YIFAN JIANG
AMY ZHANG
MELODY FENG
JASON LEE
KAI HAYDEN

THE UNIVERSITY OF CHICAGO

Beijing Air Quality

Time Series Analysis and Forecasting (MSCA 31006-1)

GROUP 7

YIFAN JIANG
AMY ZHANG
MELODY FENG
JASON LEE
KAI HAYDEN

THE UNIVERSITY OF CHICAGO

Contents

GROUP 7

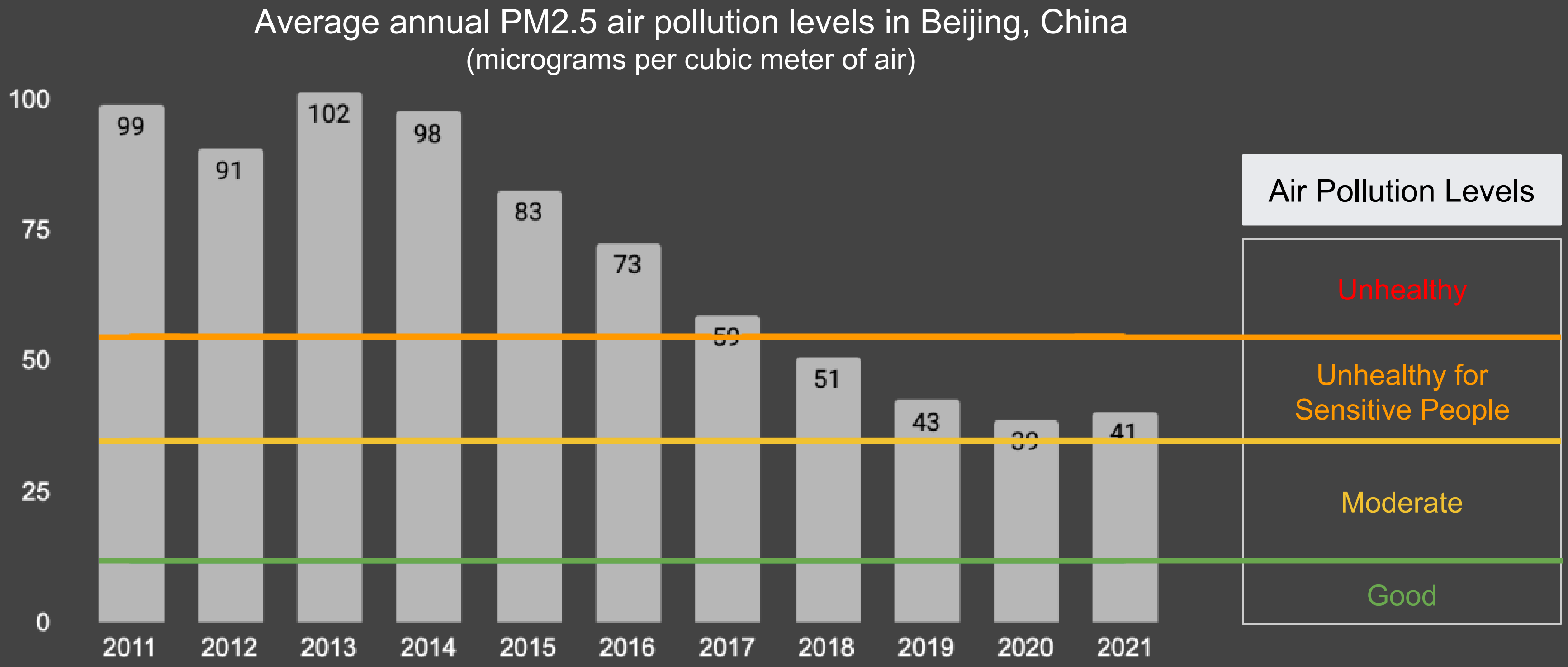
YIFAN JIANG
AMY ZHANG
MELODY FENG
JASON LEE
KAI HAYDEN

Business Problem

01

GROUP 7

Problem



Source: Statista

Goal & Business Value

Goal: Build a time series model that predicts PM2.5 daily levels

1. Obtain PM2.5 measures in Beijing from March 2013 to June 2017
2. Fit 4 models: ARIMA, ARIMA errors, VAR, and TBATS to the PM2.5 time series
3. Cross-validate best fitted model from each model type
4. Recommend the best performing model

Business Value

1. Inform citizens on days that are suitable for outdoor activities
2. Simulate air quality trend following government interventions and regulations

YIFAN JIANG
AMY ZHANG
MELODY FENG
JASON LEE
KAI HAYDEN

THE UNIVERSITY OF CHICAGO

Dataset & Cleaning

02

GROUP 7

Data Summary

- ➔ **File Type:** CSV file
- ➔ **File Size:** 2,598 kB
- ➔ **Data Shape:** 18 Columns × 35,064 Rows
- ➔ Using data from **Wanliu** station

Variables	Type	Description	Format
Year, Month, Day, Hour	Integer	Hourly data, from 2013/3 0:00 to 2017/2 23:00	2013, 3, 1, 1
PM2.5	Float	Hourly PM2.5 concentration (ug/m^3)	80.25
PM10	Float	Hourly PM10 concentration (ug/m^3)	120.25
Temp	Float	Hourly temperature measurements in Celsius	-1.1
Pres	Float	Hourly pressure measurements (hPa)	1023.2
Rain	Float	Hourly precipitation (mm)	0

Data Cleaning

Missing Values

Linear Interpolation with values before and after NA values

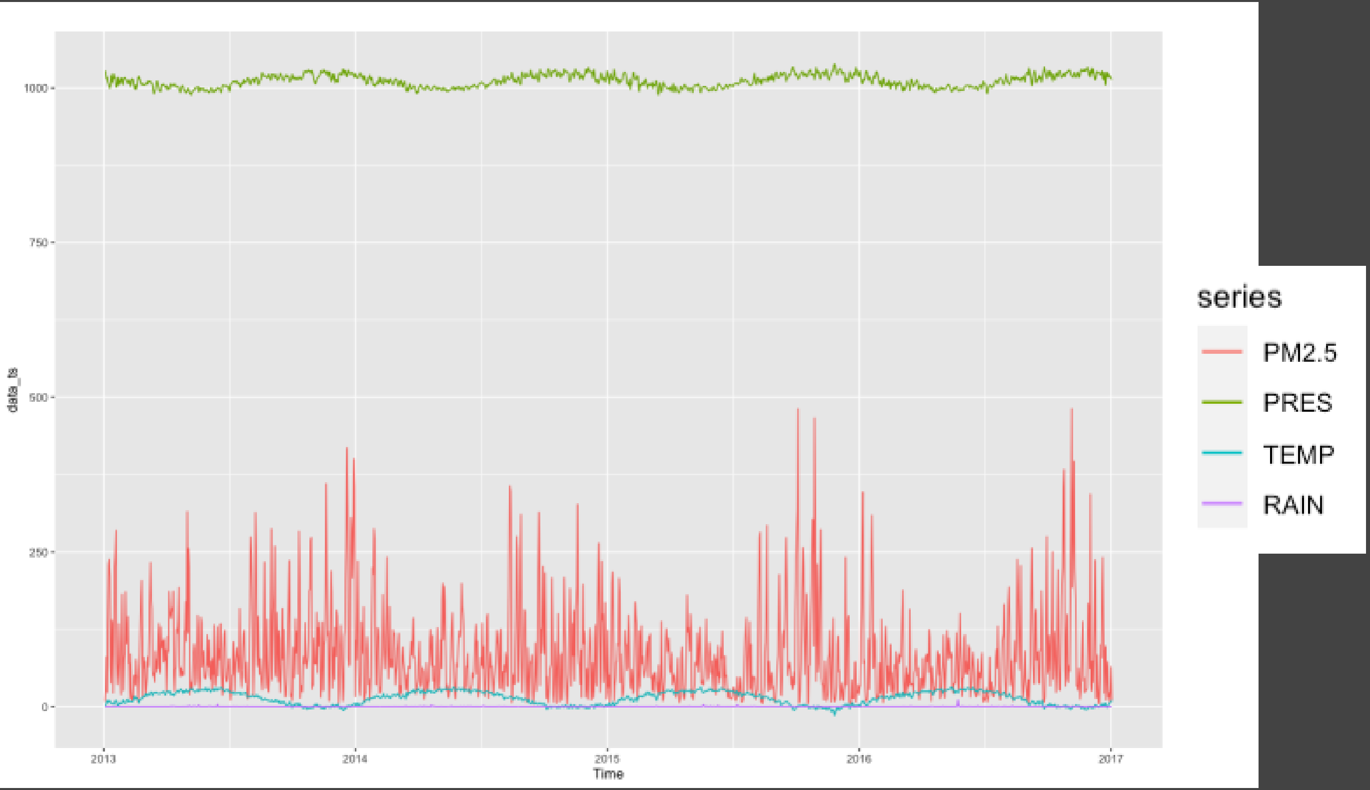
Too Many Data Points

Take average by date to transform hourly data to daily data

After Cleaning

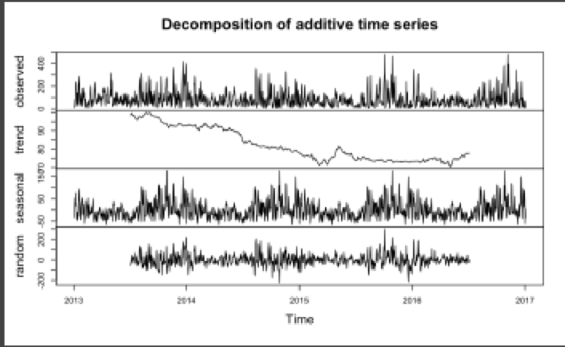
6 Columns × 1461 Rows

Data Plot of daily PM2.5, pressure, temperature and rai

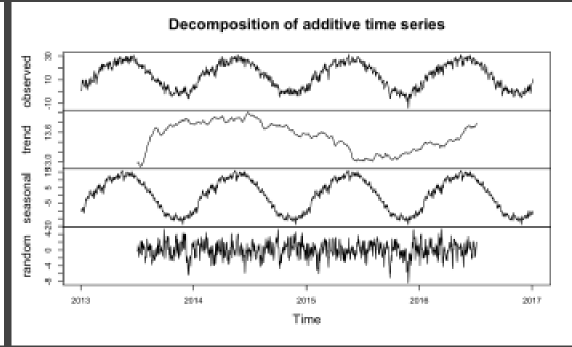


Decomposition

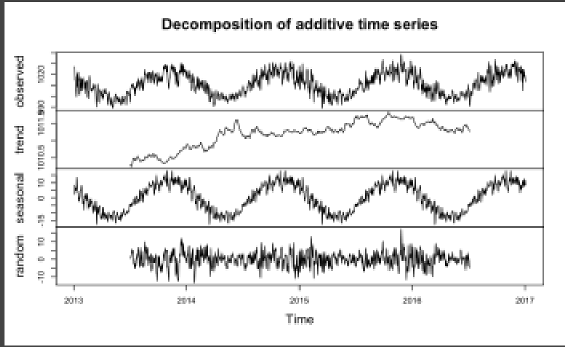
PM 2.5



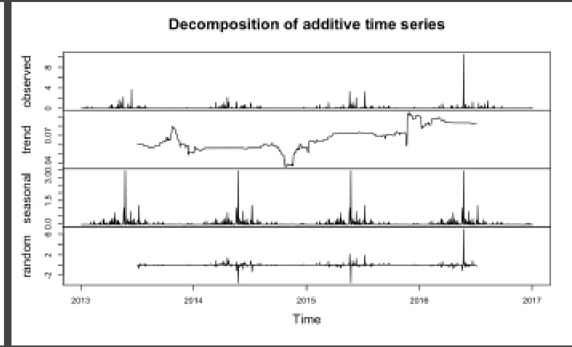
Temperature



Pressure



Rain



Hypothesis Testing

Variable	Augmented Dickey-Fuller Test	KPSS Test for Level Stationarity	KPSS Test for Trend Stationarity
PM 2.5	P-value < 0.05 Stationary	P-value < 0.05 Non-stationary	P-value > 0.05 Stationary
Temperature	P-value > 0.05 Non-stationary	P-value < 0.05 Non-stationary	P-value < 0.05 Non-stationary
Pressure	P-value > 0.05 Non-stationary	P-value < 0.05 Non-stationary	P-value < 0.05 Non-stationary
Rain	P-value < 0.05 Stationary	P-value > 0.05 Stationary	P-value > 0.05 Stationary

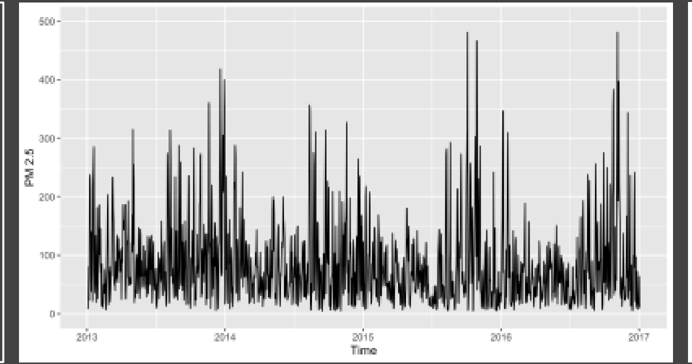
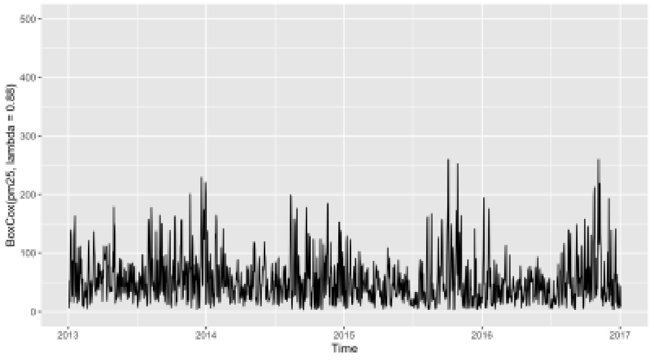
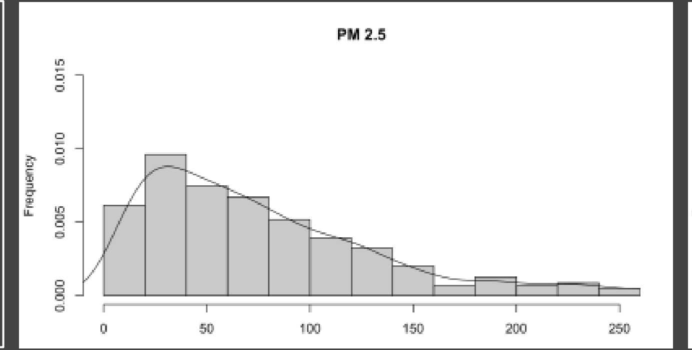
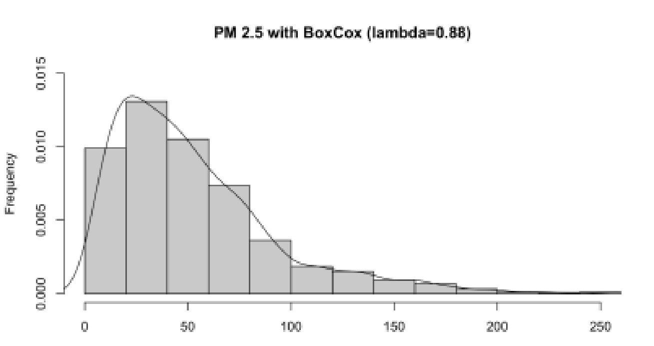
YIFAN JIANG
AMY ZHANG
MELODY FENG
JASON LEE
KAI HAYDEN

Experimental Results & Analysis

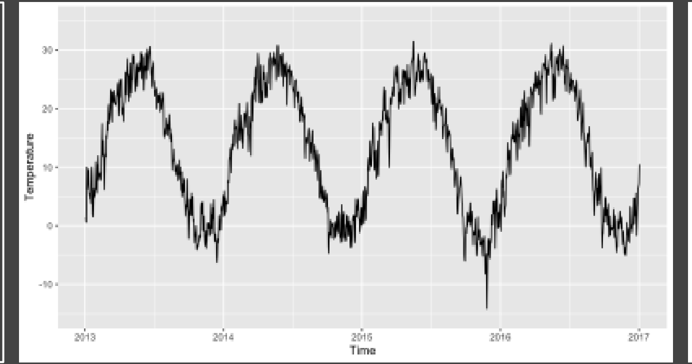
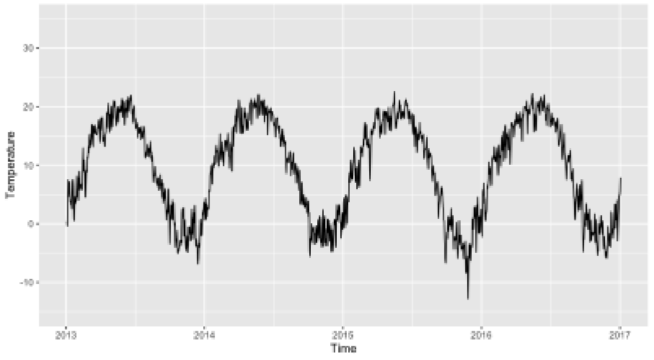
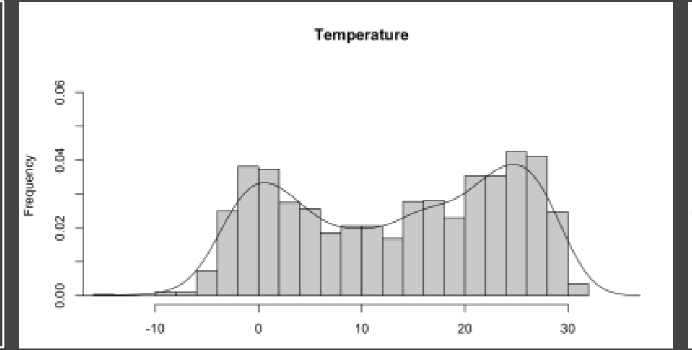
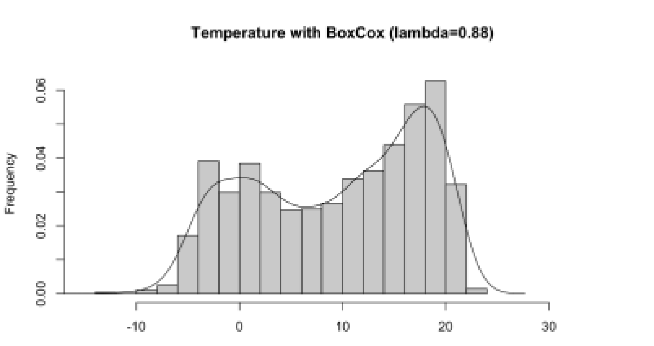
03

GROUP 7

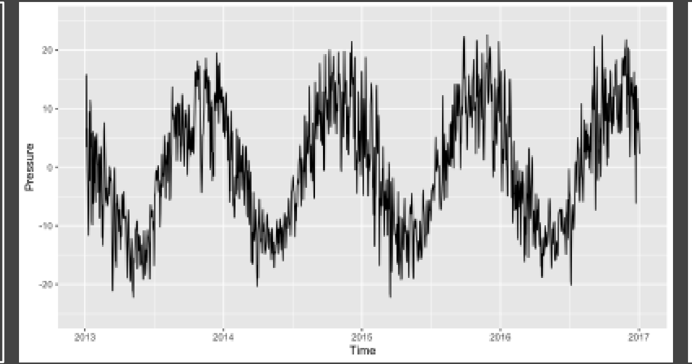
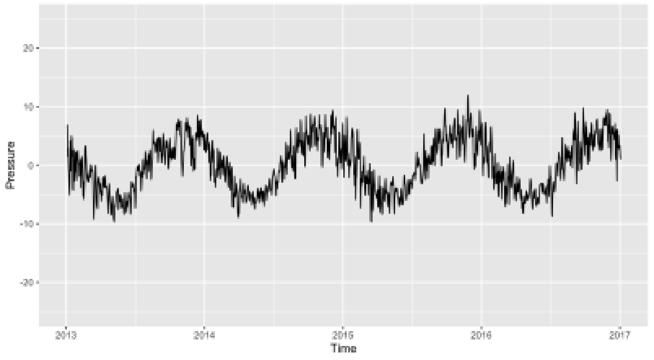
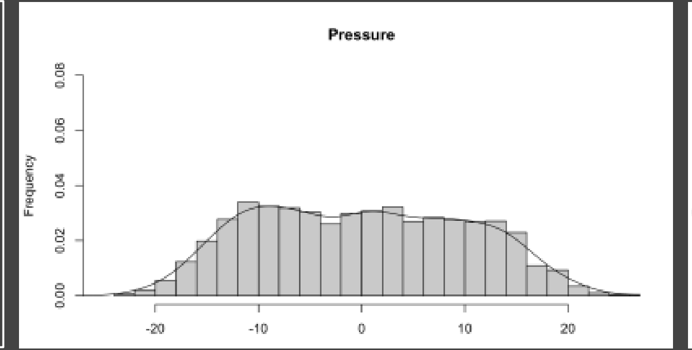
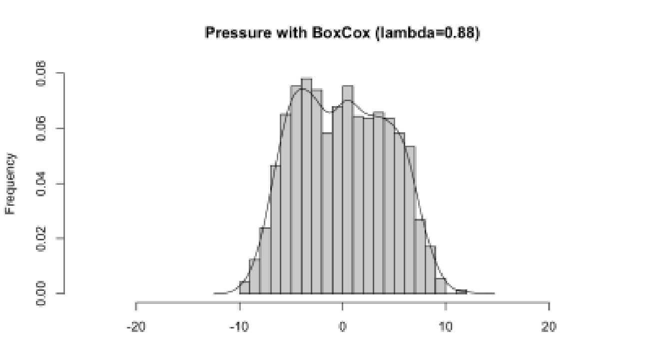
Box Cox Transform - PM2.5

	Before Transformation	After Transformation
Time Series Plot		
Histogram		

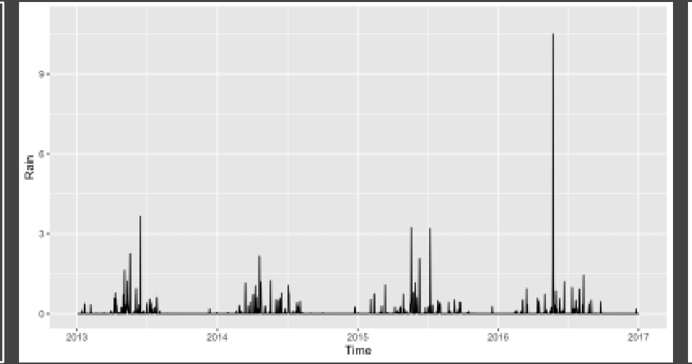
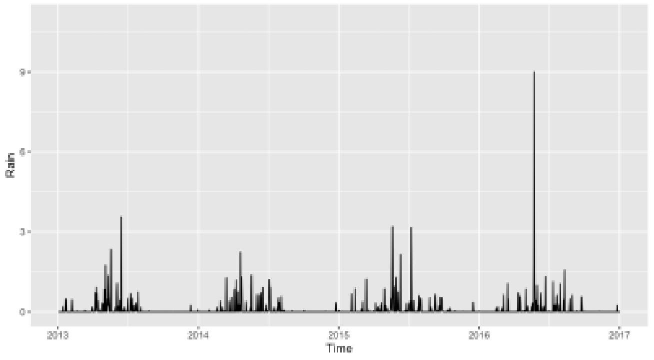
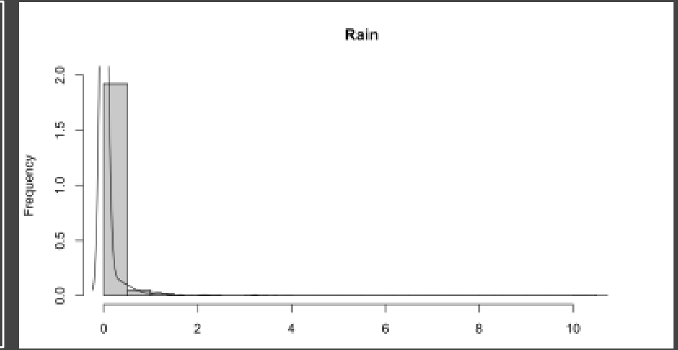
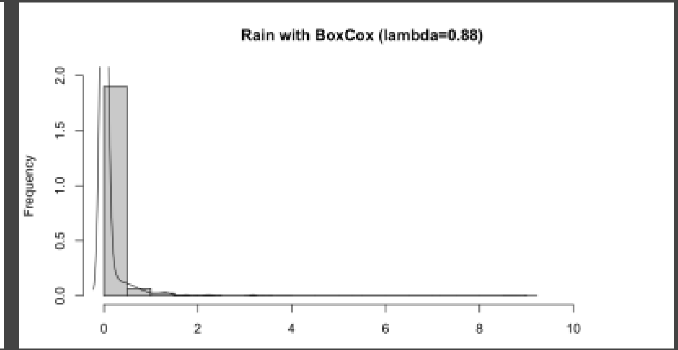
Box Cox Transform - Temperature

	Before Transformation	After Transformation
Time Series Plot		
Histogram		

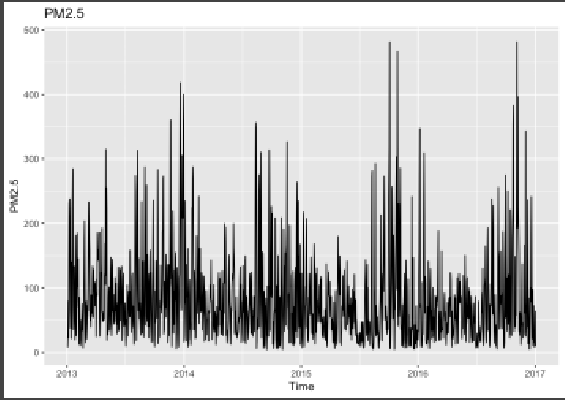
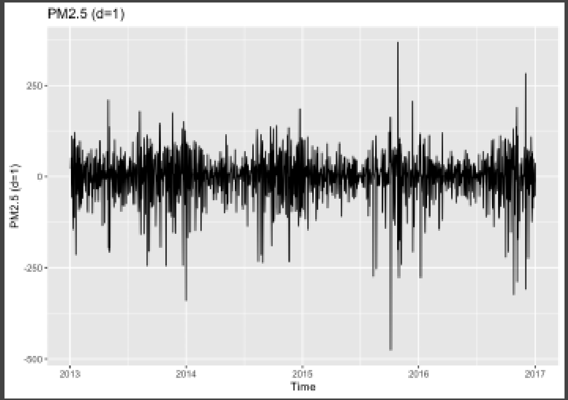
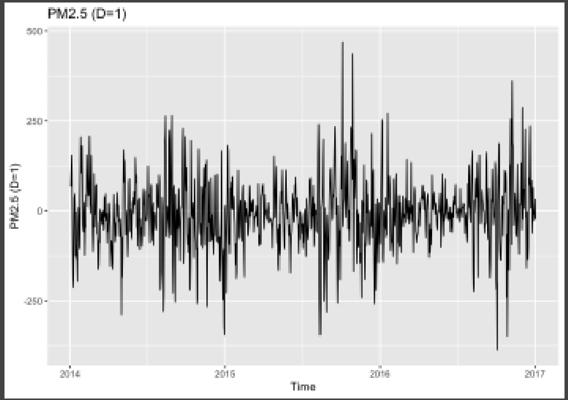
Box Cox Transform - Pressure

	Before Transformation	After Transformation
Time Series Plot		
Histogram		

Box Cox Transform - Rain

	Before Transformation	After Transformation
Time Series Plot		
Histogram		

Differencing - PM2.5

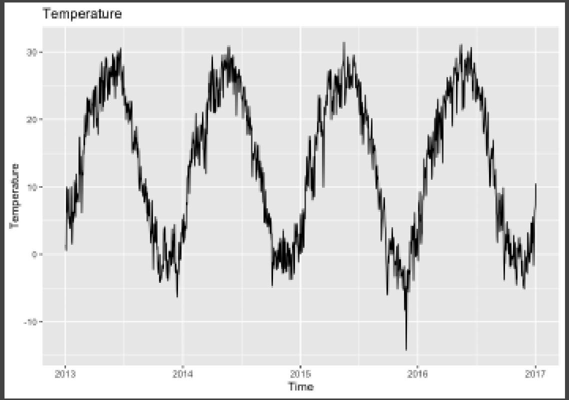
Before Differencing	First Order Differencing	Seasonal Differencing
		

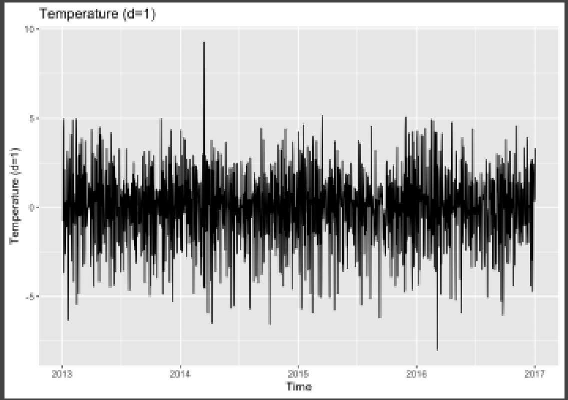
Augmented Dickey-Fuller Test Statistics

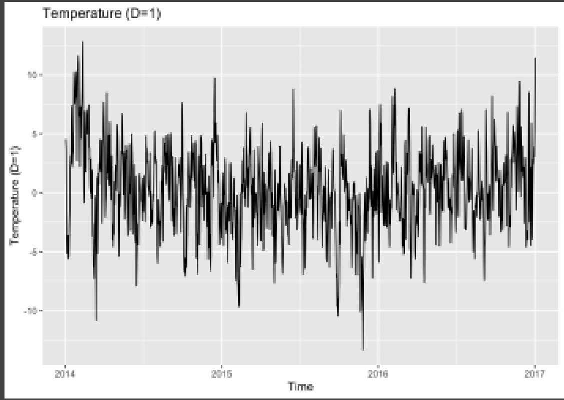
-9.029	-18.598	-10.152
Stationary	Stationary	Stationary

Differencing - Temperature

Before Differencing	First Order Differencing	Seasonal Differencing





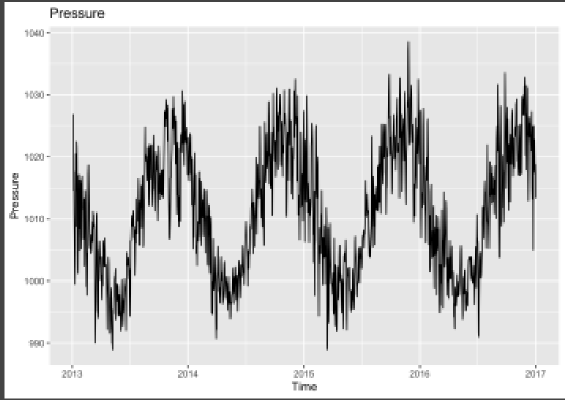


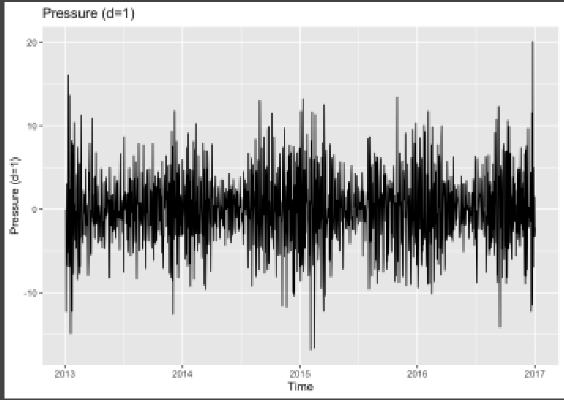
Augmented Dickey-Fuller Test Statistics

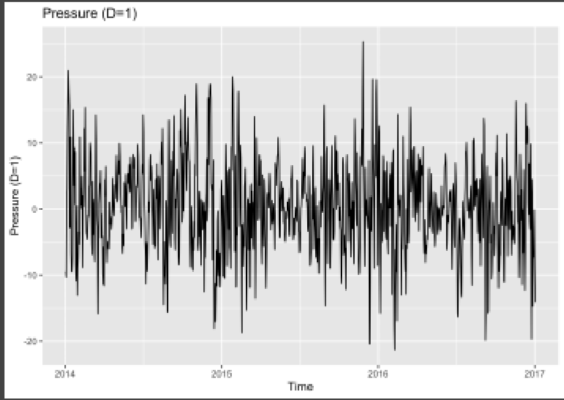
-1.944	-12.845	-7.399
Non-stationary	Stationary	Stationary

Differencing - Pressure

Before Differencing	First Order Differencing	Seasonal Differencing



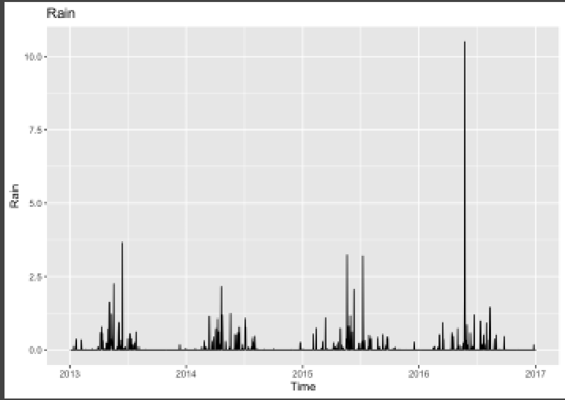
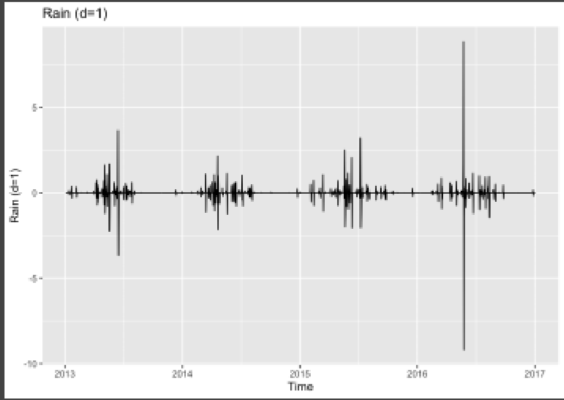
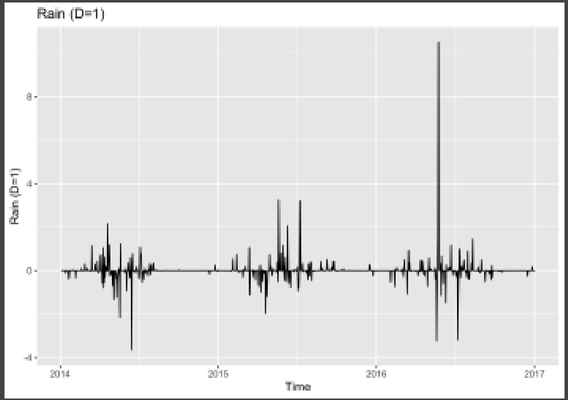




Augmented Dickey-Fuller Test Statistics

Before Differencing	First Order Differencing	Seasonal Differencing
-2.685	-17.126	-9.686
Non-stationary	Stationary	Stationary

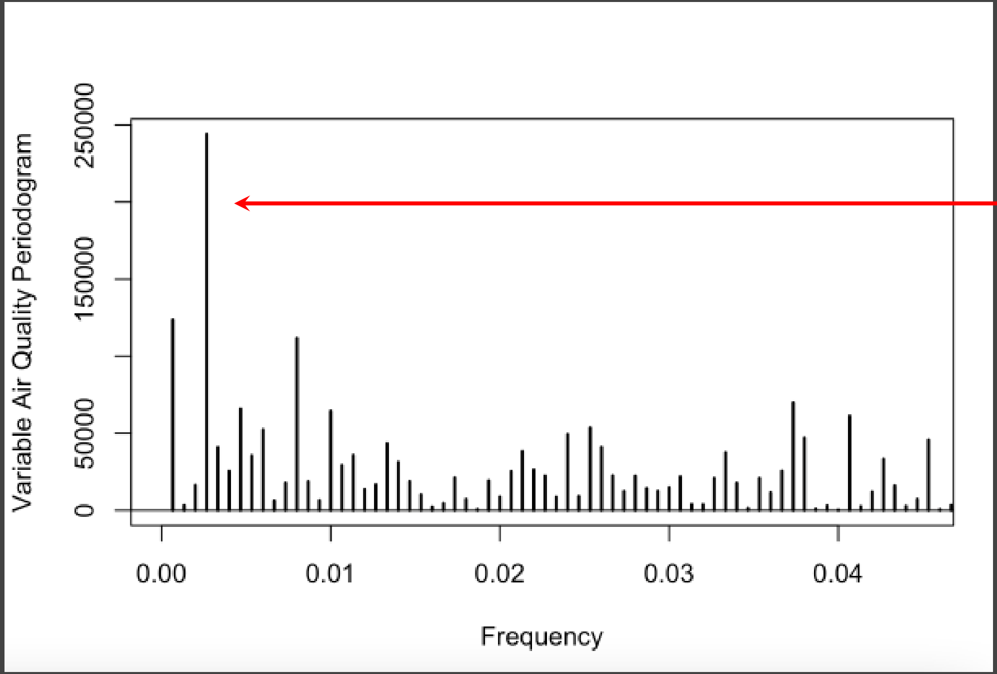
Differencing - Rain

Before Differencing	First Order Differencing	Seasonal Differencing
		

Augmented Dickey-Fuller Test Statistics

Before Differencing	First Order Differencing	Seasonal Differencing
-9.771	-18.197	-10.153
Stationary	Stationary	Stationary

Periodogram



Max Frequency = 0.00266

Seasonality = ~375 Days

The maximum frequency in the periodogram corresponds to a period of approximately 1 year

YIFAN JIANG
AMY ZHANG
MELODY FENG
JASON LEE
KAI HAYDEN

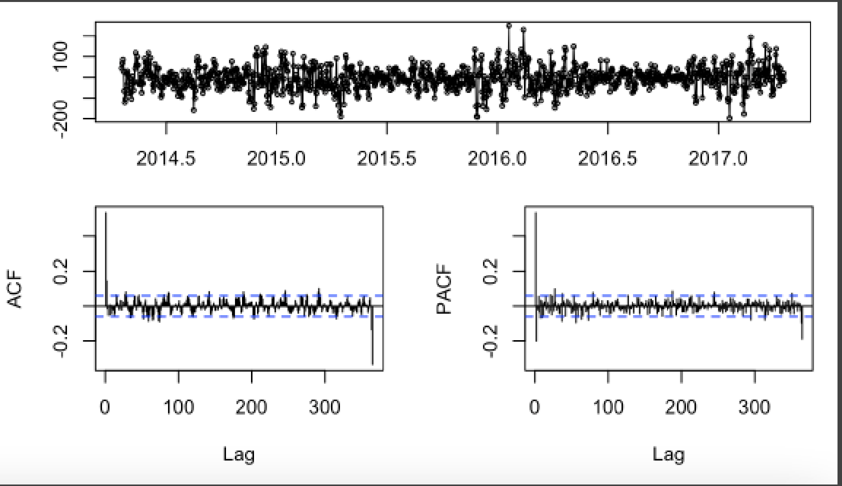
Modeling

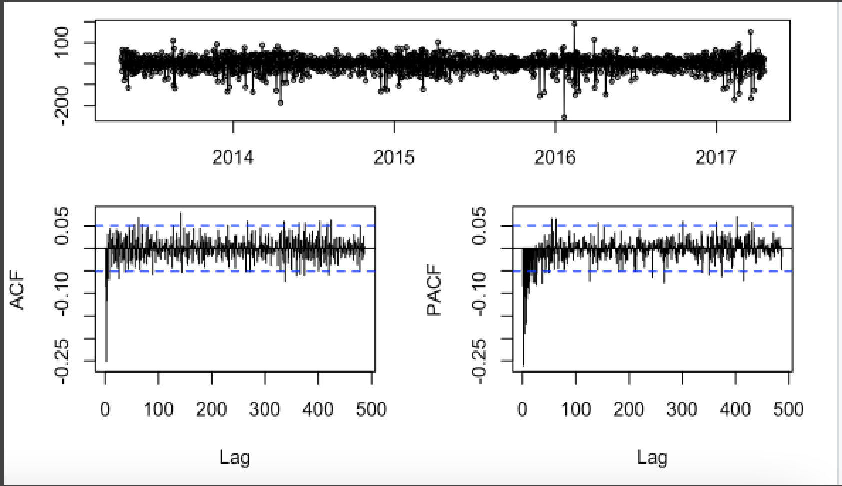
04

GROUP 7

ARIMA Model

Seasonal Differencing	1st Order Differencing





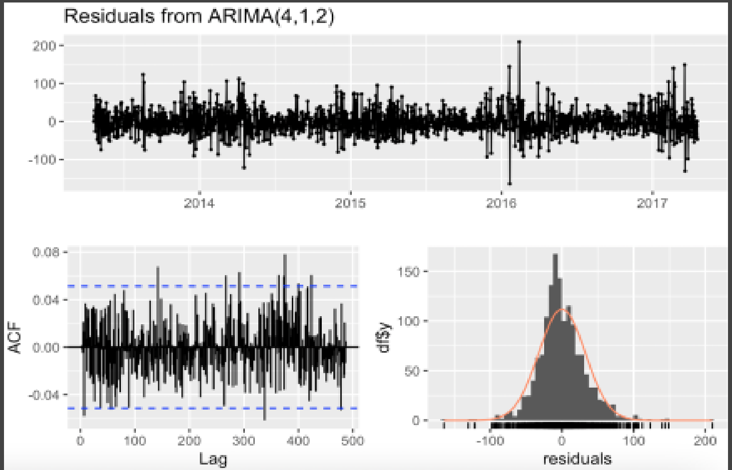
KPSS Test for Level Stationarity: 0.09
Stationary

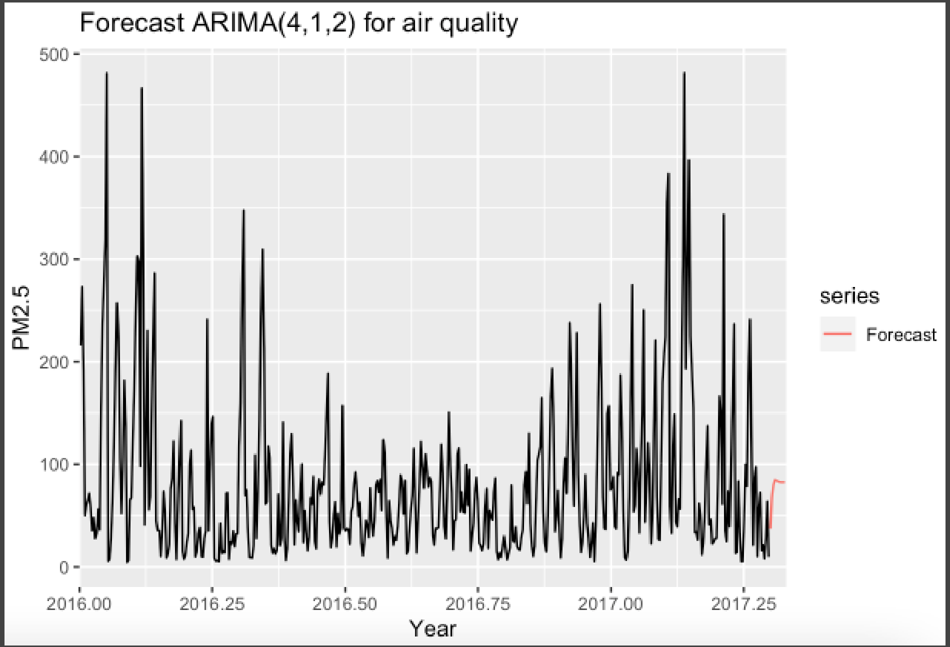
KPSS Test for Level Stationarity: 0.1
Stationary

ARIMA Model

ARIMA(4,1,2)

Lambda	AR1	AR2	AR3	AR4	MA1	MA2
0.8812	0.0412	0.1947	-0.085	-0.016	-0.370	-0.596





Ljung-Box Test

- P-value = 0.9561
- Independently Distributed

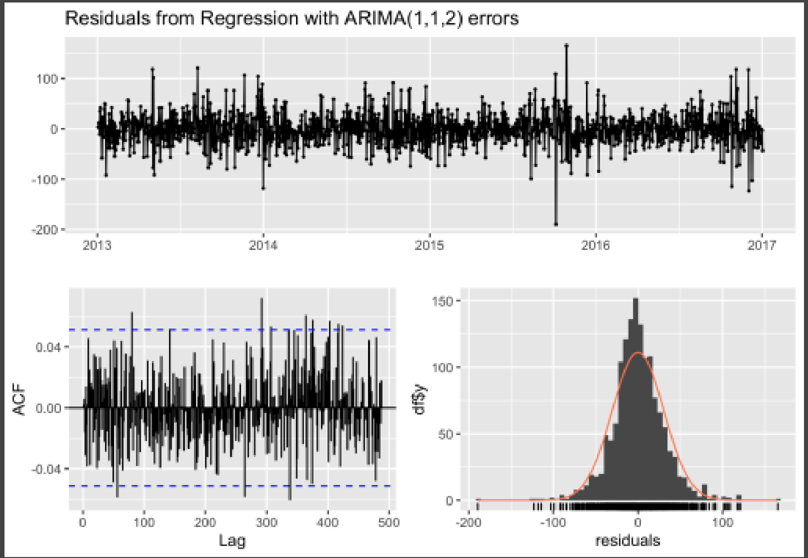
Regression with ARIMA Errors

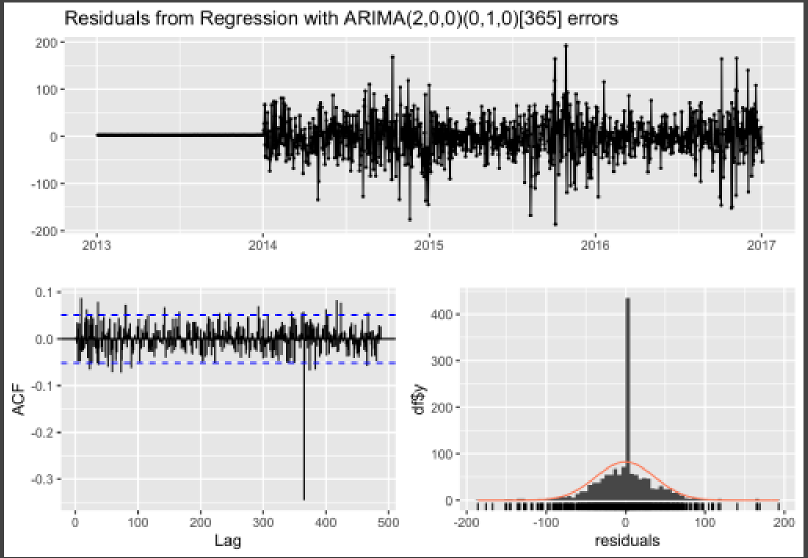
ARIMA(1,1,2) Errors

ar1	ma1	ma2	temp	pres	rain
0.4182	-0.6631	-0.2950	-2.3596	-3.3031	-0.7672

ARIMA(2,0,0)(0,1,0)[365] Errors

ar1	ar2	temp	pres	rain
0.6913	-0.1812	0.0668	-2.9515	0.5163





AICc: 14137.47, Independently Distributed

AICc: 11356.90, Serially Correlated

VAR Model

Adjusted R-Squared to check the model performance	**After taking the difference for PM 2.5, Temperature, Pressure**

VARselect(x_df, lag.max = 10, type = "both")$selection

AIC(n) HQ(n) SC(n) FPE(n)

10 9 4 10

—--

Portmanteau Test (asymptotic)

Chi-squared = 75.475, df = 9, p-value < 2.2e-16

Estimation results for equation pm_d1:
Multiple R-Squared: 0.2145, Adjusted R-squared: 0.2024

Estimation results for equation temp_d1:
Multiple R-Squared: 0.2169, Adjusted R-squared: 0.2015

Estimation results for equation pres_d1:
Multiple R-Squared: 0.1931, Adjusted R-squared: 0.1833

- ➔ The VARselect function selected the VAR(10) by the AIC, and VAR(4) by BIC.
- ➔ Rain was excluded since it does not have much contribution to the model of predicting other variables and the R^2 for itself is very low

- ➔ Adjusted R-Squared from the Summary table are low, suggesting the model has a bad fit
- ➔ Null hypothesis of no serial correlation in residuals is rejected for VAR(10) and VAR(4) ($p<0.05$)
- ➔ Pass the test to VAR(p) with $p \in [1,10]$ and all failed. So we decided to fit models to the original data

VAR Model

Fit model on original, undifferenced data

```
x = cbind(ts_pm, ts_temp, ts_pres)

VARselect(x, lag.max = 10, type = 'both')$selection
—-----------------------------------------------------------------
AIC(n)  HQ(n)  SC(n) FPE(n)
     5      5      3      5
```

- ➔ The VARselect function selected the VAR(5) by the AIC, and VAR(3) by BIC

Portmanteau test to check whether the residuals are correlated for each model

VAR(5)	VAR(3)	VAR(6)
p-value = 0.02188 < 0.05 ⇒ serial correlation	p-value = 1.609e-05 < 0.05 ⇒ serial correlation	p-value = 0.05698 > 0.05 ⇒ no serial correlation

- ➔ The null hypothesis of no serial correlation in the residuals is rejected for both a VAR(5) and a VAR(3) (p-value <0.05).
- ➔ Continued to VAR(6) and the model has passed the serial test, proving there’s little/no serial correlation, so we decided to build VAR(6)

VAR Model

- Estimation results for equation ts_pm: Adjusted R-squared: 0.3624

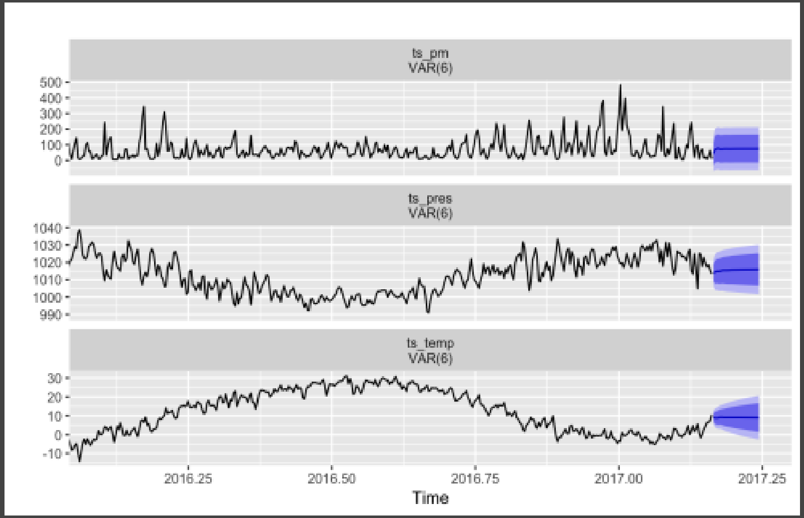
ts_pm.l1	ts_temp.l1	ts_pres.l1	ts_pm.l2	ts_temp.l2	ts_pres.l2	ts_pm.l3	ts_temp.l3	ts_pres.l3	ts_pm.l4	ts_temp.l4	ts_pres.l4	ts_pm.l5	ts_temp.l5	ts_pres.l5	ts_pm.l6	ts_temp.l6	ts_pres.l6	const	trend
6.598e-01	1.383e+00	1.309e+00	-2.194e-01	4.768e-01	-3.773e-01	5.504e-02	-6.123e-01	1.013e+00	-4.430e-02	5.675e-01	-2.325e-01	4.132e-02	-9.354e-01	-1.464e-01	-2.521e-02	-3.080e-01	1.530e-01	-1.691e+03	-1.307e-02

- Estimation results for equation ts_temp:

Multiple R-Squared: 0.9706, Adjusted R-squared: 0.9702

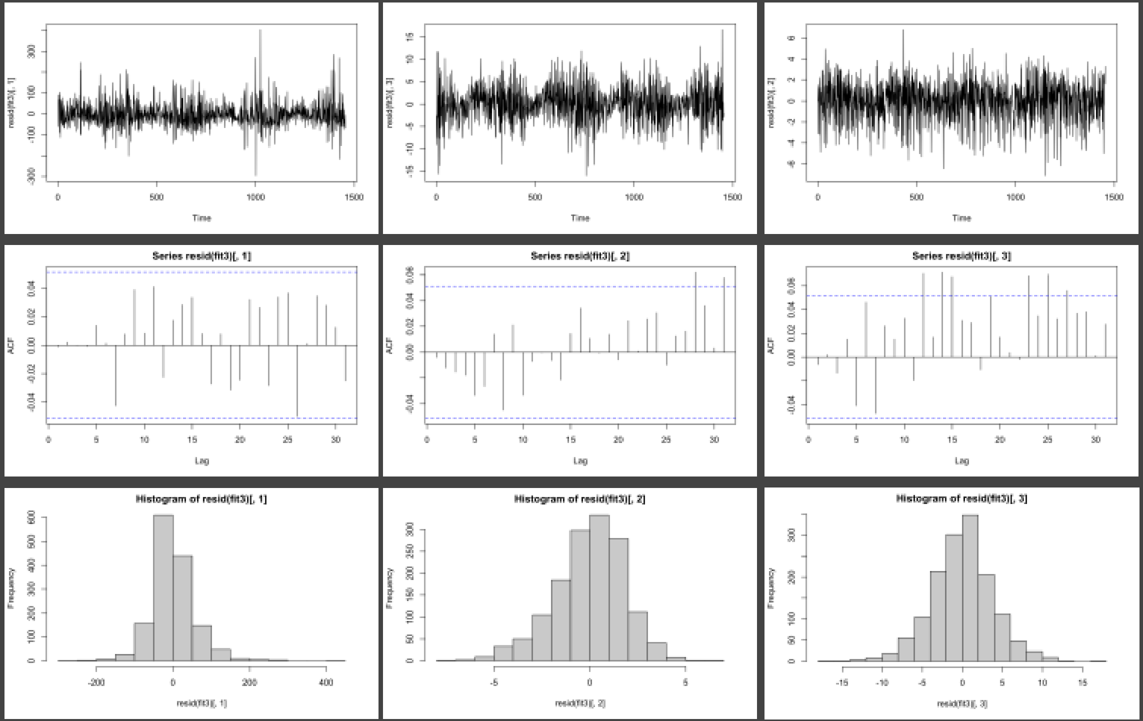
- Estimation results for equation ts_pres:

Multiple R-Squared: 0.8567, Adjusted R-squared: 0.8548



VAR Model

	PM 2.5	Temperature	Temperature
Time Series			
ACF			
Histogram			

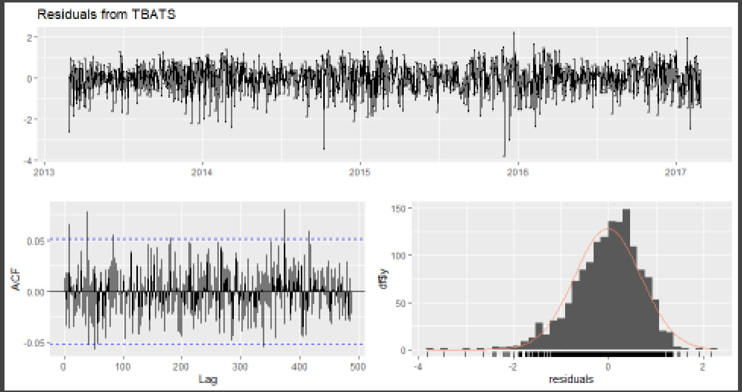


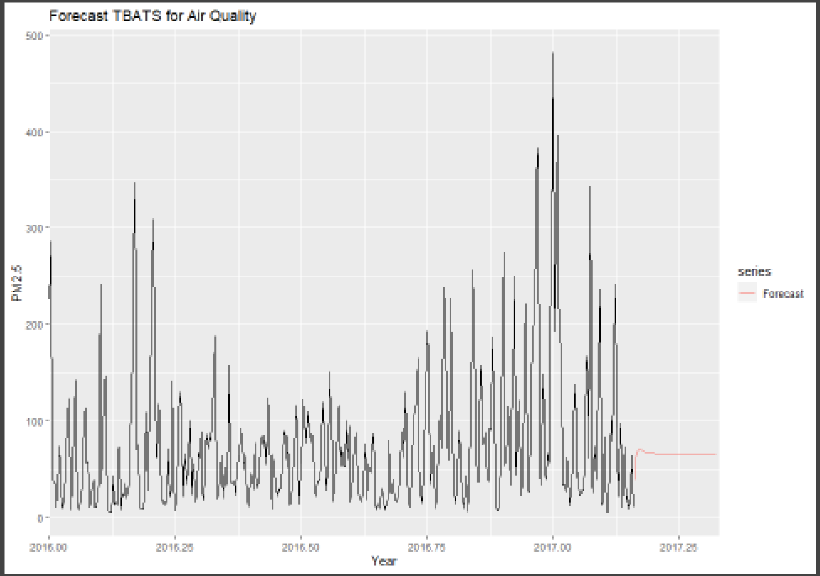
Residuals have little autocorrelation, especially for PM2.5 and temperature

Residuals generally follow the normal distribution

TBATS Model

Lambda	Alpha	Beta	Damping	AR1	AR2	MA
0.037646	-0.020774	0.014209	0.8	0.1924	0.0534	0.4085





Ljung-Box Test

- P-value = 0.02795
- Serially correlated

YIFAN JIANG
AMY ZHANG
MELODY FENG
JASON LEE
KAI HAYDEN

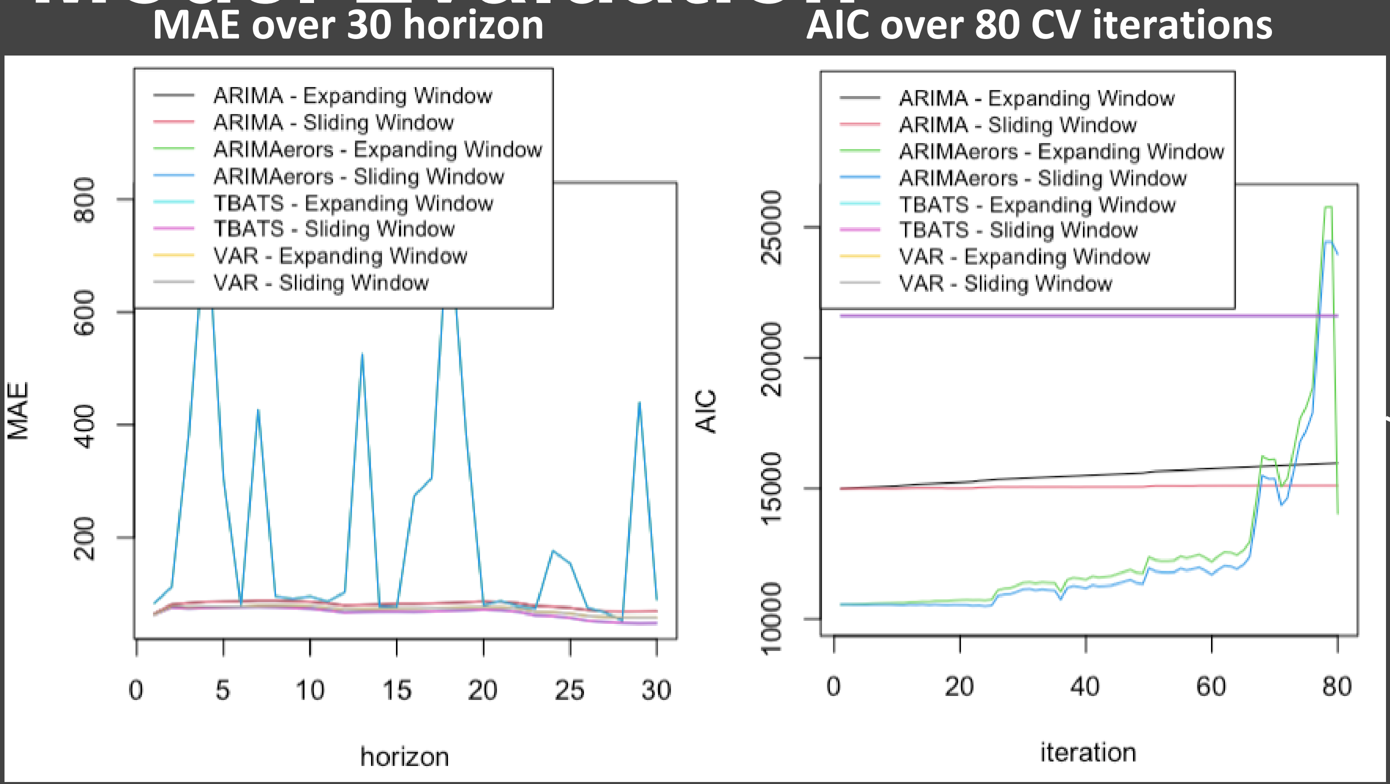
THE UNIVERSITY OF CHICAGO

Model Evaluation

05

GROUP 7

Model Evaluation



AIC for VAR(6) model is too large to display in the graph

YIFAN JIANG
AMY ZHANG
MELODY FENG
JASON LEE
KAI HAYDEN

THE UNIVERSITY OF CHICAGO

Conclusion

06

GROUP 7

Conclusion

- TBATS was the best on predicting PM2.5. It was an automated model, so almost no adjustments was made.
- Rain doesn’t have much correlation with PM2.5, temperature, and pressure, making it less useful as a predictor. The other three set of time series data can be used to predict each other due to correlation.
- ARIMA model does not capture other variables correlations with PM2.5.
- Regression with ARIMA errors better captures correlations than ARIMA but there are still patterns in the data that are not exploited by the model.
- VAR is good at predicting temperature and pressure but not the PM2.5

34

Improvement and Future Work

- Try advanced models:
  - RNN, ARCH, and GARCH
- Better data selection:
  - Try using other data combinations to train the VAR model when predict PM 2.5
  - Consider using alternative data sources
- Try using different tools:
  - Meta has an interesting library - Prophet
  - Get some experience with Python time series (for practice)

35

YIFAN JIANG
AMY ZHANG
MELODY FENG
JASON LEE
KAI HAYDEN

Thanks for Listening!

Any Questions?

GROUP 7

36

Appendix: Model Evaluation

RMSE over 30 horizons