Do I need a mask tomorrow ?

-- Predicting Degree of Pollution in Guangzhou, China

Yang Yu, Jing Wang, Chu Chu

Introduction

- **Objectives**
  - Predict the air pollution level of Guangzhou based on provided PM2.5 dataset
  - Create a warming system to alarm the public when tomorrow's air quality is predicted concerning
  - Educate the public and provide useful information for creating a brochure
- **Methods**
  - Prediction and classification with linear regression model
  - Data visualizations

Data Description

- Hourly PM2.5 data collected continuously in two sites of Guangzhou from 2011 to 2015
- Raw dataset contains 52584 observations
- 17 Variables
  - PM2.5 data
    - PM_City.Station, PM_5th.Middle.School, PM_US.Post
  - Continuous Covariates (10)
    - DEWP (Dew Point), HUMI (Humidity), PRES (Pressure), TEMP (Temperature), precipitation (hourly precipitation), Iws (Cumulated wind speed), Iprec (Cumulated precipitation)
  - Categorical Covariates (7)
    - No, Year, Month, Day, Hour, Season, cbwd (Combined wind direction)

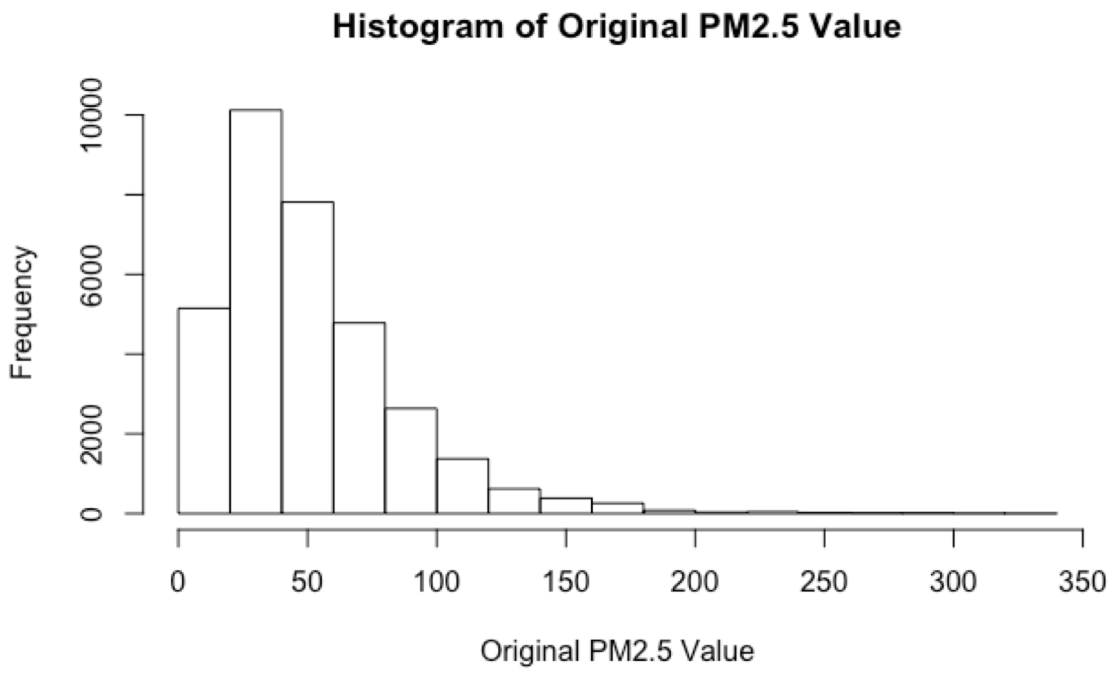
What have we done:

