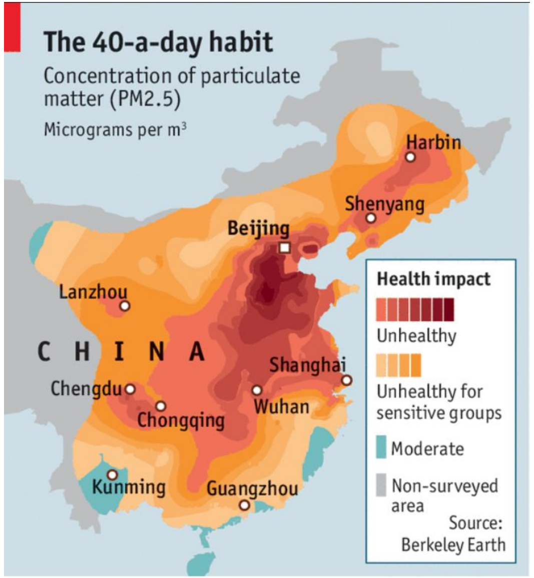
Predicting Air Pollution in Beijing

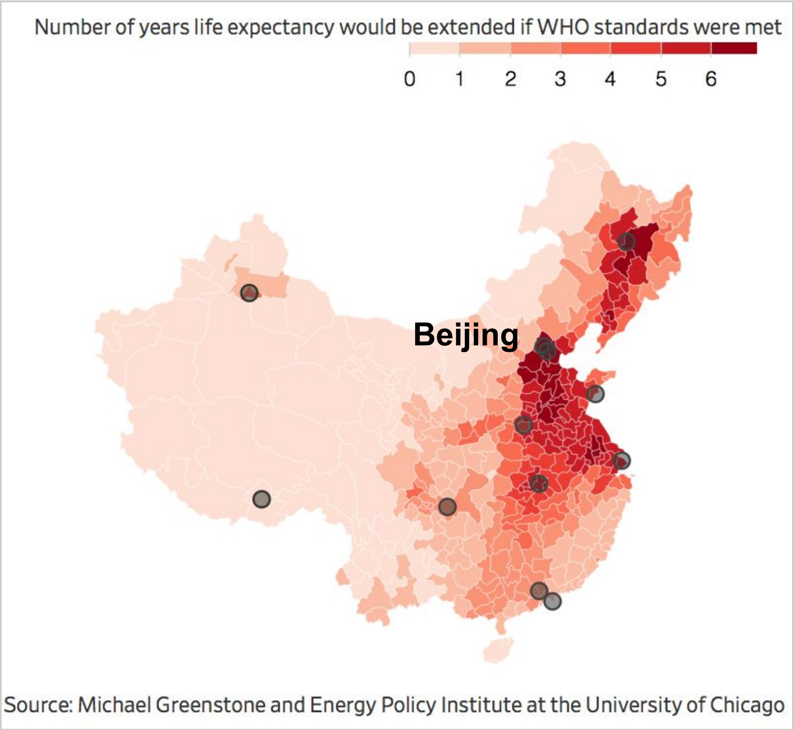
Brenton Arnaboldi and Zach Zhang



Economist.com

Introduction

- **Air pollution estimated to contribute to 1.6 million deaths/year in China** (Rohde, Muller 2015)
- **Poor air quality has decreased life expectancy in China by 3.5 years.** (Greenstone, 2017)
- A forecasting model for air quality might help people prepare more effectively for bad pollution. (Better public health outcomes).



