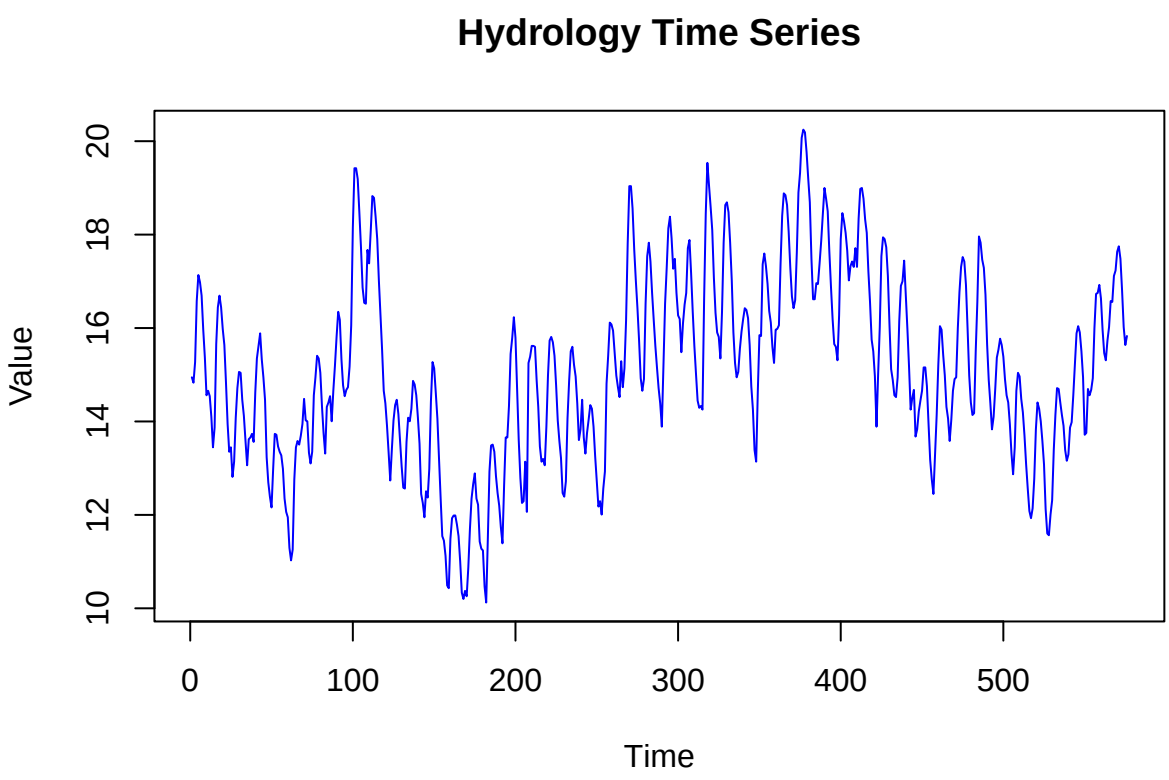
Stat 443 Project

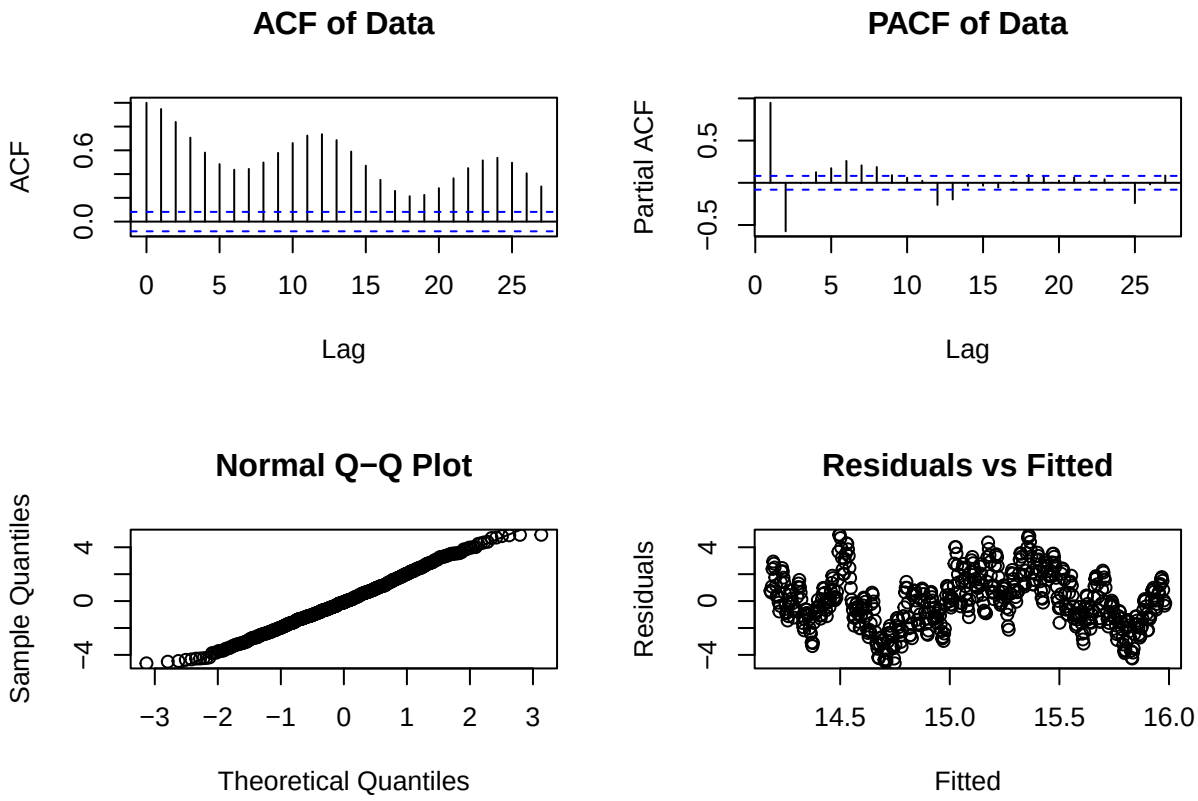
Scenario 1

Number of pages for text + R Output (excluding plots + code): 1.75 pages

We first start by plotting our hydrology data.



We will first diagnose stationarity by plotting the ACF/PACF plots of the original data with residual diagnostics:



The ACF pattern is an oscillation with a downward trend, which implies that the time series has some kind of seasonal pattern. This is clearly not stationary. The PACF cuts off at $p = 1$ The residuals vs fitted plot appears to also indicate this sinusodial pattern. So we will try using a SARIMA model with parameter $p = 1$.

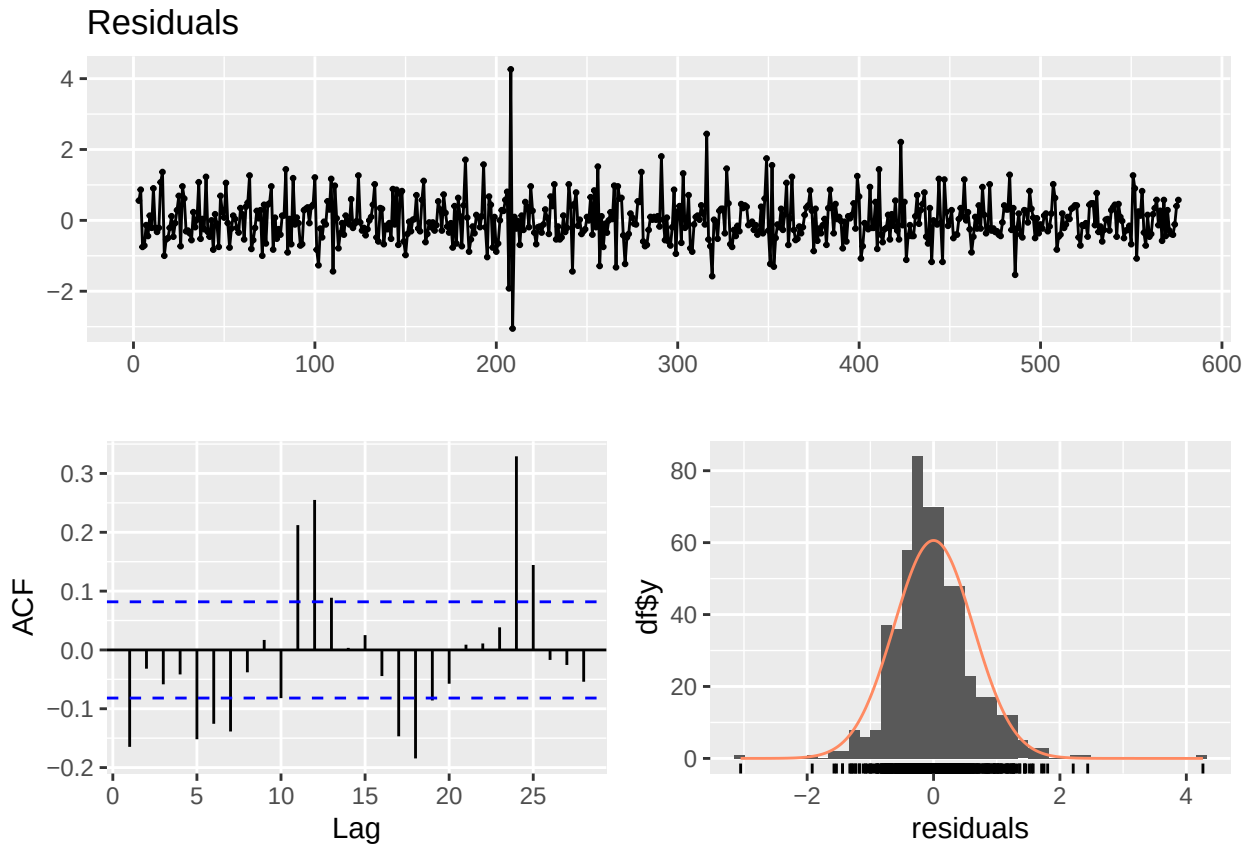
To confirm stationary, we also use the Augmented Dickey-Fuller (ADF) test and KPSS test to check for stationarity.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  hydro_data
## Dickey-Fuller = -2.3118, Lag order = 8, p-value = 0.4463
## alternative hypothesis: stationary