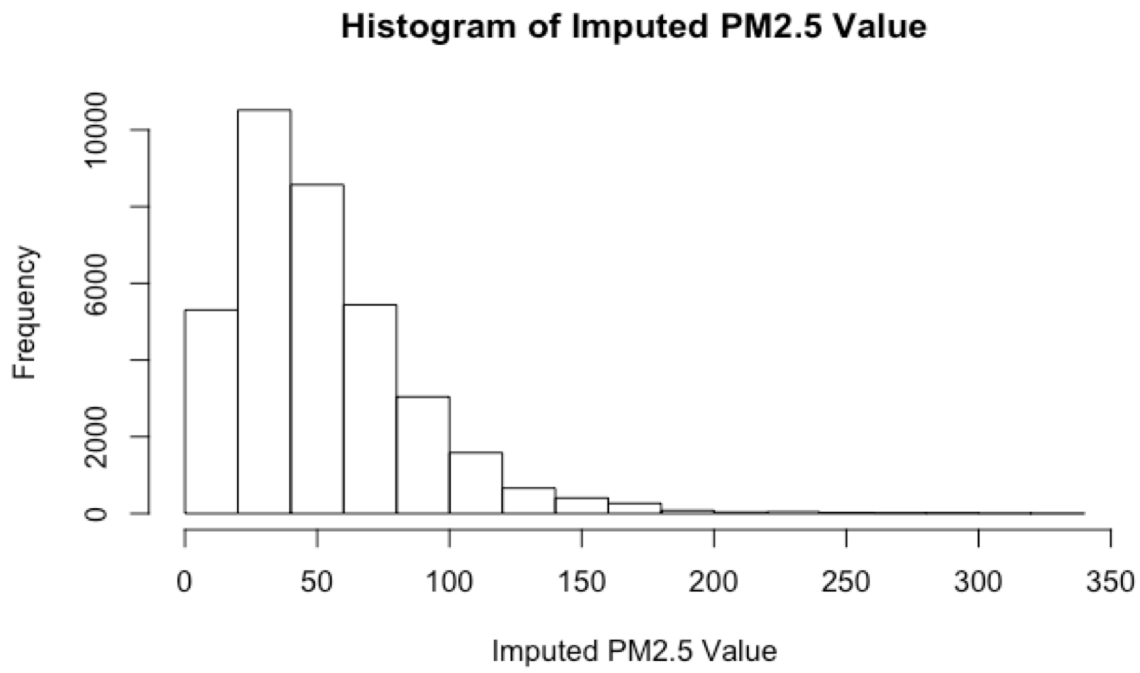
- Data cleaning & data description
- Prediction model construction, selection & evaluation
- PM2.5 level classification & high pollution warning system
- Public education

Data Cleaning

- Two collection points have identical measurements: city station and US post
- → We ignored one of them during our analyses because including both won't give extra information
- Continuous missing data for middle school station from year 2011 to early 2013
- → We ignored these messing values and used the average PM2.5 value to conduct our analyses
- Prediction doesn't allow us to use data from the same day
- →We will look at relationships between current day's PM2.5 value and factor values collected from the previous day
- Some random missing data appear in PM2.5 value column
- →We use K Nearest Neighbors method to imputed the whole data

Examine Imputed Data Distribution





Prediction model

Our Goal: Corrected predict and classify PM2.5 values into 3 groups:

- Group 1: PM2.5 < 75 (mild pollution) – no warning
- Group 2: 75<PM2.5<150 (medium pollution) – WARNING
- Group 3: 150 <PM2.5 (severe pollution) – WARNING

Models tested:

- Linear Regression
- Decision Trees
- Random Forest

Model Performance in Classification

Linear Regression	Pred Group 1 (PM2.5<75)	Pred Group 2 (75<PM2.5<150)	Pred Group 3 (150<PM2.5)
Actual Group 1 (PM2.5<75)	212	22	0
Actual Group 2 (75<PM2.5<150)	80	96	2
Actual Group 3 (150<PM2.5)	6	31	1

Tree	Pred Group 1 (PM2.5<75)	Pred Group 2 (75<PM2.5<150)	Pred Group 3 (150<PM2.5)
Actual Group 1 (PM2.5<75)	223	11	0
Actual Group 2 (75<PM2.5<150)	109	69	0
Actual Group 3 (150<PM2.5)	16	21	1

