Predicting Air Pollution in Beijing

Brenton Arnaboldi
ba1303@nyu.edu

Zach Zhang
zz1409@nyu.edu

Abstract

Air pollution in China has reached dangerous levels due to rapid industrial growth. Poor air quality has caused a public health crisis, as pollution is estimated to contribute to 1.6 million deaths annually in China. Against this backdrop, we have created a forecasting model that accurately predicts fine particulate matter ($PM_{2.5}$) concentrations one day ahead in Beijing. We compared several different approaches (Random Forests regression, LSTM, and Gaussian Process), and found that combining the Random Forests and Gaussian Process models led to the most accurate predictions.

1 Introduction

Air pollution is a serious health hazard in many developing countries, and is believed to kill more people worldwide than AIDS, malaria, breast cancer, or tuberculosis (Rohde, 2015). Pollution is a particularly acute problem in China. While Chinas economic growth in the past 30 years has been extraordinary, lifting billions of people out of poverty, the rapid industrialization of the country has led to environmental degradation and poor health outcomes. According to one study, air pollution is estimated to have contributed to 1.6 million deaths per year in China (Rohde, 2015). Another study concluded that poor air quality has decreased life expectancy in China by 3.5 years, and in Beijing by 6.4 years (Greenstone, 2017). Even though the Chinese government has prioritized clean air in recent years (e.g. imposing a nationwide cap on coal use), pollution in Beijing routinely rises above the air quality guidelines of the World Health Organization (WHO). The average concentration of $PM_{2.5}$ in Beijing in 2017 was 58 $\mu g/m^3$ (micrograms per cubic meters), well above WHOs standard of 10 $\mu g/m^3$. As such, air pollution remains a constant menace in Beijing.

This public health crisis has inspired our work on the prediction of air quality. An accurate forecasting model for pollution would allow authorities in China and Beijing to release timely warnings and therefore help protect people from polluted air. Possible interventions could include having construction workers indoors, temporarily reducing industrial output, or requiring public schools to have indoor recesses. Decreased exposure to pollutants such as $PM_{2.5}$ could result in thousands or even millions of lives saved annually in China.

For our project, we sought to predict average daily $PM_{2.5}$ one day in advance for Beijing. We first built a Random Forests model with hard-coded autoregressive features, which achieved fairly robust performance on out-of-sample data. We then improved our results by combining the Random Forests model with a Gaussian Process Matern kernel.

Related Work

Much of the prior literature on air pollution modeling uses feed-forward neural networks. The inputs to these models tend to be autoregressive features and weather variables. Corani (2005) experimented with a variety of simple neural networks to predict PM_{10} and ozone concentration in Milan. More recently, Fu, Wang, Le and Khorram (2015) constructed a feed-forward neural network with rolling mechanism and accumulated generating operation of gray model (RM-GM-FFNN) to predict daily $PM_{2.5}$ in Hangzhou, Nanjing, and Shanghai. The authors argue that traditional neural networks do

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

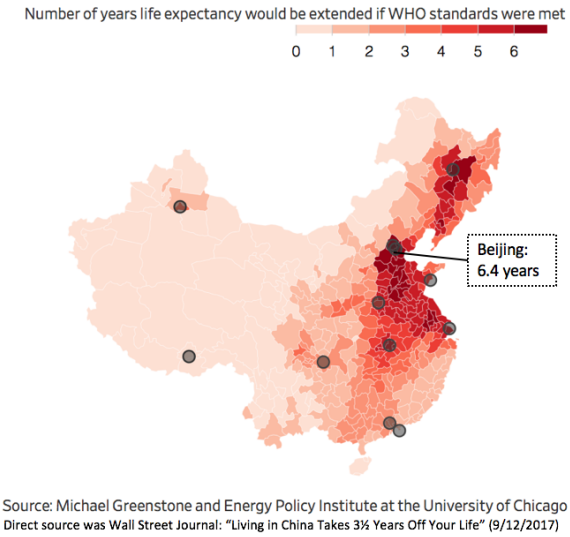

Figure 1: Estimated loss in life expectancy from air pollution in China

not sufficiently assess possible correlation between different input variables. Finally, other studies have proposed hybrid models. Diaz-Robles et al. (2008) implement a model combining ARIMA and neural networks to predict PM_{10} in Chile. The authors found that the hybrid model achieved better performance than the two separate models.

To our knowledge, Gaussian Processes have not been used for predicting future values of pollution. Interestingly, however, researchers have used Gaussian Processes to predict pollution in areas far away from monitoring stations (Jutzeler, 2014). Other methods in the literature include the use of Kalman filtering (Hoi, 2008) and Hidden Markov Models (HMMs) with non-Gaussian distributions to forecast high pollution days (Sun, 2012). But the majority of research has used feed-forward neural networks.

Given the severity of pollution in China, we expected there to be more literature on predicting air quality in Beijing. One study (Ni, Huang and Du, 2016) used an autoregressive integrated moving average (ARIMA) model to predict hourly $PM_{2.5}$ in Beijing. The study achieved an R^2 score of 0.984, but with one major caveat. The model only made predictions one hour in advance. (Their final fitted model was ARIMA (1, 1, 3)). The authors also built a neural network model to predict daily $PM_{2.5}$. Comparing their results with ours is difficult because they took data from a different source (Beijing's municipal system) and different time period (Jan-Dec 2014). Furthermore, while their model used concentrations of other pollutants (NO_2, CO, SO_2) as input features, we felt that including such variables was not methodologically sound. Another recent study (Feng et al, 2015) attempted to predict $PM_{2.5}$ in Beijing and Tianjin by combining neural networks with "air mass trajectory analysis and wavelet transformation", essentially using advanced meteorological data as inputs to a neural network. The data used for this paper came from September 2013 to October 2014 (our data ranged from May 2014 to December 2016). Overall, most research on Beijing pollution has focused directly on health effects rather than predictive modeling.

2 Problem Definition and Algorithm

2.1 Task

The goal of this project is to forecast the concentration of airborne particulate matter less than 2.5 micrometers in diameter ($PM_{2.5}$) in Beijing. $PM_{2.5}$ is widely considered the most harmful pollutant, as it can travel deep into the lungs and enter the bloodstream. Various studies have linked $PM_{2.5}$ concentrations to higher risks of heart disease, stroke and lung cancer (Health Effects Institute, 2016), and so as a result this was the pollutant we decided to forecast. Our external inputs for the model were weather-related features (wind speed, wind angle, temperature, humidity, air pressure) and recent

$PM_{2.5}$ values from other cities across China. Ultimately, we attempt to predict the next day's average $PM_{2.5}$ in Beijing.

2.2 Algorithm

2.2.1 Random Forests

We first implemented a Random Forests regression model to predict $PM_{2.5}$. Since Random Forests is not a "traditional" time series algorithm, we needed to explicitly define features that captured the temporal dependencies in Beijing's $PM_{2.5}$. As a first step, we analyzed the autocorrelation of daily $PM_{2.5}$ values. Based on the plot below, it appears that the correlation is significant for t-1 and t-2 days, but by t-3 days the correlation is no longer significant. As such, when we introduced autoregressive features to the Random Forests model, we only added Beijing's $PM_{2.5}$ values from the previous two days (t-1 and t-2).

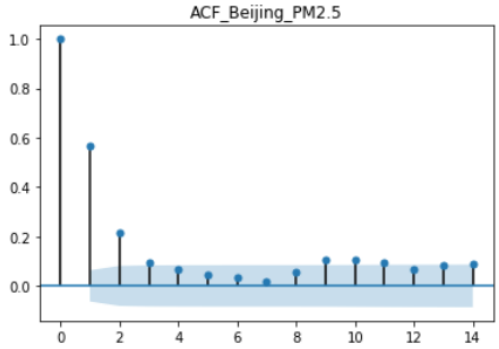

Figure 2: Autocorrelation Plot for Daily $PM_{2.5}$ in Beijing

In terms of feature engineering, we also incorporated pollution values from other cities in northern and central China. We chose four cities: Xian, Harbin, Baotou, and Qingdao. We chose these particular cities because they are all: i) relatively close to Beijing but ii) sufficiently far away from Beijing to avoid redundancy with the existing autoregressive variables. Furthermore, they lie in different directions from Beijing. A map of the cities is below. When we looked at the cross-correlation between Beijing's $PM_{2.5}$ and the $PM_{2.5}$ of the other cities, we generally found significant relationships after 1 day but not much after 2 days. As such, our final model only included features for $PM_{2.5}$ in Baotou, Xi'an, Qingdao, and Harbin at t-1 days.

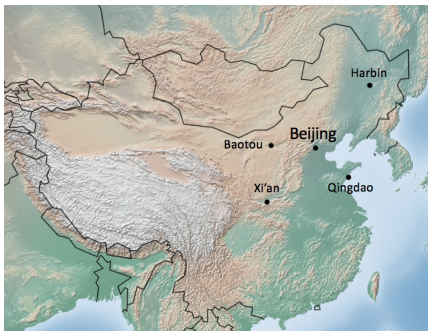

Figure 3: Cities used in model

2.2.2 LSTM

Another approach that we experimented with was to use a Long Short Term Memory (LSTM) neural network. LSTMs are recurrent neural networks (RNNs) that use memory gates to model long term

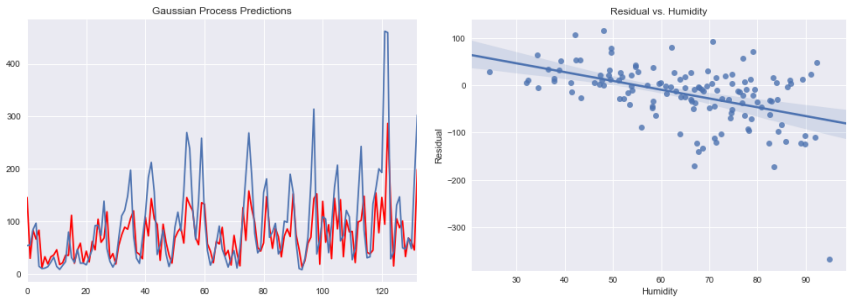

Figure 4: Left: Predictions made by Gaussian Process on test set. Right: plot of residuals vs. humidity. Significant relationships between the residuals and weather variables motivated us to try an ensemble of random forests and Gaussian process models.

dependencies. These models have achieved state of the art results in many sequential modeling tasks in natural language processing. Here we implement it by feeding in the same features as the Random Forest to the LSTM at each time step. The motivation for this was to model longer range temporal dependencies in the pollution itself as well as with the weather data. However, we found that the LSTM architecture performed poorly on this task. In general, deep learning methods are very data hungry in that they need a large amount of training data to learn well. We believe that the size of our dataset (only 964 daily records) limits the effectiveness of RNNs and motivates us to make different model choices.

2.2.3 Gaussian Process

For a Gaussian process, the key assumption is that all data points are drawn from a multivariate Gaussian distribution with mean function $\mu(x)$ and covariance $k(x_i, x_j)$. Gaussian processes tend to learn well from less data; as such, we thought GPs would be well suited for our problem. Furthermore, they offer a lot of flexibility in that we can encode our knowledge about the problem into the covariance kernel k. In our experiments we use the standard RBF kernel whose covariance is shown below.

$$k(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2) \tag{1}$$

We compare this kernel to a rational quadratic kernel as well as a Matern kernel. The Matern kernel is a generalization of the RBF kernel with additional parameter ν that can control the smoothness of the kernel. This approach is able to outperform the LSTM; however, it is difficult to encode some of the other external variables (such as weather features). When analyzing the predictions of this model (shown in fig 4), we find that the model learns the normal behavior very well. However, the model tends to be worse on high pollution days, missing many of the peaks illustrated above. We found that the residuals are significantly correlated with weather variables, which motivated us to combine the GP with the static random forests model.

2.2.4 Ensemble Model

After several unsuccessful attempts to incorporate other features into the Gaussian process, we ultimately settled on an ensemble approach. First, we trained a GP with the temporal data (pollution-only). We then provided the predicted mean and covariance as two additional features to the Random Forest model. Essentially, this configuration allowed the Random Forest to use the GP prediction when the GP was confident (low covariance) and to use other features (e.g. weather) when the GP model was uncertain (high covariance). We found that this strategy resulted in slight improvements in performance.

3 Experimental Evaluation

3.1 Data

To build our models, we used pollution data released by the Chinese government as well as weather data collected from Weather Underground.

3.1.1 Pollution Data

The main dataset for our analysis was curated by researchers from the Center for Geographic Analysis at Harvard University, but originally came from the China National Environmental Monitoring Center. The dataset contains hourly pollution data from 1497 monitoring stations across China from May 13, 2014 to December 31, 2016. Rather than keep the target variable ($PM_{2.5}$) in hourly form, we decided to aggregate the values to the day level. We transformed the target variable to daily form for a number of reasons. First, air quality guidelines (such as those from the World Health Organization) tend to specify annual and 24-hour average limits, as opposed individual hours. Second, most of the literature on predicting pollution looks at day-level concentrations. Third, modeling a Gaussian process is far more convenient for a day-level target variable (964 records) than for an hour-level target (23136 records). Finally, the issue of missing values was far less serious when our target variable was daily. To calculate daily $PM_{2.5}$, we averaged the hourly $PM_{2.5}$ values for each day. Each day in the dataset had at least 12 hours' worth of $PM_2.5$ data. As such, missing a few hours of data did not markedly impact our results when the target variable was daily. Missing values (at the hourly level) were imputed by taking a weighted average of the most recent observations.

Of the 1497 stations in the dataset, 12 are in the Beijing metropolitan area. For our target variable, we chose to track the $PM_{2.5}$ values from Station 1012A, which had the fewest number of non-missing values among stations in Beijing. In hindsight, we probably should have taken an average of the 12 stations in Beijing (rather than just tracking 1012A), but we were concerned that some of the stations might be in suburban areas and thus underestimate pollution levels in the actual city.

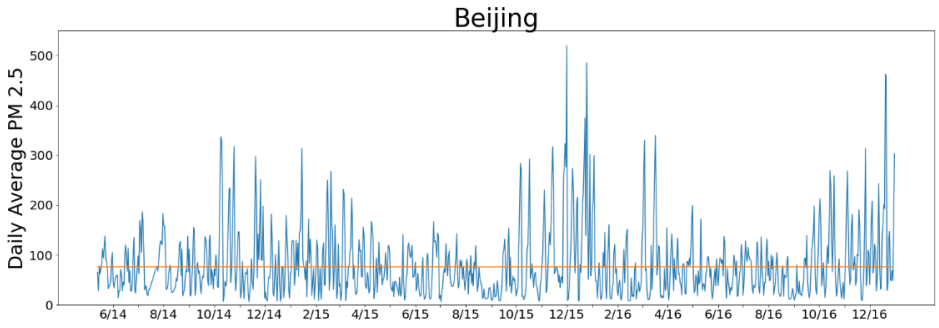

Figure 5: Daily Average PM 2.5 in Beijing. The orange line (75 mg/m^3) represents China's 24-hour limit for urban areas.

The plot above illustrates the daily average $PM_{2.5}$ in Beijing over the 964 days in our dataset. The series is volatile and has many rapid spikes. Most of these spikes occur in the winter months, but the data is fairly noisy. Furthermore, the distribution of $PM_{2.5}$ values is highly skewed, as most observations are between 0-60, but some values exceed 500. The skewed distribution of $PM_{2.5}$ values was a problem because for our Gaussian Process models, we assume that the distribution of the target variable is Gaussian. In response, we transformed the target variable by taking its logarithm, then shifted the distribution to center it at $\mu = 0$. A comparison of the distributions (before and after the log transformations) is below:

3.1.2 Weather Data

In addition we have found weather data to be important for pollution prediction. The weather naturally affects air flow and will relate to how much pollution is in the air. We utilize weather data collected from Weather Underground's API. Weather Underground maintains as database of historical weather readings recorded from the station in Beijing. The readings are recorded each hour. We fill missing

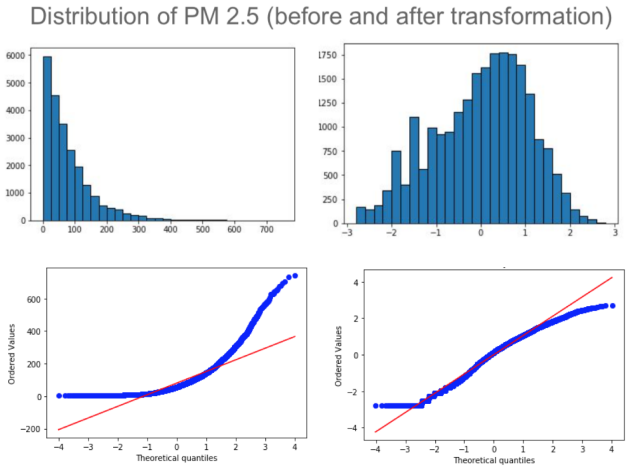

Figure 6: Distributions of PM 2.5 before and after transformation, with Q-Q plots below

Feature	Mean	Standard Deviation	minimum	25%	50 %	75%	maximum
Temperature	57.60	20.04	4.77	37.55	61.85	75.95	90.72
Humidity	57.23	19.28	8.91	42.12	58.2	72.5	97.95
Wind Speed	6.40	3.24	1.75	4.36	5.58	7.35	22.45
Wind Direction (deg.)	155.9	51.6	25.31	120.94	159.37	194.06	273.75
Pressure	30.01	0.301	29.37	29.75	30.00	30.26	30.83

Table 1: Summary statistics for daily weather data. Note that for wind direction, 0 degrees is set to West, 90 degrees to North, 180 to East.

values by carrying forward the previous values. We use temperature, humidity, wind speed, wind direction, as well as pressure. Summary statistics of the variables are given below.

Our original weather data was hourly, but we converted the features to the day level by taking the average of 24 hourly values per day. For temperature, humidity, wind speed, and air pressure, this process was trivial. However, for wind angle, the process was far more complicated because of the variable's circular nature (for example, a 359-degree angle is very similar to a 1-degree angle). As such, using the simple averaging method from before, a day with winds blowing out of the northwest (315 degrees) half the time and from the northeast (45 degrees) the other half of the time would be computed as having an average wind angle of 180 degrees (south), when really the average direction should have been 0 degrees (north). We considered dividing the wind angle variable into two parts: an east-west component (cosine of angle) and north-south component (sine of angle). However, we were concerned that this approach might diminish the significance of the wind direction variable in our model. Instead, we kept averaging hourly wind-angle values to obtain a daily average wind angle, but rather than indexing 0 degrees to the North, we indexed 0 degrees to the West. Based on an analysis of the hourly data, we noted that wind in Beijing rarely comes from the west. Of the 20273 hourly observations, approximately 10.8% of records listed the wind as coming from the West, versus 39.5% from the North, 23% from the East, and 27% from the South. By indexing 0 degrees to the West (where winds are least frequent) instead of North, we reduced the noisiness of the day-level wind direction variable.

3.2 Methodology

For our experiments, we trained the models on data from May 2014-August 2016 (842 records), leaving the final four months (September 2016-December 2016) as the test set (122 records). To evaluate our models we use two different metrics. Firstly, we use R^2 as it is a standard metric for regression problems. However, this metric doesn't capture the fact that we care most about high pollution days. We would like another metric that penalizes models for failing to predict high pollution days, regardless of their overall R^2 score. To accomplish this we also measure precision and

recall for identifying days where PM 2.5 is greater than 150 $\mu g/m^3$. While the Chinese government has set an official daily average limit of 75 $\mu g/m^3$ (GB 3095-2012), this threshold is eclipsed 30-40% of the time in Beijing, so we wanted to use a more selective benchmark. In 2012, the Chinese government divided PM 2.5 values into six classes: 035 ("excellent"), 3575 ("good"), 75115 ("light pollution"), 115150 ("medium pollution"), 150250 ("heavy pollution") and larger than 250 ("severe pollution"). (Technical Regulation on Ambient Air Quality Index). Based on these standards, we decided to predict days with PM 2.5 greater than 150 (days with "heavy" to "severe" pollution). In the 4-month test set, there were 20 such days (16.4 percent of records).

It is important to clarify that we trained our models using the logarithm of $PM_{2.5}$, but evaluated the models according to the original $PM_{2.5}$ values. Transforming the target variable to "log space" condenses the data into a tighter window and can obscure large differences in pollution. For example, the values of 0.3 and 1.2 might not seem far apart (if the distribution is centered at 0), but when re-converted back to the "original range" of $PM_{2.5}$ values – under the formula $\exp(x + 3.89)$, they correspond to 66 $\mu g/m^3$ ("good") and 162 $\mu g/m^3$ ("heavy pollution"). As such, when evaluating the effectiveness of the models, it is important to analyze the performance in terms of the original $PM_{2.5}$ values.

3.3 Results

Random Forests

The first random forests model achieved an R^2 score of 0.691, and identified high-pollution days with 60% recall and 80% precision. For a feature importances table, please refer to the appendix.

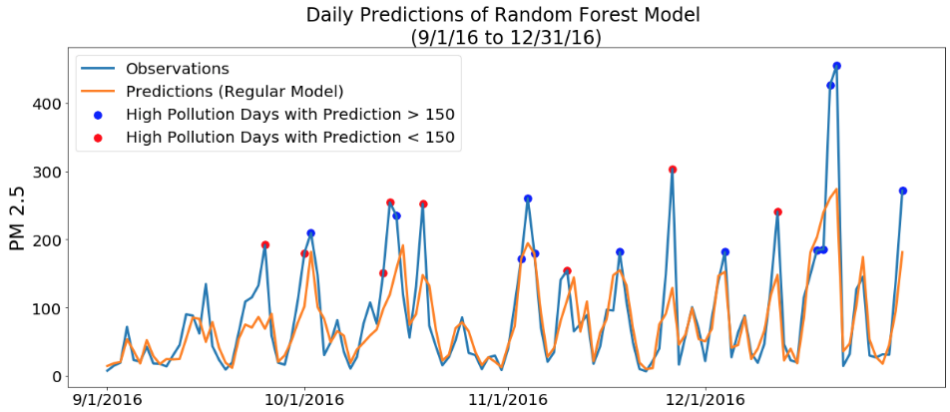

Figure 7: Predictions from Random Forests Model for Test Set. The blue dots represent observations over 150 $\mu g/m^3$ where the model's prediction was also above 150 ("success"), while red dots represent observations where the model's prediction failed to be above 150.

While the model does a fairly good job of predicting $PM_{2.5}$, it often underestimates the spikes in pollution. This trend is reflected in the relatively low recall rate (60%). Ultimately, from a real world perspective, the primary goal of our model should be to forecast bad pollution. Identifying days with high $PM_{2.5}$ is more important than predicting days with clean air, so we considered ways to improve the model's recall of 150 mg days.

One of our ideas was to place higher weights on training examples with high $PM_{2.5}$ values. The specific weighting formula is below. (x is the transformed log($PM_{2.5}$) value for the observation – note that the mean of the transformed data is 0).

$$Weight(x < 0) = 0.01 \tag{2}$$

$$Weight(x \geq 0) = x^3 + 0.01 \tag{3}$$

Essentially, if x is less than 0, we set a weight of 0.01, whereas if the x is greater than 0, we cube it and add 0.01. Under this weighting strategy, the model basically ignores the bottom 50 percent

of $PM_{2.5}$ values and places higher weights on high-pollution days. Furthermore, because of the cubed term, the weighting function places particularly high weights for observations of $x \geq 1$ (or equivalently, $PM_{2.5}$ values above 133 $\mu g/m^3$).

The new model's R^2 score and precision were slightly lower (0.641 and 68%), but recall increased from 60% to 75%. The plots below illustrate that the new model generally does a better job of predicting the peaks in $PM_{2.5}$. The bottom plot directly compares the predictions of the two models, with the green line representing the weighted RF model and the orange line indicating the original RF model. Because the sample size of high-pollution days in the test set (20) is small, we also analyzed the recall of bad pollution days in the training set to confirm the effectiveness of the weighted model. The regular RF model achieved a recall of only 59% (44 of 75) on the training set; by contrast, the weighted RF model attained a recall of 73% (55 of 75).

Recall was 75% (15 of 20) in the test set, but the first plot below illustrates that the model was very close to identifying at least three of the five missed points. Of all the mis-classifications, the only egregious error was the spike on November 26, 2016. We dug through that data point more closely, and we noticed a few anomalies. For one, the wind speed was average (about 5.3 miles per hour), whereas pollution tends to spike when there is little wind. Furthermore, while pollution in Beijing had been fairly high the previous day (around 148 $\mu g/m^3$), a day-to-day increase of 150 (to 303 $\mu g/m^3$) was unusual. Finally, the $PM_{2.5}$ values in other cities (Xi'an, Qingdao, Harbin, and Baotou) were only slightly higher than average the previous day. The combination of these factors made it very difficult for the model to predict the spike in pollution on this date.

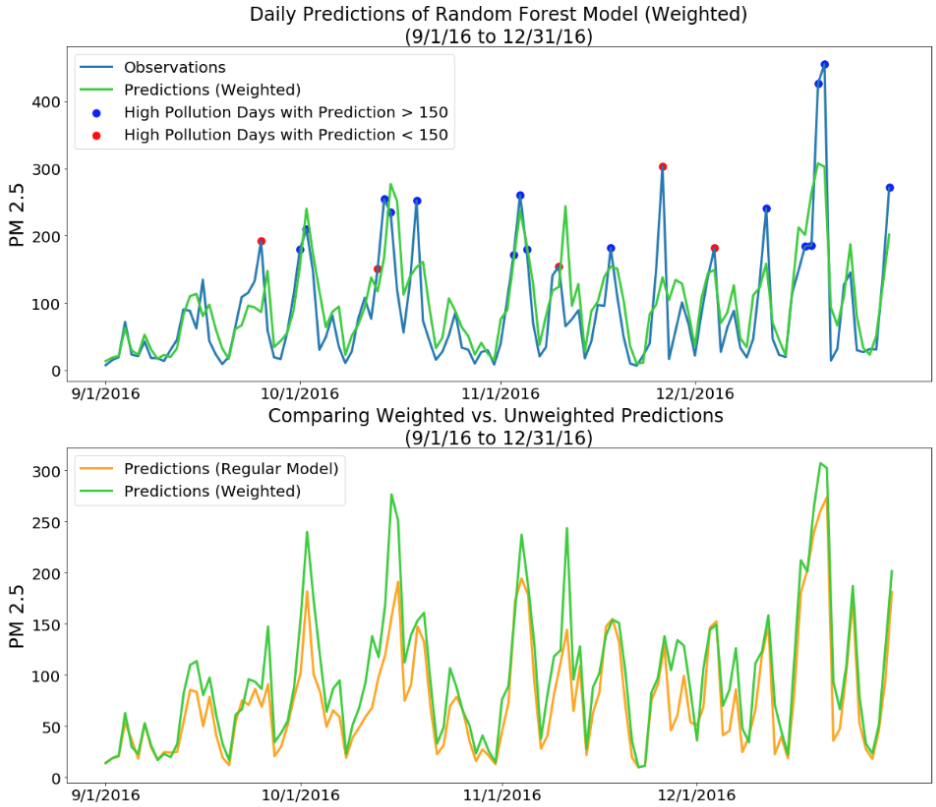

Figure 8: Predictions from Weighted Random Forests Model for test set.

Unfortunately, we did not have time to implement the weighted Random Forests model for the ensemble method listed in the performance table below. This extension is an area for future research.

Model	R^2	Precision	Recall
Random Forest	0.691	0.8	0.6
Random Forest (Weighted)	0.641	0.62	0.75
LSTM	0.477	0.357	0.25
GP RBF kernel	0.591	0.525	0.45
GP Rational Quadratic kernel	0.598	0.542	0.40
GP Matern kernel	0.605	0.57	0.50
GP + Random Forest	**0.71**	**0.82**	**0.65**

Table 2: Results for each model. The Gaussian Process and Random Forest together achieve the best performance in terms of R^2.

Other Models

During our evaluation we found that the LSTM performs very poorly on this task. The model tends to overfit fairly quickly as deep learning models are very data hungry. The GP performs better but still inferior to the static random forest model. This is to be expected as the GP doesn't account for any weather data. We also found the Matern kernel to give the best results. Our theory is that the wider tails of the kernel allow it to consider more information. The overall best model was combining the GP and the static random forest model.

4 Conclusion

Increasing pollution has caused millions of premature deaths in China. Theoretically, our model could be used by government officials in China to alert the general public about bad pollution. If citizens are better informed about future air quality, they will be able to prepare more effectively for the adverse effects of smog and therefore experience better health outcomes. Our work in modeling pollution in Beijing is very general and can be applied to other areas in China as well as other developing countries facing similar issues. Our best model combines a Gaussian Process with a static random forests model. While this model is able to perform well in terms of precision and recall, a single model that could capture both aspects of the data would be a more desirable solution. Another direction of future work is to create a model that performs inference. Whenever a machine learning system is used to make decisions, it is important to quantify our uncertainty as well.

Adding features from other data sources (besides the weather) might also improve the model's performance. For example, knowing the traffic patterns or industrial output of factories in Beijing would be very useful for predicting $PM_{2.5}$. Before hosting major international events (such as the 2008 Beijing Olympics), the Chinese government often resorts to last-ditch measures to reduce pollution levels, such as temporarily closing factories or banning certain cars from the roads. Our model would have been stronger had we considered these additional features. Finally, to keep in line with existing literature, we could have experimented with feed-forward neural networks rather than a random forests model.

Contributions

Brenton preprocessed the data and built the random forests model. Zach extracted weather data and built the Gaussian Process models, experimenting with a variety of kernels.

References

Berman L., 2017. National AQI Observations (2014-05 to 2016-12), Harvard Dataverse. Accessed via: https://doi.org/10.7910/DVN/QDX6L8

Chinese Ministry of Environmental Protection, 2012. Technical Regulation on Ambient Air Quality Index. (http://kjs.mep.gov.cn/hjbhbz/bzwb/jcffbz/201203/W020120410332725219541.pdf)

Chinese Ministry of Environmental Protection, 2012. GB-3095-2012: Ambient Air Quality Standards. (http://kjs.mep.gov.cn/hjbhbz/bzwb/dqhjbh/dqhjzlbz/201203/W020120410330232398521.pdf)

Corani G., 2005. Air quality prediction in Milan: feed-forward neural networks, pruned neural networks, and lazy learning. Ecological Modelling 185: 513-529.

Diaz-Robles L., Ortega J., Fu J., Reed G., Chow J., Watson J., Moncada-Herrera J., 2008. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. Atmospheric Environment 42, 8331-8340.

Feng X., Li Q., Zhu Y., Hou J., Jin L., Wang, J., 2015. Artificial neural networks forecasting of $PM_{2.5}$ pollution using air mass trajectory based geographic model and wavelet transformation. Atmospheric Environment 107, 118-128.

Fu M., Wang W., Le Z., Khorram M., 2015. Prediction of particular matter concentrations by developed feed-forward neutral network with rolling mechanism and gray model. Neural Computing & Applications 26, 1789-1797.

GBD MAPS Working Group, 2016. Special Report 20: Burden of Disease Attributable to Coal Burning and Other Major Sources of Air Pollution in China. Health Effects Institute.

Greenstone M., Ebenstein A., Fan M., He G., Zhou M., 2017. New evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River Policy. Proceedings of the National Academy of Sciences for the United States of America 201616784; DOI: 10.1073.

Hoi K.I., Yuen K.V., Mok K.M., 2008. Kalman Filter Based Prediction System for Wintertime PM 10 Concentrations in Macau. Global NEST Journal 10(2), 140-150.

Ni X.Y., Huang H., Du W.P., 2017. Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data. Atmospheric Environment 150, 146-161.

Ordieres J.B., Vergara E.P., Capuz R.S., Salazar R.E, 2005. Neural network prediction for fine particulate matter on the US-Mexico border in El Paso (Texas) and Ciudad Juarez (Chihuahua). Environmental Modelling & Software 20, 547-559.

Rohde R.A., Muller R., 2015. Air Pollution in China: Mapping of Concentrations and Sources. PLOS ONE 10(8), 1-14.

Sun W., Zhang H., Palazoglu A., Singh A., Zhang W., Liu S., 2013. Prediction of 24-hour-average PM 2.5 concentrations using a hidden Markov model with different emission distributions in Northern California. Science of the Total Environment 443, 93-103.

World Health Organization (WHO), 2018. Ambient (outdoor) air quality and health. (http://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health)

Appendix: Feature Importances

Below is a table of feature importances from the Random Forests model.

	Feature	Gini Coefficient
0	Wind Speed	0.231486
1	Beijing t-1	0.201913
2	wind_angle	0.201045
3	Baotou t-1	0.16234
4	Humidity	0.0786482
5	Harbin t-1	0.0315862
6	Xian t-1	0.0212367
7	Pressure	0.0161154
8	Temp.	0.0142671
9	Beijing t-2	0.014101
10	month	0.0119078
11	Qingdao t-1	0.00996284
12	weekday	0.00539194

Figure 9: Feature importances

The most significant features were wind speed, wind direction, and Beijing $PM_{2.5}$ from one day ago. Low wind speed is associated with higher pollution. Interestingly, pollution tends to be higher when

wind blows from the east and south; the mountains north and west of Beijing trap airborne pollutants when the wind comes from these directions. Based on our analysis of the data, Beijing's average $PM_{2.5}$ with a north wind is 70.5 $\mu g/m^3$; 114.7 for an east wind; 90.6 for a south wind; and 50.5 for a west wind.

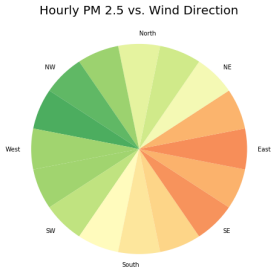

Figure 10: Hourly PM 2.5 vs. Wind Direction. Values range from 30 (dark green) to 120 (red)