##
## ######################
## # KPSS Unit Root Test #
## ######################
##
## Test is of type: mu with 6 lags.
##
## Value of test-statistic is: 1.3229
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

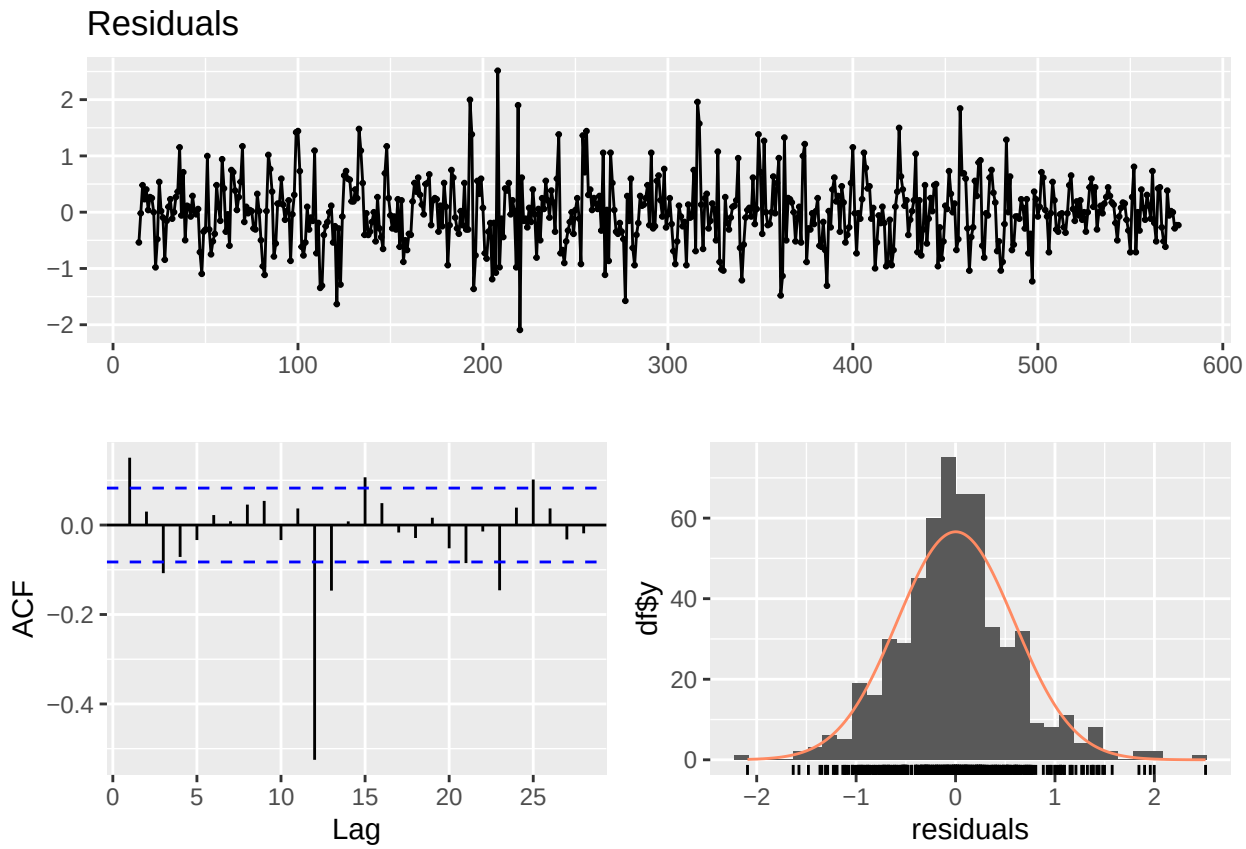
By the KPSS test for stationary, for the original data, our KPSS value of 1.3229 exceeds all cutoffs that reject stationary. This is also supported by the ADF test having p-value > 0.05.

Because we want to achieve stationarity, we will to determine the difference parameters (for seasonal and non-seasonal components). To do so, we can continuously difference them (separately) until we achieve an ADF p-value of < 0.05.