Random Forest	Pred Group 1 (PM2.5<75)	Pred Group 2 (75<PM2.5<150)	Pred Group 3 (150<PM2.5)
Actual Group 1 (PM2.5<75)	218	16	0
Actual Group 2 (75<PM2.5<150)	99	79	0
Actual Group 3 (150<PM2.5)	15	23	0

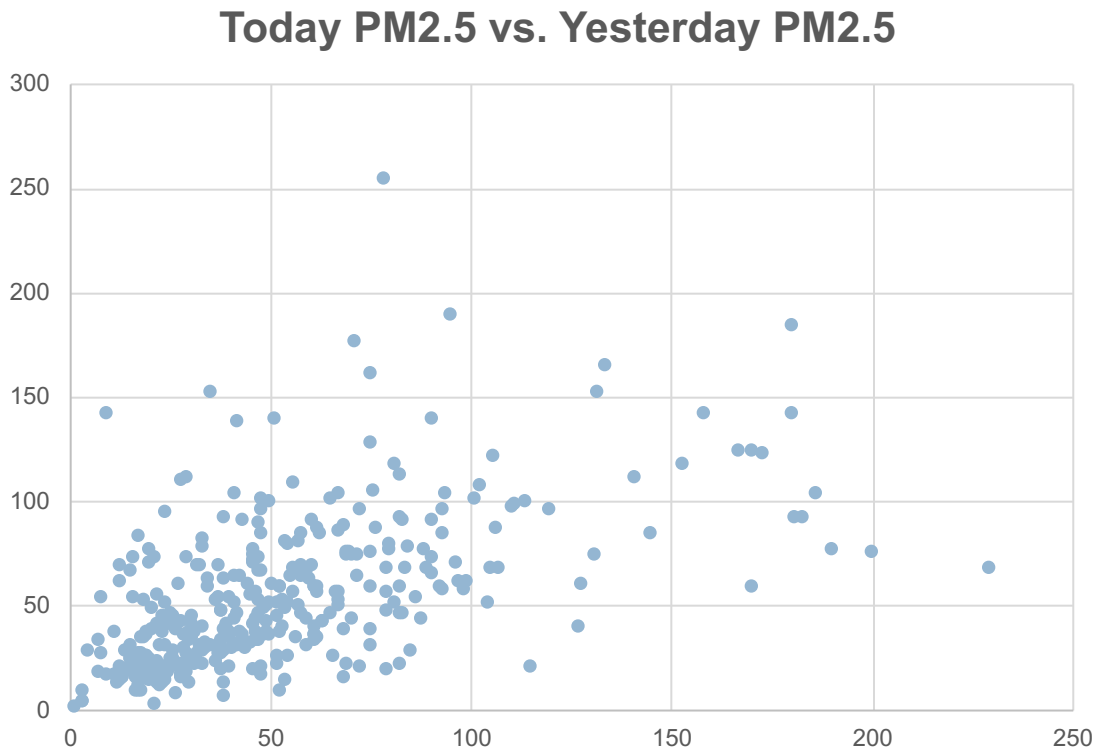
Problem: Hard to predict Group3 data

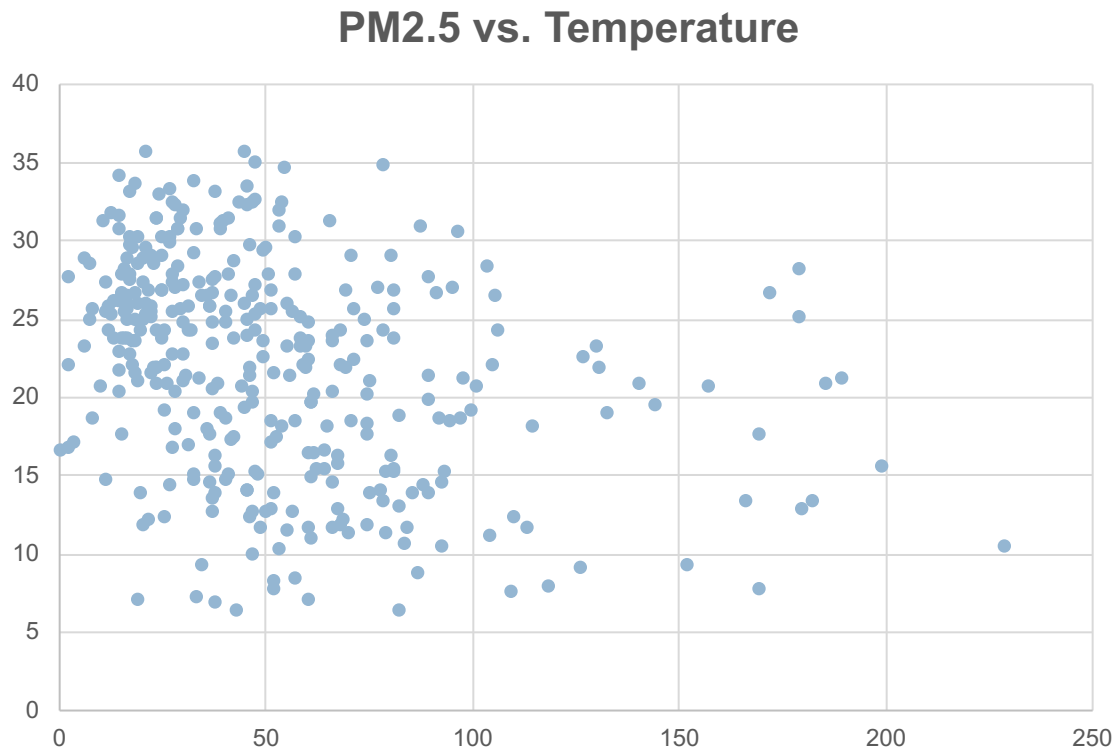
Model Comparison

Model Type	Redundant	Accuracy	Adjusted Accuracy	Warning Error
Linear Regression	0.05	0.69	0.76	0.19
Tree	0.02	0.65	0.70	0.28
Random Forest	0.04	0.66	0.71	0.25

- **Redundant**: Gave warning when no warning is needed
- **Accuracy**: All correct classification
- **Adjusted Accuracy**: Sent out warnings when needed, no matter the type of warning
- **Warning Error**: All incorrect classification

Why isn’t our prediction desirable?





- Large values are too spread out for our model to predict
- Only about 2% observations are in group 3 (PM2.5 > 150)

Create A Warning System

- To improve our classification accuracy and lower our classification error using our predicted data, we adjusted our classification cutoffs to create a more sensitive warning system:

- Group 1: PM2.5 < 60 (mild pollution) – no warning
- Group 2: 60<PM2.5<130 (medium pollution) – WARNING
- Group 3: 130 <PM2.5 (severe pollution) – WARNING

Final Model Performance

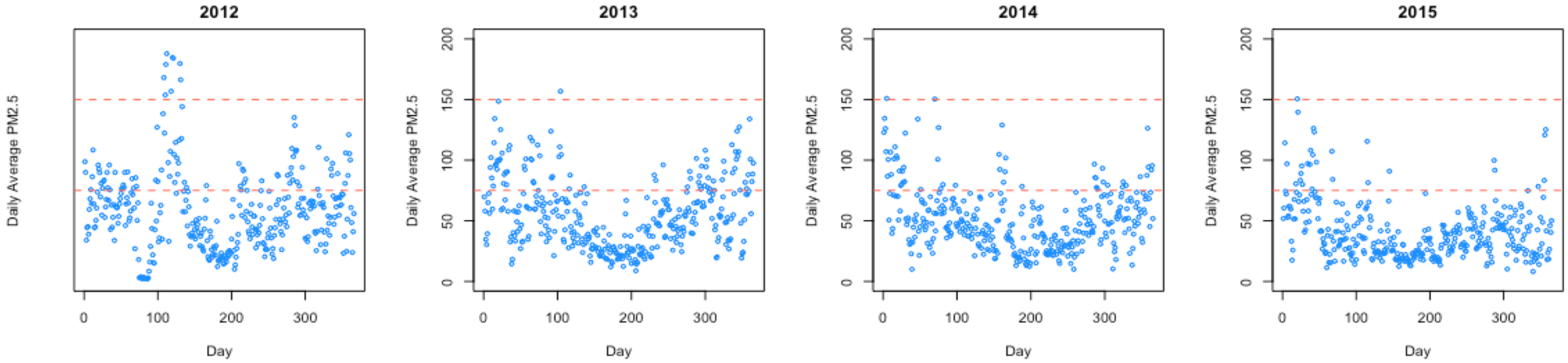
Model Type	Redundant	Accuracy	Adjusted Accuracy	Warning Error
Linear Regression	0.14	0.71	0.79	0.07

- **Redundant**: Gave warning when no warning is needed
- **Accuracy**: All correct classification
- **Adjusted Accuracy**: Sent out warnings when needed, no matter the type of warning
- **Warning Error**: All incorrect classification

Educate the Public

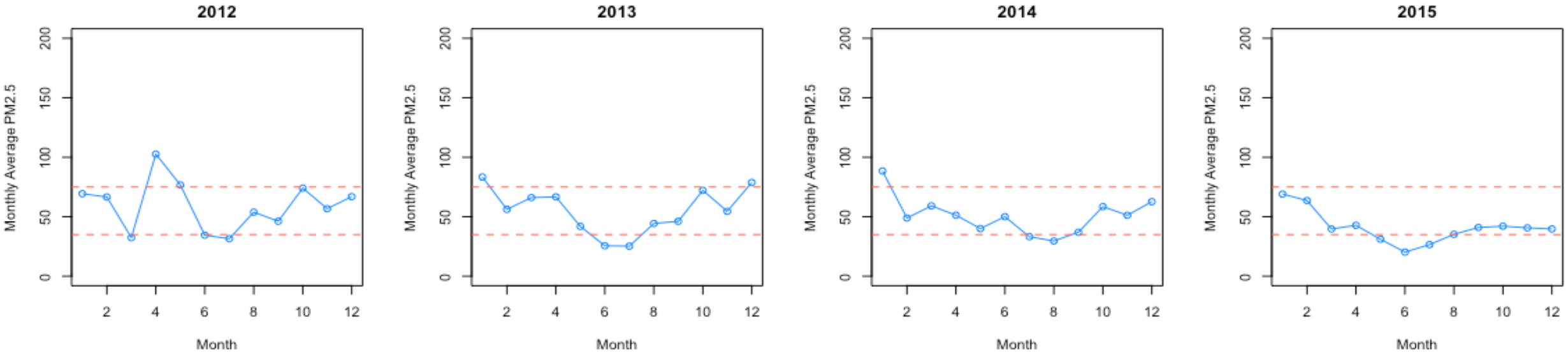
- 1) We create data visualizations to illustrate general trends in PM2.5 values
- 2) We give suggestions on creating an educational pamphlet for the residents of Guangzhou

Guangzhou Air Pollution Trend, 2012-2015



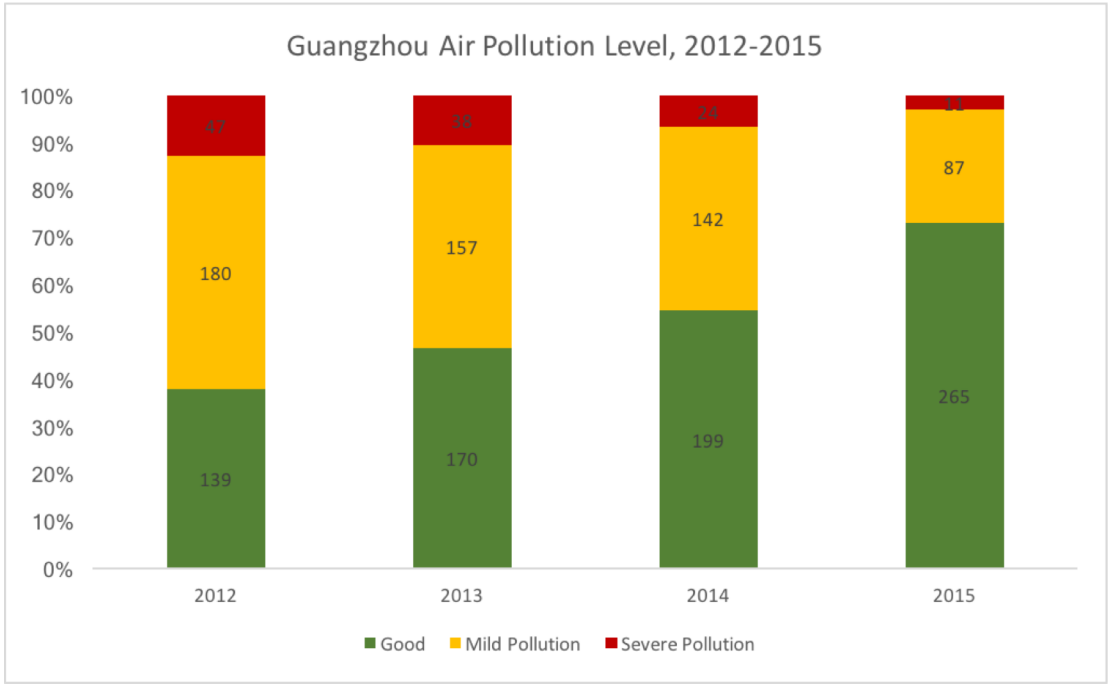
- **Trend over 4 years & Trend within a year**
  - U-shape patterns from 2013 to 2015
    - Lower PM2.5 during summer and higher PM2.5 during winter
  - Extremely severe pollution during the spring of 2012

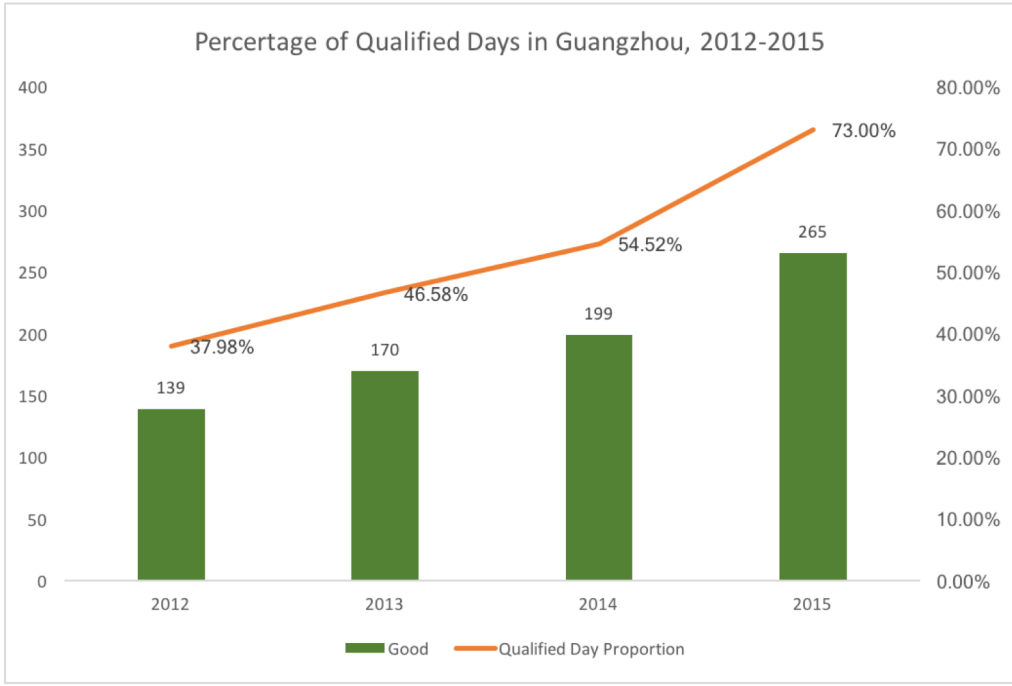
Guangzhou Air Pollution Monthly Trend, 2012-2015



- **Monthly trend over 4 years & Monthly trend within a year**
  - The air quality is getting better over the most recent four years on average
  - The air quality is best during summer (May, June, July, August and September)
  - Average PM2.5 decline over time for most months except for January and February from 2013 to 2015
    - Perhaps more traffic around the Spring Festival

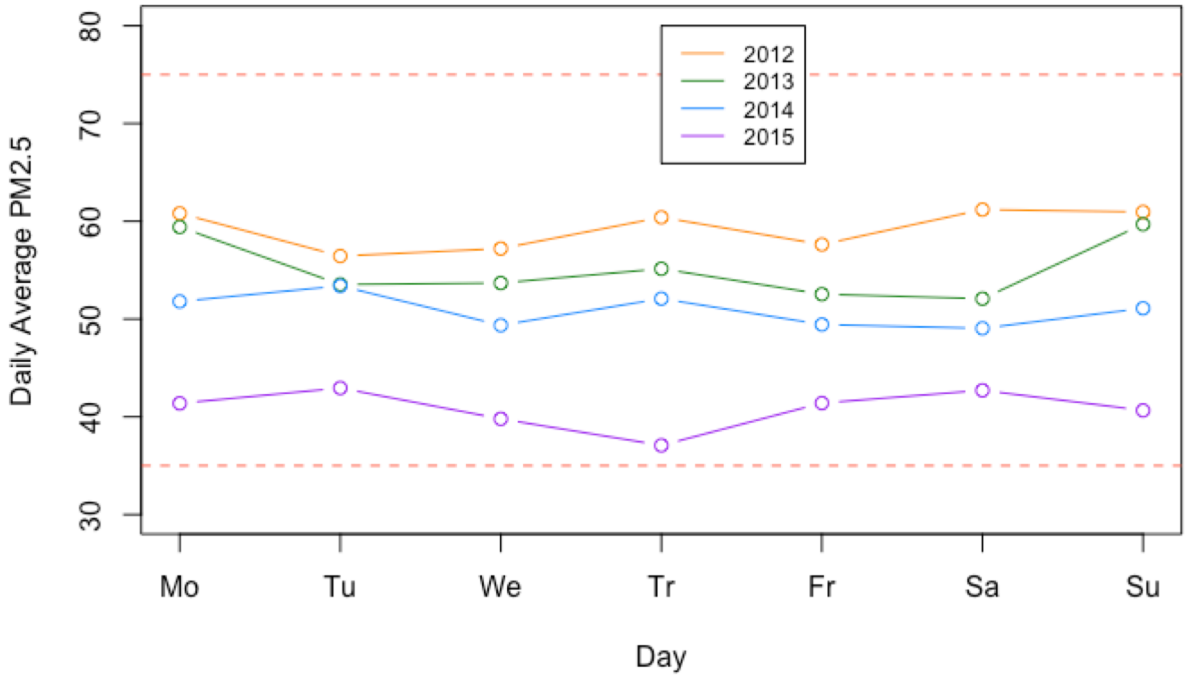
Guangzhou Air Pollution Trend, 2012-2015