Problem Definition

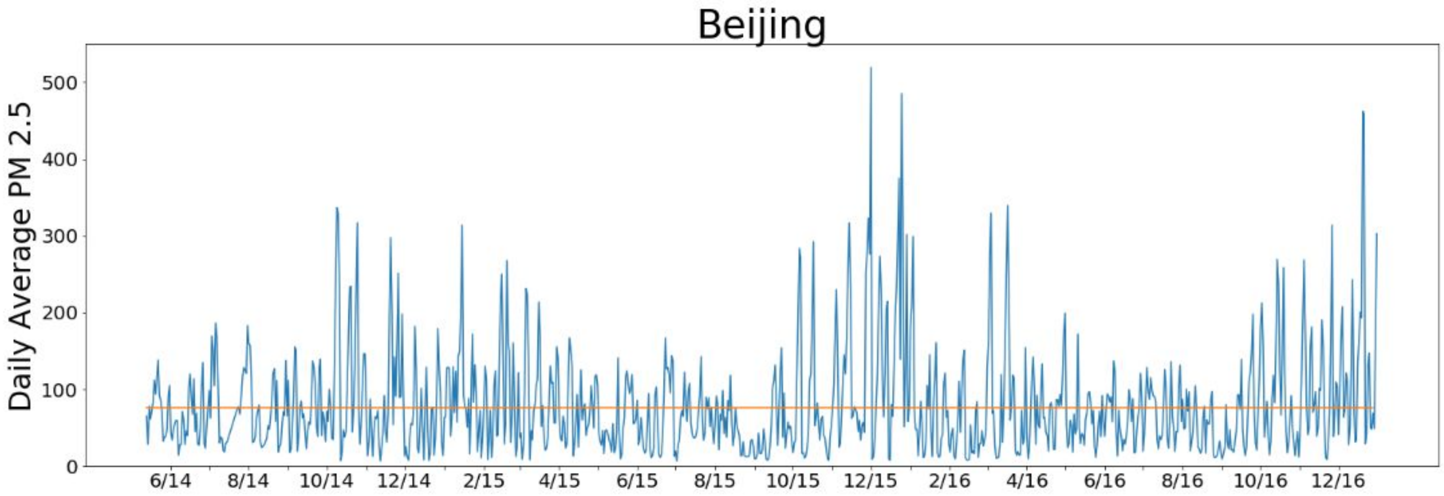
Predict daily average PM 2.5 concentration in Beijing

- Predictions are made one day in advance.
- PM 2.5 is particulate matter smaller than 2.5 micrometers (small but dangerous pollutant)

Models Tested: 1) Random Forests, 2) LSTM, 3) Gaussian Process

Data

- Air quality readings from government monitoring stations in China
  - Hourly data from May 13, 2014 to December 31, 2016 (964 days)
- Weather data
  - Scraped from Weather Underground
  - Temperature, Humidity, Wind Speed, Wind Angle, Pressure



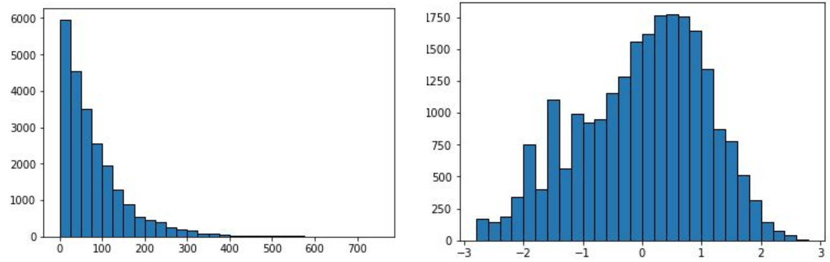
China's standard for PM 2.5 is 75 micrograms per cubic meter.

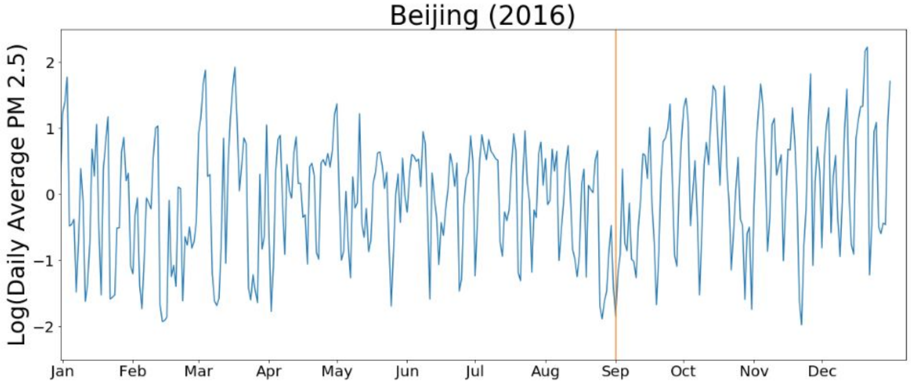
Note seasonal variation: pollution spikes in winter.

Data Preprocessing

- **Missing Data**: took weighted average of most recent observations
- Converted hour-level data into day-level data
- **Log Transform:** to transform daily PM 2.5 to a Gaussian distribution, we took the logarithm of PM 2.5, and then shifted the distribution to a mean of 0.
- **Training/Test Split**: ~**87% training, ~13% test**. (Test data from 9/1/2016 to 12/31/2016.)

Distribution of PM 2.5 (before and after transformation)

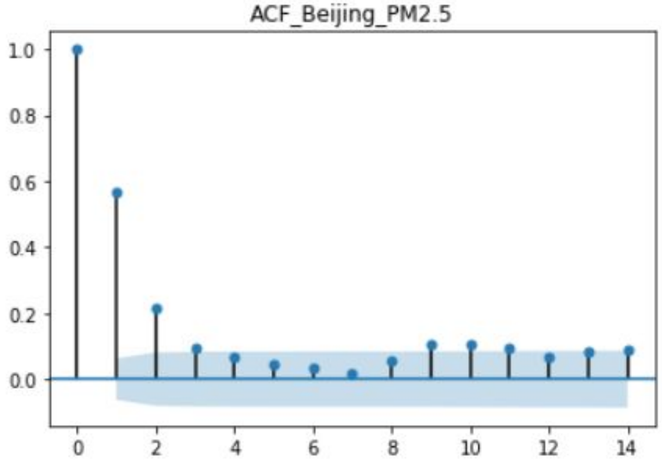




Model - Random Forests

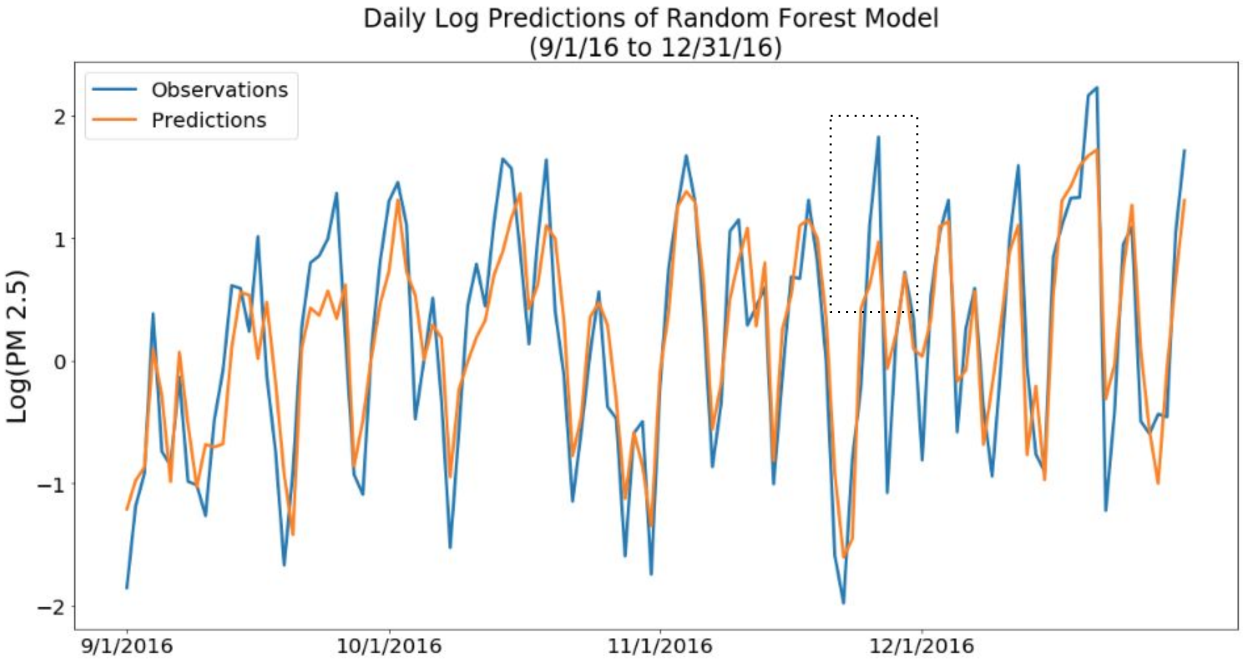
Features:

- ***Weather***
  - Temperature, humidity, wind speed and direction
- ***Autoregressive***
  - PM 2.5 in Beijing in the previous 2 days (d-1, d-2)
- ***Pollution from Other Cities***
  - PM 2.5 from Harbin, Xian, Qingdao, and Baotou from the previous day.



Significant correlation for d-1 and d-2.
By d-3, little correlation exists.

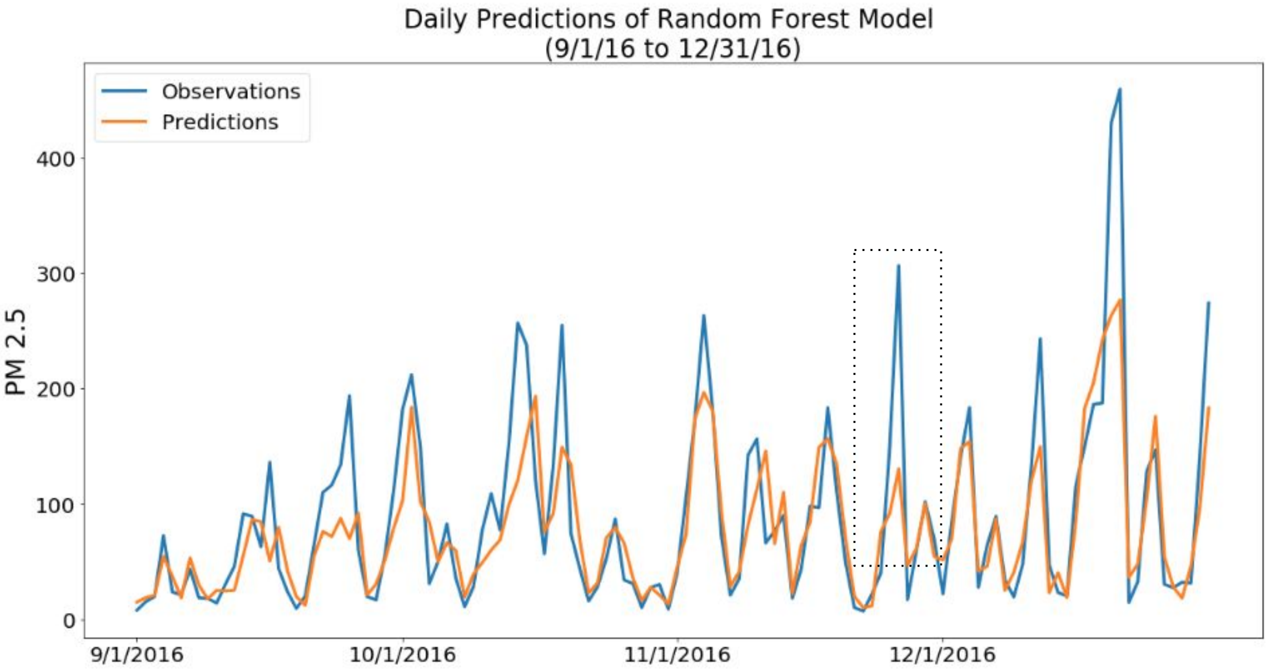
Model - Random Forests



	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

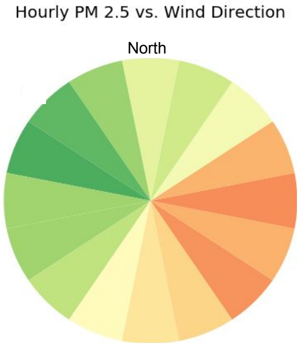
Train R^2: 89.5%
Test R^2: 78.3%

Model - Random Forests



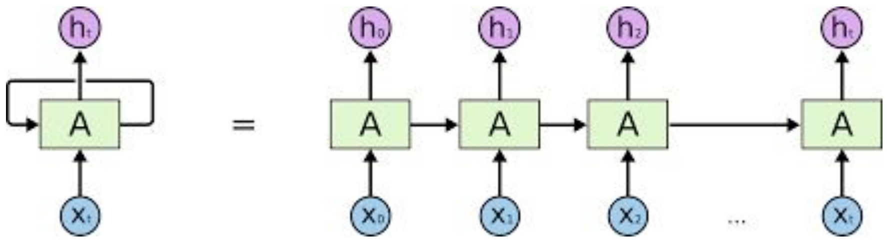
Train R^2: 85.8%
Test R^2: 69.1%

	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



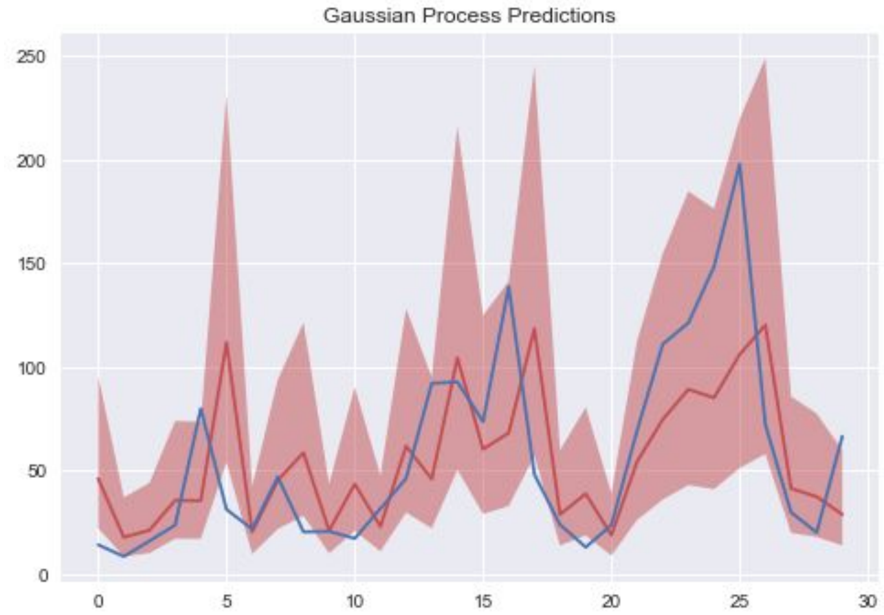
Model - LSTM

- Recurrent Neural Network
- Takes Time, AR, and Weather features at each timestep
- Very data hungry
    - Quickly overfits

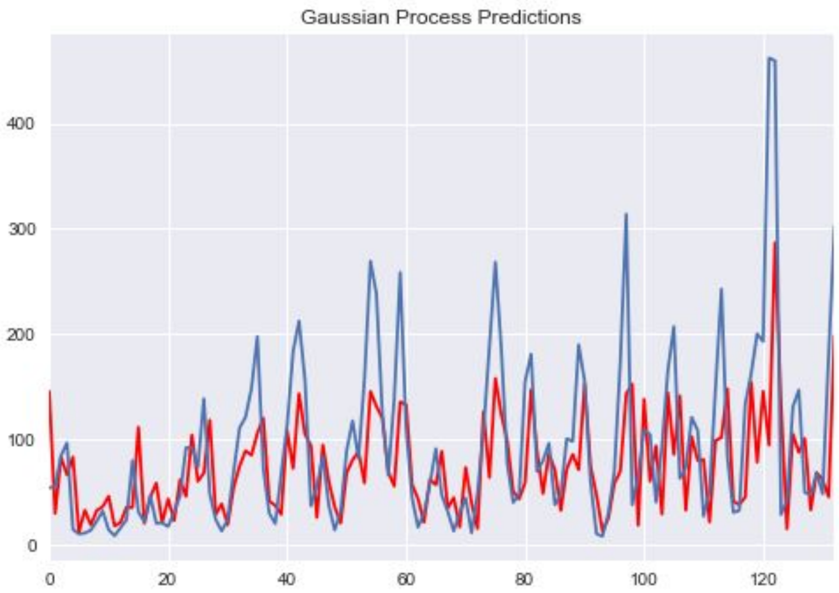


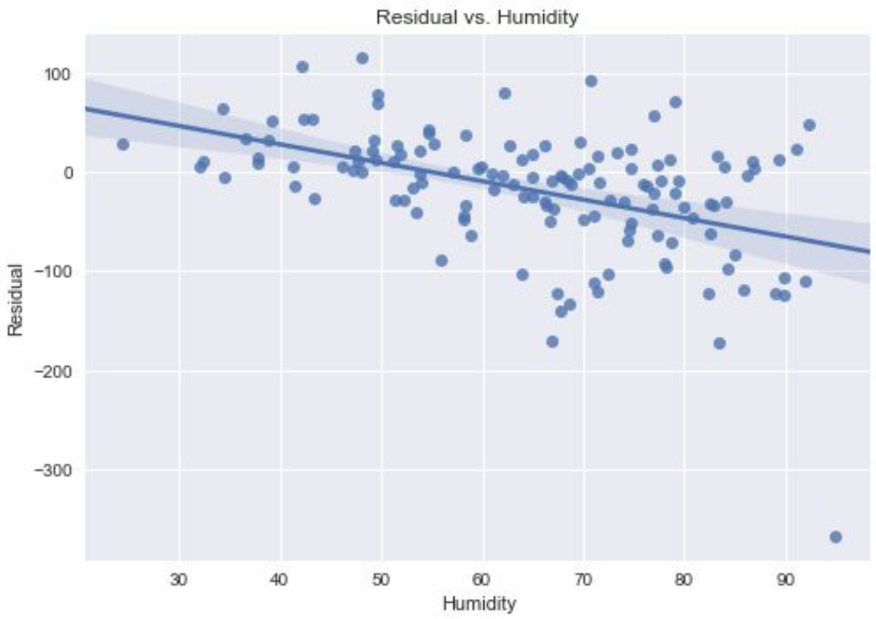
Model - Gaussian Process

- RBF kernel
- Non-parametric
  - Requires much less data
- Quantify Uncertainty
- Difficult to handle weather features



Model - Gaussian Process





Model - GP + Random Forest

- Provide predictions of GP to random forest as additional features
- Include the point estimate as well as the uncertainty
  - Model learns to rely on other features when GP is uncertain
- Exceeds all individual models

Results

- Last 4 months of data for testing

Model	R squared
Random Forest	0.691
LSTM	0.477
GP	0.591
Random Forest+GP	0.703

Conclusion and Next Steps

- We have developed a successful model for predicting air pollution in Beijing
- Government can issue warnings the day before to minimize exposure
- Get more training data
    - Only 964 records in dataset - 2017 data should now be available
- Experiment with more sophisticated kernel designs