The final output with residuals diagnostics is displayed below:




```
##
##  Ljung-Box test
##
## data:  Residuals
## Q* = 57.974, df = 10, p-value = 8.742e-09
##
## Model df: 0.   Total lags used: 10
```

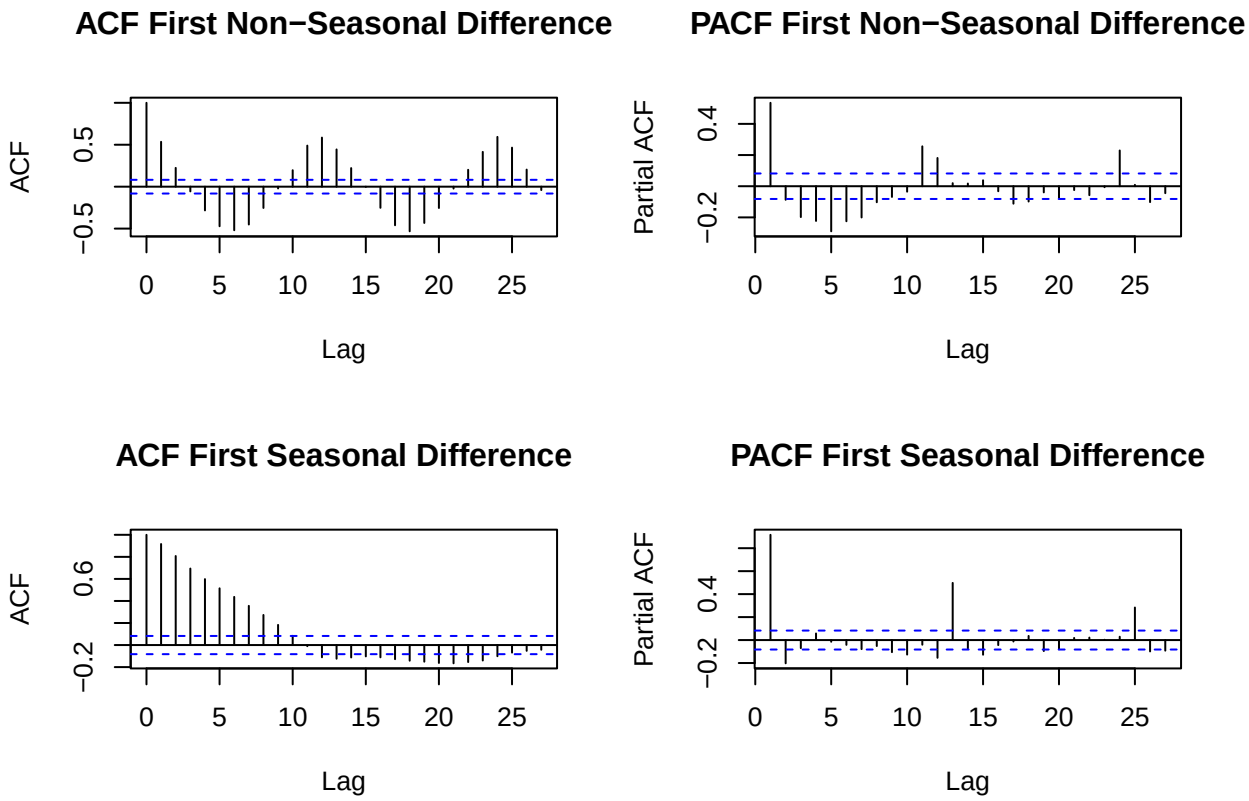



```
##
##  Ljung-Box test
##
## data:  Residuals
## Q* = 27.281, df = 10, p-value = 0.00235
##
## Model df: 0.   Total lags used: 10

## Non-seasonal differencing (d): 1

## Seasonal differencing (D): 1
```

It looks like the best values for the differences is $d = 1$ for non-seasonal difference and $D = 1$ for seasonal difference

Confirming this with our residual diagnostics, both ACF's appear to have lost the sinusodial property and now somewhat appears stationary: - The residuals appear random in both plots - The ACF's only have spikes at the seasonal lags at 12 - The residual bell-curves are both symmetrically normal.

Now we want to find the best AR and MA parameters. To do this, we identify the ACF and PACF plots of both non-seasonal and seasonal models.

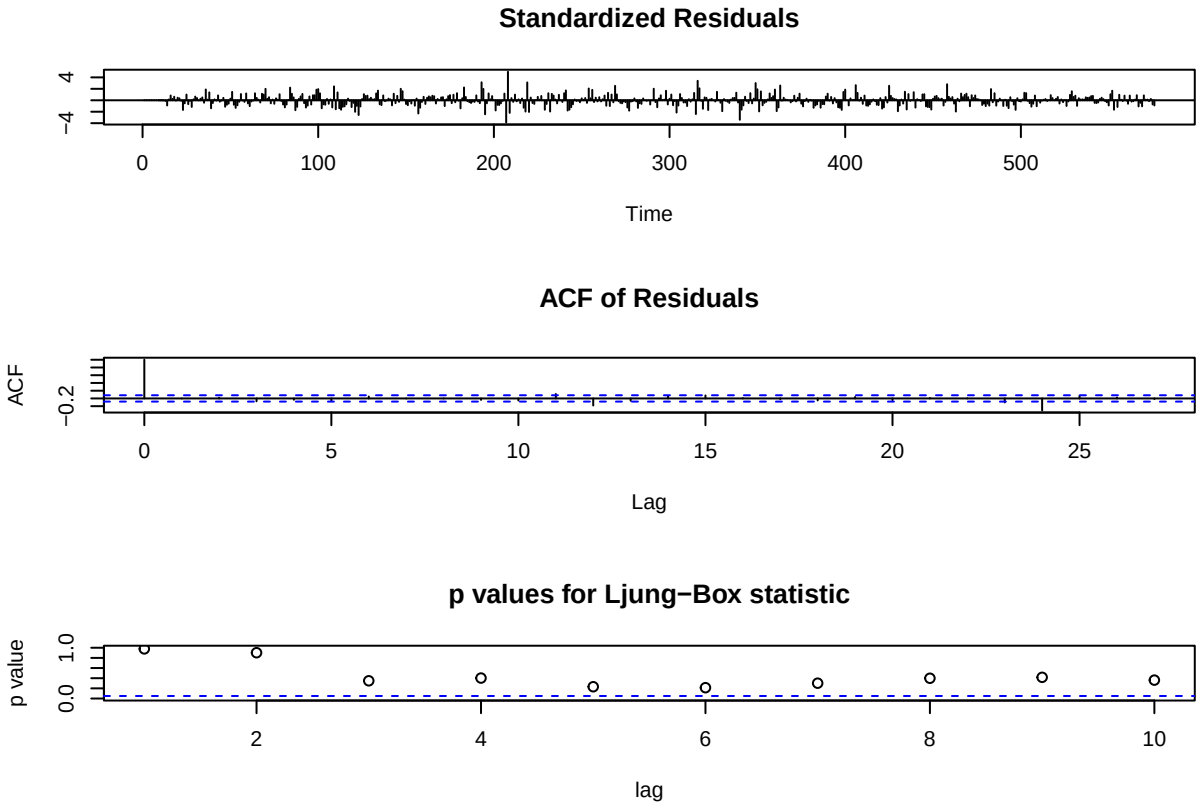


For the non-seasonal difference model, the ACF still appears to draw a sinusodial pattern (so perhaps no q value), whereas the PACF cutsoff looks a bit unclear, so it is worth exploring different p values.

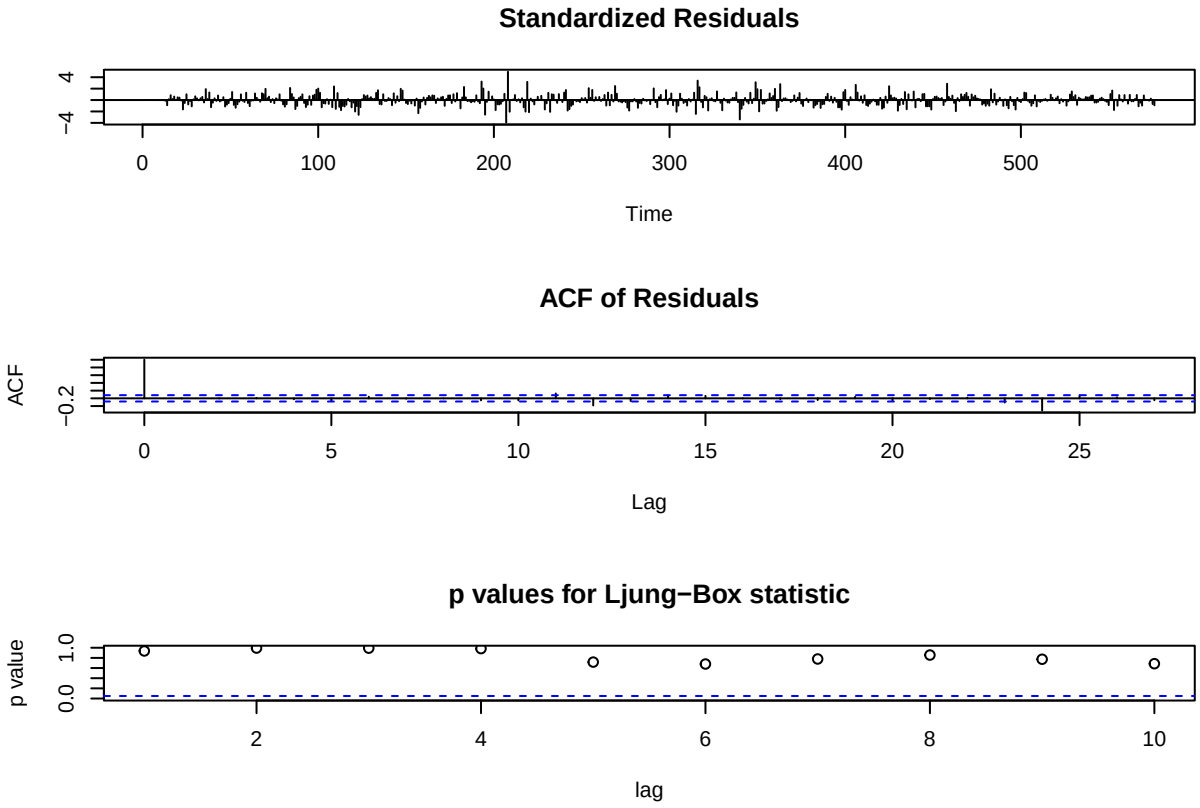
For the seasonal difference model, the ACF is slowly decreasing (it is worth exploring $q = 0$ and $q = 1$) and the PACF has a cutoff at $p = 1$, along with spikes at the seasonalities every 12 lags.

So we want to test out a few SARIMA models with what we have learned:

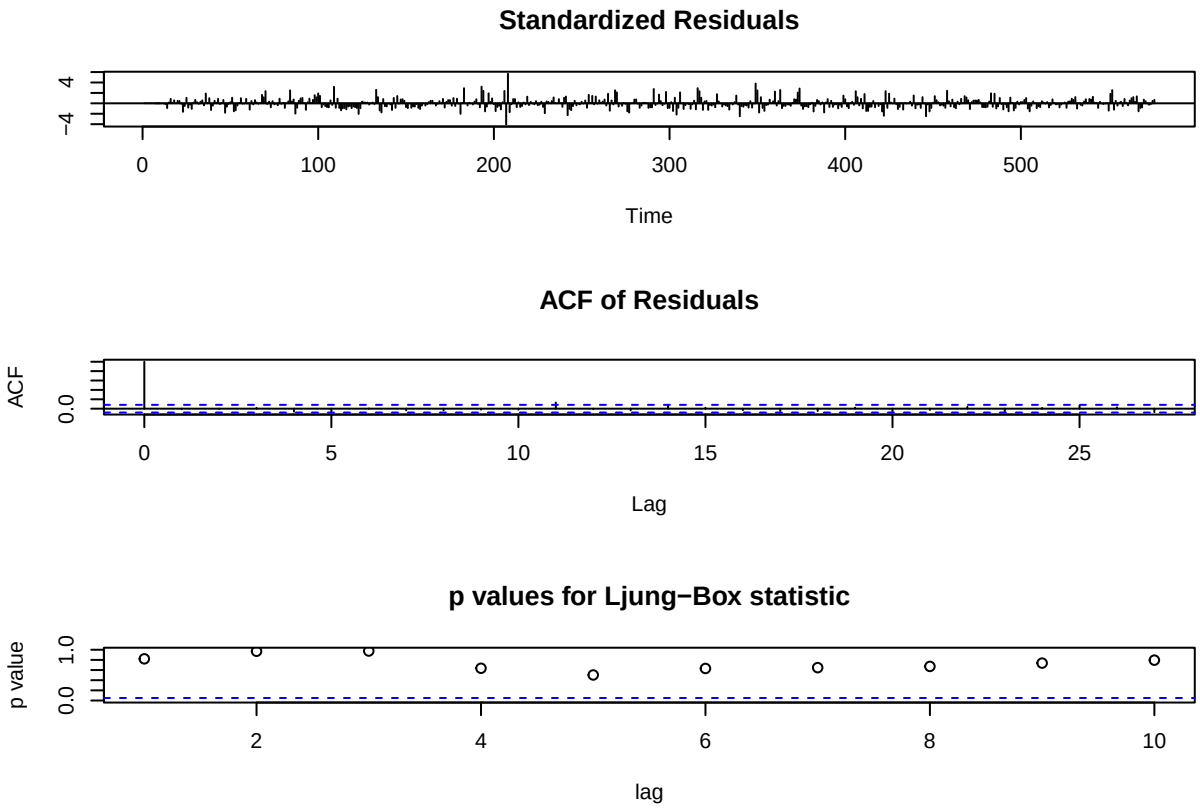
First we try out the SARIMA(1,1,0,1,1,0) model, by using the ARIMA function in R:



