Time Series Analysis of Air Quality

Kaiyang LIU	1830004016
Ganrui CAI	1830005020
Sihong LIN	1830005002
Yiqi SHAN	1830005028
Yunzhi XIAO	1830006195

CONTENTS

Background Information

Background Information

What is the AQI

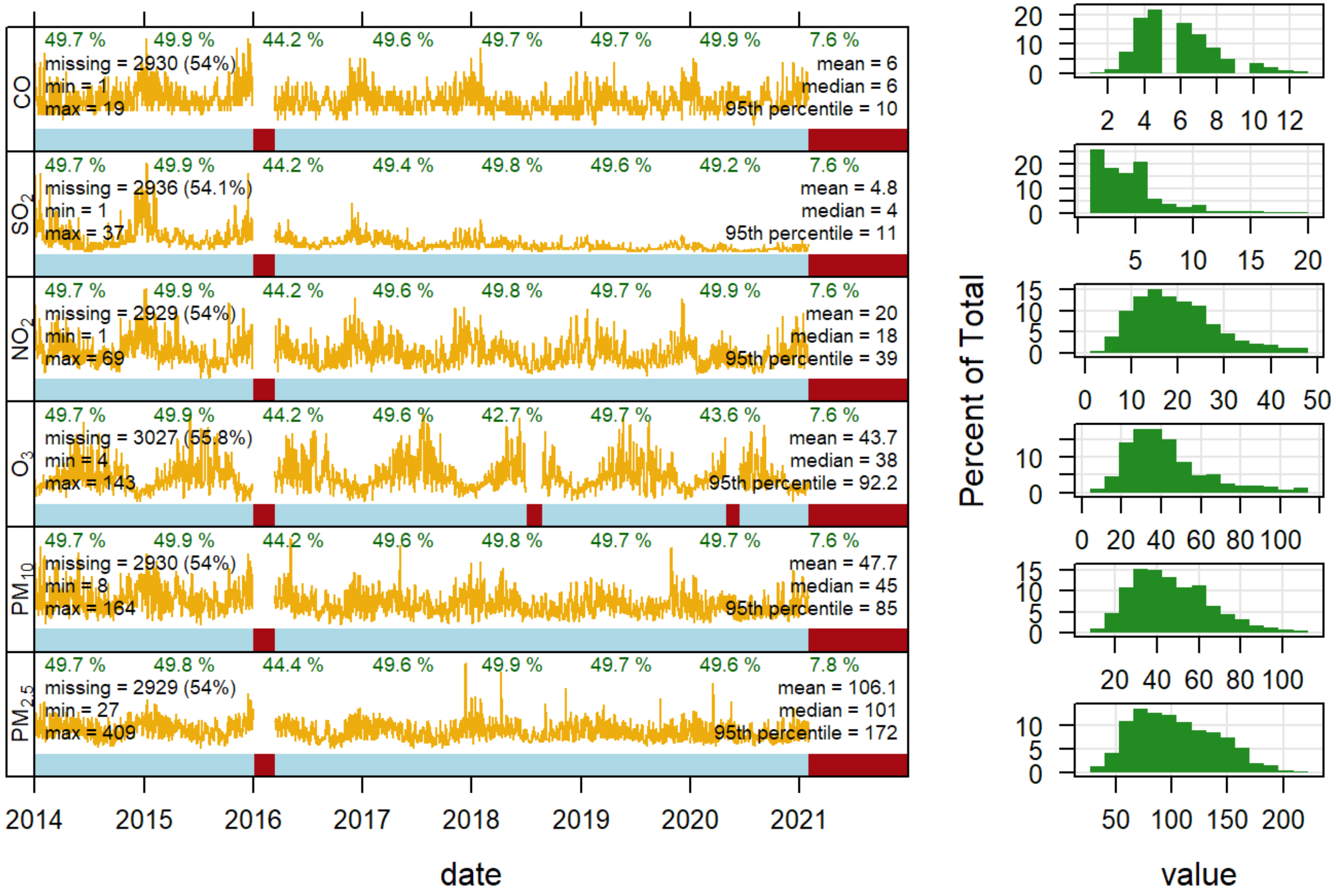
The full name of AQI is Air Quality Index, is an index for reporting daily air quality.

Is calculated for four major air pollutants regulated by the Clean Air Act:

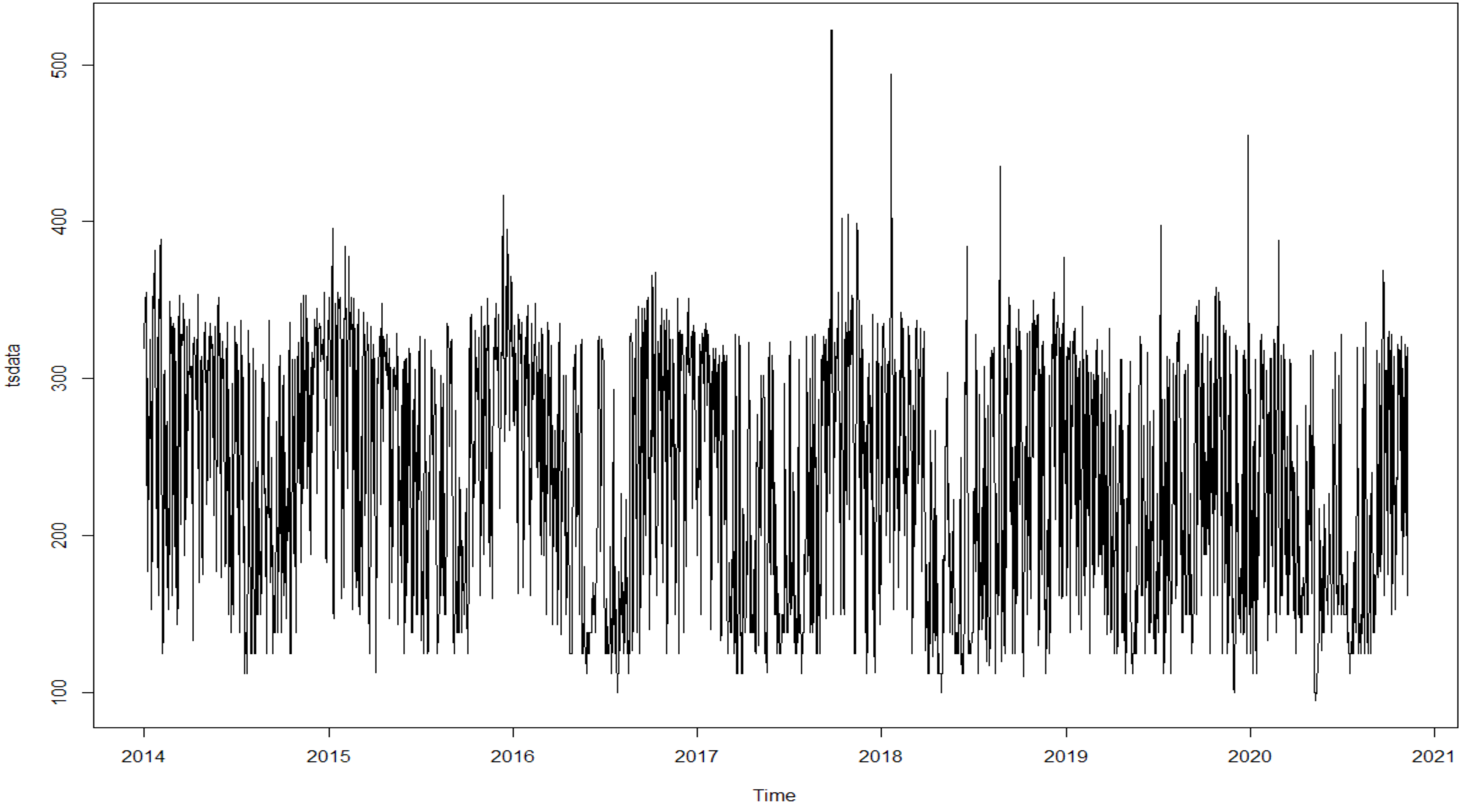
- **Ground level ozone**
- **Particle pollution**
- **Carbon monoxide**
- **Sulfur dioxide**

Air Quality Index (AQI) Values	Levels of Health Concern	Colors
When the AQI is in this range:	*...air quality conditions are:*	*...as symbolized by this color:*
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

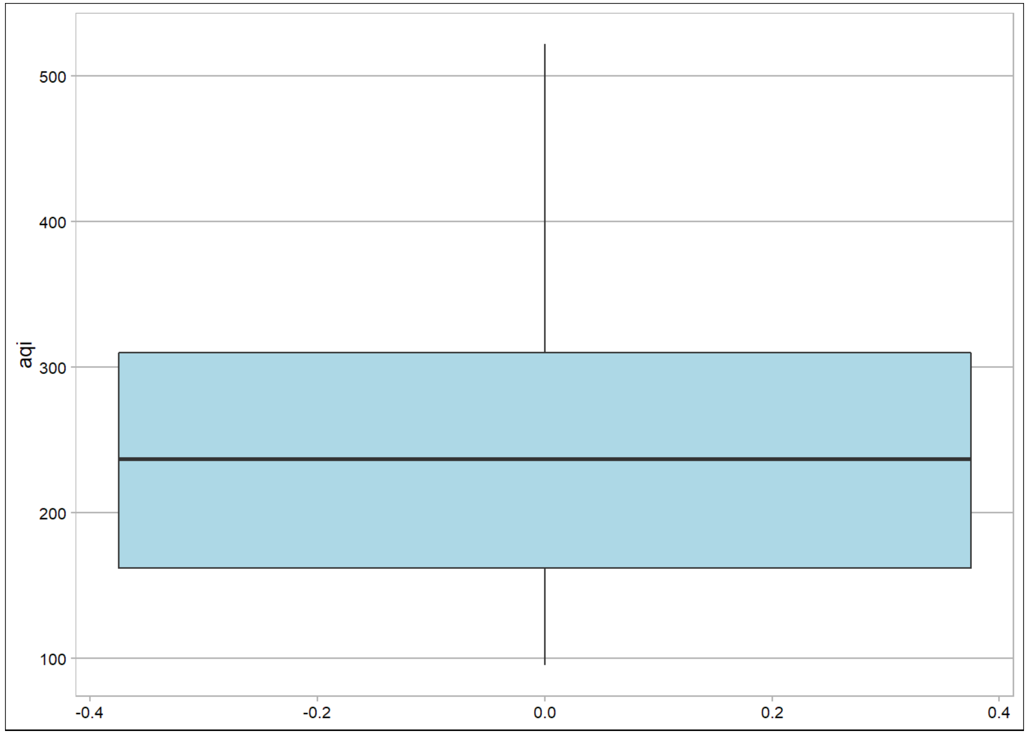
Data Overview

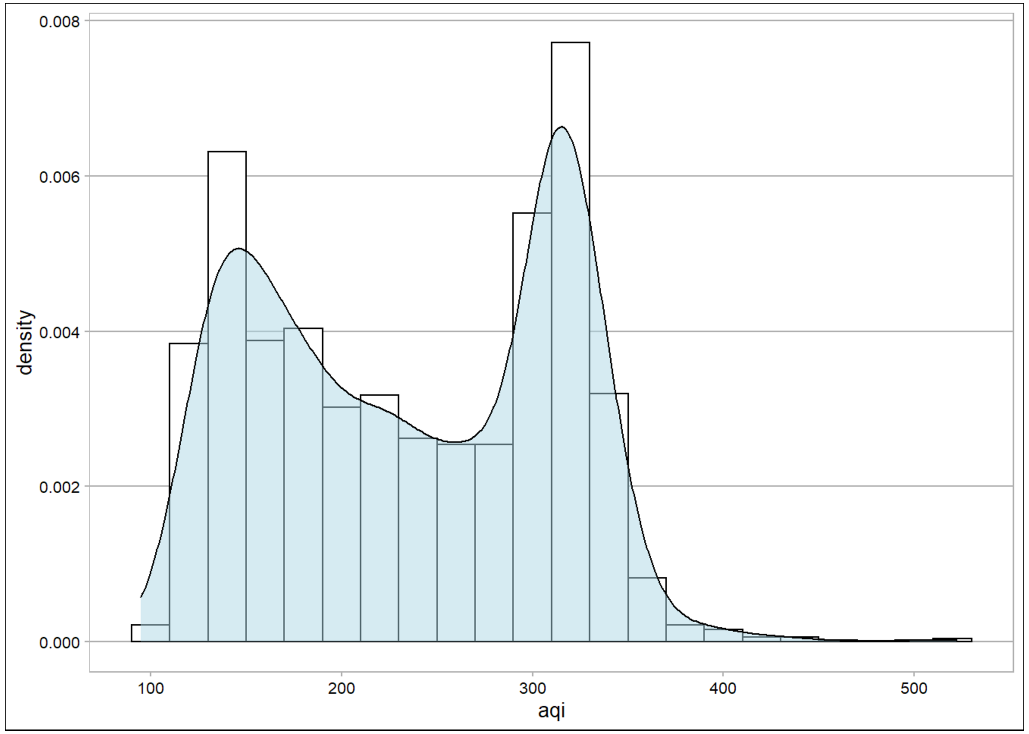


Data Overview

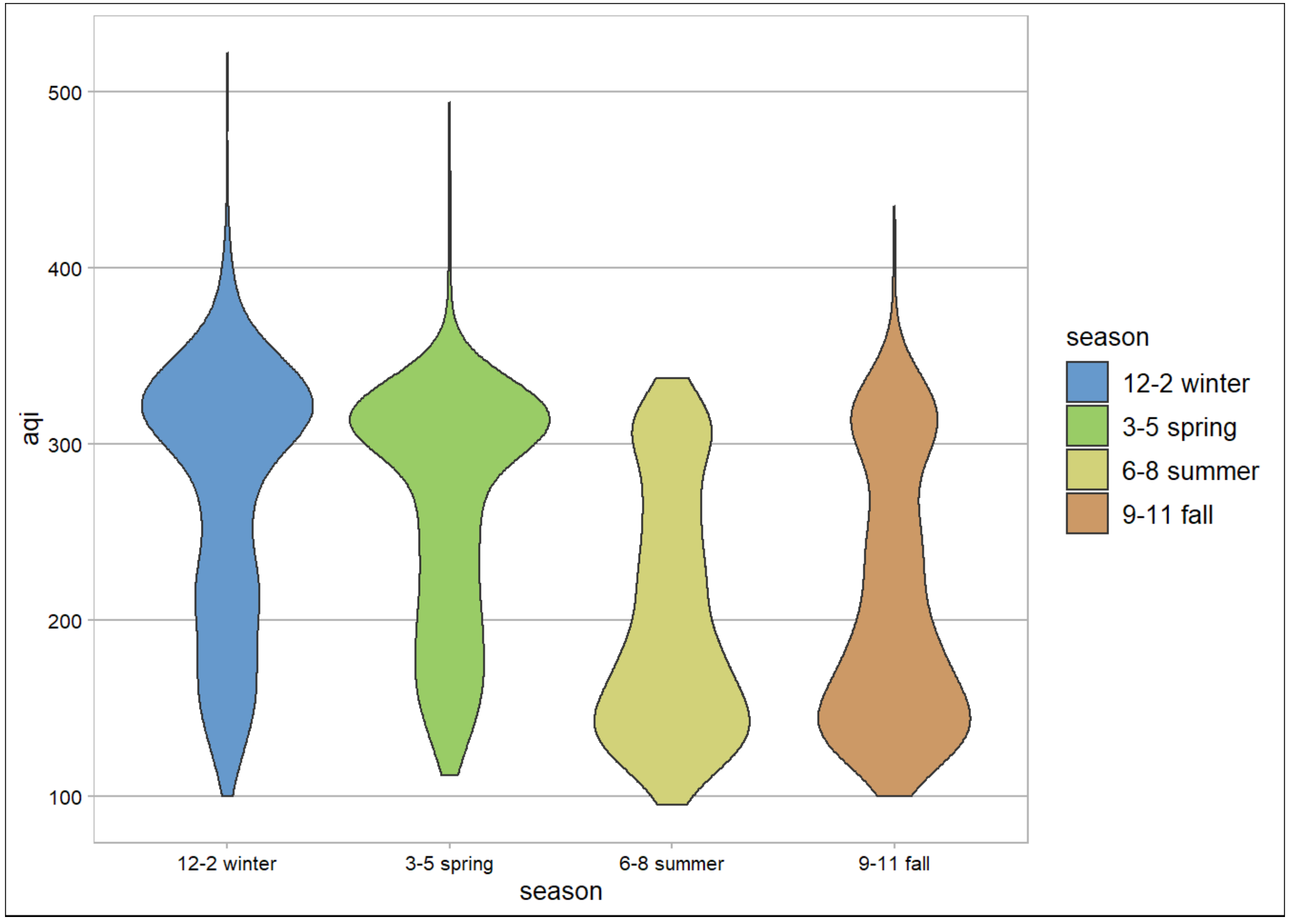


Data Overview

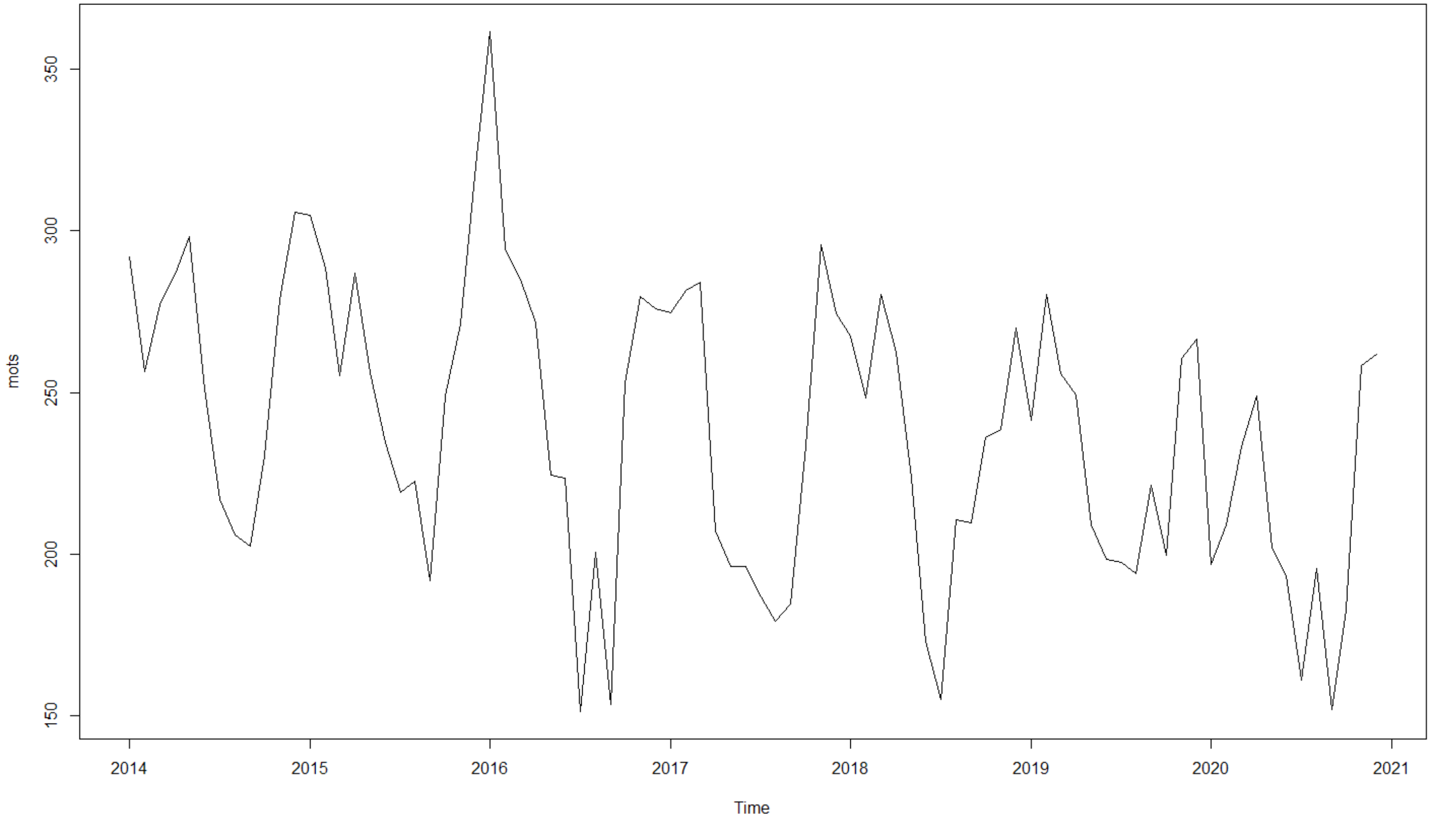




Data Overview



Data Overview



Exponential Smoothing

Exponential Smoothing

Holt-Winters method

- Holt (1957) and Winters (1960) extended Holt's method to capture seasonality.

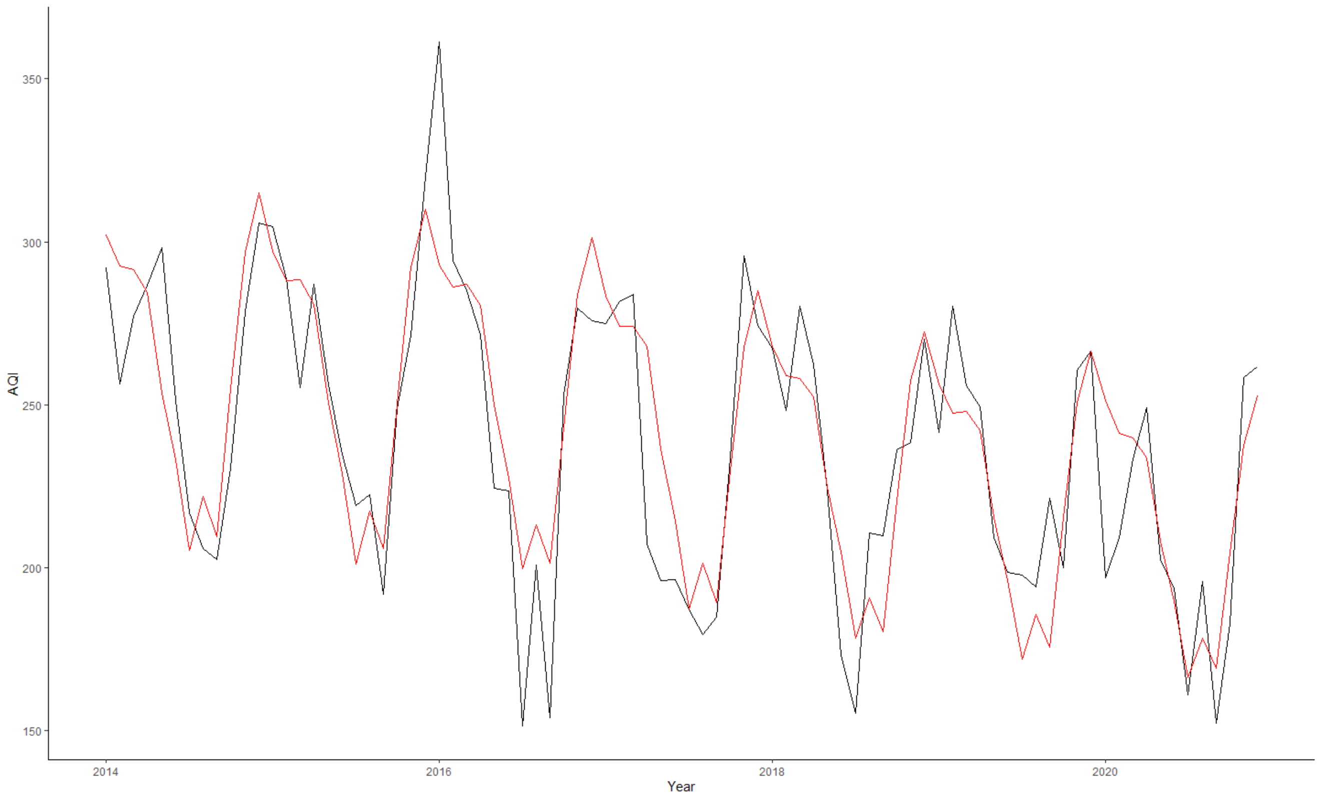
- The Holt-Winters seasonal method comprises the forecast equation and three smoothing equations — one for the level l_t, one for the trend b_t and one for the seasonal component s_t, with corresponding smoothing parameters α, β, γ.

Multiplicative Holt-Winters

Theorem (Multiplicative Holt-Winters Method)

$$\hat{y}_{t+h|t} = (l_t + hb_t)s_{t+h-m(k+1)}$$
$$l_t = \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1})$$
$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma\frac{y_t}{l_{t-1} + b_{t-1}} + (1-\gamma)s_{t-m}$$

Multiplicative Holt-Winters



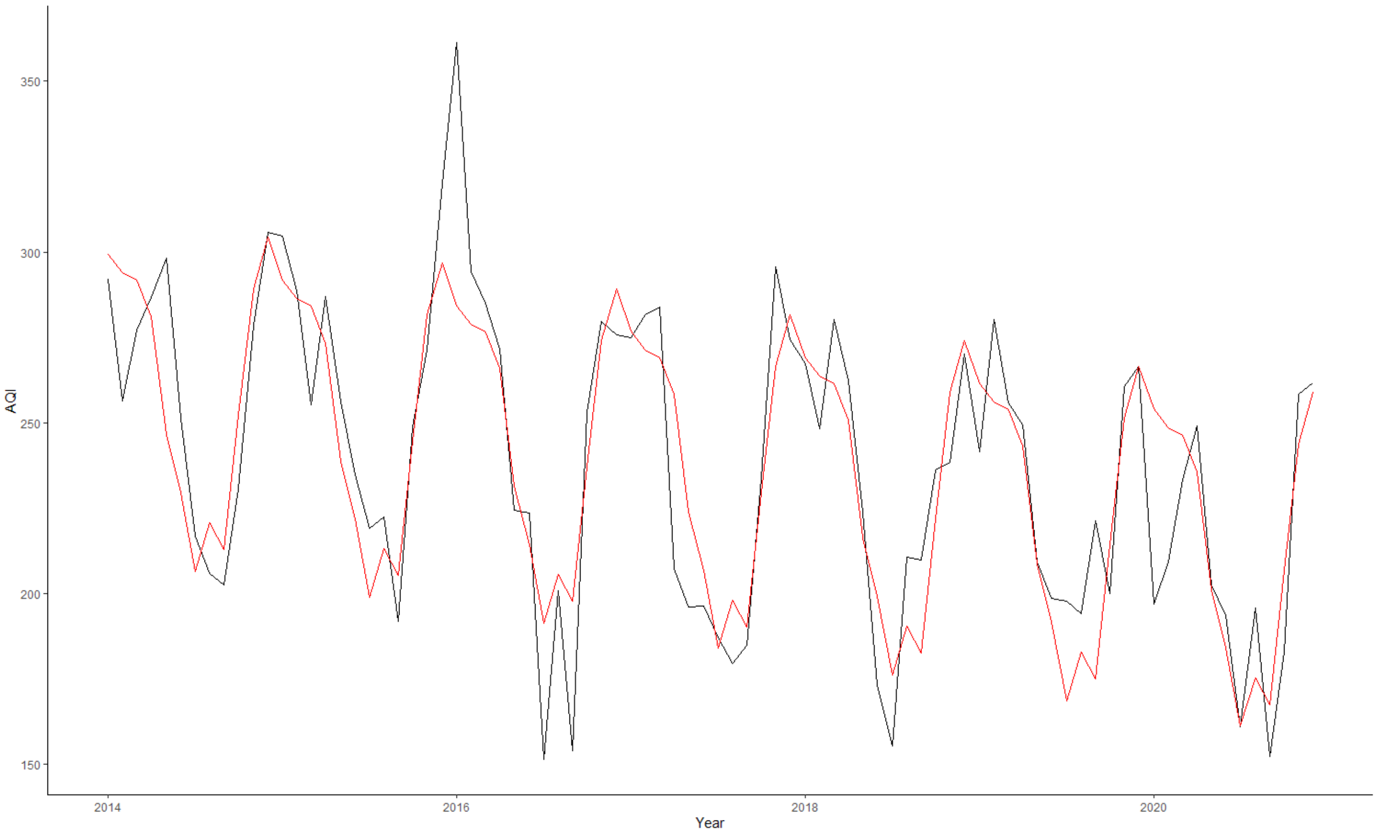

```
> fi2$model
Holt-Winters' multiplicative method

Call:
 hw(y = mots, seasonal = "multiplicative")

  Smoothing parameters:
    alpha = 0.0335
    beta  = 0.0023
    gamma = 1e-04

  Initial states:
    l = 265.3479
    b = -0.1847
    s = 1.2053 1.1322 0.9681 0.7937 0.8381 0.7773
           0.8837 0.9666 1.0835 1.1073 1.1052 1.1389

  sigma:  0.1056

     AIC     AICc      BIC
929.9375 939.2102 971.2613
```

Additive Holt-Winters

Theorem (Additive Holt-Winters)

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)}$$
$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1})$$
$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$$

Additive Holt-Winters




```
> fi1$model
Holt-Winters' additive method

Call:
 hw(y = mots, seasonal = "additive")

  Smoothing parameters:
    alpha = 1e-04
    beta  = 1e-04
    gamma = 1e-04

  Initial states:
    l = 264.7526
    b = -0.6348
    s = 47.1914 31.5059 -6.2318 -46.2376 -39.0541 -53.835
           -31.563 -15.1653 18.7463 28.8623 30.4688 35.3119

  sigma:  24.0593

     AIC     AICc      BIC
922.7664 932.0391 964.0903
```

