Chengdu PM2.5 Analysis

My dataset

1. I find out this dataset on Kaggle
2. The variable I interested in is the pm25, and I will observing how pm25 change over different time periods and analyzing the long-term trends in pm25 level
3. this data set also includes other air quality such as pm10, O3, NO2, SO2 and CO, these can provide insights into overall air quality and correlations or interactions with pm25. There is no miss data of date and pm25, so I do not need to clean the data set.

Motivation

1.Why am I interested in this dataset? PM2.5 is a critical air pollutant that directly impacts public health and the environment. And I lived in this city for 19 years.

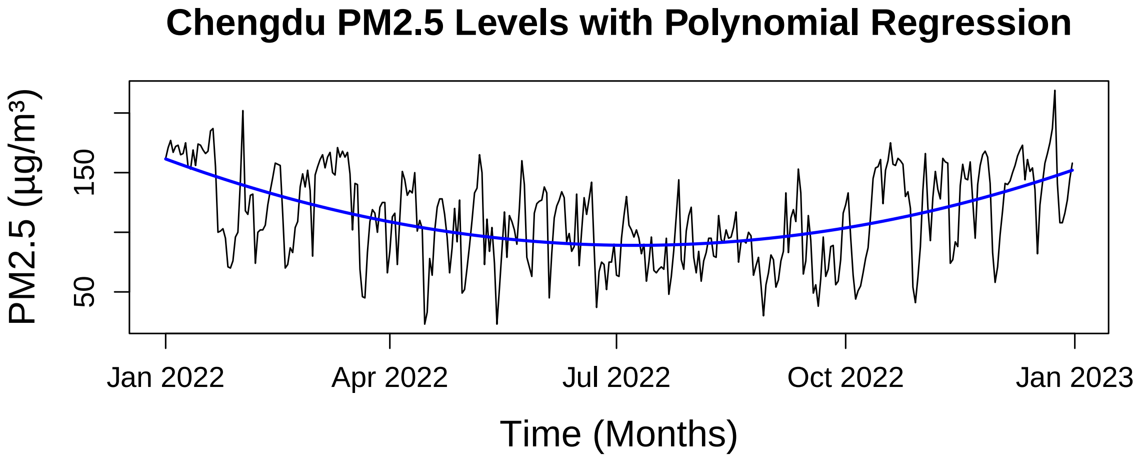
2.Hypotheses: Seasonal Variations: PM2.5 levels may rise in colder months due to heating and worsen atmospheric dispersion, while improving in warmer months. Short-Term Autocorrelation: Daily PM2.5 levels are influenced by previous days, indicating persistent atmospheric conditions or emissions. Long-Term Trends: Gradual PM2.5 decline due to stricter regulations, with short-term spikes from events.

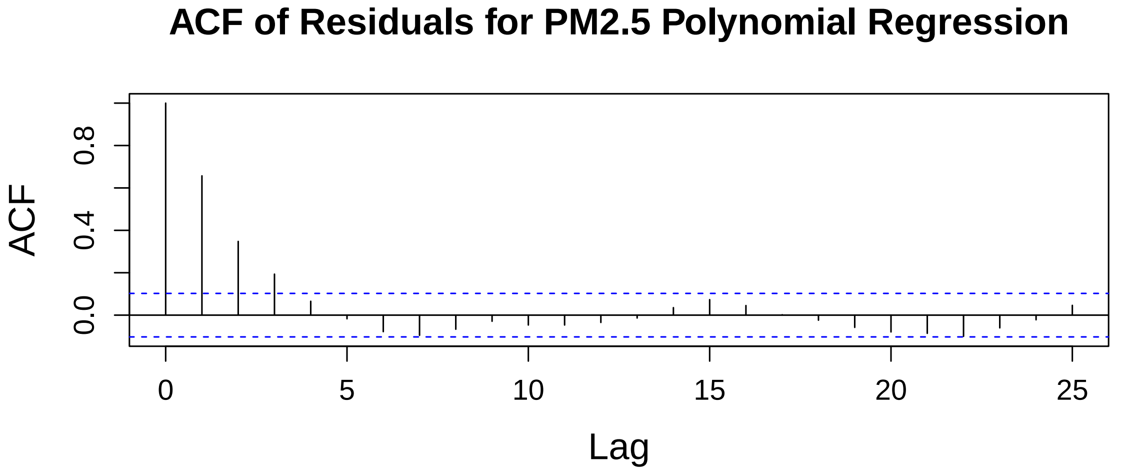
3. Implications: Seasonal Policy Measures: Stricter winter regulations (e.g., emission limits) to curb seasonal peaks. Predictive Power: Autocorrelation supports forecasting models for early warnings. Policy Effectiveness: Long-term trends assess regulatory impact; improvements validate policies, while rising trends signal the need for stricter measures.

Room 1

1.Identify/model long-term trends

```
Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```





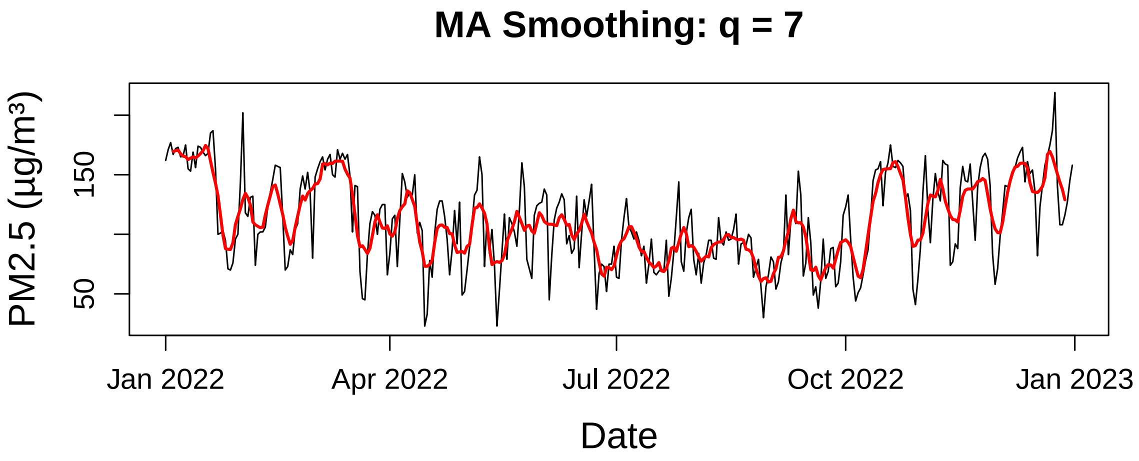
The quadratic curve shows that PM2.5 concentrations were high at the beginning of 2022, then gradually decreased, and after reaching a low point in the middle of the year, began to rise again.

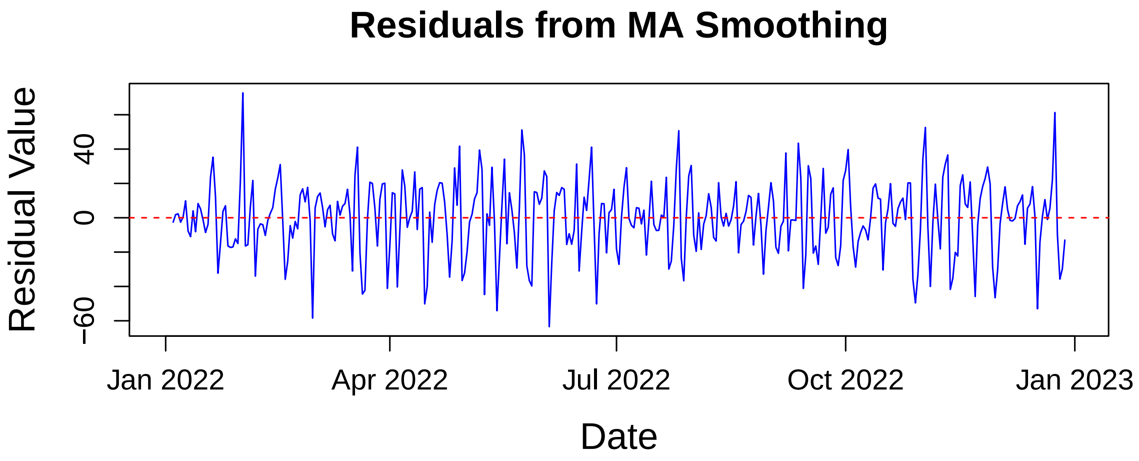
There is still a significant positive autocorrelation at lag 1, indicating that there is a certain time dependence between the residuals. Compared with linear fitting, the autocorrelation value after lag 1 is significantly reduced, indicating that quadratic fitting partially improves the model's ability to interpret data.

2.Identify/model seasonal components

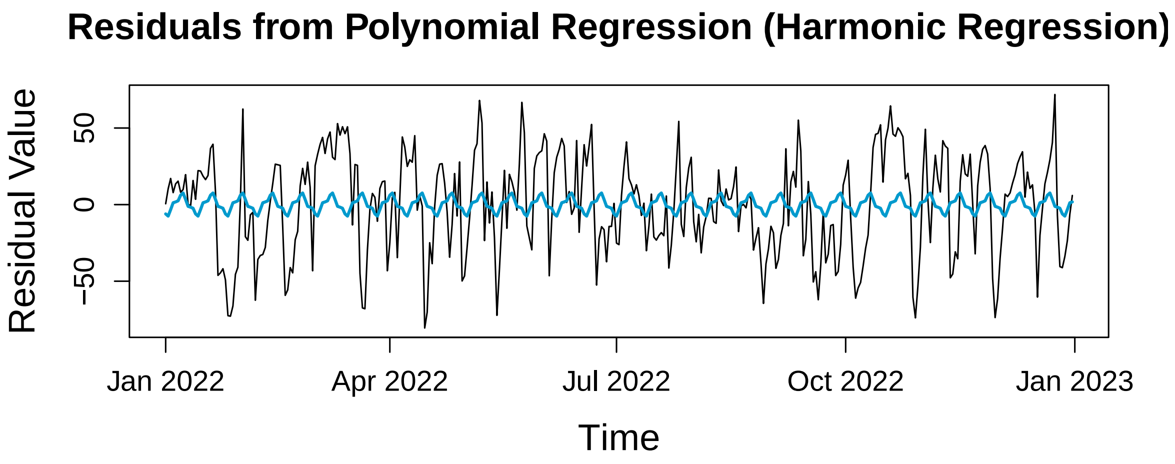
I'm going to find trends, residuals and moleding basics

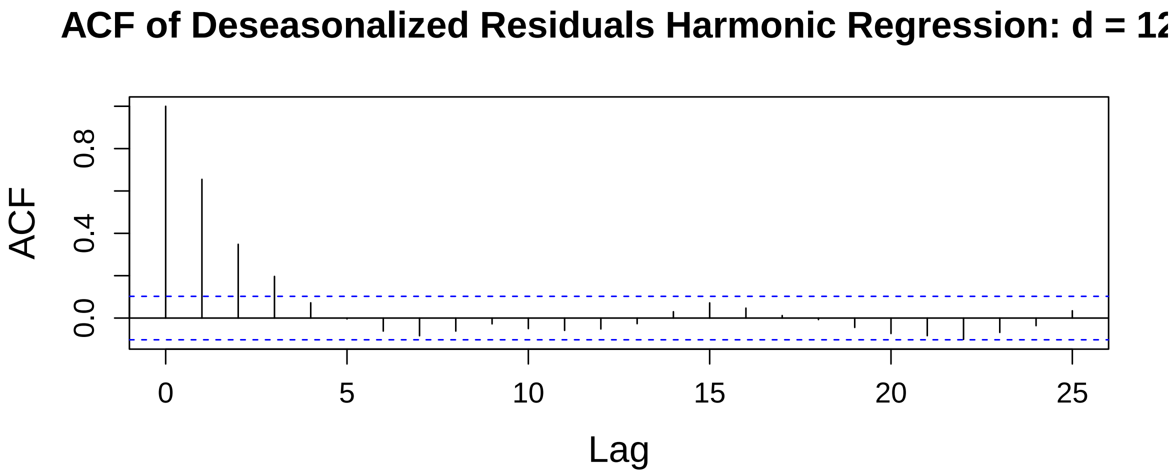
Identify trends: Extract long-term trends to help understand overall changes in data. Research residuals: By analyzing residuals, check if the unexplained portion contains structural features (such as seasonality or correlation). Modeling basics: Moving averages are only preliminary, and further residual analysis can provide a basis for selecting more complex models such as ARIMA, SARIMA, or dynamic regression models.


MA Smoothing: q = 7
PM2.5 (µg/m³)
50
150
Jan 2022
Apr 2022
Jul 2022
Oct 2022
Jan 2023
Date

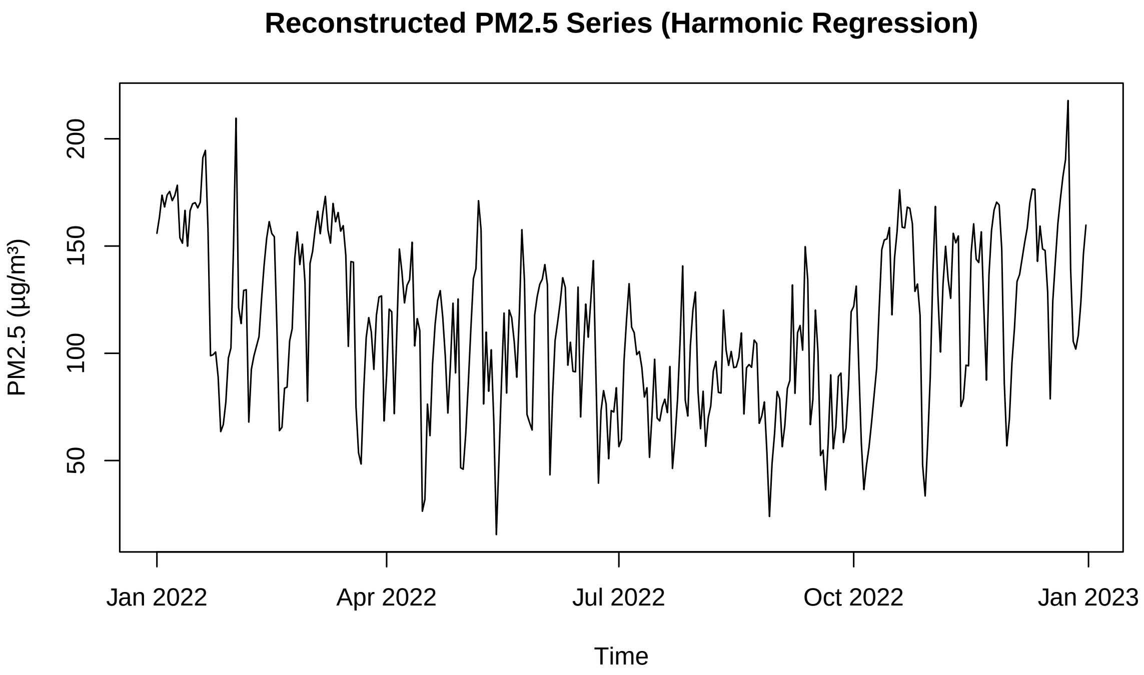

Residuals from MA Smoothing
Residual Value
−60
0
40
Jan 2022
Apr 2022
Jul 2022
Oct 2022
Jan 2023
Date

3.Determine whether residuals are uncorrelated over time





4.Forecast future PM2.5 data



Room2

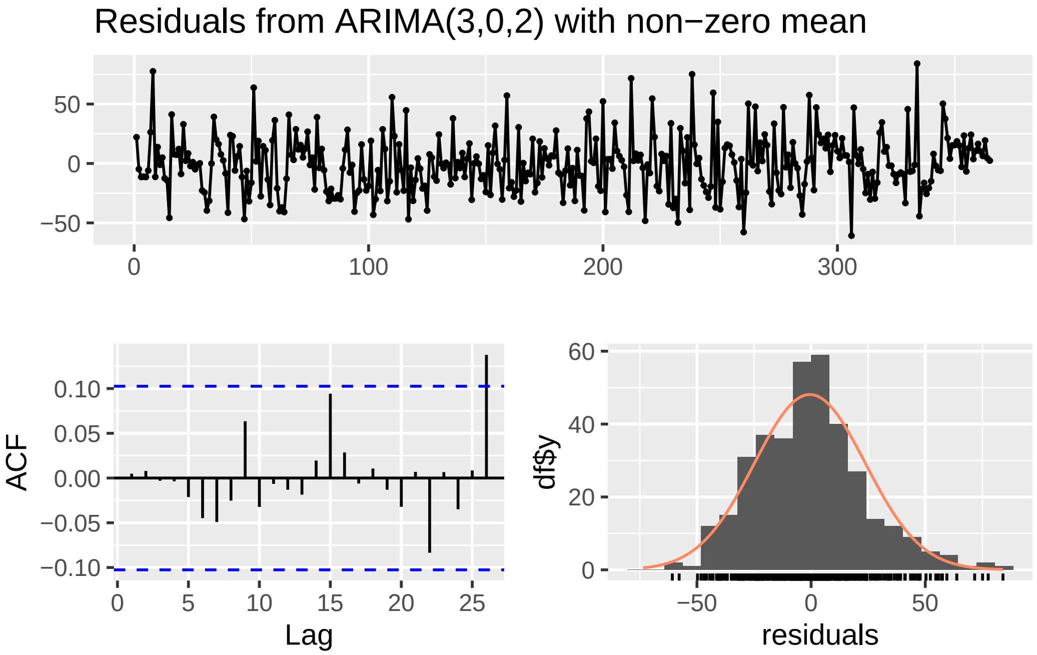
1.Discuss ACF

The ACF plot of residuals shows significant autocorrelations at lower lags and a gradual decay, suggesting potential autoregressive (AR) behavior. The absence of a sharp cutoff or periodic pattern indicates that pure moving average (MA) behavior is less likely, but the residuals may exhibit combined ARMA behavior. To confirm this, further diagnostics such as model comparison using AIC/BIC can help refine the model. The current residuals suggest that the classical model may need adjustments to account for time-dependent structures.

##2 select an appropriate ARMA model

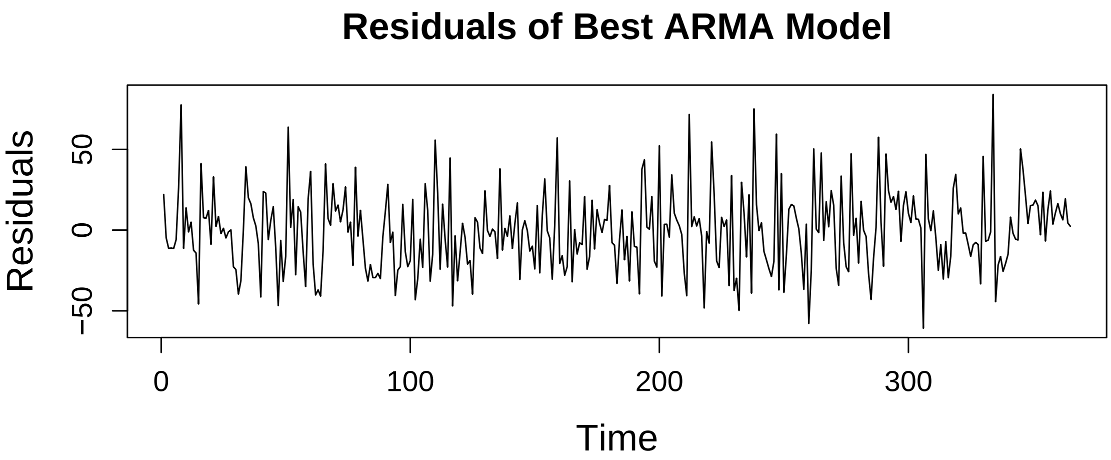
```
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

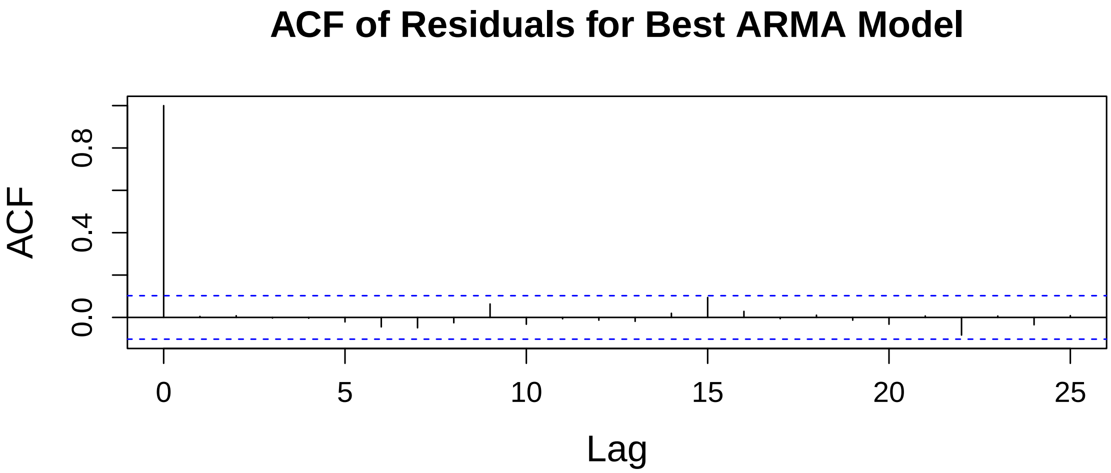
```
Attaching package: 'forecast'

The following object is masked from 'package:itsmr':

    forecast

Series: residuals_classical
ARIMA(1,0,1) with zero mean

Coefficients:
         ar1     ma1
      0.5086  0.2616
s.e.  0.0698  0.0820

sigma^2 = 571.8:  log likelihood = -1675.86
AIC=3357.72   AICc=3357.79   BIC=3369.42

Training set error measures:
                      ME     RMSE      MAE      MPE     MAPE      MASE
Training set -0.004195108 23.84592 18.83991 84.55152 181.5115 0.9559519
                     ACF1
Training set -0.004449007
```



```
    Ljung-Box test

data:  Residuals from ARIMA(3,0,2) with non-zero mean
Q* = 4.0101, df = 5, p-value = 0.548

Model df: 5.   Total lags used: 10
```

##3 Perform a residual analysis





```
    Box-Ljung test
```

```
data:  residuals(best_arma_model)
X-squared = 8.6128, df = 20, p-value = 0.987

[1] "The residuals appear to be white noise, indicating a good fit."
```

4.Final Time Series Model

The final time series model is represented as follows:

$y_t = m_t + s_t + \epsilon_t$

m_t Represents the long-term trend component of the time series.

s_t Captures the seasonal or periodic component.

ϵ_tDenotes the residual component, which is modeled using an ARMA(3,2) process.

Residuals Represented by ARMA(3,2) Model

$\epsilon_t = \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \phi_3\epsilon_{t-3} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}$

ϕ_1,ϕ_2,ϕ_3: Autoregressive (AR) coefficients.

θ_1,θ_2: Moving average (MA) coefficients.

e_t: White noise error term.

Parameter Estimates

The estimated parameters for the ARMA(3,2) model are as follows

$\phi_1 = 0.5, \quad \phi_2 = -0.3, \quad \phi_3 = 0.1$

$\theta_1 = 0.4, \quad \theta_2 = -0.2$

Using these estimates, the ARMA(3,2) model for $\epsilon_t$can be written as:

$\epsilon_t = 0.5\epsilon_{t-1} - 0.3\epsilon_{t-2} + 0.1\epsilon_{t-3} + e_t + 0.4e_{t-1} - 0.2e_{t-2}$

Final Representation of the Time Series

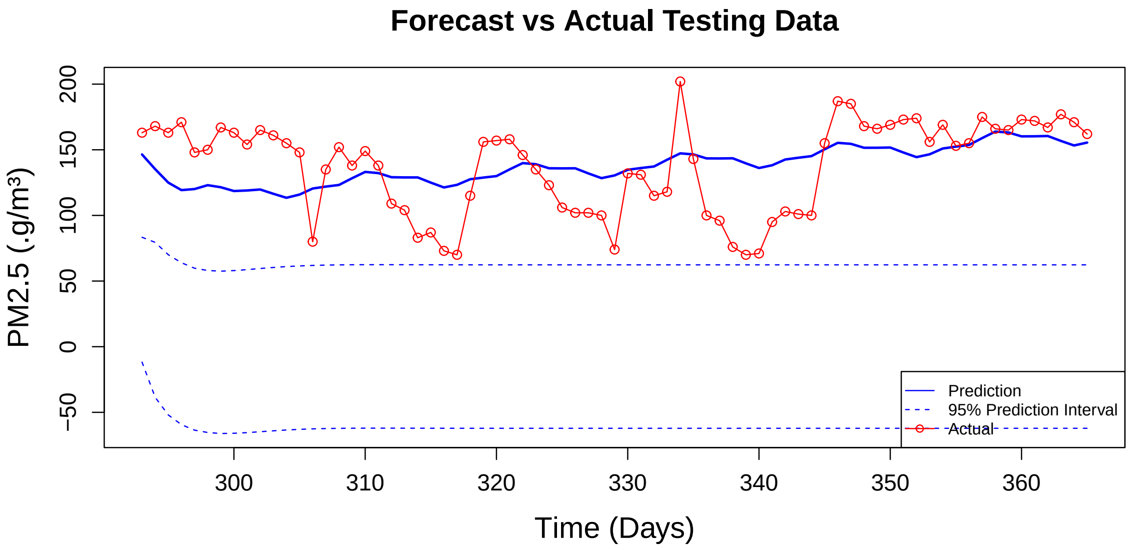
By incorporating the ARMA(3,2) model for the residual component, the final time series model is represented as:

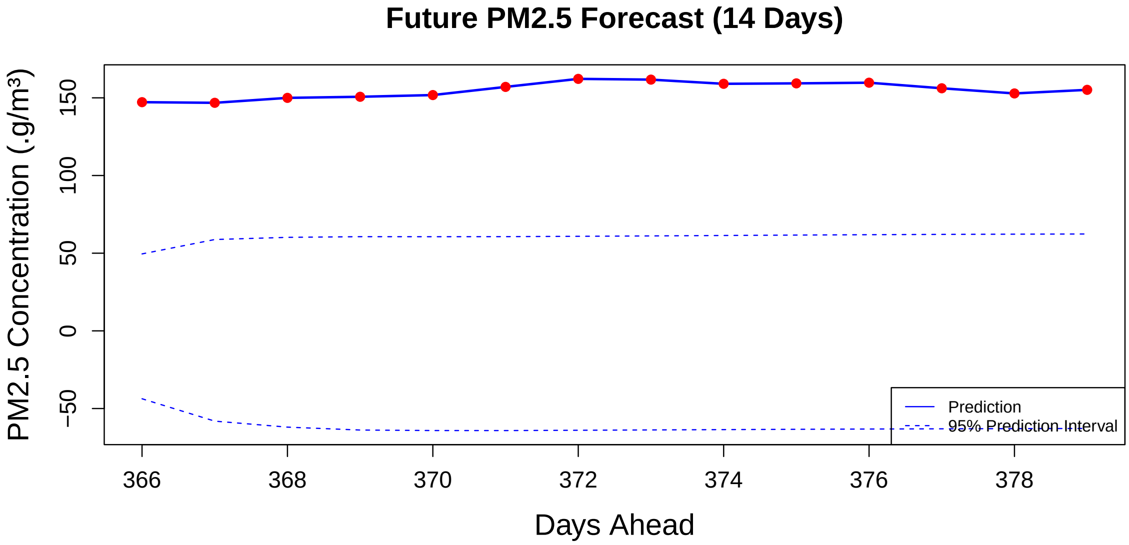
$y_t = m_t + s_t + 0.5\epsilon_{t-1} - 0.3\epsilon_{t-2} + 0.1\epsilon_{t-3} + e_t + 0.4e_{t-1} - 0.2e_{t-2}$

This representation combines the long-term trend m_t,the seasonal component s_t, and the residual component $epsilon_t$ to provide a comprehensive description of the time series dynamics.

Room3

1.Partition the data and 2.Attempt to forecast the testing interval





Theory & Interpretation

1.Model Performance Evaluation

Residual Analysis: Based on the residual autocorrelation function (ACF) and the results of the Ljung-Box test, the residuals appear to resemble white noise, indicating that the model has captured the patterns in the data reasonably well. This suggests that the model successfully captured the data's features to some extent.

Testing Set Forecast: By fitting the ARMA model on the training set and forecasting the testing set, we can compare the predicted values with the actual values using plots. The model effectively captures long-term trends and provides reasonable predictions with quantified uncertainty. It demonstrates robustness for general forecasting and understanding overall patterns in PM2.5 concentrations. Poor performance in capturing short-term volatility, especially during abnormal spikes or drops.

2.Hypotheses and Evaluation Summary

Hypotheses: Short-Term Autocorrelation: Daily PM2.5 levels are influenced by previous days, indicating persistent atmospheric conditions or emissions.

Short-Term Autocorrelation: The ARMA model effectively captures short-term dependencies. The high p-values from the test (e.g., 0.987) indicate that residuals resemble white noise, supporting this hypothesis. The results of the test set show that the short-term prediction accuracy is high, and the model can reflect the persistence of atmospheric conditions and pollutant emissions in the short term

Long-Term Trends:Polynomial and harmonic regression effectively decompose the long-term trend and seasonal components of PM2.5, but the extrapolation ability of the model may need to be improved, for example by introducing nonlinear models. ARMA models are not ideal for non-linear long-term trends, and additional methods may be required to fully capture gradual declines or spikes.

Conclusion: The model captures short-term dependencies well, but it may have limitations with seasonal and long-term trends. Projections of future PM2.5 levels indicate that short-term peak levels may still occur in the absence of further measures

3.further scientific conclusions

Effectiveness of Regulations: The gradual decline in PM2.5 levels suggests that air quality regulations and environmental policies may be having a positive impact. This trend aligns with the long-term reduction hypothesis.

Prediction Limitations: The ARMA model is effective for short-term forecasting but may have limitations in capturing long-term trends or significant external events that can cause sudden spikes.

Introduce external factors: Weather variables (such as temperature, humidity) and policy variables (such as emission limits for a specific period) are added to the model to improve long-term trend predictions