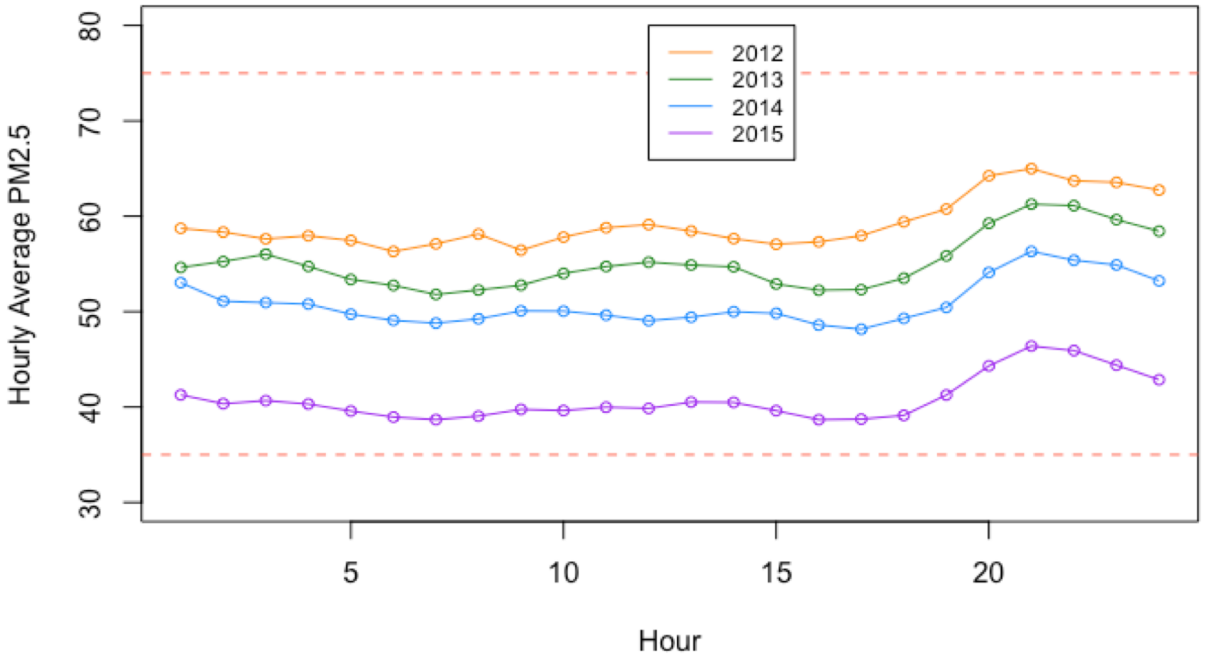
- **Trend over 4 years**
  - The proportion of qualified days is increasing from 2012 to 2015
  - The number of polluted (75<PM2.5 <150) and heavily-polluted (PM2.5 >150) days is declining over time
  - Polices on mitigating air pollution in Guangzhou have effects

Guangzhou Air Pollution Weekly Trend, 2012-2015



- Trend within a week
- No significant difference among days in a week

Guangzhou Air Pollution Daily Trend, 2012-2015



- **Trend within a day**
  - Air pollution level increases from morning to evening
  - Reaches peak during early night and decreases over the night
  - Traffic peak hours are usually 6-9 pm
- Average PM2.5 level is decreasing from 2012 to 2015

Conclusion

- The air quality of Guangzhou is getting better based on PM2.5 data from 2012 to 2015
  - The policies and regulations for mitigating air pollution in Guangzhou are helpful
- Air pollution is most severe during winter
  - Recommendation: reduce travelling and outdoor activities and be prepared with masks in January and February
- Air quality is best during summer
  - Recommendation: increase outdoor activities during May to September
- Air pollution is most severe during early night of a day
  - Recommendation: increase outdoor activities before evenings and be prepared with masks when going out during nights

Suggestions

- Conduct more research on severely polluted days (PM2.5>150)
- Collect data for other air quality measurements (e.g. SO2, NO)
- Reasons for missing records

- Data source
  - https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities