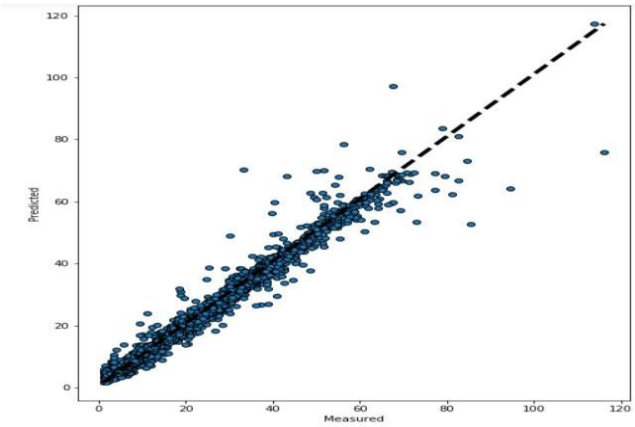

Figure 20: PM2.5 test dataset prediction vs actual value for final NN model

Table 4: PM2.5 prediction values RMSE using neural network.

	No activation /SGD	**No activation /Adam**	**Activation /SGD**	**Activation /Adam**
RMSE	23.42	25.69	2.28	3.76

Another model that was tested on the training dataset was 5-fold cross validation. The predicted data was plotted versus the actual data and RMSE was calculated. The RMSE in this case is 21.87 compared to the average of 35.31. From this result cross validation was not chosen as the model to proceed with.

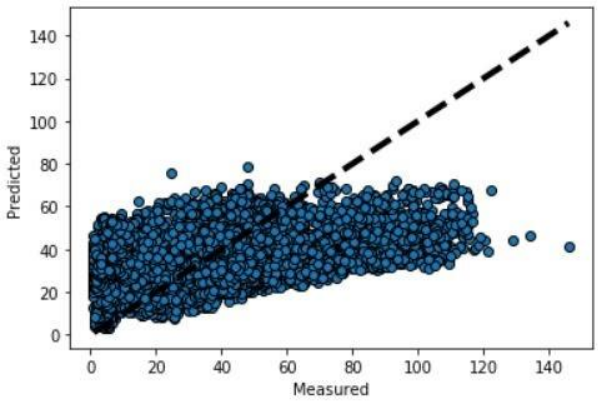

Figure 21: PM2.5 test dataset prediction vs actual value for 5-fold CV model

Then, the prediction from the final selected model were compared to the ground truth. The performance metric for this project was NRMSE. The results are illustrated below.

Table 5: PM2.5 prediction values RMSE using neural network.

Test Type	**Zero-Noise Variance**	**0.5σ Noise Variance**	**1.0σ Noise Variance**
Long-term	0.3970	0.6402	1.5225

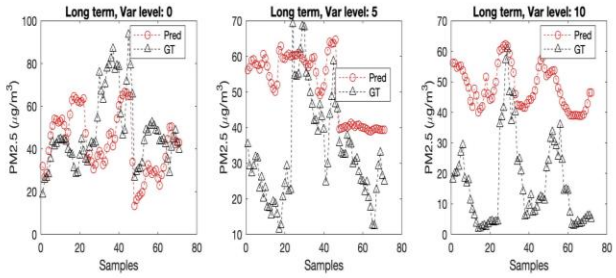

Figure 22: PM2.5 prediction vs ground truth using final NN

From the above result, it is observed that the model yielded predictions that were approximately within the range of the ground truth especially in the 0 var case. The prediction in all cases follow the same pattern as the actual data, however they may have a slight shift in comparison to actual PM 2.5 values. The overall, NRMSE for all long-term is 0.8532 which is in an acceptable range.

6.3 Interpolation Predictions

The GPR model was selected to interpolate the data missing between the second day between 11:30 am and 12:30 pm. The results of the interpolation are an output of 5-minute predictions

between the stated interpolation time for a total of 12 predictions. The robustness test consisted of the dataset with zero-mean Gaussian noise of 0.5σ and 1.0σ. The interpolation prediction results with the time series data are plotted below for zero-noise level, 0.5σ and 1.0σ level.

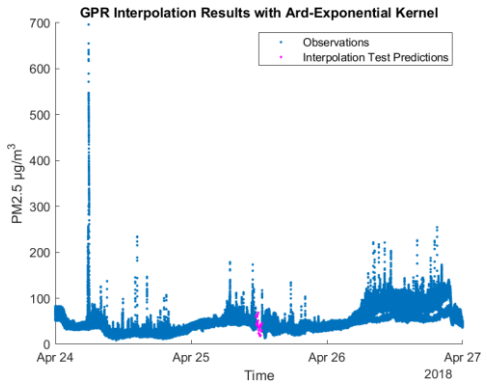

Figure 23: Zero-Noise Variance Added

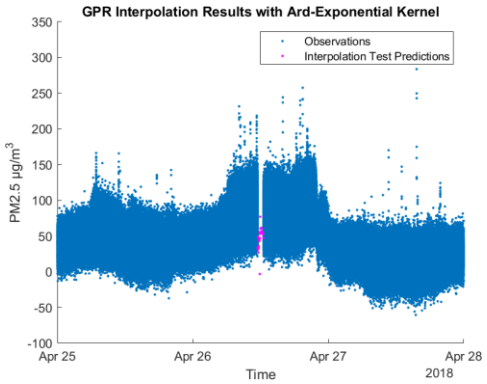

Figure 24: 0.5σ Noise Variance Added

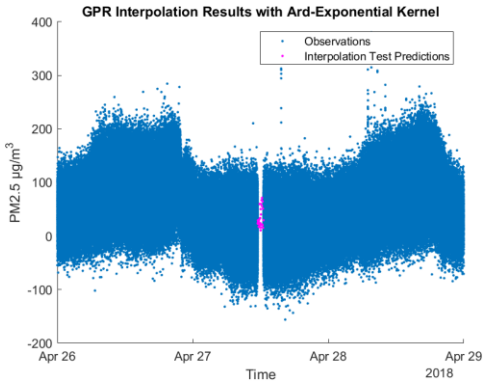

Figure 25: 1.0σ Noise Variance Added

The predictions from all three tests were compared to the ground truth to evaluate the effectiveness and accuracy of the model. The performance testing metric used was the normalized root mean squared error (NRMSE). The results are illustrated below in the following table:

Table 6. NRMSE Test Results for Interpolation

Test Type	Zero-Noise Variance	0.5σ Noise Variance	1.0σ Noise Variance
NRMSE	0.3046	0.4758	2.2474

The deviation of the Interpolation predictions from the ground truth can be visualized in the following plot:

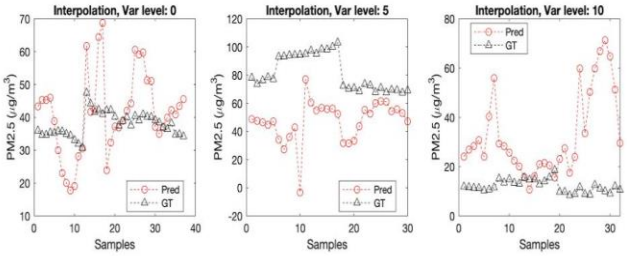

Figure 26: Interpolation Results Compared to Ground Truth

From the illustrated table and figures, the GPR model overall performs well in predicting and interpolating the missing PM 2.5 measurements. The GPR model performs best with zero noise level. The NRMSE value for this test was 0.3046. The predictions for this test result fluctuate above and beneath the ground truth as illustrated in the figure. This is different from what is observed in the figure for 0.5σ Noise Variance. The predictions follow a consistent pattern and trend similar to the ground truth. The NRMSE value for this test was 0.4758 which is an indication that the model does well in the predictions. However, there is a discrepancy in the predictions for 0.5σ Noise Variance which is a P.M 2.5 prediction below 0 µg/m3. This could be due to the variance of the model and the model failing to generalize to future predicting datasets. The last test that was performed (1.0σ Noise Variance) for interpolation illustrated the poorest results, compared to the previous test, with an NRMSE value of 2.2474. The deviation of the predictions from the ground truth can be observed in the figure for 1.0 σ Noise Variance. There is a consistent high error towards the final third of the hour predictions. This discrepancy could be due to the noise addition impacting the performance and ability of the model to adjust and predict accordingly for unforeseen data. However, the GPR demonstrated to be a robust modeling method across the three interpolation tests.

7 Improvements for the Future

7.1 Data Collection

To improve the data collection procedure and be able to obtain better and more reliable data, enhanced approaches can be used to analyze data from various sources and be able to merge them together in a sensible way. Also, improvement in the technology lead to sensors that provide more accurate data. Another critical issue in data collection is missing PM 2.5 values throughout the time-series data. Reducing and minimizing the number of missing values can be achieved using various techniques like improving the source, sensors or analytical approaches.

7.2 Data Analysis

Data analysis is an integral part of the development of the model. Therefore, improving the techniques and widen the approach is essential for the enhancement and accuracy of the model. This could be achieved by identifying data that is representative to train an accurate model. The exclusion of data has to be appropriate to provide a reliable training source for model training. This is also essential to prevent a bias dataset that will skew the performance and results of the model. Another aspect that can be enhanced in data analysis is the correct identification of different parameters to limit noise in the training dataset. Improving this aspect will be important as experienced in the project and results, noise will impact the performance of the model.

8 Conclusion

Prediction of PM 2.5 values for short-term seems to be challenging as the fluctuation of data in short periods is high. However, utilizing several models allowed to present the best model for this task. The 5-fold cross validation with polynomial regression was selected as the most appropriate model for short-term prediction. The final model contains speed, temperature, and humidity as the main components to predict PM 2.5 values. The values of NRMSE have been reported as 0.3046 and 0.4758 for zero and 0.5σ variance level, respectively. These statistics show the good performance of the model for short-term prediction. Further improvement of the prediction can be achieved by implementing Ridge and Lasso regression in cross validation model.

Although different methods such as cross validation, GPR and polynomial curve fitting were tested for long-term prediction, Neural network with activation developed the best result for predicting hourly PM 2.5 data for 24 hours at a specified location based on one-week data. The NRMSE for this model was 0.8532 which indicates the well performance of the predictions.

The interpolation predictions were successfully achieved with the GPR model. The GPR model yielded to the most accurate and robust method for interpolating the 5-minute PM 2.5 measurements interval based on three days of data. The model used temperature, humidity, and time as features to train accordingly. With low NRMSE values such as 0.3046 and 0.4758 for zero and 0.5σ variance level, respectively, the model performed well overall in the different tests. However, improvements can be performed to minimize the performance impact due to noise in the system.

Overall, our conclusion was that each type of predictions required a different approach to do the best prediction. Although all methods were tested for all three tasks, Cross-validation, Neural network, and GPR was used for short-term, long-term and interpolation task, respectively.

REFERENCES

[1] EPA, 2020. *Particulate Matter (PM) Basics*. Environmental Protection Agency. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics