PM 2.5 Prediction in Beijing

Prepared For MSCA 31006
5/27/2022

By: Naibo Hu (Ray), Weijia Wang (Joyce), Wen Zhang, Rujue Du, Xinyi Zhang

Agenda

01 **Introduction**
Business Problem
Project Goal

02 **Data profile & processing**
Data Description
Handle Missing Values

03 **Exploratory Data Analysis (EDA)**
Understand Data Characteristics

04 **Experiment Results & Analysis**
Models Application & Analysis

05 **Models Evaluation & Selection**
Compare Model Performances

06 **Conclusion & Future Work**
Key Takeaways and Potential Improvements

Business Problem

While China's economy and population have grown dramatically in the past three decades, air pollution has become a serious issue. **Of the twenty cities with the worst air pollution worldwide, sixteen are located in China. Beijing**, the capital of China, **suffers most from air pollution**.

Project Goal

In this project, our team aims to **utilize time series forecasting to predict the levels of PM 2.5 in Beijing.**

Accurate and accessible air pollution forecasts can raise public awareness, allow for sensitive populations to plan ahead, and provide governments with information for public health alerts.

Data Profile

Data Description

Overview:
Hourly data set contains the PM2.5 data of US Embassy in Beijing.

Dimension :
43, 824 instance, 13 variables

Year data covered:
2010 - 2014

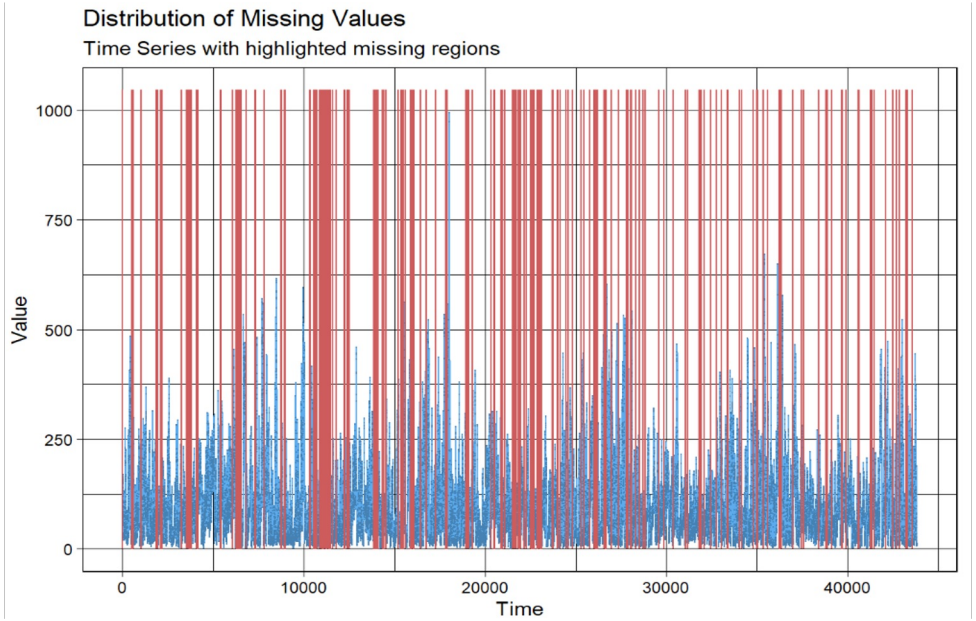
Source: UCI

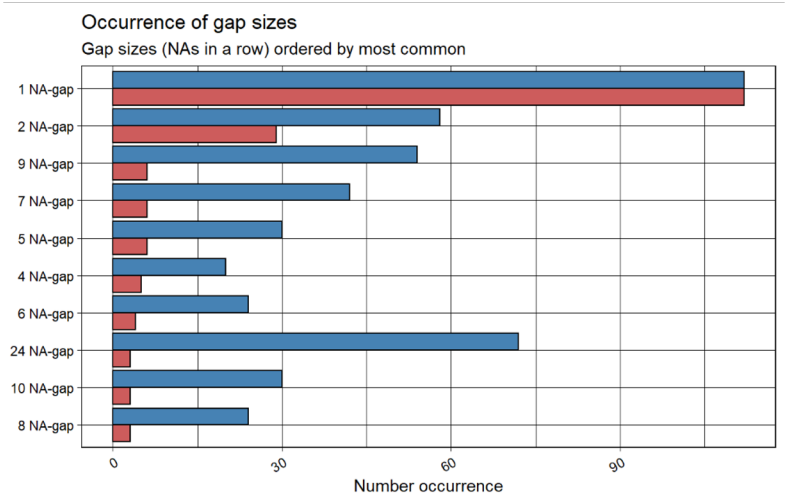
Attribute Information

No: Row Number
Year: year of data
Month: month of data
Day: day of data
Hour: Hour of data
Pm 2.5: PM2.5 concentration
DEWP: Dew Point
Temp: Temperature
PRES: Pressure
Cbwd: Combined wind direction
Iws: Cumulated wind speed (m/s)
Is: Cumulated hours of snow
Ir: Cumulated hours of rain

We undersampled hourly data and imputed missing values with weighted moving average

- The variable PM2.5 has 2,067 missing values.
- All missing values are randomly distributed, and most of them have 1 NA gap.

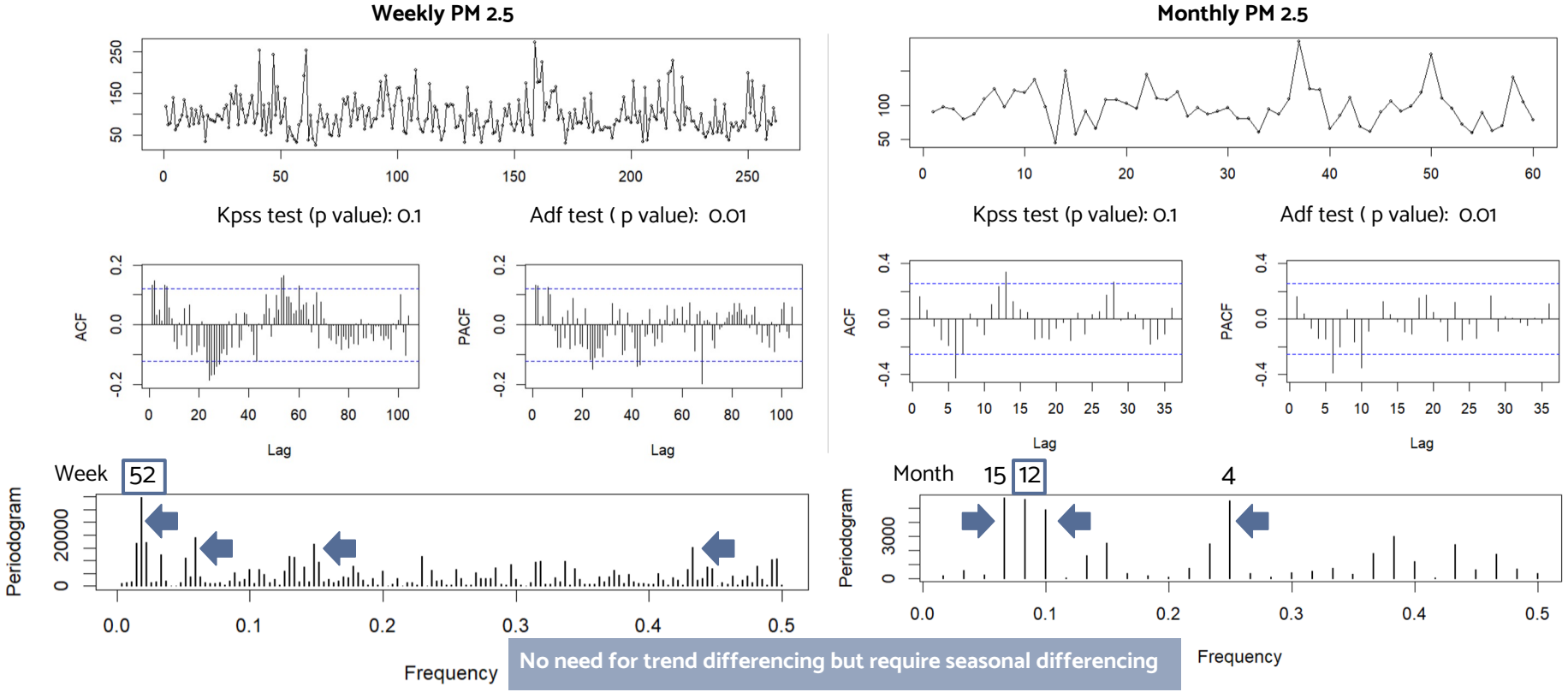




Since there are too many hourly observations, we decided to first **undersample the dataset by taking weekly and monthly mean of PM2.5** and then impute missing values with weighted moving averages.

After undersampling, we end up with 262 weekly and 60 monthly observations.

Both weekly and monthly data do not have trend or level but contain several dominant periodic components

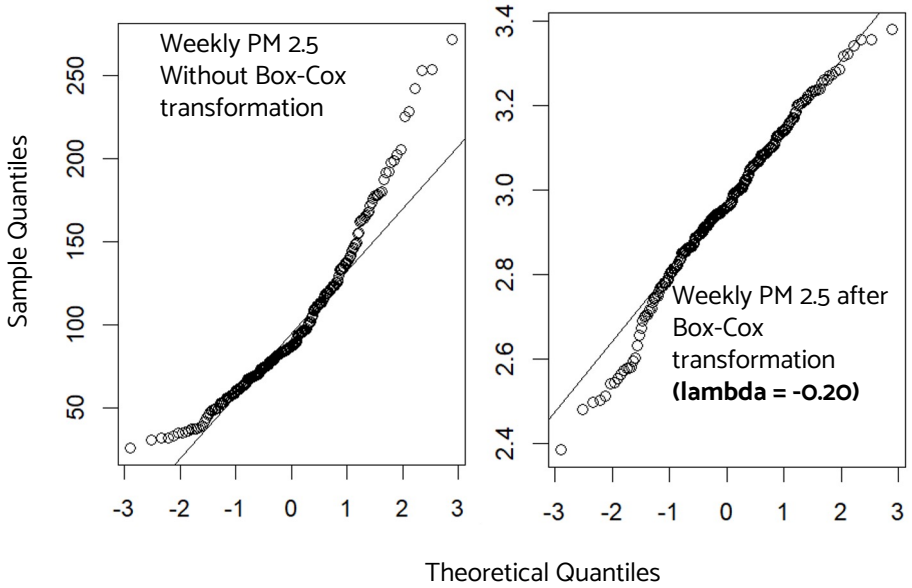


Why we do not use Box-Cox transformation?

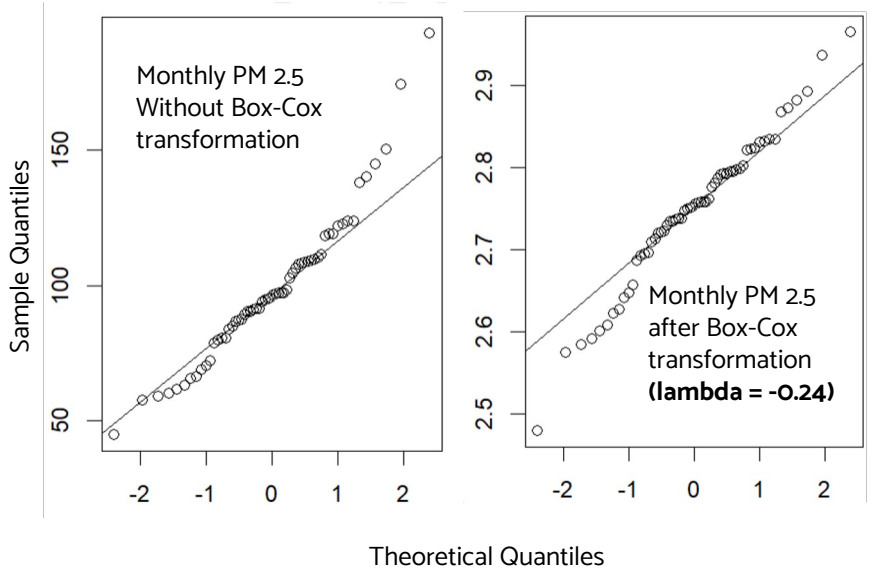
We compute the Box-Cox lambdas and transform both weekly and monthly data. QQ plots show that Box-Cox transformation makes the data distribution closer to the normal distribution.

However, when fitting different models, we find that data without Box-Cox transformation actually generate better training and prediction accuracy. Thus, we decide to not use Box-Cox transformation.

Normal QQ Plots: Weekly Data



Normal QQ Plots: Monthly Data



We utilized both univariate and multivariate approaches for PM 2.5 prediction

We applied single train-test split, having **80% training data and 20% testing data**. We decided not to use cross validation, as it could introduce leakage from future data to the model. Also, it can potentially cause information loss problem.

Univariate Analysis

1. Simple Forecasting Models
   - Naive
   - Seasonal Naive
   - Drift
   - Mean
1. Seasonal ARIMA
2. Dynamic Harmonic Regression
3. TBATs Model

Multivariate Analysis

1. ARMA with error
2. Vector Autoregressions (VAR model)

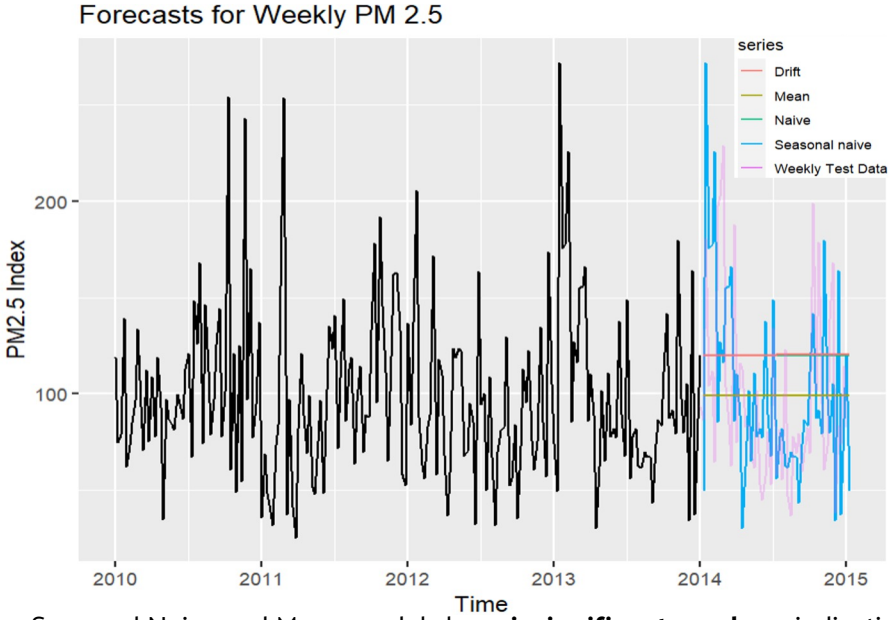
Weekly Data

Train: 2010/1/3 - 2013/12/29
Test: 2014/1/5 - 2014/12/31
(h = 53 weeks)

Monthly Data

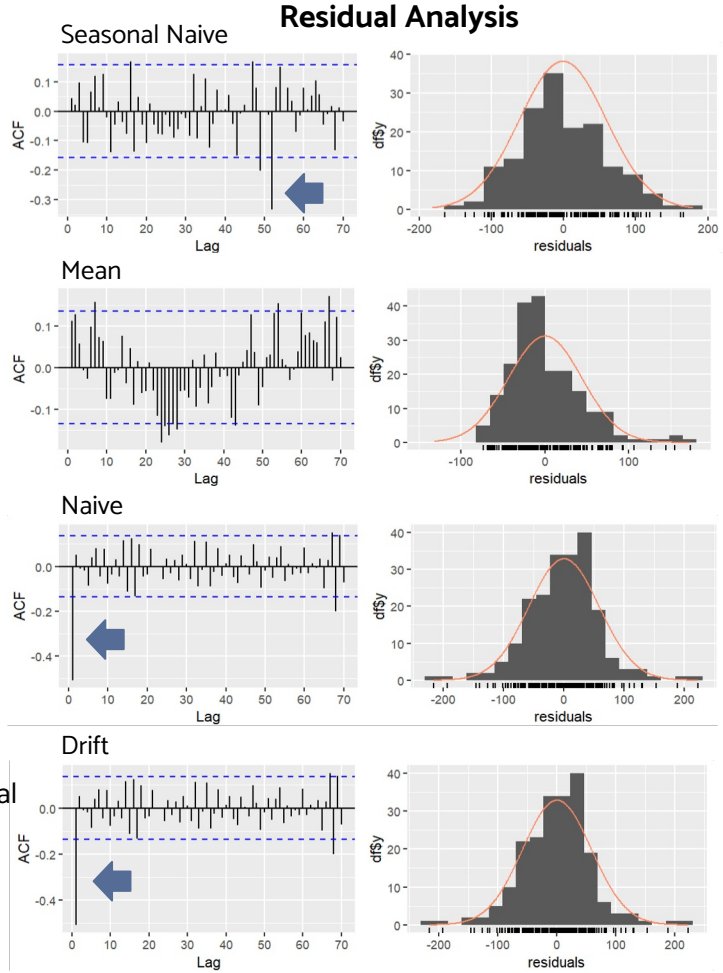
Train: 2010/1 - 2013/12
Test: 2014/1 - 2014/12
(h = 12 months)

Mean model captured the most information on weekly data



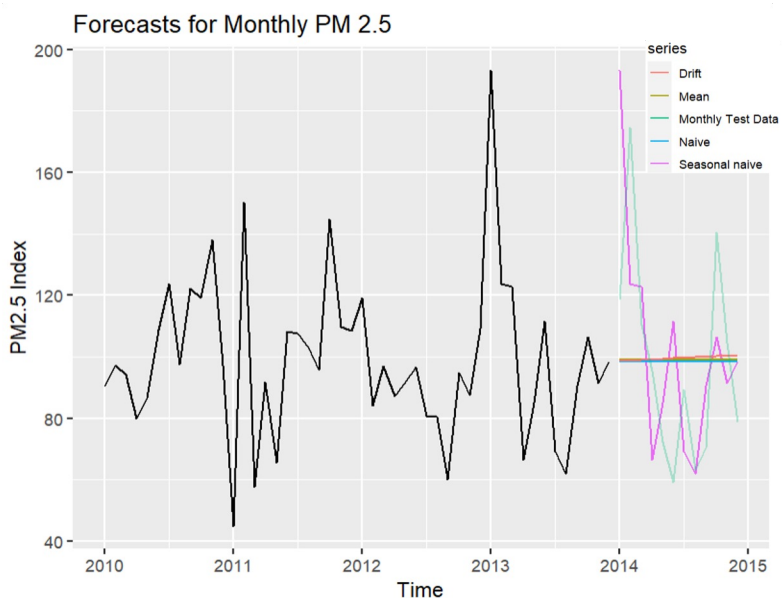
Seasonal Naive and Mean models have **insignificant p-values,** indicating that the **residuals are independently distributed** and do not exhibit serial correlation.

All models except the mean model have **significant lags** in ACF residual plot, which means that **not all information is captured** by the model.



Ljung_Box test (p value)
0.5749
0.1006
1.721e-13
1.721e-13

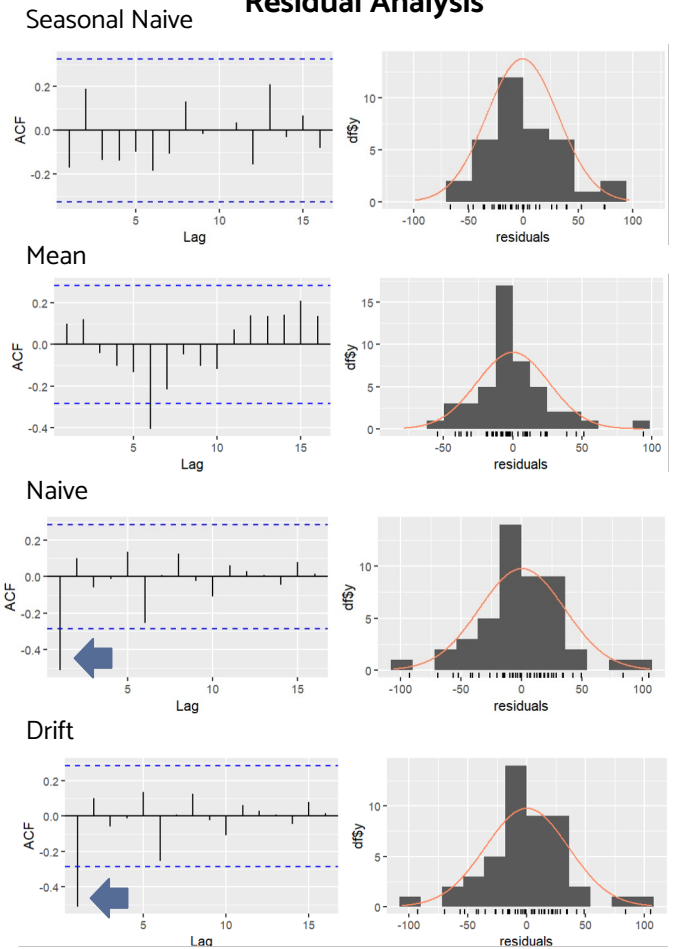
Mean and Seasonal Naive models outperform others on monthly data



Seasonal Naive and Mean models have **insignificant p-values,** indicating that the **residuals are independently distributed** and do not exhibit serial correlation.

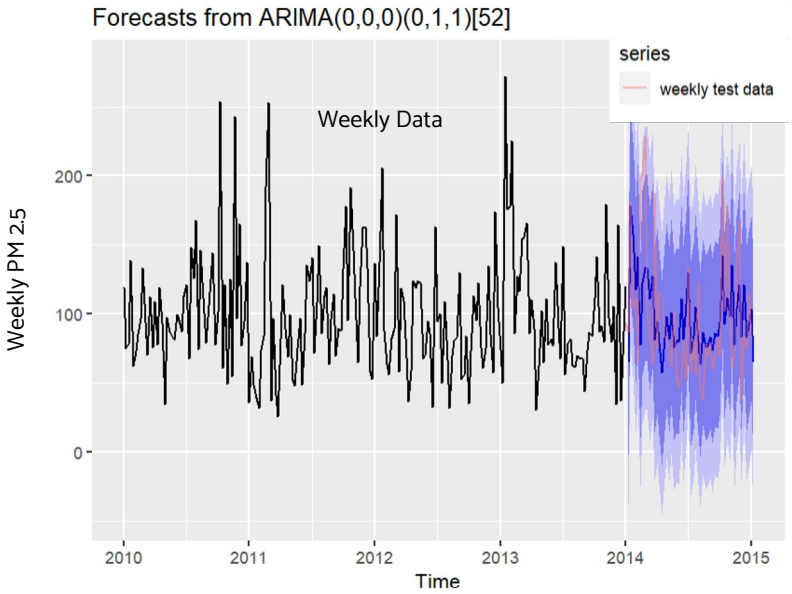
Naive and Drift models have **significant lags** in ACF residual plot, which means that **not all information is captured** by the model.

Residual Analysis

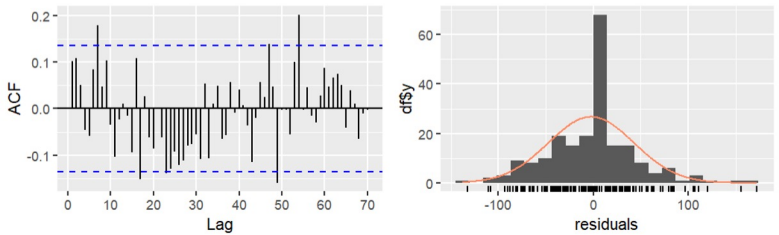


Ljung_Box test (p value)
0.29
0.4824
0.00
0.00

For SARIMA models, we applied seasonal differencing at lag 52 and 12



Residual Analysis



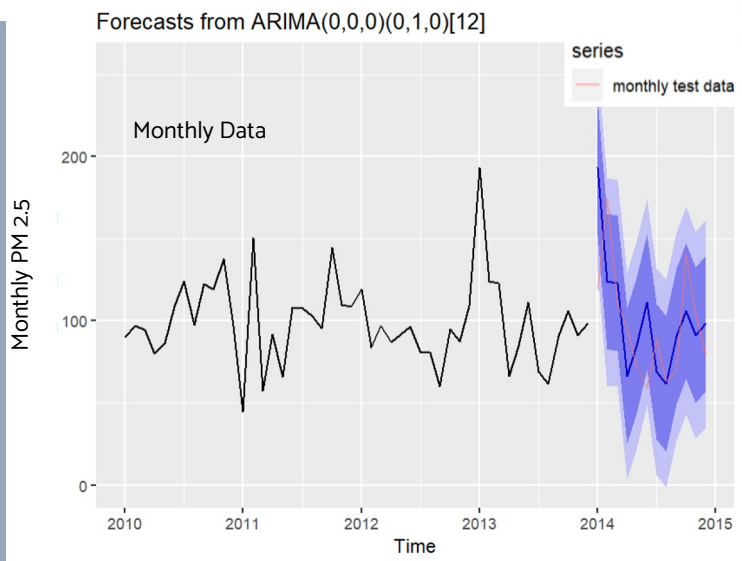
Weekly Model:
AIC= 1708.76
AICc= 1708.83
BIC= 1714.87

sma1 = -0.5457

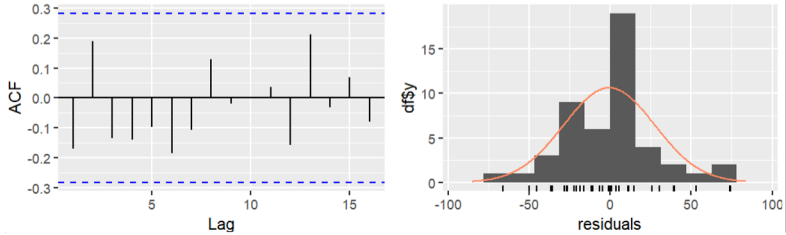
Ljung-Box (p value): 0.1357

Monthly Model:
AIC= 354.12
AICc= 354.24
BIC= 355.7

Ljung-Box (p value): 0.2236

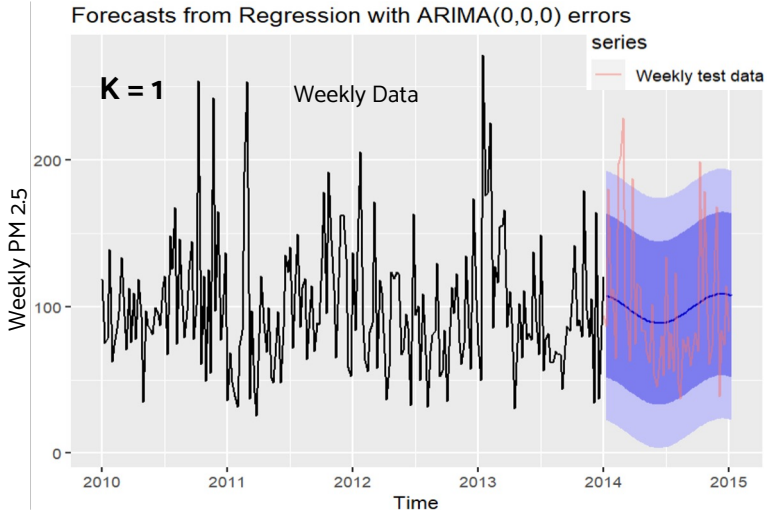


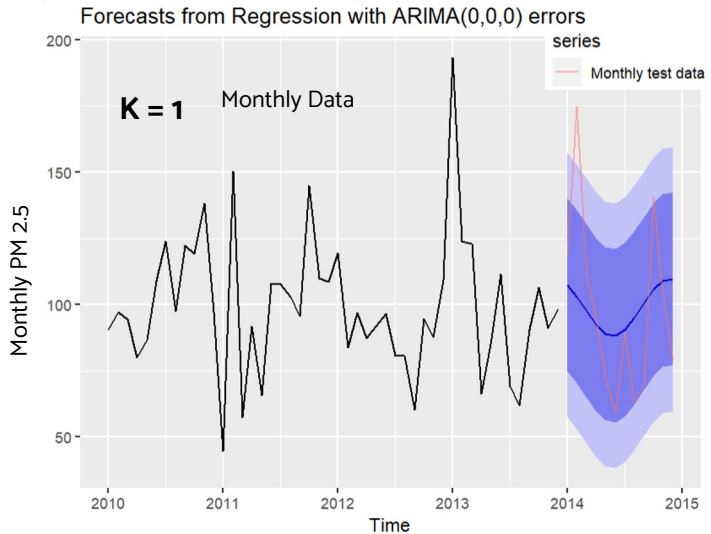
Residual Analysis



For dynamic harmonic regression, we selected k = 1

Select the number of fourier pairs (K) that minimizes the AICc value





Weekly Model:

AIC=2174.96
AICc=2175.16
BIC=2188.33

Intercept = 98.9221
S1-52 = -2.2217
C1-52 = 9.7836

Ljung-Box (p value): 0.1917

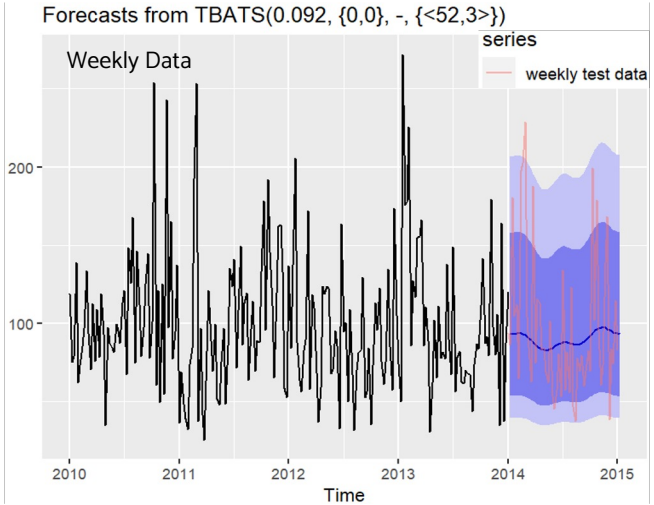
Monthly Model:

AIC=451.87
AICc=452.8
BIC=459.35

Intercept = 99.0184
S1-12 = -1.5262
C1-12 = 10.6963

Ljung-Box (p value): 0.8969

TBATS Model automatically applied Box-Cox transformation



Weekly Model

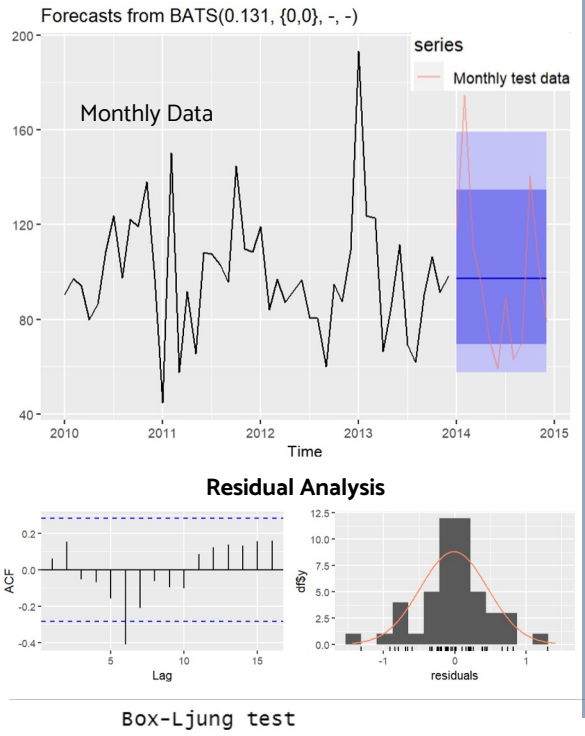
Box-Cox transformation: lambda = 0.092

No ARMA error

No Damping parameter

Seasonal period: 52

Fourier terms: 3



Monthly Model

Box-Cox transformation: lambda = 0.131

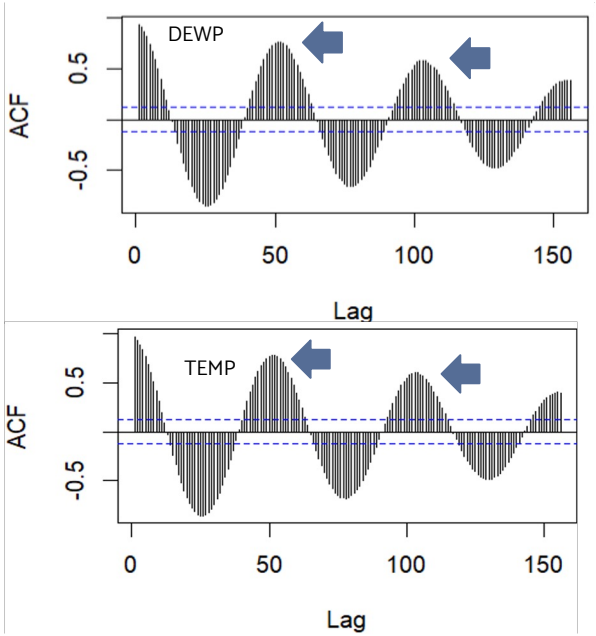
No ARMA error

No Damping parameter

No seasonal period

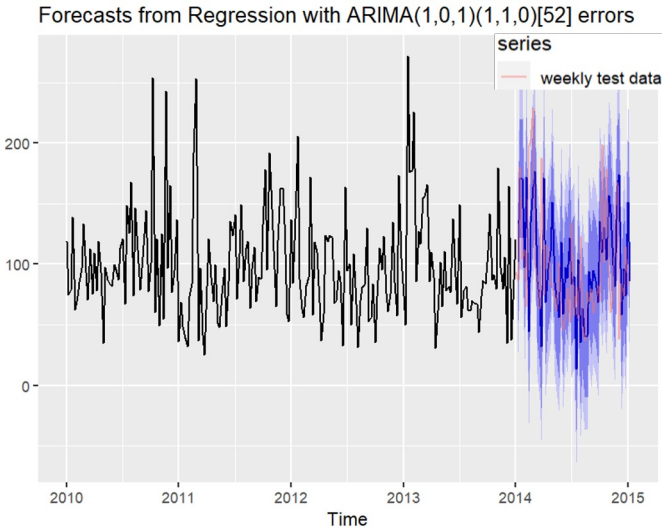
No Fourier term

We took seasonal differencing on weekly data before fitting regression with ARMA error model

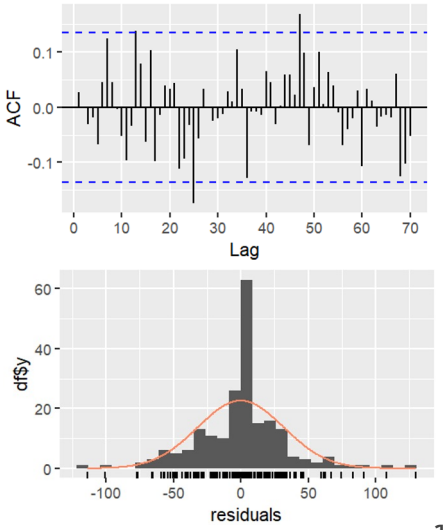



```
Series: air_train_week_ts_sub[, "pm2.5"]
Regression with ARIMA(1,0,1)(1,1,0)[52] errors

Coefficients:
                                          DEWP      TEMP                lws       ls       lr
         ar1      ma1     sar1      dewpoint  temperature  windspeed     snow     rain
      0.9022  -0.7781  -0.4491        9.8851      -2.9779    -0.1159  -5.4972  -11.5732
s.e.  0.0862   0.1215   0.0838        0.9359       1.4722     0.1304   7.6562    4.9273

sigma^2 = 1434:  log likelihood = -795.18
AIC=1608.35   AICc=1609.58   BIC=1635.86
```

Ljung- Box test (p-value): 0.6935

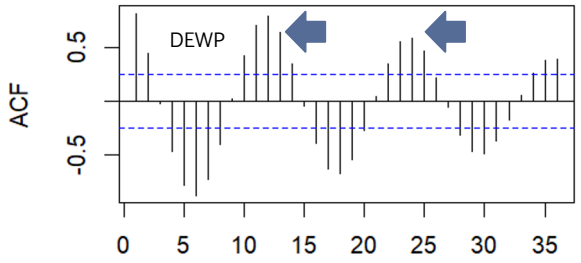


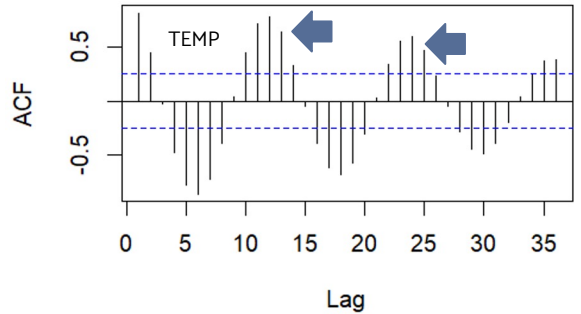
Residual analysis



- Variables dew point and temperature have strong seasonality at **lag 52, 104, etc.** Thus, Take seasonal differencing at lag 52 (D = 1) removed seasonality
- DEWP, Is, Ir have strong impact on pm2.5.

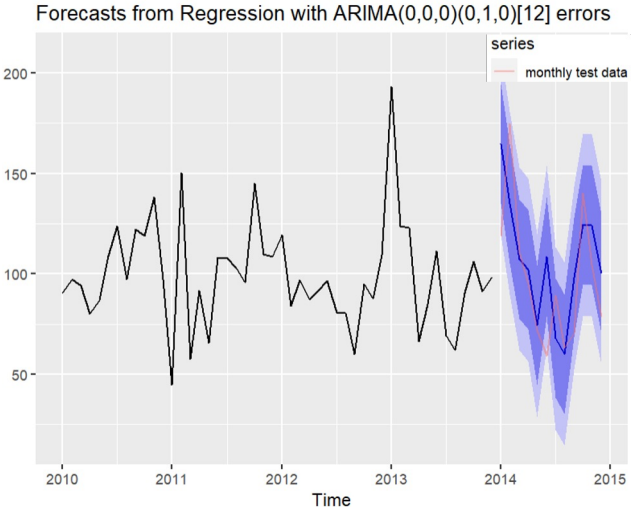
We took seasonal differencing on monthly data before fitting regression with ARMA error model



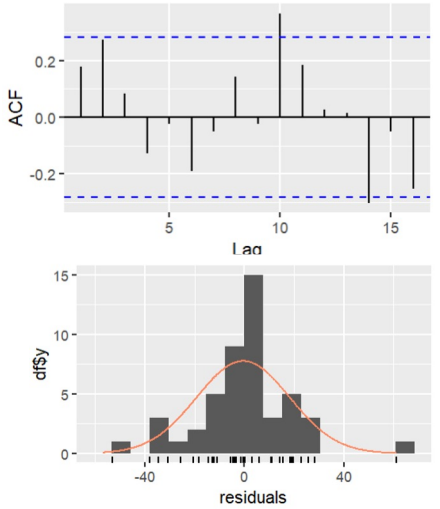



```
Series: air_train_month_ts_sub[, "pm2.5"]
Regression with ARIMA(0,0,0)(0,1,0)[12] errors

Coefficients:
                TEMP      Iws      Is       Ir
DEWP  dewpoint  temperature  windspeed  snow  rain
      5.4342    -0.6111   -0.2393  41.7496  -9.8309
s.e.  2.1843     2.4153    0.3205  26.4569  15.1046

sigma^2 = 536.5:  log likelihood = -161.52
AIC=335.04   AICc=337.94   BIC=344.54
```

Ljung- Box test (p-value): 0.2037



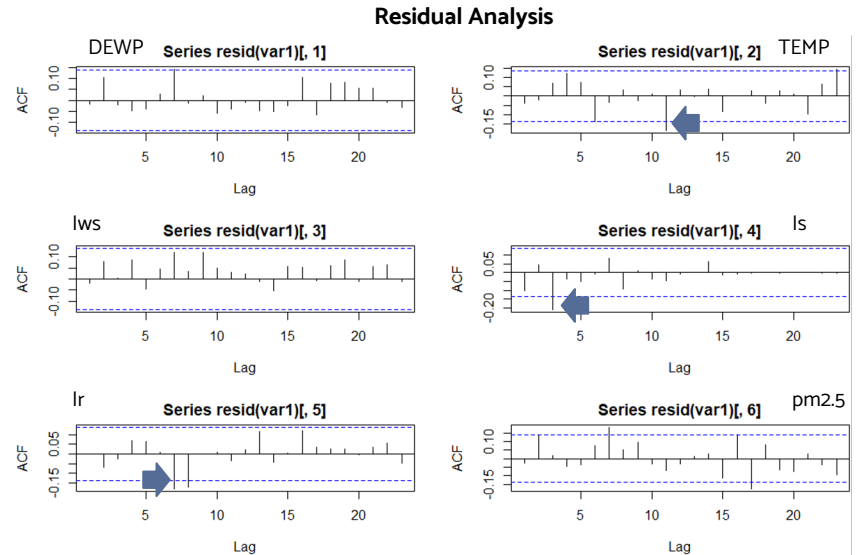
Residual analysis



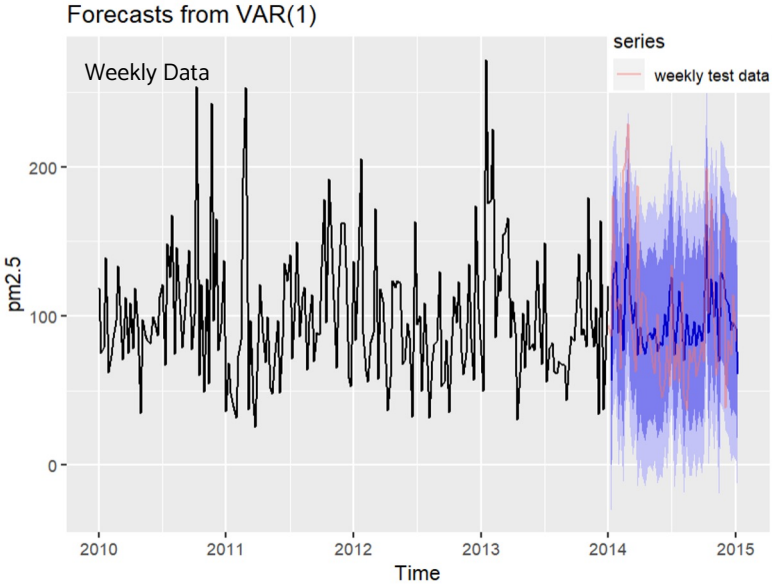
- Variables dew point and temperature have strong seasonality at lag 12, 24, etc. Thus, take seasonal differencing at lag 12 (D = 1) removed seasonality
- DEWP, Is, Ir have strong impact on pm2.5

We selected VAR(1) for weekly data

- VARselect suggested var(1) model based on AIC and var(6) model based on BIC. We selected var(1) because of its simplicity.
- According to Portmanteau test, Var(1) model has p-values less than 0.05. Thus, the null hypothesis of no serial correlation in the residuals is rejected.




```
AIC(n)  HQ(n)  SC(n) FPE(n)
     6       1      1      1
```



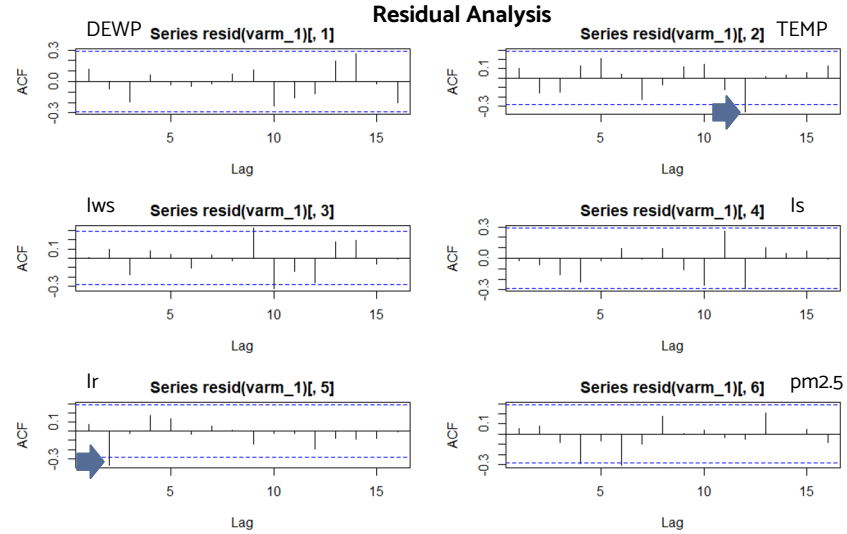

```
        Portmanteau Test (asymptotic)

data:  Residuals of VAR object var1
Chi-squared = 409.47, df = 324, p-value = 0.0008841
```

Pairs of variables with cross- correlation

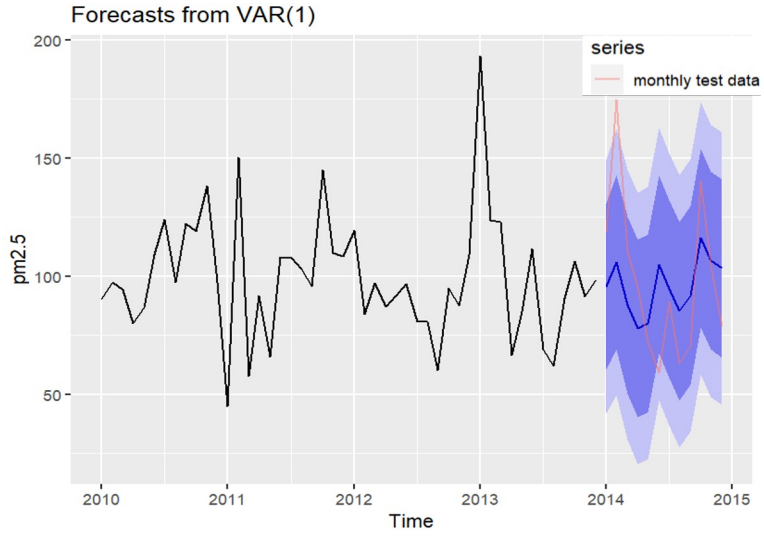
DEWP & TEMP,
DEWP & IWS,
DEWP & Is,
DEWP & PM2.5

We selected VAR(1) for monthly data

- VARselect suggested var(7) model based on AIC and BIC. However, we believed var(7) model is way too complicated. Thus, we selected var(1) due to its simplicity.
- According to Portmanteau test, Var(1) model has p-values less than 0.05. Thus, the null hypothesis of no serial correlation in the residuals is rejected.




```
AIC(n)  HQ(n)  SC(n) FPE(n)
     7       7      7      7
```




```
        Portmanteau Test (asymptotic)

data:  Residuals of VAR object varm_1
Chi-squared = 376.48, df = 324, p-value = 0.02354
```

Pairs of variables with cross- correlation

DEWP & IWS,
IWS & PM2.5

DEWP & PM2.5

VAR Model is the optimal model for weekly data

Model	ME	RMSE	MAE	MPE	MAPE
Naive	-22.71	52.23	46.48	-49.65	62.13
Seasonal Naive	-4.46	58.56	45.73	-20.26	50.18
Drift	-22.87	52.32	46.58	-49.86	62.30
Mean	-1.65	47.06	37.23	-23.39	43.70
Seasonal ARIMA	-1.46	44.06	33.40	-18.84	38.85
Dynamic Harmonic Regression	-1.60	45.24	35.75	-22.13	41.66
TBATs	7.35	46.28	33.76	-11.48	36.04
Regression with ARMA error	-4.31	60.38	47.55	-25.04	55.64
VAR model	-1.03	41.92	32.27	-19.45	39.14

We use **RMSE** to select the best model, because the **RMSE number is in the same unit as the projected value, making it easier to interpret.**

Other metrics have several disadvantages. For example, MAE does not reveal the proportional scale of the error, which makes it difficult to distinguish between large and little errors. Also, MAPE takes an extreme value if this value is exceedingly tiny or huge.

According to the RMSE results, **VAR model** is the optimal model for weekly data. It is possibly due to the fact that VAR model has the ability to **capture the cross-correlation among multiple variables.**

Regression with ARMA error model is the best performing model for monthly data

Model	ME	RMSE	MAE	MPE	MAPE
Naive	-0.47	32.79	26.45	-10.94	28.74
Seasonal Naive	-3.66	34.88	28.39	-9.47	29.60
Drift	-1.58	33.04	26.79	-12.26	29.42
Mean	-0.97	32.80	26.53	-11.51	28.97
Seasonal ARIMA	-3.66	34.88	28.38	-9.47	29.60
Dynamic Harmonic Regression	-0.97	29.82	23.18	-10.12	25.10
TBATs	0.69	32.80	26.26	-9.63	28.21
Regression with ARMA error	-7.57	26.62	21.28	-12.75	23.55
VAR model	2.22	29.42	23.76	-6.54	25.66

For monthly data, **Regression with ARMA error** outperforms other models. This could be due to the fact that our time series data contain correlated regression errors. This problem can be well resolved by the regression with ARMA error model.

We also see that **VAR** is the second-best performing model. Therefore, VAR model has a stable and robust performance for both weekly and monthly data.

We identified several insights and potential ways of improvement

Conclusion

VAR model performs best in weekly data, while the Regression with arma error model works better in monthly data. Both models involve multivariate analysis, which can capture the impact of other variables on pm2.5.

Mean model also has relative good performance on both weekly and monthly data. It is computationally inexpensive and time saving compared to other models. So, governments can consider using mean model for pm 2.5 prediction.

Future Work

Since our project only applied memoryless models, we can think of applying memory models, such as Neural Network Autoregression (NNAR) and Recurrent Neural Network (RNN), which can remember things over time.

We can also include other air quality features such as carbon monoxide level and nitrogen oxides to further improve the forecast accuracy. In addition, we would like to extend the forecasting model to develop a real-time forecasting service.

THANKS

APPENDIX

Group Task Distribution

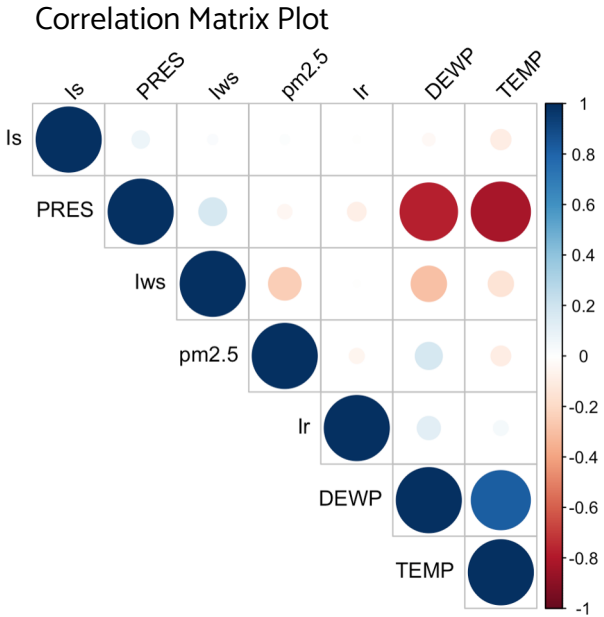
Naibo Hu (Ray), Weijia Wang (Joyce)

- Data Processing
- EDA
- SARIMA Model
- Dynamic Harmonic Regression Model
- TBATs Model
- Presentation slides

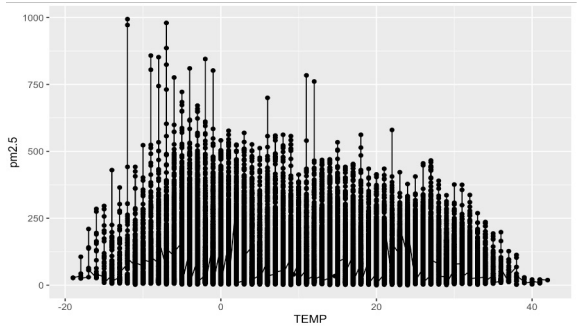
Wen Zhang, Rujue Du, Xinyi Zhang

- Data Processing
- Simple Time Series Models
    - Naive, Drift, Mean, Seasonal Naive
- Var Model
- Machine Learning Models
- Presentation slides

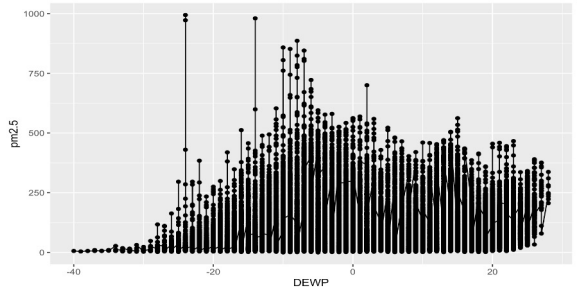
EDA others



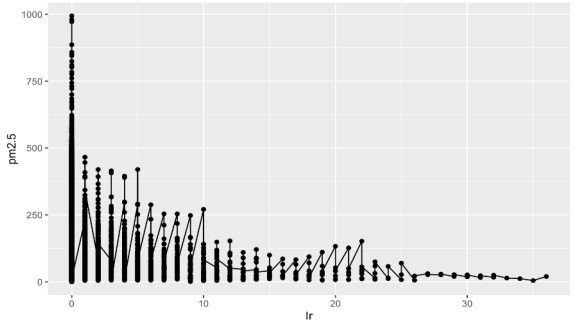
- PRES, DEWP and TEMP have strong correlations with each other



PM2.5 vs Temperature



PM2.5 vs Dew Point



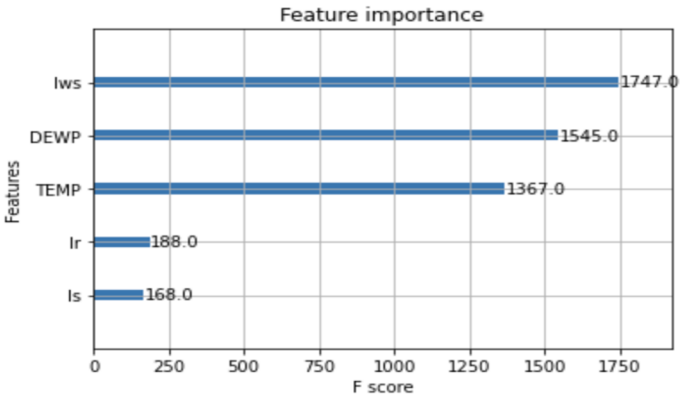
PM2.5 vs Cumulated hours of rain

Machine Learning Model

	feature	VIF
0	No	8.281290e+05
1	year	7.958164e+05
2	month	3.292615e+04
3	day	2.298788e+02
4	hour	1.545492e+00
5	pm2.5	1.381283e+00
6	DEWP	5.202275e+00
7	TEMP	6.393889e+00
8	PRES	3.793980e+00
9	cbwd_1	5.602537e+11
10	cbwd_2	3.438519e+11
11	cbwd_3	5.184000e+11
12	cbwd_4	1.828427e+11
13	Iws	1.290308e+00
14	Is	1.029040e+00
15	Ir	1.053893e+00

- Drop columns with VIF greater than 5, including 'cbwd_1','cbwd_2','cbwd_3','cbwd_4','year','month','day','hour','No','PRES'

Model	RMSE with CV	RMSE on Train	RMSE on Test
Linear Regression	80.409	80.397	82.840
XGBoost	73.677	67.352	76.473
KNN	79.092	66.530	82.151
Random Forest	79.979	50.223	82.267



Iws, DEWP, TEMP are the most important features.

According to the results, XGBoosting model with low testing RMSE and less overfitting result is the best fitted model in this case.