4 in 1 Plot

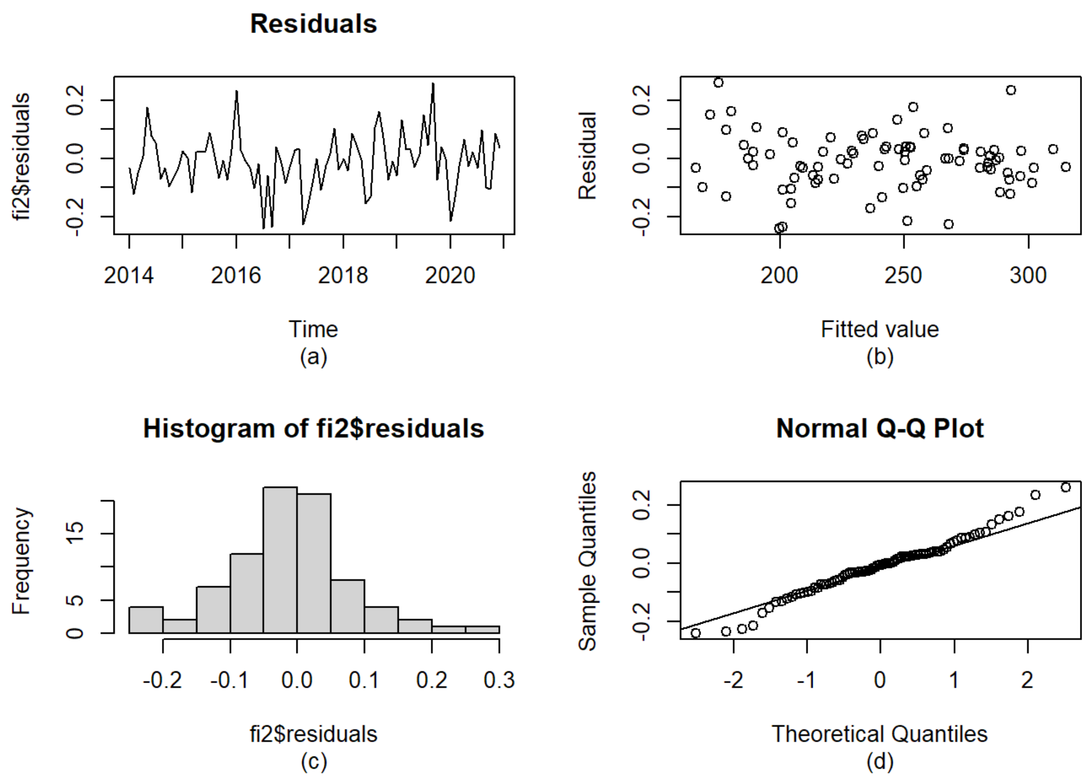


Figure: Multiplicative Holt-Winters

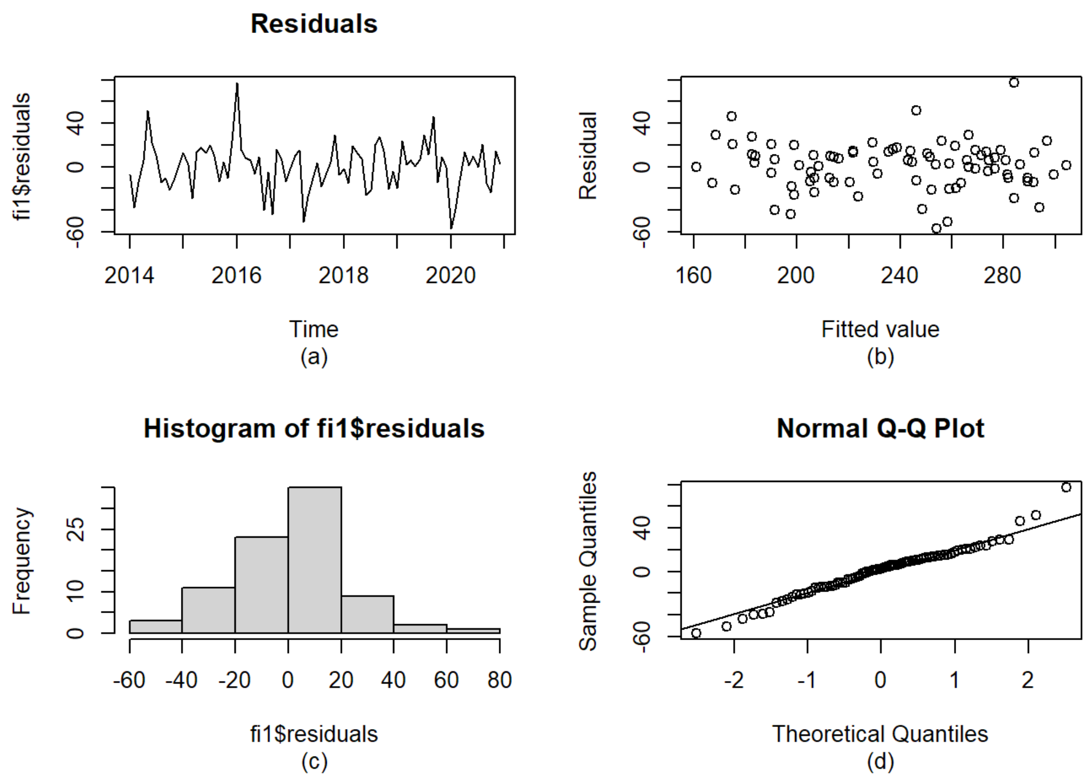


Figure: Additive Holt-Winters

Model Testing

The Ljung-Box test

The Ljung-Box test uses the following hypotheses:

- H_0: The residuals are independently distributed.
- H_A : The residuals are not independently distributed; they exhibit serial correlation.

```
accuracy(fi1)
```

```
##                     ME     RMSE      MAE        MPE     MAPE      MASE      ACF1
## Training set 0.374896 21.64702 16.40522 -0.6840305 7.264644 0.6839709 0.2353069
```

```
accuracy(fi2)
```

```
##                      ME     RMSE      MAE       MPE     MAPE     MASE      ACF1
## Training set -2.868968 22.01595 16.56271 -2.067032 7.438094 0.690537 0.2277724
```

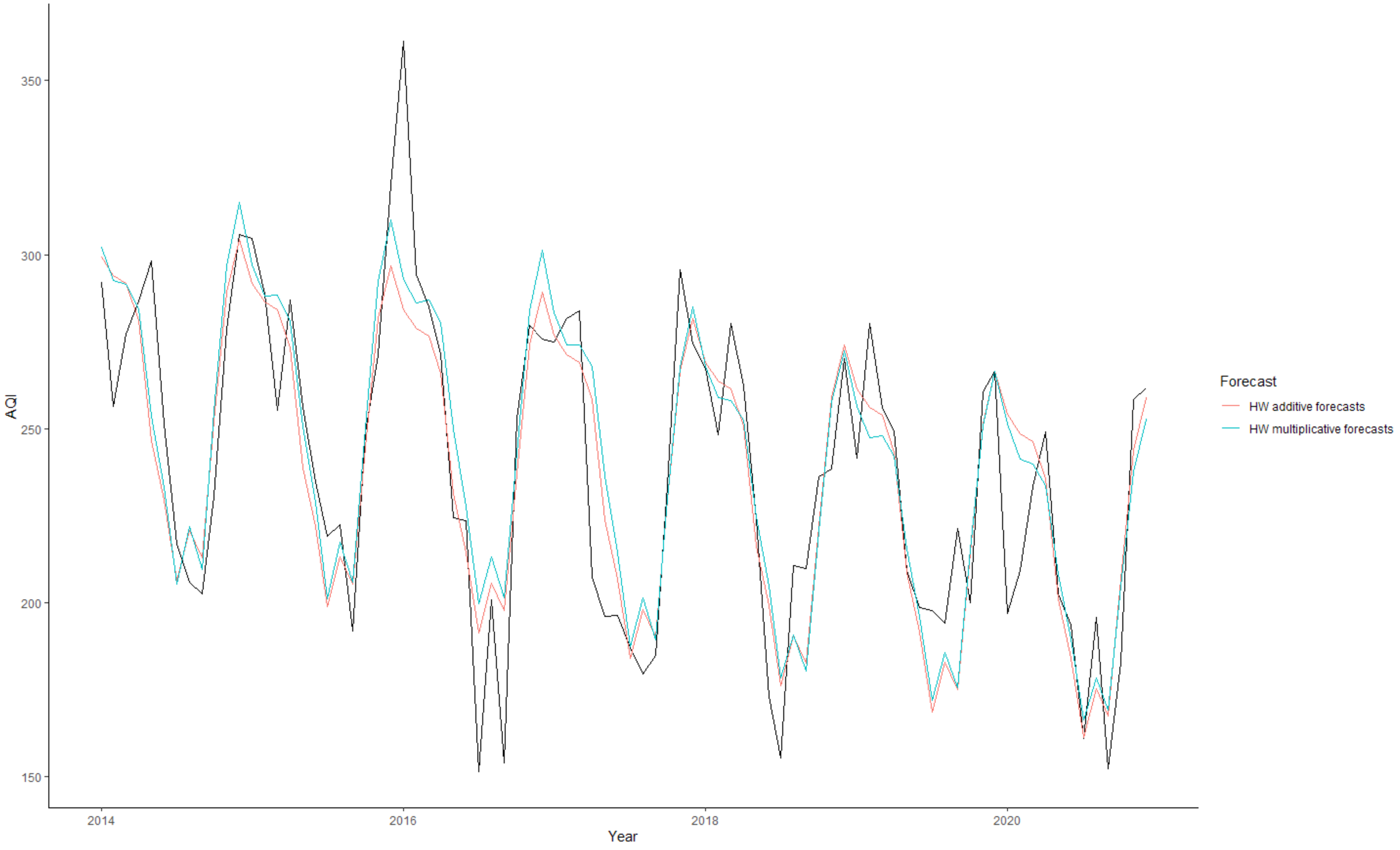
```
Box.test(fi1$residuals,lag = 1,type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  fi1$residuals
## X-squared = 4.8191, df = 1, p-value = 0.02815
```

```
Box.test(fi2$residuals,lag = 1,type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  fi2$residuals
## X-squared = 4.1404, df = 1, p-value = 0.04187
```

Model Comparison



Forecasting

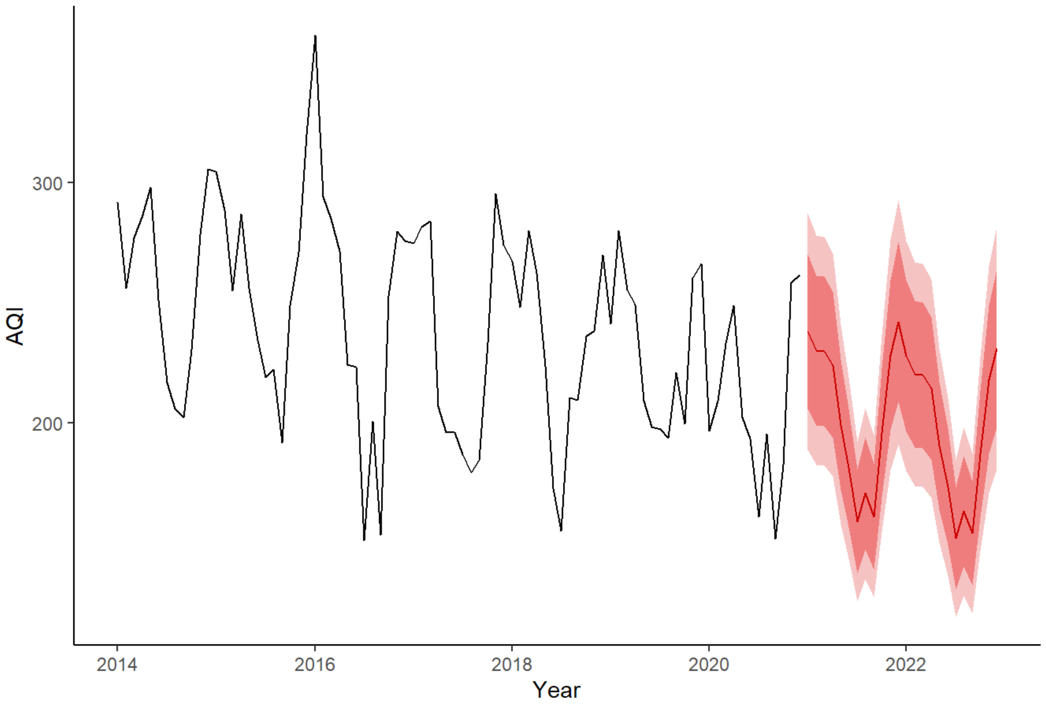


Figure: Multiplicative Holt-Winters

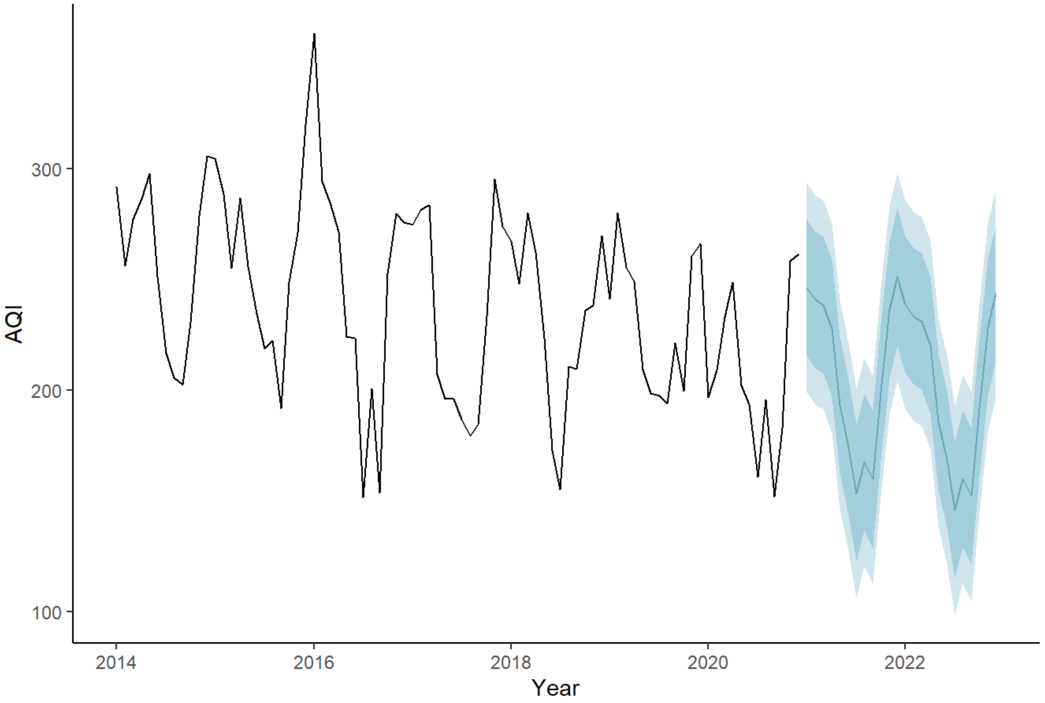
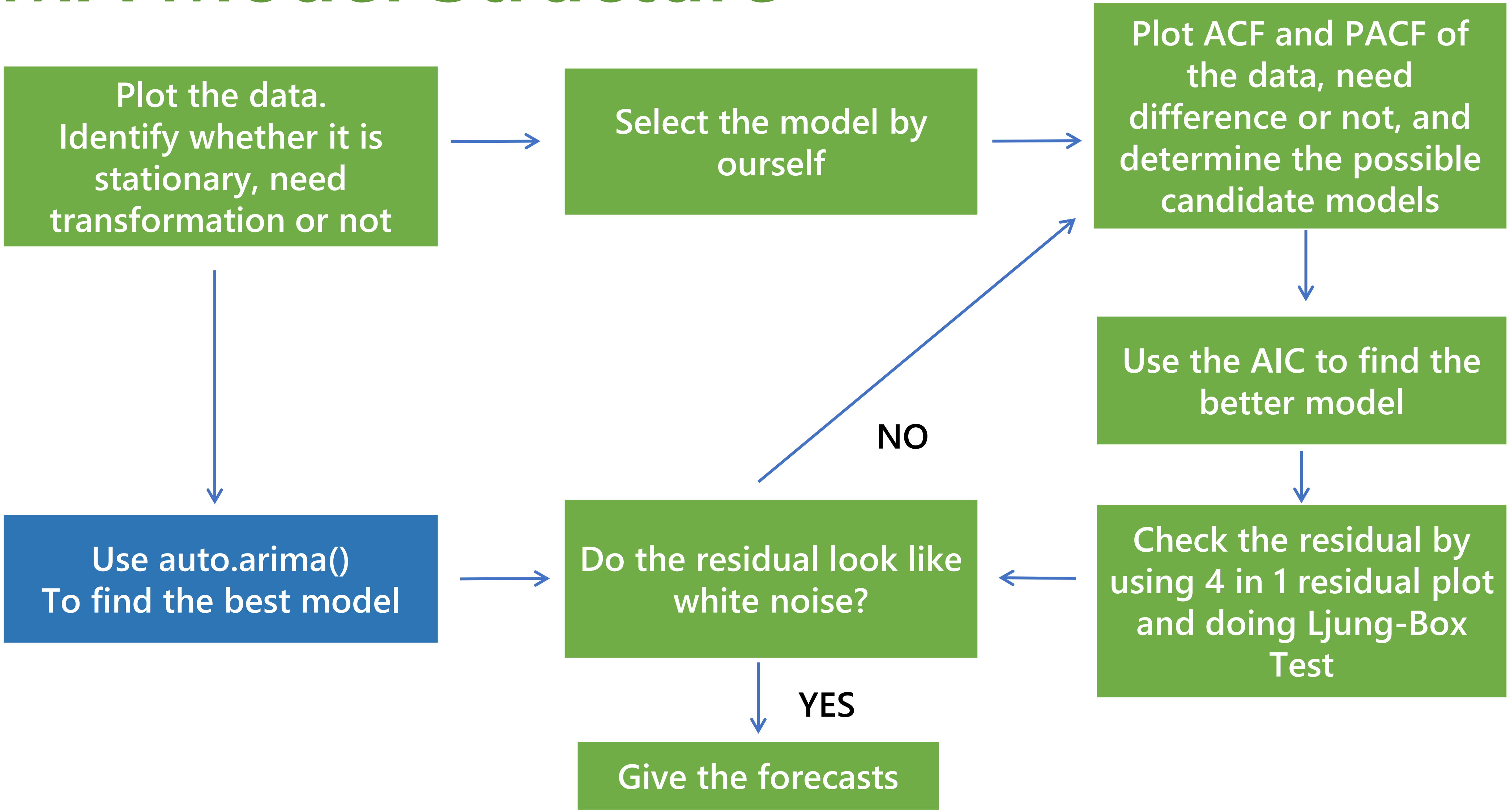


Figure: Additive Holt-Winters

Autoregressive Integrated Moving Average

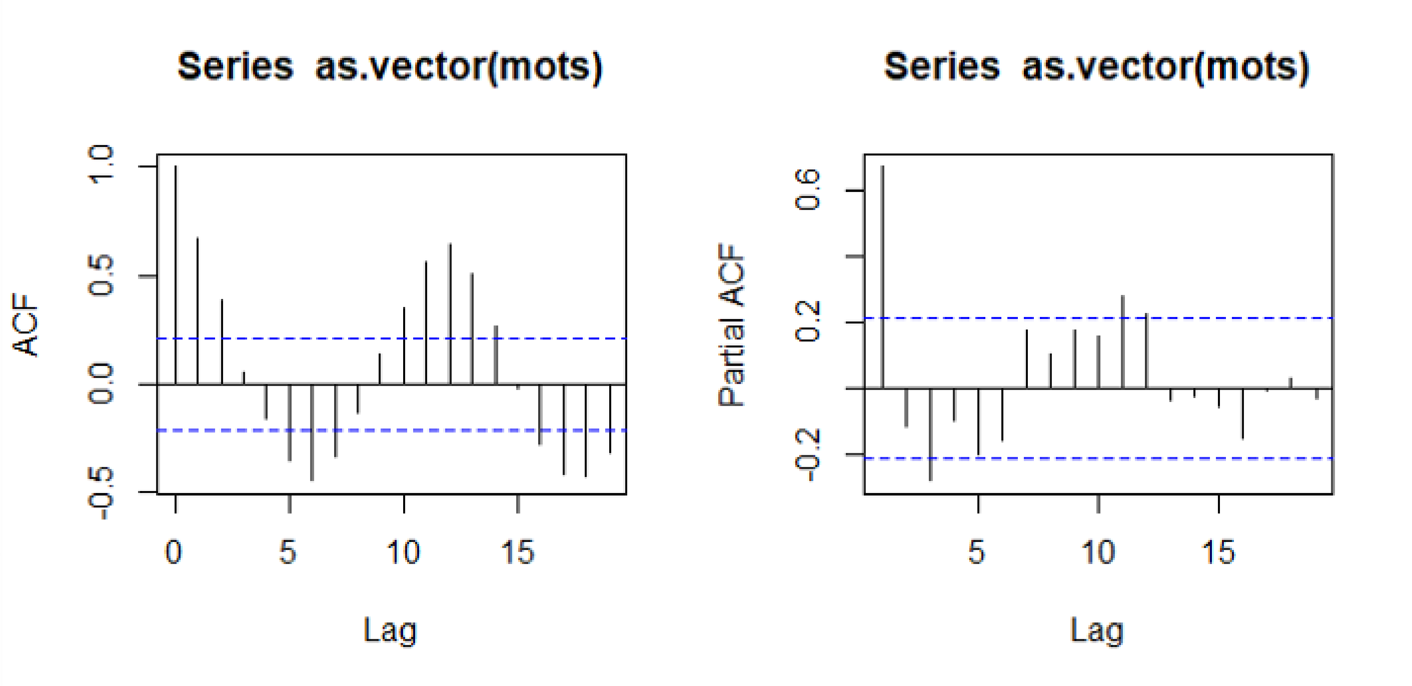
ARIMA Model Structure



Data Detection

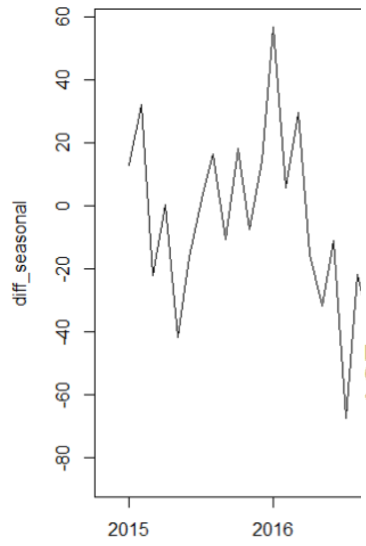
The ADF test

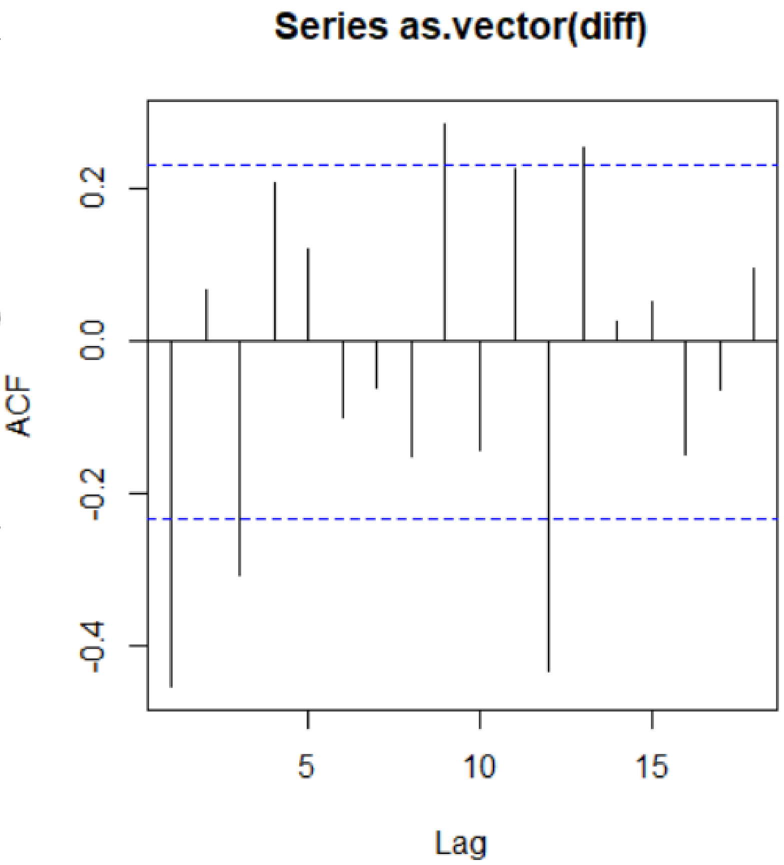
Augmented Dickey Fuller test (ADF Test) is a common statistical test used to test whether a given Time series is stationary or not.

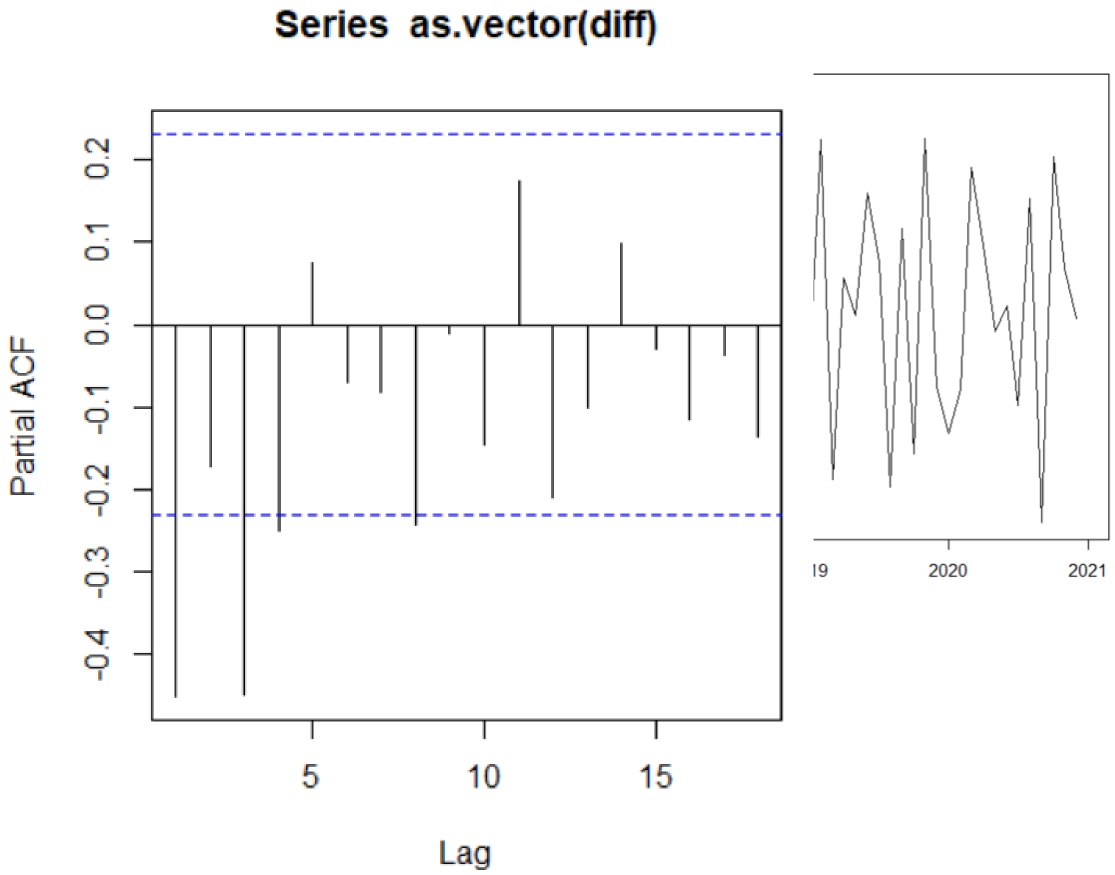



```
> adf.test(mots,alt = "stationary")

        Augmented Dickey-Fuller Test

data:  mots
Dickey-Fuller = -6.5421, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Remove the seasonal then take first difference







Model construction

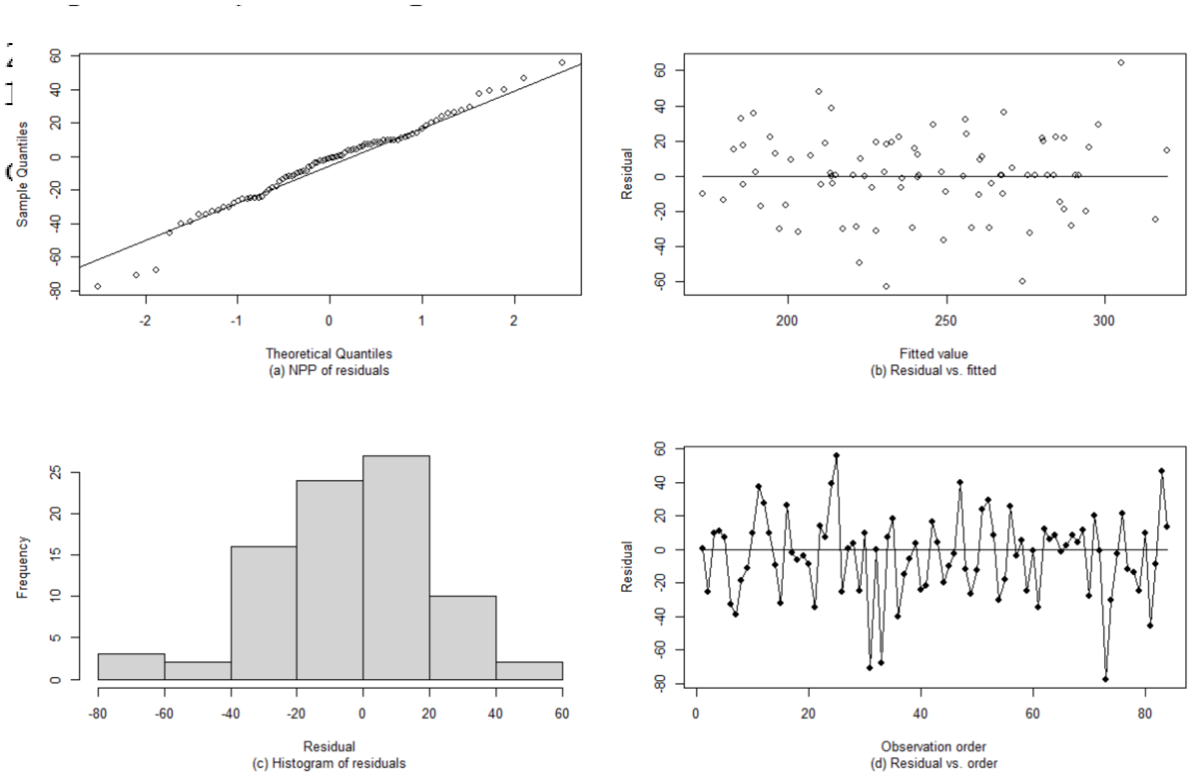
- Model 1 : optimal algorithm

```
aics <- matrix(0,6,6,dimnames = list(p = 0:5, q = 0:5))

for (q in 1:5){
  aics[1,1+q] <- arima(mots,c(0,1,q))$aic
}

for (p in 1:5){
  for (q in 0:5){
    aics[1+p,1+q] <- arima(mots,c(p,1,q))$aic
  }
}
```

```
> aics
   q
p          0        1        2        3        4        5
  0   0.0000 827.7217 829.3625 815.6023 816.5706 818.4856
  1 827.6985 829.0729 830.2993 816.4716 818.4614 820.2657
  2 829.2668 830.6150 829.4969 818.4545 819.9830 821.9217
  3 828.8085 812.8813 798.0615 820.4983 822.3339 821.1103
  4 830.1852 813.2300 815.8662 801.8518 795.6890 791.7156
  5 831.3815 809.0995 793.9694 791.7252 793.6138 794.4770
```

Model construction

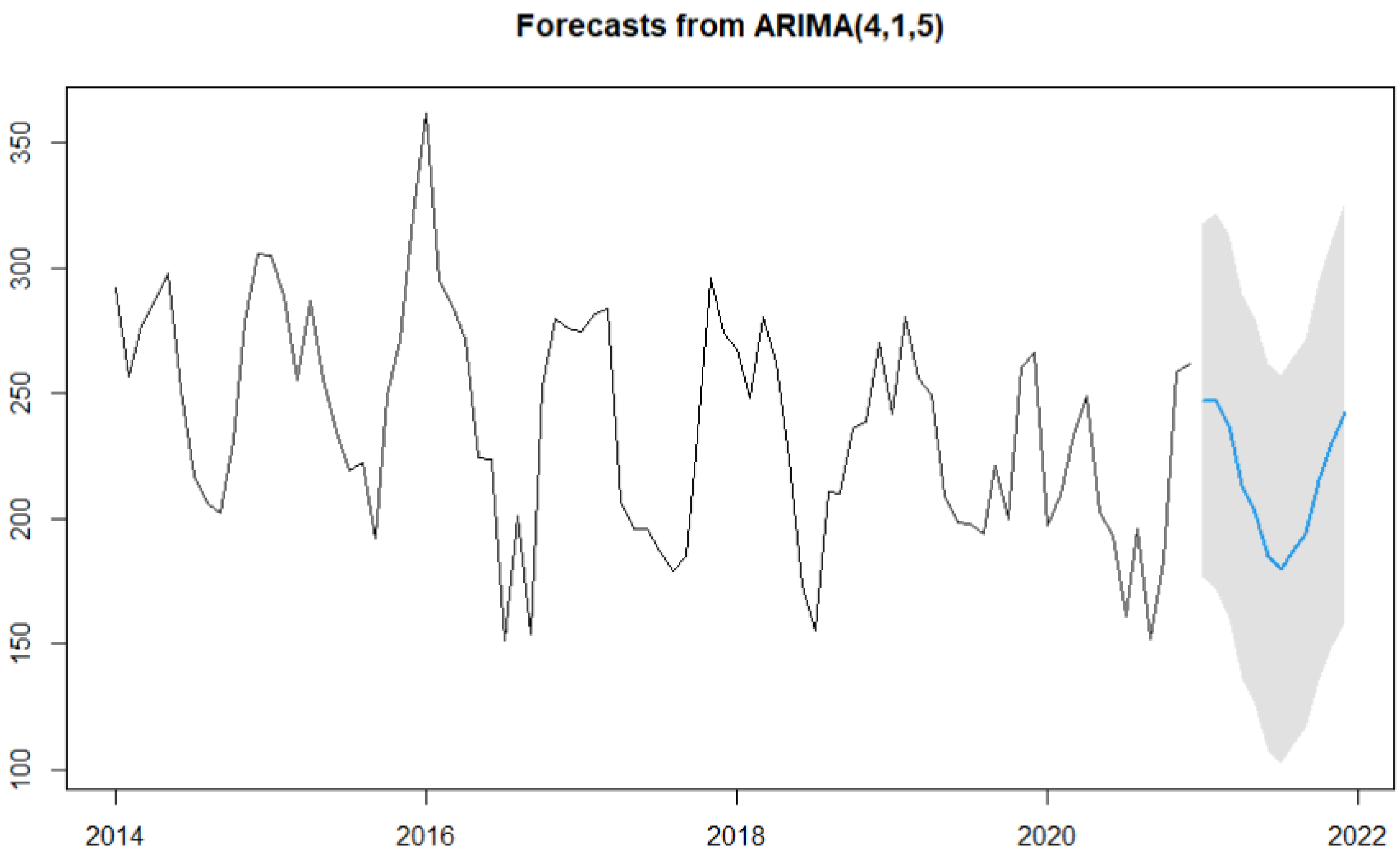
```
> Model1

Call:
stats::arima(x = mots, order = c(4, 1, 5), method = "ML")

Coefficients:
         ar1     ar2     ar3     ar4      ma1      ma2
      0.3051  0.6115  0.0328  -0.820  -0.9419  -0.5777  0.
s.e.  0.0916  0.1034  0.0900   0.074   0.1284   0.1150  0.

sigma^2 estimated as 625.9:  log likelihood = -388.91,  ai
```



Model Testing

```
> Box.test(my.residual, lag = 12, type = c("Ljung-Box"))

        Box-Ljung test

data:  my.residual
X-squared = 9.9142, df = 12, p-value = 0.6235

> dwtest(my.residual ~ fitted(Model1))

        Durbin-Watson test

data:  my.residual ~ fitted(Model1)
DW = 1.8728, p-value = 0.2487
alternative hypothesis: true autocorrelation is greater than 0
```

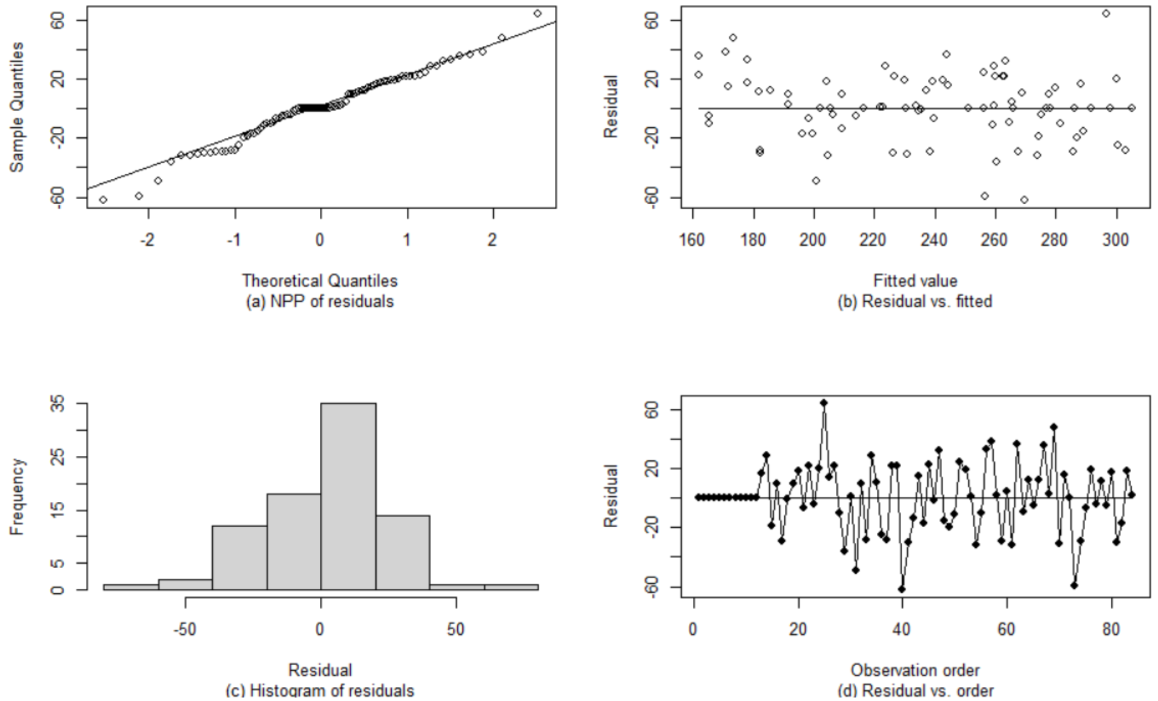
Forecasts


Forecasts from ARIMA(4,1,5)
350
300
250
200
150
100
2014
2016
2018
2020
2022

Model construction

- Model 2 : auto.ARIMA

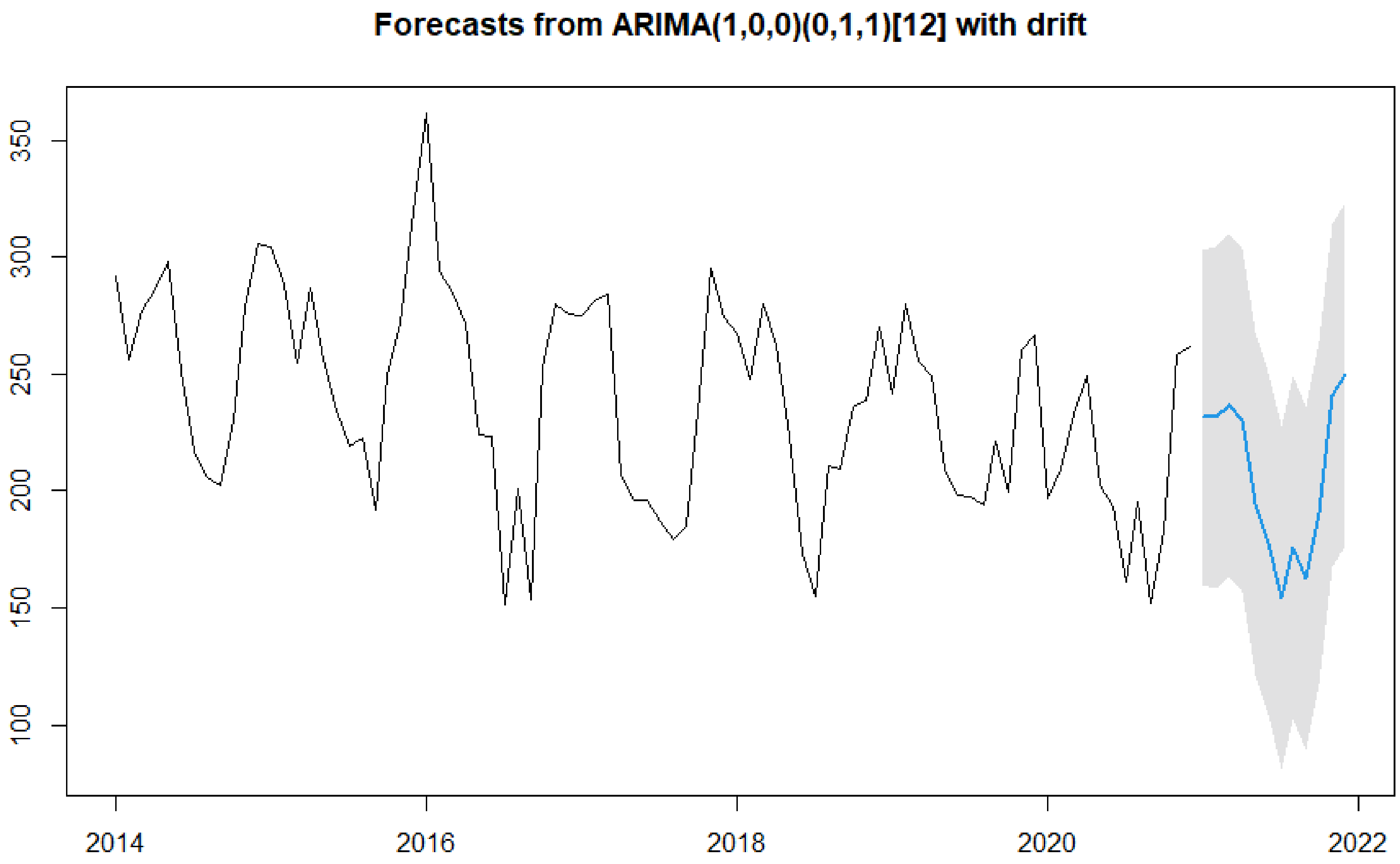
```
> Model2
Series: mots
ARIMA(1,0,0)(0,1,1)[12] with drift

Coefficients:
         ar1     sma1    drift
      0.2260  -0.7008  -0.6573
s.e.  0.1149   0.1969   0.1552

sigma^2 estimated as 644.9:  log likelihood=-337.56
AIC=683.11   AICc=683.71   BIC=692.22
```



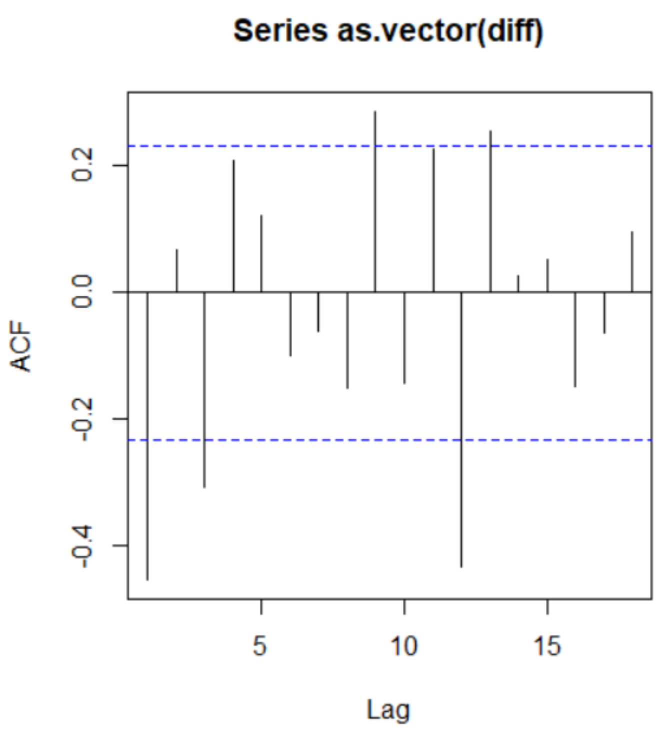
Model Testing

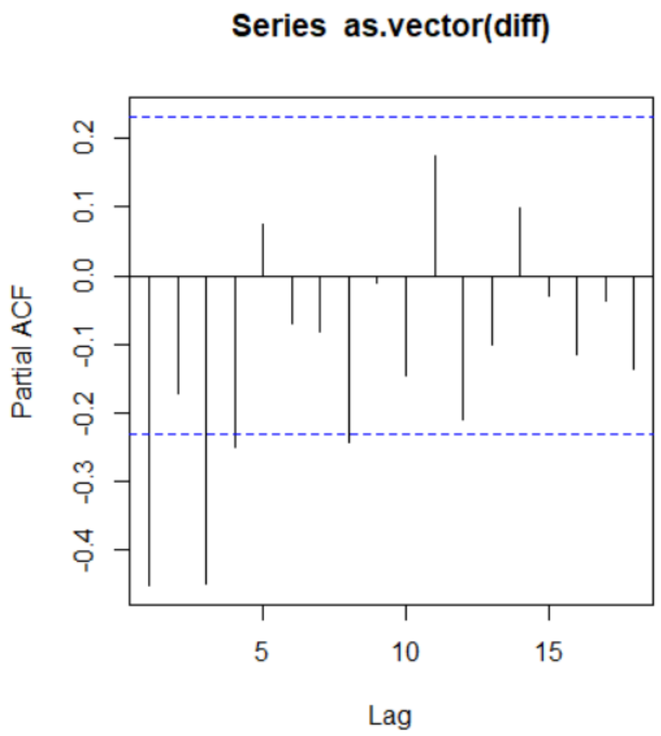
```
> Box.test(my.residual, lag = 12, type = c("Ljung-Box"))

        Box-Ljung test

data:  my.residual
X-squared = 11.054, df = 12, p-value = 0.5243

> dwtest(my.residual ~ fitted(Model2))

        Durbin-Watson test

data:  my.residual ~ fitted(Model2)
DW = 1.9465, p-value = 0.3696
alternative hypothesis: true autocorrelation is greater than 0
```

Forecast


Forecasts from ARIMA(1,0,0)(0,1,1)[12] with drift
100
150
200
250
300
350
2014
2016
2018
2020
2022

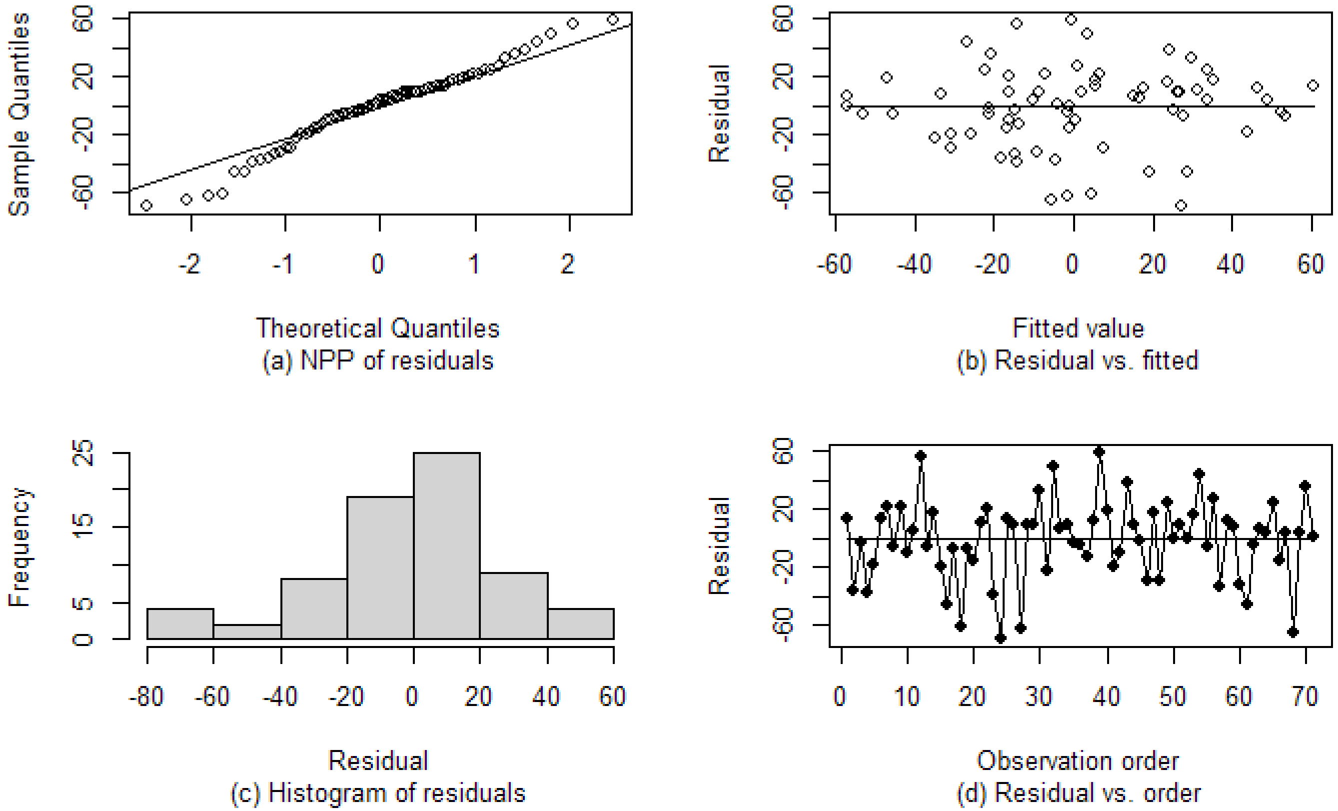
Model construction






```
> Model3

Call:
arima(x = diff, order = c(4, 0, 3), method = "ML")

Coefficients:
         ar1      ar2      ar3     ar4      ma1     ma2      ma3  intercept
      0.4381  -0.6093  -0.0851  0.2843  -1.2957  1.0190  -0.7234    -0.1667
s.e.  0.1831   0.2430   0.1403  0.1359   0.1672  0.2937   0.2552     0.2192

sigma^2 estimated as 742.3:  log likelihood = -337.54,  aic = 691.09
```

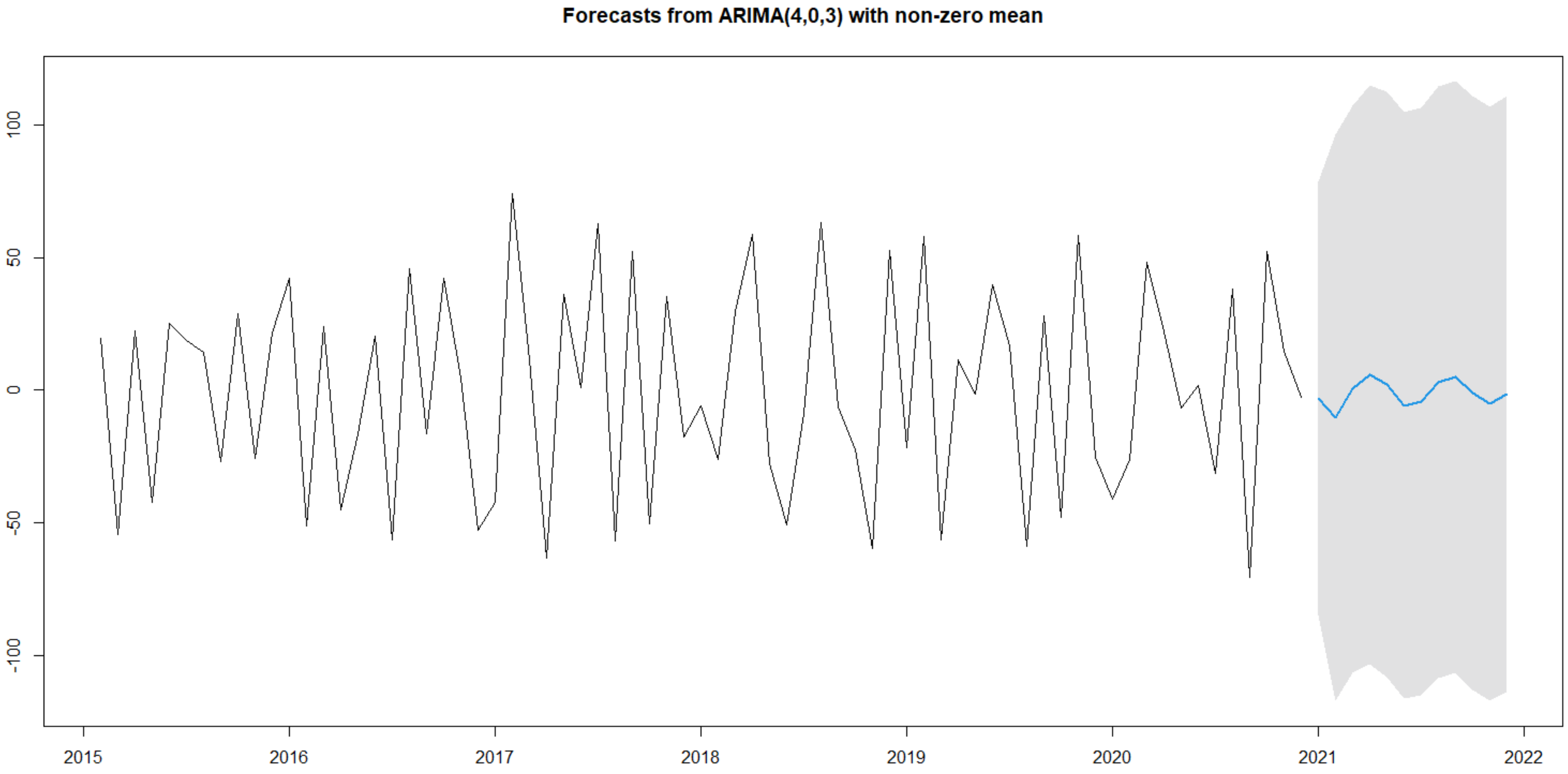
Model construction



Model Testing

```
> Box.test(my.residual, lag = 12, type = c("Ljung-Box"))

        Box-Ljung test

data:  my.residual
X-squared = 9.0362, df = 12,
p-value = 0.6998

> dwtest(my.residual ~ fitted(Model3))

        Durbin-Watson test

data:  my.residual ~ fitted(Model3)
DW = 1.9201, p-value = 0.3726
alternative hypothesis: true autocorrelation is greater than 0
```

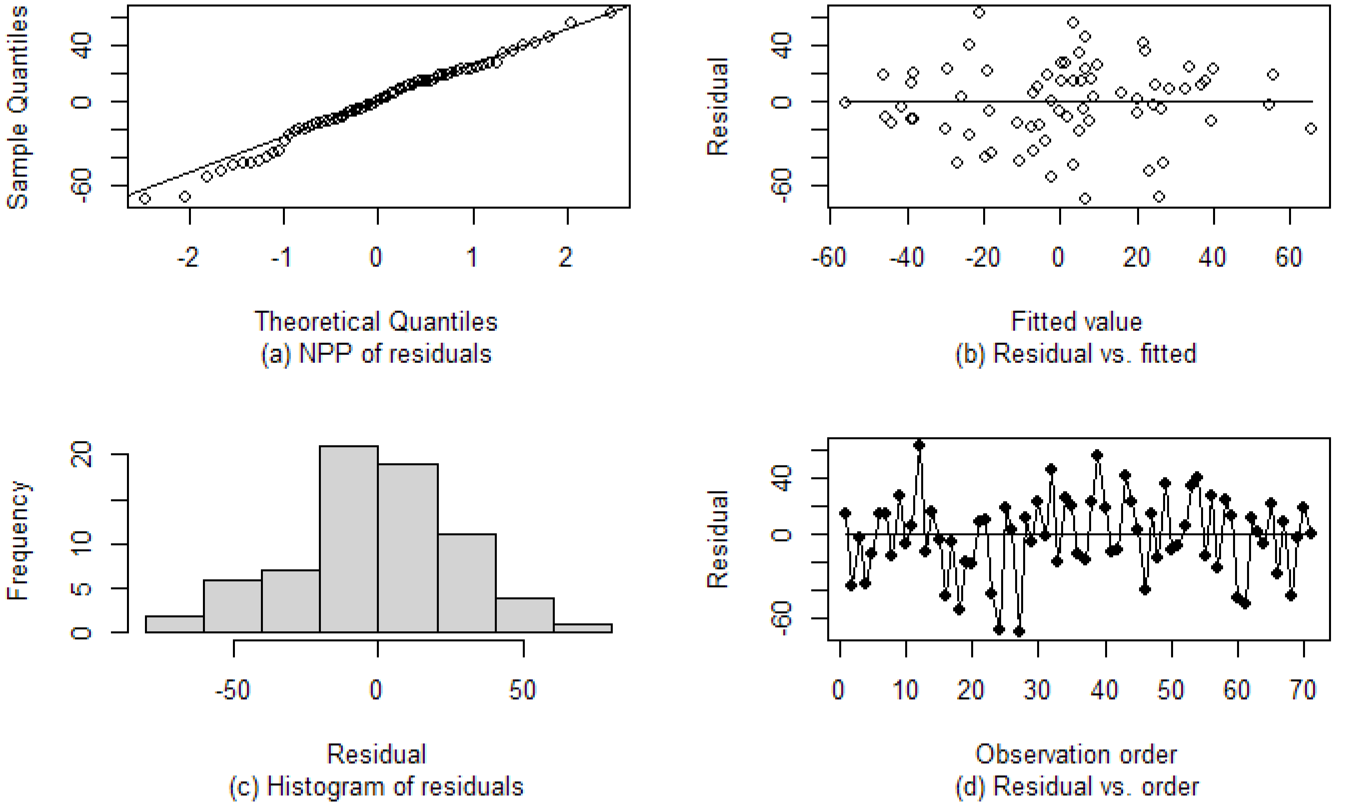
Forecast



Forecasts from ARIMA(4,0,3) with non-zero mean
100
50
0
-50
-100
2015
2016
2017
2018
2019
2020
2021
2022

Model construction

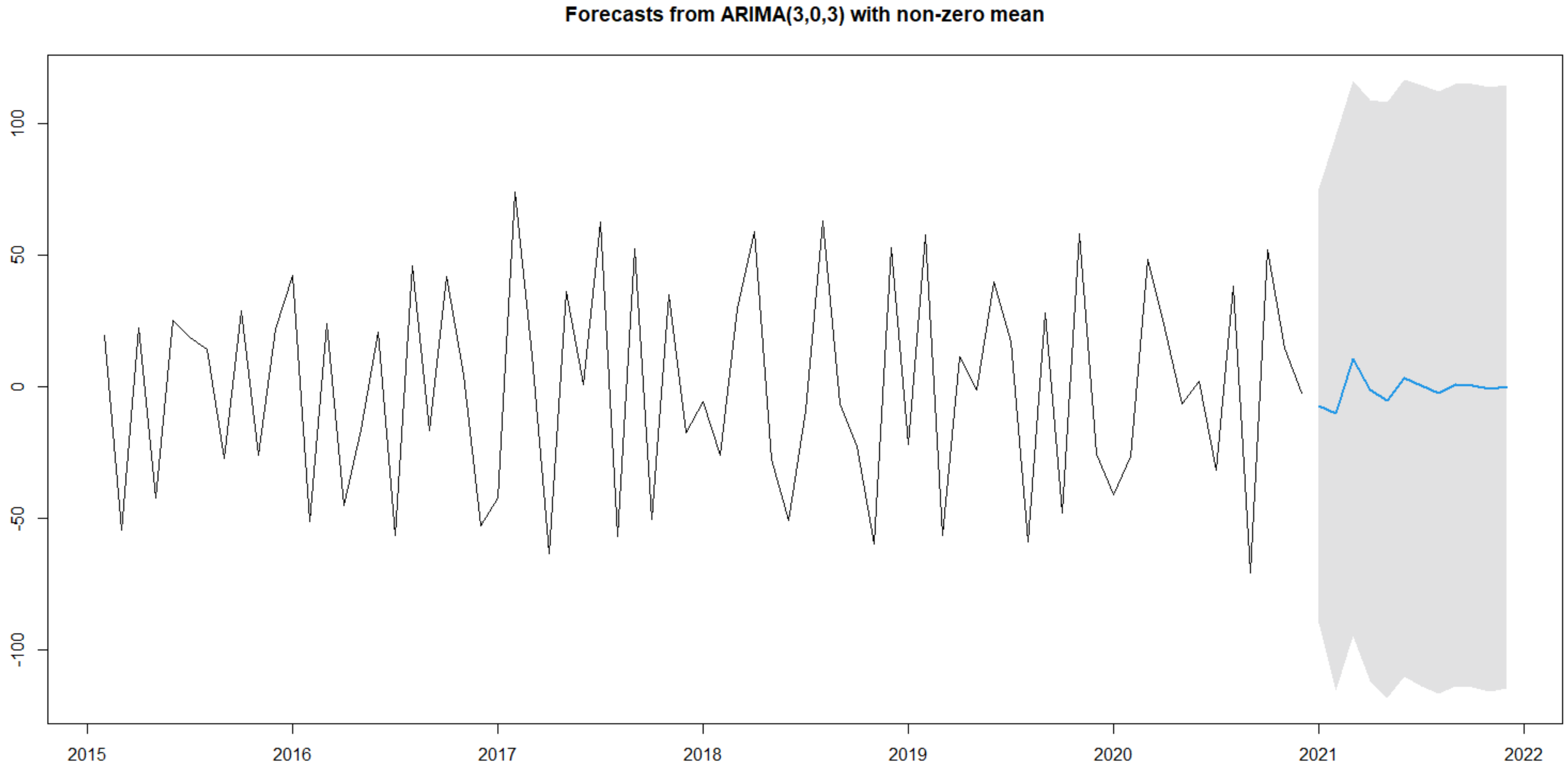
```
> Model4

Call:
arima(x = diff, order = c(3, 0, 3), method = "ML")

Coefficients:
          ar1      ar2      ar3      ma1      ma2
      -0.6531  -0.5644  -0.0460  -0.1527  -0.0136
s.e.   0.1912   0.1703   0.1468   0.1630   0.1299

sigma^2 estimated as 770.6:  log likelihood = -339.(
```



Model Testing

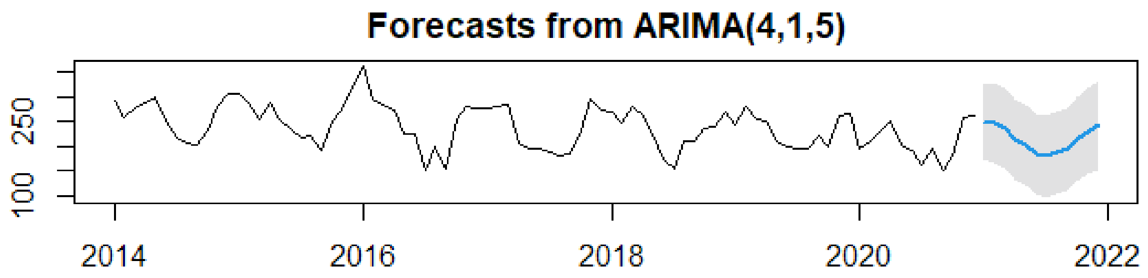
```
> Box.test(my.residual, lag = 12, type = c("Ljung-Box"))

	Box-Ljung test

data:  my.residual
X-squared = 13.711, df = 12, p-value = 0.3195

> dwtest(my.residual ~ fitted(Model4))

	Durbin-Watson test

data:  my.residual ~ fitted(Model4)
DW = 1.9415, p-value = 0.408
alternative hypothesis: true autocorrelation is greater than 0
```

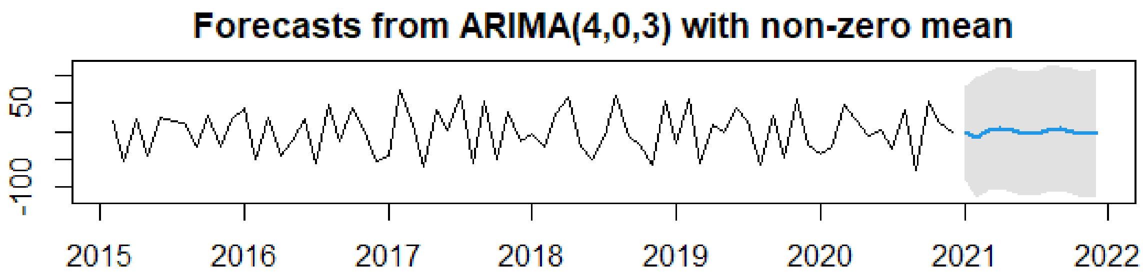
Forecast

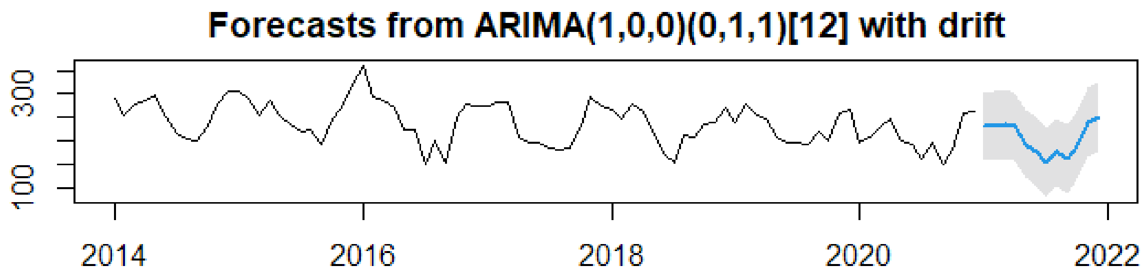


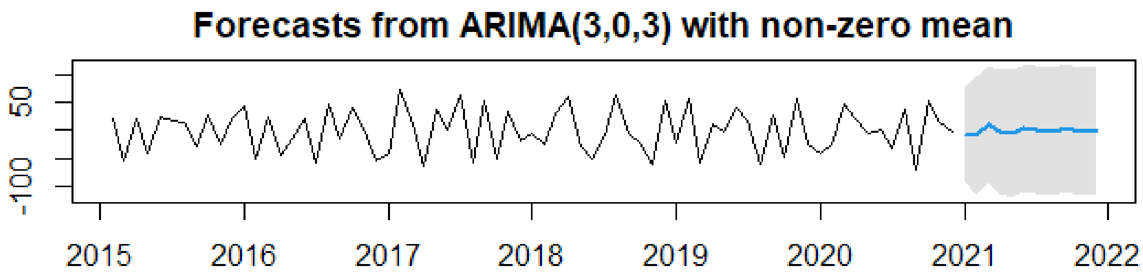
Forecasts from ARIMA(3,0,3) with non-zero mean
100
50
0
-50
-100
2015
2016
2017
2018
2019
2020
2021
2022

Forecast Comparision





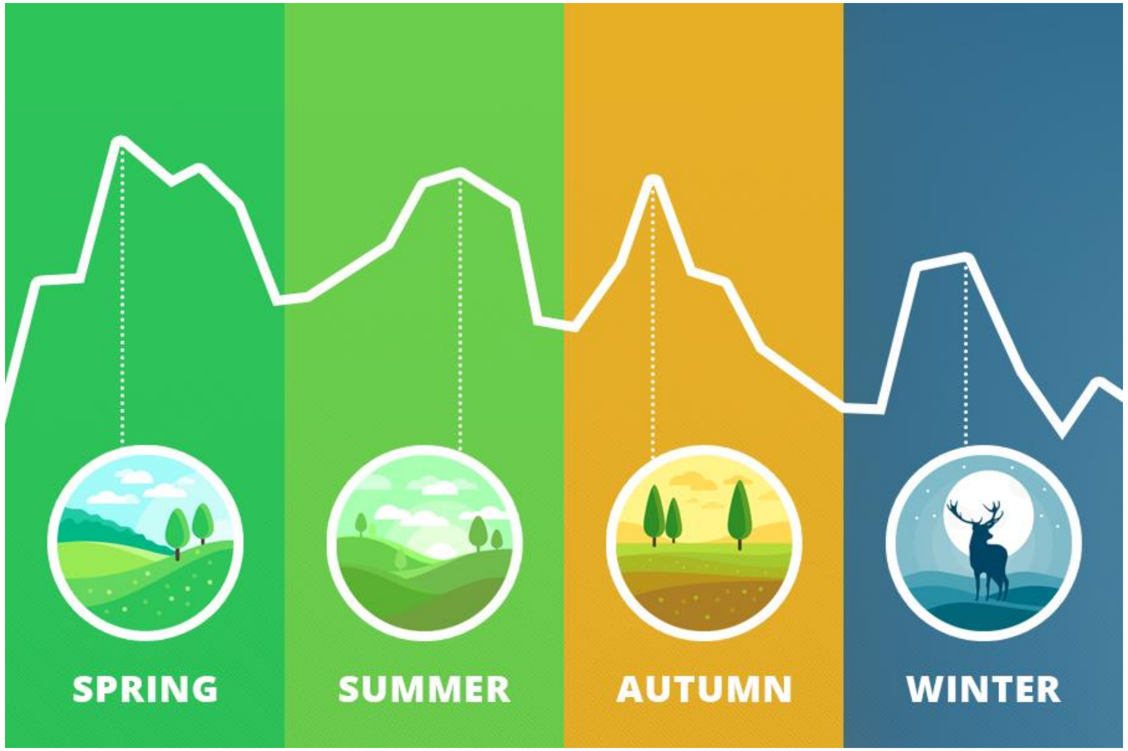


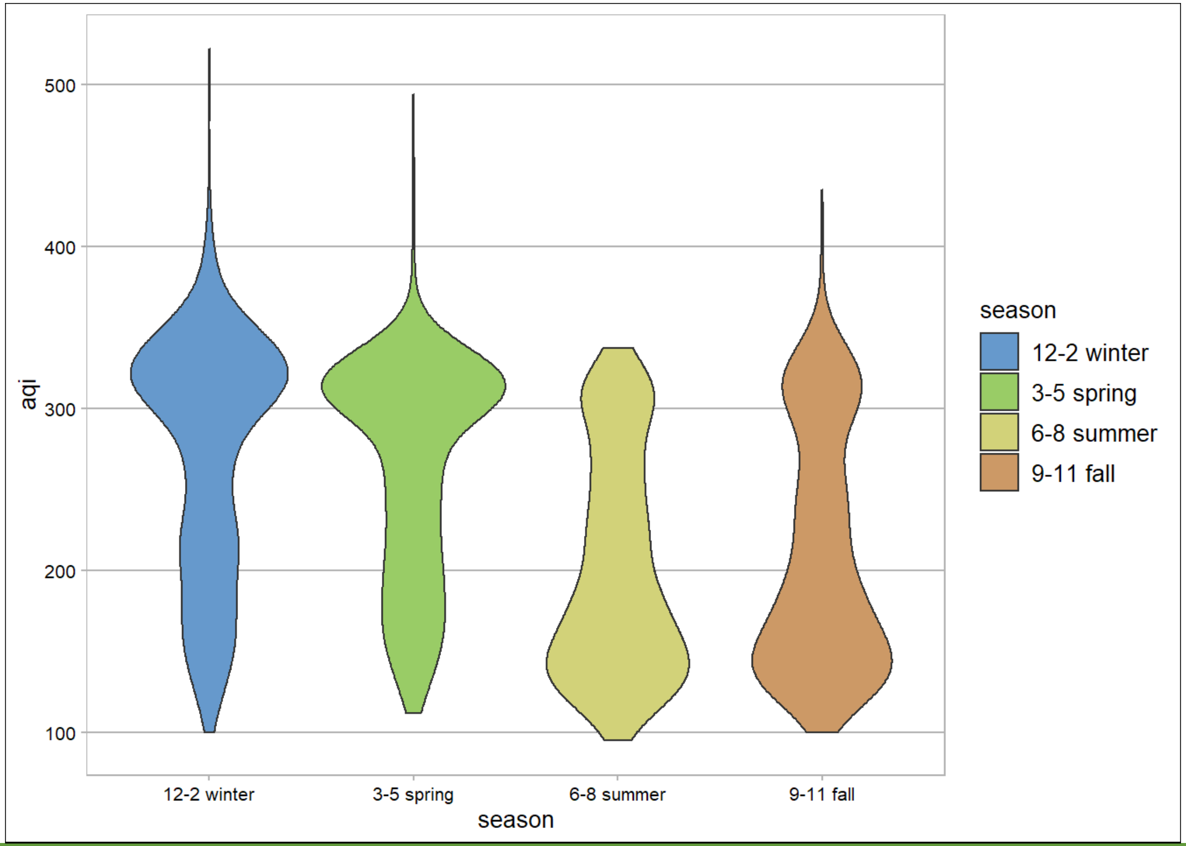


Summarize

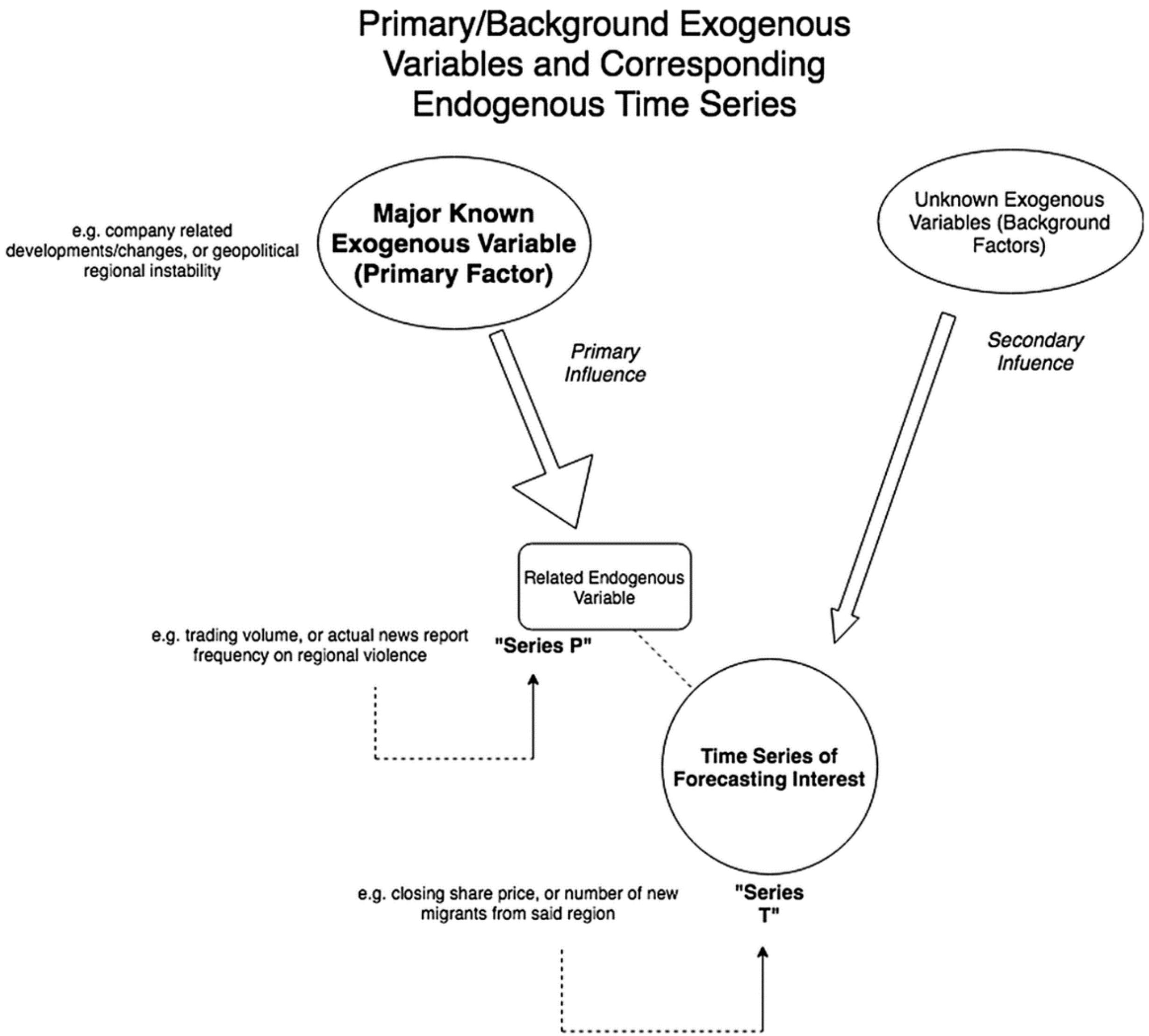
For exponential smoothing

$$\hat{y}_{t+1|t} = \alpha y_t + (1-\alpha)\hat{y}_{t|t-1}$$



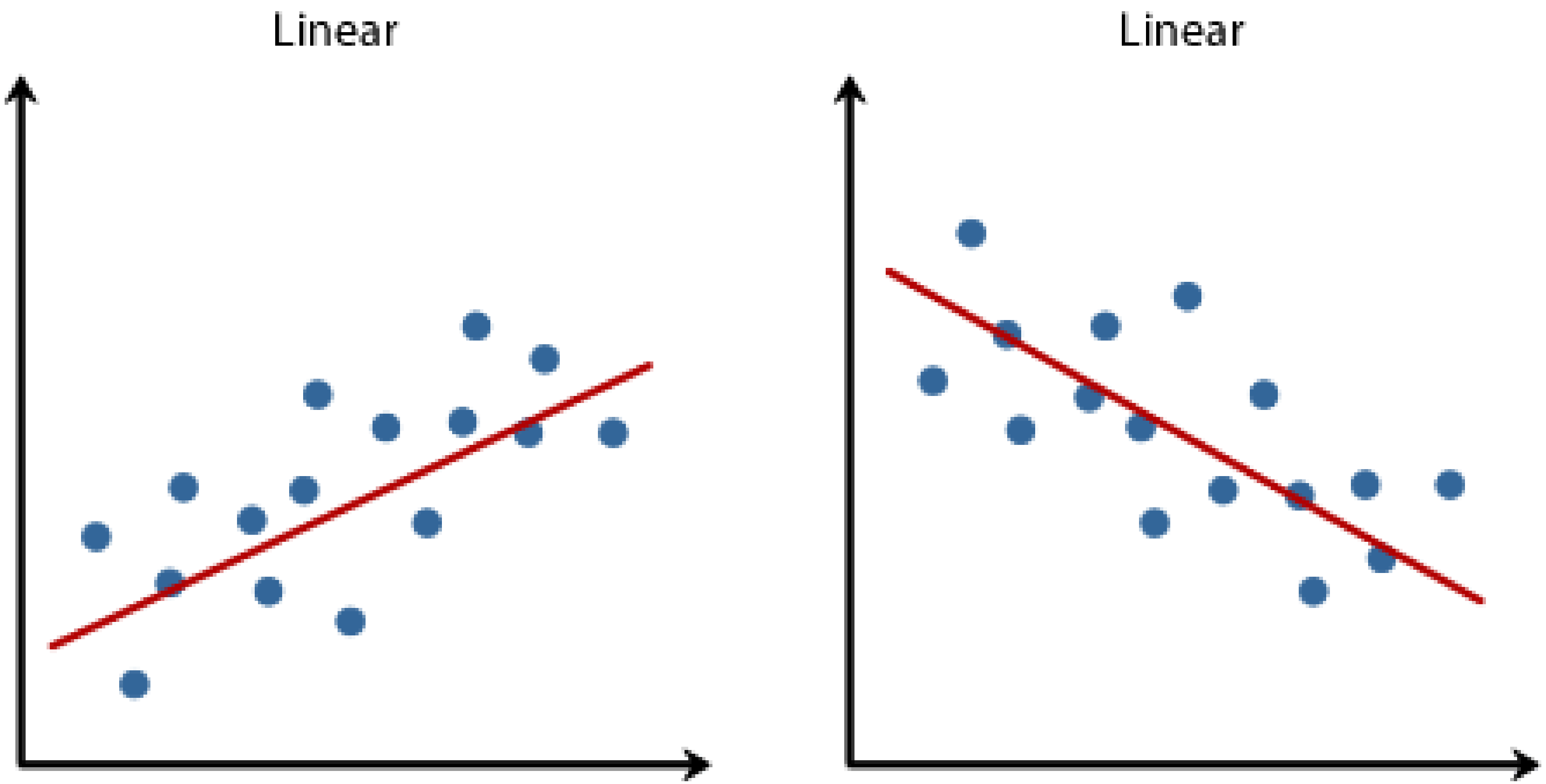


For ARIMA

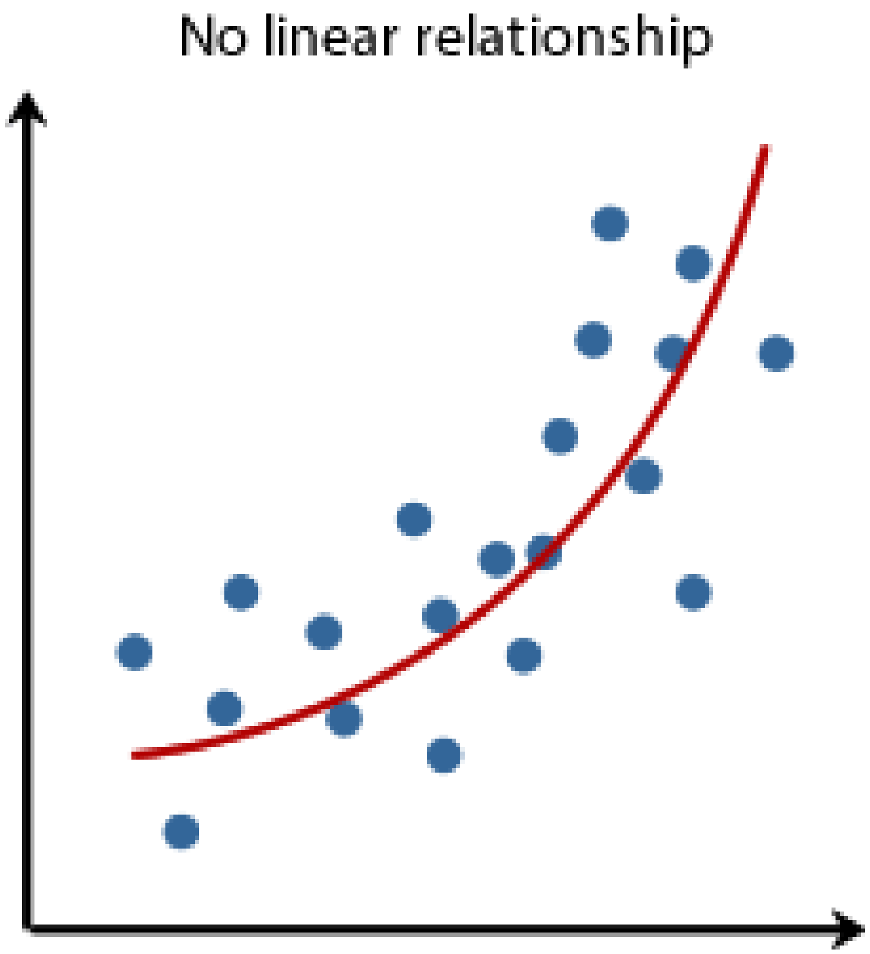


Primary/Background Exogenous Variables and Corresponding Endogenous Time Series
e.g. company related developments/changes, or geopolitical regional instability
Major Known Exogenous Variable (Primary Factor)
Unknown Exogenous Variables (Background Factors)
Primary Influence
Secondary Infuence
Related Endogenous Variable
e.g. trading volume, or actual news report frequency on regional violence
"Series P"
Time Series of Forecasting Interest
e.g. closing share price, or number of new migrants from said region
"Series T"

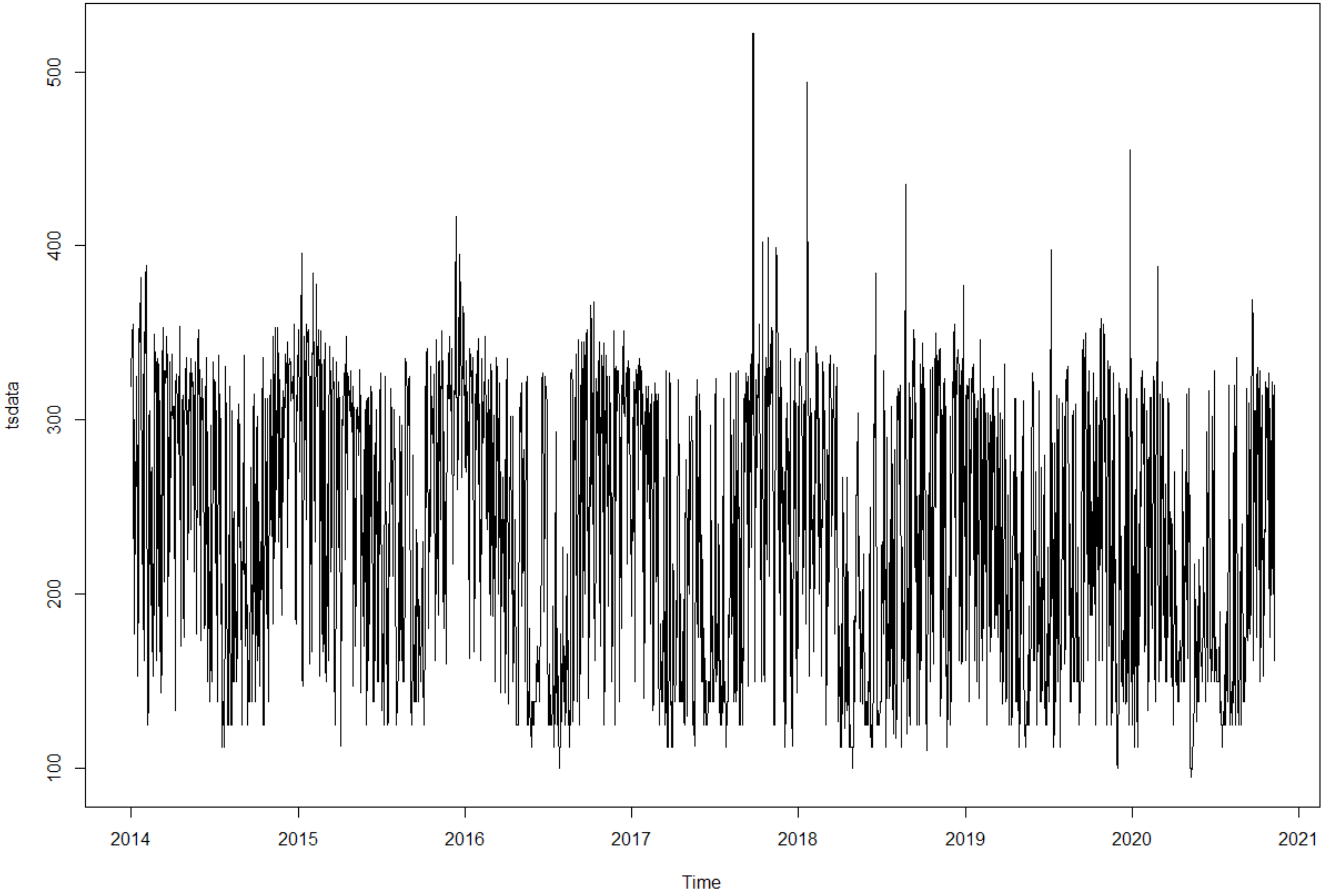
For ARIMA



For ARIMA



For ARIMA

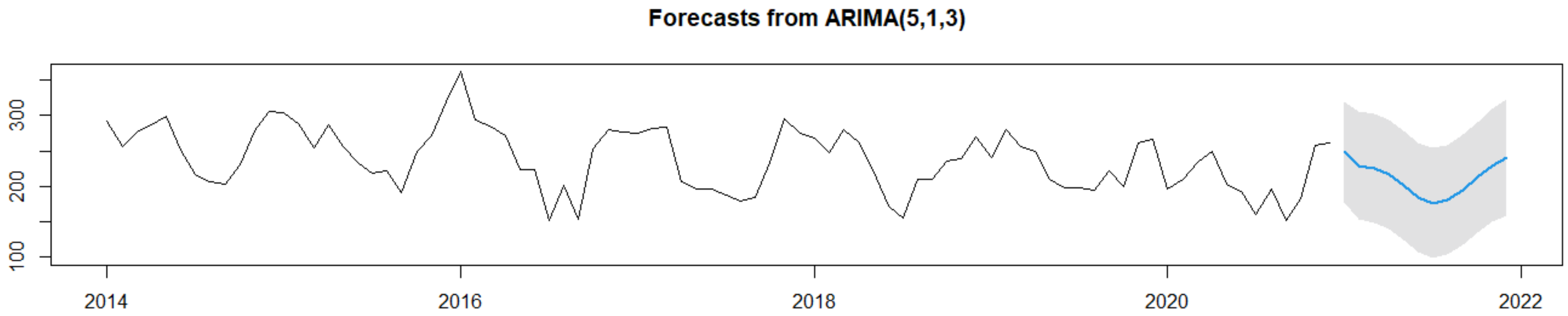


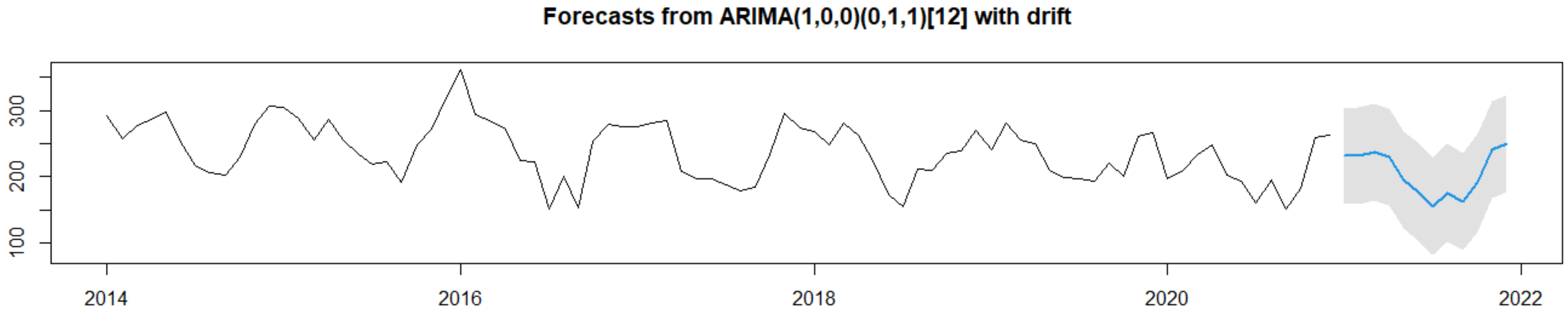
500
400
300
200
100
tsdata
2014
2015
2016
2017
2018
2019
2020
2021
Time

Model Improvement

$$SARIMA\underbrace{(p,d,q)}_{non-seasonal}\underbrace{(P,D,Q)_m}_{seasonal}$$

Model Improvement





Final interpretation

- The AQI and Shanghai's air quality have a decreasing trend with seasonal fluctuating.
- In future, the AQI will around 150-200. Whether this index still imply unhealthy, we believe the whole environment tends to get better gradually.