PM 2.5 Spatial-Temporal Prediction: an Ensemble Learning Method with Dynamic Weighting Scheme

Ben Flanagan*
bflanag2@stanford.edu
Stanford University
Stanford, California, USA

Chenbo Wang*
wangcb@stanford.edu
Stanford University
Stanford, California, USA

ABSTRACT

This study proposes an ensemble learning model with dynamic weighting scheme to predict the PM2.5 concentration at locations of interest. The model can perform short-term and long-term prediction, and interpolation tasks. The ensemble method is an amalgamation of six individual models of different levels of flexibility. Through bootstrapping, 100 samples are drawn from 28 days of data collected by both mobile and static sensors in Tianjin and 7 days of static and mobile data collected in Foshan. Six individual models are evaluated on the 100 samples to quantify uncertainties associated with each model. Weights of individual models consist of two parts: preliminary weights and dynamic weights. The preliminary weights are preassigned according to model performance on the 100 bootstrapped samples. The remaining weights are dynamically decided by either Cross-Validation error or error on the 20% held-out data during the learning process. The proposed ensemble learning model is expected to perform well on all three tasks for which it was designed.

KEYWORDS

ensemble learning, bootstrap, PM2.5 prediction, mobile sensor

ACM Reference Format:
Ben Flanagan and Chenbo Wang. 2020. PM 2.5 Spatial-Temporal Prediction: an Ensemble Learning Method with Dynamic Weighting Scheme. In *Proceedings of CEE254 Data Analytics: PM2.5 Air Pollution Forecasting Competition (CEE254 Fall 2020).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 BACKGROUND

Urban air pollution from fine particulate matter can have serious health and environmental consequences for communities. As the dominant contributor to regional haze in urban cities in China, high PM2.5 concentration is known to correlate well with atmospheric invisibility. PM2.5 is also closely associated with the risk of morbidity and mortality from cardio-vascular and respiratory diseases in China [3]. In order to effectively issue health advisories and inform policy decisions, communities need a prompt and thorough understanding of the air pollution. Existing sensors and data acquisition systems are sparse, only offering limited resolution of the air pollution within a city. Recent advancements include mobile sensing networks but resolution remains sparse. The data acquired from these existing networks is characteristically noisy. Sensors can report a wide range of fine particulate air pollution readings over a very short time period. This variation is natural for this system and can result from gusting wind and passing vehicles, but presents a challenge when forecasting future air pollution levels. This study aims to propose a model that can accurately predict PM2.5 concentration at given locations. The proposed model should also be robust to noises and computationally efficient so that health advisories could be timely issued.

2 LITERATURE REVIEW

Creating accurate and robust forecasting methods for predicting PM2.5 air pollution is a topic of current research. Many methods have been explored. Due to the complex factors influencing PM2.5 air pollution levels, many of the proposed forecasting methods offer some insight but none offer a perfect solution. A brief review of existing research revealed: "prediction methods based on machine learning technologies are becoming increasingly common" [4]. This shift towards machine learning methods is due in part to the flexibility of many machine learning models to represent complex non-linear behavior. In a study on deploying an ensemble method for traffic state prediction, it was suggested that "the prediction accuracy can also be improved" for other grid based spatial-temporal data such as fine grained air quality when using an ensemble method [2]. One study explored just this topic: using an ensemble model to predict spatial-temporal air pollution data [5]. For this study, the ensemble method is paired with several other machine learning models.

3 DATA DESCRIPTION

The sample data used in this project was collected from Tianjin, China in the spring of 2018, and Foshan, China in the autumn of 2018. The data collected in Tianjin spans 28 days and includes observations from both static and mobile sensors. The data from Foshan spans 7 days and includes the same combination of static and mobile sensor data. The recorded data contains a large amount of useful information but also presents several challenges.

3.1 Sensor Data

The data collected from both the static and mobile sensors includes the same type of information. The database includes timestamps

*Both authors contributed equally to this research.

CEE254 Fall 2020, November 16–18, 2020, Stanford, CA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

for each sample and a record of the humidity, temperature, vehicle speed, pm2.5 air pollution, latitude and longitude. The sensors recorded data every three seconds. A summary review of the data reveals a stark contrast between the two sensor types, despite similar recorded information between the sensors.

The static sensors are far more consistent in recording data. The static sensors generally report a measurement every three seconds; however, multiple deviations from this trend were observed. While the static sensors were the more consistent sensor type, the inconsistencies in reporting measurements would present challenges to interpolation methods relying on uniform time steps between samples.

The mobile sensors were less consistent. They were generally active during the same hours each day and would have interruptions in reporting during the night. The large gaps in recorded data every evening from the mobile sensors would provide a formidable challenge to executing methods that rely on uniform time steps.

The samples of a three-second time-step are characteristically noisy. The air pollution measurements tend to follow long term trends, but each sensor is susceptible to momentary spikes. These spikes are likely the result of a high polluting moving source passing by a sensor location. One example would be a high emission diesel truck stopped adjacent to a vehicle equipped with a mobile sensor at a traffic light. In this case, the mobile sensor would record large PM2.5 air pollution levels for a few seconds before returning to the baseline after passing the diesel truck. While these peaks in the recorded data are interesting to observe, they represent noise in the data when making predictions on the scale of a few minutes to a few hours.

3.2 Re-sampling method: bootstrap

The collected data spans over 28 days. However, the objective model should only use 7 days of training data for the long-term prediction task (or 3 days for the short-term prediction and interpolation tasks). Such conflict necessitate re-sampling. Bootstrap is a powerful statistical tool which can be used to quantify the uncertainty associated with an estimator or a statistical learning method. A simple example of application is that bootstrap can be used to estimate the standard deviations of the coefficients of a linear regression fit [1]. In this project, bootstrap is used as a re-sampling method to evaluate different individual learning methods. 7 consecutive days (or 3 consecutive days depending on the task type) worth of data are randomly selected as training data, and one of the static sensors is randomly selected to provide test data. 100 sets of samples are bootstrapped for training and testing. Other than bootstrap, validation set approach and 5-fold cross validation are also adopted as re-sampling methods during the training of individual models and determining dynamic weights. Details are given in section 5.

3.3 Data Processing

For the models employed within this project, the primary challenges with the provided PM2.5 data were the large volume of data and the noise in the data. Both of these challenges were addressed through down-sampling the data with a space-time averaging method. Figure 1 shows an example of how the geographic limits of the input data are split into a finite grid. At every time step, the average PM2.5

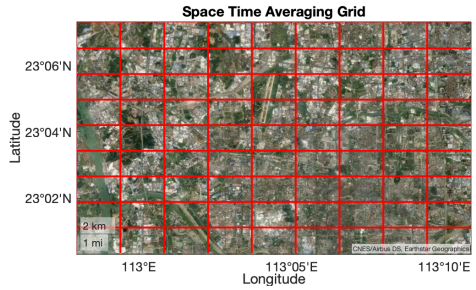

Figure 1: Spatial Grid for Down-sampling Data

air pollution measurement is computed for all measurements taken from within each grid section. Often, many of the grid sections have no active sensor within their bounds for a given time step. For this reason, after down-sampling the number of data points at each time step is much smaller than the total number of grid sections. The number of rectangles represents the total number of possible observations at a given time step after down-sampling. In addition to averaging the PM2.5 air pollution, the associated variables such as humidity and temperature were also averaged following the same procedure.

4 INDIVIDUAL METHODS

The ensemble method is an amalgamation of several individual models. For this project, six individual models were considered, each offering increasing levels of complexity. All of the individual methods considered perform well, at least on a subset of the training data. Each of the six individual models offers unique insight into the PM2.5 air pollution data, but each of these individual models may also perform poorly on some portion of the input data. With an ensemble approach, the goal is to always have a majority of the models performing well to balance the potential poor behaviors by certain models. The individual models considered for this project include the following models: Simple Linear, Simple Linear LASSO, Polynomial LASSO, Sine Ridge, Sine LASSO, and Gaussian Process Regression (GPR).

4.1 Simple Linear

As suggested by its name, Simple linear model is fit by simply including all available variables as predictors: time, humidity, speed, temperature, latitude, and longitude. Figure 2 shows the Simple Linear model fit on data collected by one of the static sensors. The model fit follows the ground truth quite closely, capturing the overall trend reasonably well. Linear regression is known as a less flexible method compared to other more complicated regression methods such as Gaussian Process Regression. Nonetheless, overfitting remains a concern, as suggested by the wiggly model fit shown in Figure 2.

4.2 Simple Linear LASSO

To limit the flexibility of the model, shrinkage method Least Absolute Shrinkage Selection Operator (LASSO) is applied to multivariate linear regression. By tuning the parameter λ in the penalty term, the model could be either as flexible as Simple Linear model

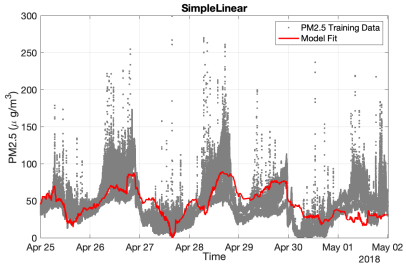

Figure 2: Simple Linear Fit Model on Sample Data

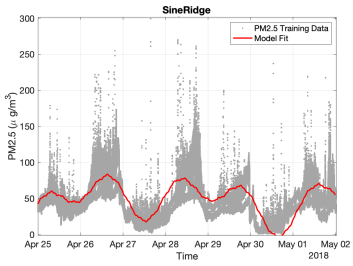

Figure 3: Sine Ridge Fit Model on Sample Data

or more inflexible. λ is selected by minimizing the cost function. With proper λ, LASSO shrinks the coefficients of certain predictors to zero, thereby performing feature selection. For this particular dataset, the model fit highly resembles the Simple Linear fit in Figure 2. For simplicity, it is not included here. It is noted that such resemblance does not necessarily hold for other data, e.g., mobile data and static data collected by other sensors.

4.3 Polynomial LASSO

To allow for more flexibility, quadratic and cubic terms of all six variables are included in addition to six original variables in the Polynomial LASSO model. As suggested by its name, Polynomial LASSO model applies LASSO as its shrinkage method. Similar to Simple Linear LASSO, the tuning of parameter λ is achieved by minimizing the cost function. Again for this dataset, the Polynomial LASSO fit shares great similarity with the Simple Linear LASSO fit in Figure 2 and Simple Linear fit, but is slightly more jagged due to increased flexibility.

4.4 Sine Ridge

Preliminary work on additive decomposition of a subset of the data reveals some periodic patterns existing in the PM2.5 time-series. Sinusoidal terms of time representing yearly, quarterly, monthly, weekly, and hourly patterns are included. In addition, polynomial terms of latitude and longitude are included to capture the spatial variation in PM2.5 data. Ridge is applied as a shrinkage method. Figure 3 shows the Sine Ridge model fit. Figure 3 presents obvious periodicity as a result of the sinusoidal terms. The Sine Ridge successfully captures the overall trend of PM2.5 time-series, and gives smoother fit compared to aforementioned models.

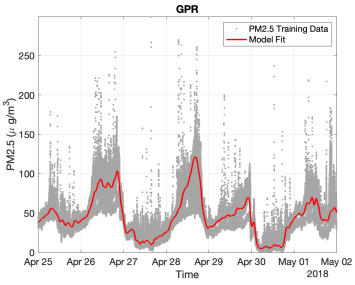

Figure 4: Gaussian Process Regression Fit Model on Sample Data

4.5 Sine LASSO

Compared to Sine Ridge, LASSO replaces Ridge as the shrinkage method in Sine LASSO model. LASSO shrinks the coefficients of some predictors to exactly zero, thereby selecting features for the model, whereas Ridge only shrinks some coefficients to small values, but never exactly zero. The model fit highly resembles that of Sine Ridge in Figure 3. Despite resemblance, due to application of LASSO the model fit is a perfectly smooth curve without small fluctuations seen in the Sine Ridge model fit.

4.6 Gaussian Process Regression

Gaussian Process Regression (GPR) is applied on all six variables to construct a continuous multivariate regression model. The "matern32" kernel is selected for the GPR model due to its superior performance on bootstrapped samples. Figure 4 shows that the resulting model is highly flexible, capturing variances of the ground truth that other models fail to attend. It is noted such high flexibility raises the concern of overfitting.

5 ENSEMBLE LEARNING MODEL

The ensemble learning method combines the individual methods into one aggregate method using a linear combination technique. Each individual model is multiplied by an associated weight. The sum of all the weights is one. Using a combination technique contributes to an overall prediction which tends to be more robust. On any given set of training data one of the individual models may preform poorly. This poor performance on some of the training data is evidenced in the large upper tails of the error distributions shown in Figures 5, 6 and the left panel of Figure 7. If only a single method were used to forecast PM2.5 air pollution, the method would occasionally exhibit large errors. The weighting scheme permits more accurate predictions by limiting the influence of these large error predictions on the ensemble prediction. This combination to reduce error relies on the assumptions that the individual models are independent and unbiased. The models must be independent such that if one model produces a poor prediction the remaining models should not also be expected to produce a poor prediction. Additionally the individual models should be unbiased to ensure the combination tends to produce a cancelling of errors. For example if all of the models tend to predict large PM2.5 values, the combination will also tend to predict large values. If, however, the models are unbiased, some will overestimate and others will underestimate

for any particular training data. When averaging these unbiased predictions the result will tend to be unbiased and have lower error.

The weights used in this ensemble model were decomposed into two components. The first part of the weighting factors represents a preliminary weight. The second part of the weighting factors is a dynamically adjusted modification. Each of the prediction types (short-term prediction, long-term prediction, and interpolation) was assigned slightly different weighting factors according to the same principles. For the interpolation prediction problem only three individual models were selected as components of the ensemble method; the Polynomial LASSO, the Sine Ridge, and the Gaussian Process Regression. These three models were selected because they consistently outperformed the other three models. These models were generally consistent and robust during this initial testing. The forecasting problems were perhaps more challenging considering the statistical regression models are better suited to interpolation than to extrapolation. For these two problems, all six individual models are trained as components of the ensemble method because all six models demonstrated good performance on some of the training data. No particular model was consistently superior for the extrapolation problem.

5.1 Preliminary Weights

The preliminary component of the ensemble method weights was predetermined. During the testing of the individual models, distributions of the error (RMSE) were obtained. Histograms of these error distributions are shown in Figures 5, 6 and the left panel of Figure 7. These error distributions display similar patterns across all models and prediction types. For each prediction type, the mean error across 100 samples is not that different between the different individual models. As expected, some models perform better than others. The standard deviation of the error tends to be quite large and the distributions show large upper tails. These histograms of errors show the individual models are generally good predictors but occasionally have large errors.

The error distributions assisted in the selection of the preliminary weights. Intuition and experimentation led to refinement of the preliminary weights to values which led to a robust and accurate ensemble prediction. Table 1 lists the preliminary weights selects for each of the prediction types. For each prediction type the weights must sum to 1. The two extrapolation methods show weights assigned to all six of the individual models demonstrating all six models are components of the ensemble while the interpolation method only shows three weights corresponding to the three models used in the interpolation prediction. The percentages shown on the left side of the table represent the contribution from the preliminary weights to the final weights. A larger percentage represents less model flexibility during training. A smaller percentage for the preliminary weights signifies a greater degree of adaptability is possible during training. Less modification was permitted on the long-term prediction model and the interpolation model. For the long-term model this constraint improved the reliability of the prediction and for the interpolation model the constraint made the results more predictable. Greater flexibility was permitted on the short-term prediction.

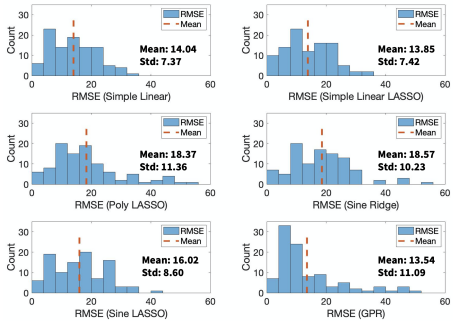

Figure 5: Short Term Prediction Errors for 100 Samples

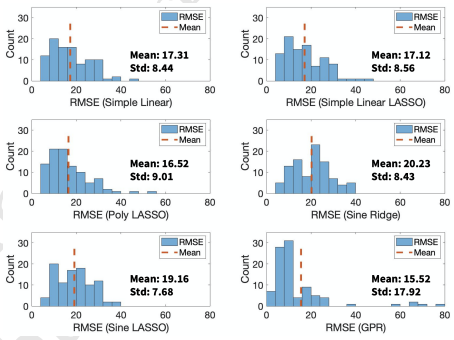

Figure 6: Long Term Prediction Errors for 100 Samples

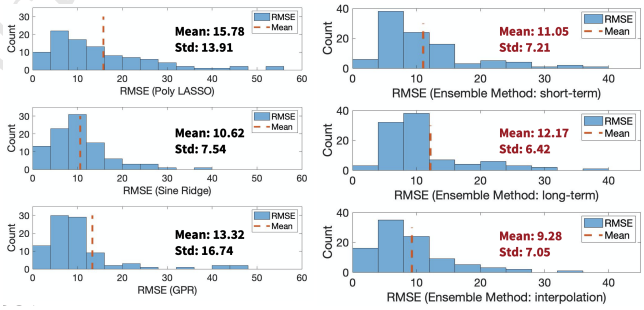

Figure 7: Left: Interpolation Prediction Errors for 100 Samples. Right: Short Term, Long Term and Interpolation Prediction Errors for 100 Samples Using Ensemble Method.

5.2 Dynamic Weight Modification

The preliminary weights are selected to represent weights which are appropriate for the majority of the training data. For each set of training data, however, the relative errors for each of the individual models will be different. For example, the Gaussian Process Regression fit is often one of the superior models but it also sometimes has large errors as shown in the long upper tail in the error distribution histogram of Figure 5, 6 and the left panel of Figure 7. In a case where one particular model is unsatisfactory, it is desirable to reduce its relative contribution to the ensemble prediction. Additionally, when one model is superior to all the other models it

Table 1: Preliminary Weights for Each Prediction Type

Prediction Model	SimpleLinear	LinearLASSO	PolyLASSO	SineRidge	SineLASSO	GPR
Short-Term Weights (60%)	0.3	0.0	0.0	0.0	0.3	0.4
Long-Term Weights (80%)	0.0	0.35	0.35	0.0	0.0	0.3
Interpolation Weights (80%)	-	-	0.35	0.3	-	0.35

would be advantageous to increase its relative contribution to the combined model.

In this project, the weights of the individual models are adjusted according to the inverse of the relative errors for each model. Individual models with large errors are assigned smaller weights and models with small errors are assigned larger weights. Two different techniques were used to estimate the relative error of each of the individual models. For the interpolation and long-term prediction problems, a traditional 5-fold cross-validation is performed using the training data to compute the interpolation error for each of the individual models. For the short-term prediction model, a modified cross-validation is performed to estimate the extrapolation error associated with each of the individual models.

The 5-fold cross validation estimates the interpolation error for each of the individual models. To perform the cross validation, the training data is randomly segmented into 5 groups. The models are each trained 5 separate times with a different group set aside as test data each trial. The average error is then computed. The weights are then adjusted according to these computed errors. The individual models are retrained on the complete training data set to ensure all input data is used in the prediction.

While this traditional 5-fold cross-validation is a great descriptor of the interpolation error for each of the individual models, it does not accurately portray the extrapolation error for the different models. The 5-fold cross validation method was modified to better estimate the extrapolation error by assigning the input data into special training and testing groups. This extrapolation error measure was named "Future Validation" within the context of this project. Future validation uses the same proportion of training and testing data as used in the 5-fold cross validation with 80% of the data as training data and 20% of the data as test data. The data is not randomly assigned to these bins but sorted such that the first 80% of measurements become training data and the final 20% of observations become test data. Using these specially assigned training and testing bins, the future validation method estimates the extrapolation error for each of the individual models. This estimate of the error is used to adjust the weights similar to the modification based on the cross validation error. Each of the individual models is then retrained on all of the input data. This retraining step is crucial for the extrapolation problem as the error tends to increase with the number of observations extrapolated in the future.

The retraining step following the weight modification is an important step for the extrapolation problems but doing so has significant ramifications on the errors computed to adjust the weights. It is assumed that the relative errors of each of the individual models will remain similar between the validation models and the final models. The majority of the training data for the final individual models (80%) is the same as the data which trained the validation models and therefore the models are expected to be similar. The weight modification does not rely on the absolute value of the error for any particular model, only the relative proportions. It seems reasonable to conclude the relative proportions of the individual model errors remain similar if the models (and training data) remain similar.

One final step in the model weighting is a check for reasonable results. Occasionally some models predict unrealistic values for the PM2.5 air pollution. One of the more frequent causes of this occurrence is cubic growth of the polynomial regression fit for the extrapolation problem. The function has several built in checks to ensure the predictions for the individual models are reasonable. If any model seems unreasonable or has any errors, the weight is reduced to zero to remove it from the final model. When one model is removed the remaining weights are proportionally adjusted to maintain a total of 1. This final check helps ensure the prediction results remain reasonable even if individual models have issues.

6 RESULTS

Testing the ensemble method revealed improved performance relative to all of the individual component models. These results confirm the observations from the literature review and demonstrate the advantage of an ensemble method. The results from tests performed on 100 sets of re-sampled data are displayed in the right panel of Figure 7. The error distributions for each of the prediction types show a smaller average error relative to any of the individual models. The standard deviation of the error is also smaller than for the individual models. These results suggest this implementation of the ensemble approach generally makes predictions closer to the ground truth and also tends to have fewer predictions that are far from the ground truth. The ensemble method required additional computational expense to train all of the models. This added training time leads to more accurate predictions and fewer predictions which are far from correct.

One of the goals of this project was to make several predictions about PM2.5 data. Figure 8 displays these prediction results. The testing described within this report lends confidence to theses predictions. A visual inspection of the prediction quantities does not reveal any abnormal results. The long-term predictions have a greater spatial variation which corresponds directly to the greater spatial variation present in the training data. For each of the 0-variance cases, the predicted values seem to follow the observable temporal trends in the data. The cases with added noise become more challenging to visually inspect because the signals with large amounts of added noise appear as one solid block of data. Even in these cases, the ensemble method makes predictions which seem plausible. With the goal being to forecast average PM2.5 values, it seems reasonable for the predictions to follow the center-line of the PM2.5 time histories.

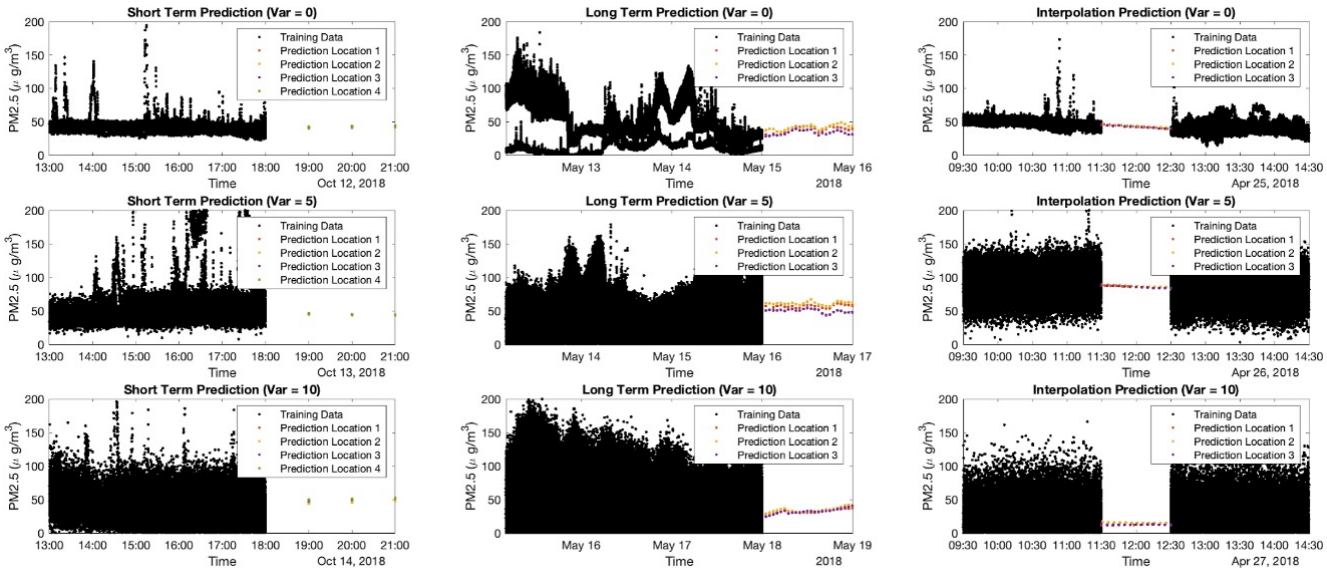

Figure 8: Prediction Results

7 CONCLUSION

The proposed ensemble model is expected to perform well on all three tasks for which it was designed. Testing described within this report demonstrates the effectiveness of the ensemble approach on the set of generated testing data. The proposed model also seems to work well even with sparse data or data with large amounts of zero mean noise added to the signal. It is believed the down-sampling, individual statistical models and ensemble approach all contribute to the proposed model's satisfactory performance with noisy data. The ensemble learning method is able to improve the prediction accuracy as well as reducing the variance of predictions. While the method requires more computational time that any individual model, the proposed model remains computationally efficient, requiring only a few minutes to complete each type of prediction. This time is directly related to the down-sampling rate and amount of input data.

8 FUTURE WORK

Multiple avenues exist for possible improvement on this ensemble method. One of the best methods to improving prediction accuracy may be refining the spatial and temporal resolution. Computational constraints limited the minimum resolution within this project but future work could explore implementations with superior models trained on faster computers. The ensemble method framework is highly flexible, allowing for the selection of any component models and weighting schemes. More advanced individual models could improve the resulting prediction. This report showed how an ensemble approach improved the prediction accuracy to be greater than any of the component models. If the models were all improved, a similar result could still be expected with the ensemble method. The re-sampling scheme in this study selects one of the static sensors to provide test data. As convenient and pragmatic as it is, limitation on the test location might introduce sample bias. Improvement is expected if the re-sample scheme is refined to include more possible test-locations. Further exploration of the weighting scheme may reveal improved methods for selecting the preliminary weighting and setting the degree of modification.

ACKNOWLEDGMENTS

We would like to express our gratitude towards Professor Haeyoung Noh and Mr. Jingxiao Liu for providing guidance throughout this project.

REFERENCES

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112. Springer.

[2] Yang Liu, Zhiyuan Liu, Hai L. Vu, and Cheng Lyu. 2020. A spatio-temporal ensemble method for large-scale traffic state prediction. *Computer-Aided Civil and Infrastructure Engineering* 35 (Dec. 2020), 26–44. https://doi.org/10.1111/mice.12459

[3] David YH Pui, Sheng-Chieh Chen, and Zhili Zuo. 2014. PM2. 5 in China: Measurements, sources, visibility and health effects, and mitigation. *Particuology* 13 (2014), 1–26.

[4] Xiaotong Sun and Wei Xu. 2019. Deep Random Subspace Learning: A Spatial-Temporal Modeling Approach for Air Quality Prediction. *Atmosphere* 10 (2019). https://doi.org/10.3390/atmos10090560

[5] Junshan Wang and Guojie Song. 2018. A Deep Spatial-Temporal Ensemble Model for Air Quality Prediction. *Neurocomputing* 314 (Nov. 2018), 198–206. https://doi.org/10.1016/j.neucom.2018.06.049