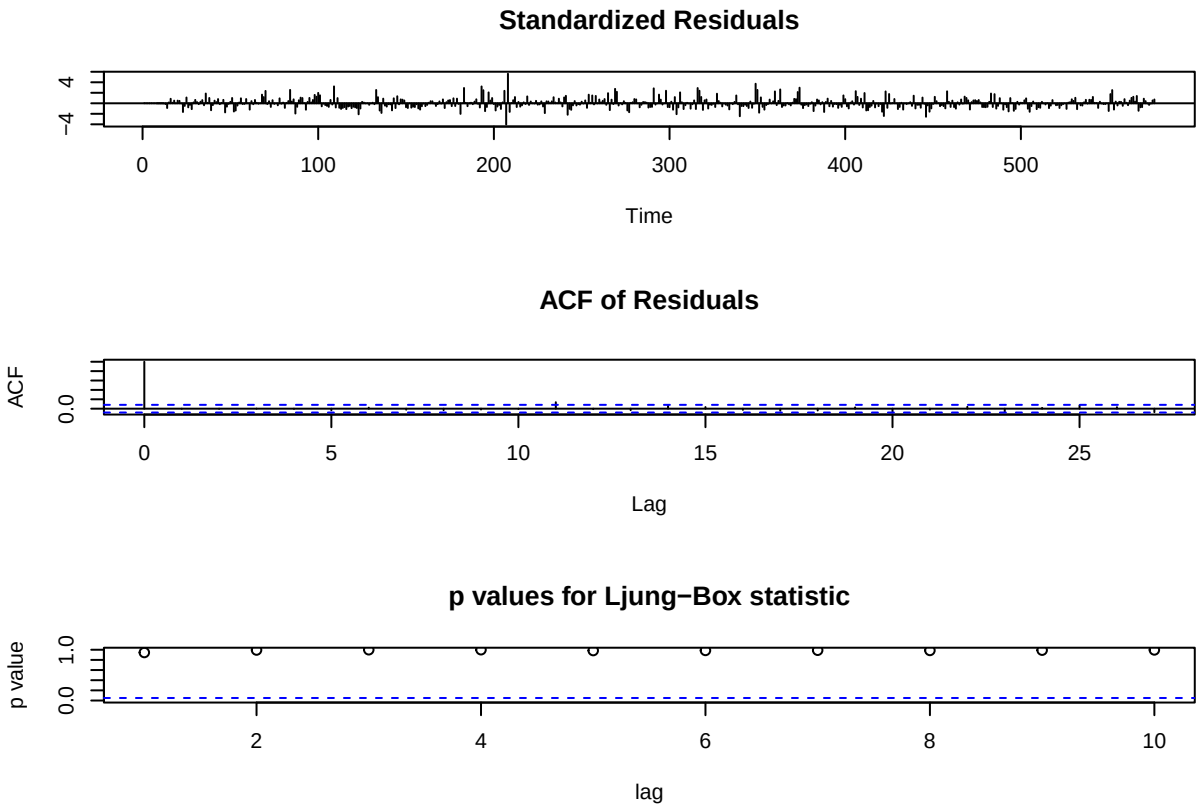
Then try out the SARIMA(3,1,0,1,1,0) model:



Then we try out the SARIMA(3,1,0,1,1,1) model:



Then we try out the SARIMA(5,1,0,1,1,1) model:



All the standardized residuals appear to be random, all the ACF values appear to be stationary (no real spikes outside), and the p-values are above the line for the Ljung-Box statistic, then all of these models are decent.

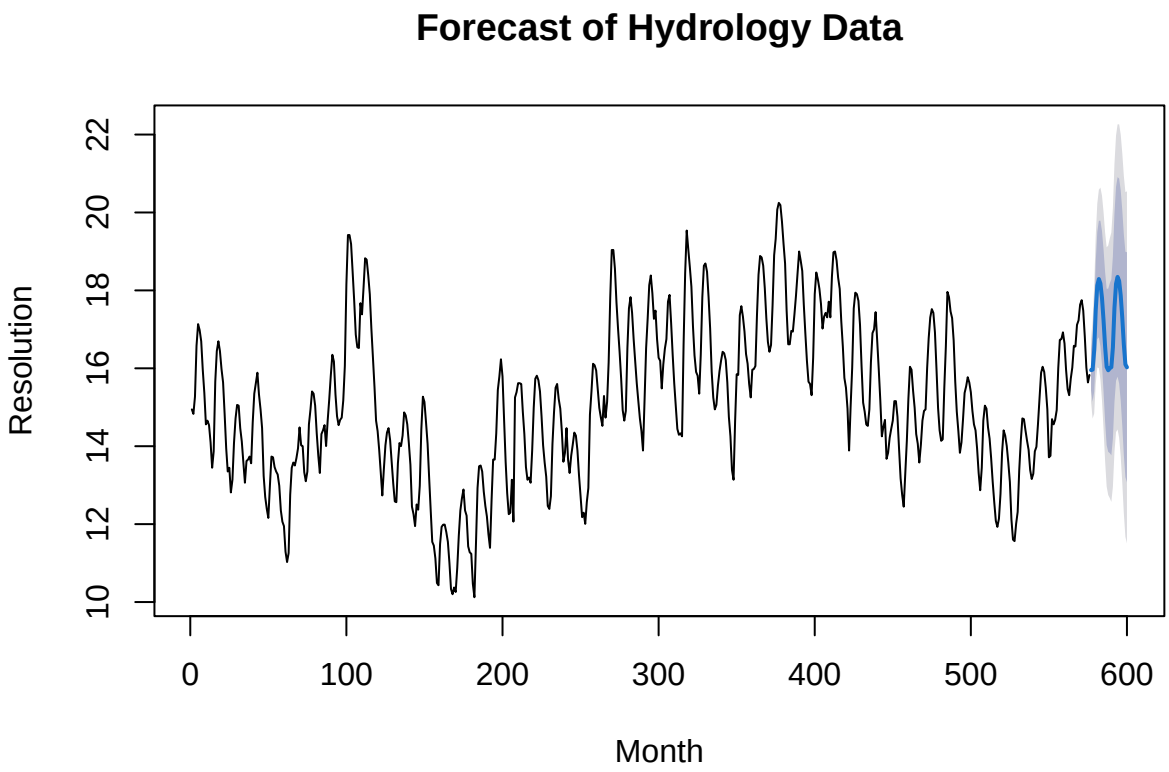
To find the best model, we pick the model with the lowest AIC/BIC value:

Table 1: AIC and BIC of Different Models

Model	AIC	BIC
SARIMA(1,1,0,1,1,0)	806.1721	819.1720
SARIMA(3,1,0,1,1,0)	806.1702	827.8366
SARIMA(3,1,0,1,1,1)	629.7440	655.7437
SARIMA(5,1,0,1,1,1)	629.9890	668.9886

With the lowest AIC/BIC values, the best model is the SARIMA(3,1,0,1,1,1) model!

Forecasting our model for the next 24 months:



For the next 24-month ahead forecasts, this looks like a reasonable forecast for our hydrology data (the blue line). The 95% prediction intervals is indicated by the light-shaded grey area around the forecast, whereas the 80% prediction intervals are the dark grey shaded area.

The 95% prediction intervals indicate that if we forecast even further, the intervals get wider. They do not look too wide for a 24-month forecast so this forecast overall looks good.

Scenario 2

Number of pages for text + R Output (excluding plots + code): 2

We first create a function to read all 40 stock text files using a loop and storing it within a data frame.

Then we fit different kinds of GARCH models for each set of stock.

We first want to test 3 models: GARCH(1,1), Exponential GARCH(1,1) (eGARCH) and Gjr-GARCH(1,1).

- The GARCH model describes the most basic time series model where variance changes over time, which is useful for modelling volatility in stocks.
- The eGARCH is designed to handle asymmetries in volatility, and it can capture leverage effects where negative reutrns might have a different impact on volatility compared to positive returns.
- GJR-GARCH incorporates an additional term to the variance formula to account for the asymmetry in volatility (similar to eGARCH with different approach)

We test these models individually for each stock, which means that we run $3 \times 40 = 120$ tests in total.

We find out which model is best by comparing the AIC Values of each of the 3 models for each stock and take the lowest one. We decide to store the index of the best model within a vector called "decision". The first index refers to GARCH, the second index refers to eGARCH, and the third index refers to GJR-GARCH.

```
# Add decision vector to decide which model best fits the stock
decision <- c()

for (i in 1:40){
  AIC <- c()
  single_stock <- stock_data[i]

  garch_model <- ugarchfit(garch, single_stock)
  egarch_model <- ugarchfit(egarch, single_stock)
  gjrgarch_model <- ugarchfit(gjrgarch, single_stock)

  AIC[1] <- infocriteria(garch_model)[1,]
  AIC[2] <- infocriteria(egarch_model)[1,]
  AIC[3] <- infocriteria(gjrgarch_model)[1,]

  # Find the index of the smallest AIC Value:
  decision[i] = which.min(AIC)
}
```

An example of this is presented below with Stock 40, having the lowest AIC value for the eGARCH model, therefore storing 2 as the last element of the decision vector.

Table 2: AIC of different models for Stock 40

Model	AIC
GARCH(1,1)	-6.429487
EGARCH(1,1)	-6.489611
GJR-GARCH(1,1)	-6.445428

We can output the decision indices and the counts of them below, where the index for each decision pertains to the stock number. For example, the 11th decision is the decision for which is the best model for the 11th stock, being the GARCH model (with value 1)

```
##  [1] 3 3 2 2 2 2 3 2 2 2 1 3 2 1 2 2 2 2 2 2 2 2 3 2 1 3 2 3 3 2 2 2 2 2 1 2 2 1
## [39] 2 2
```

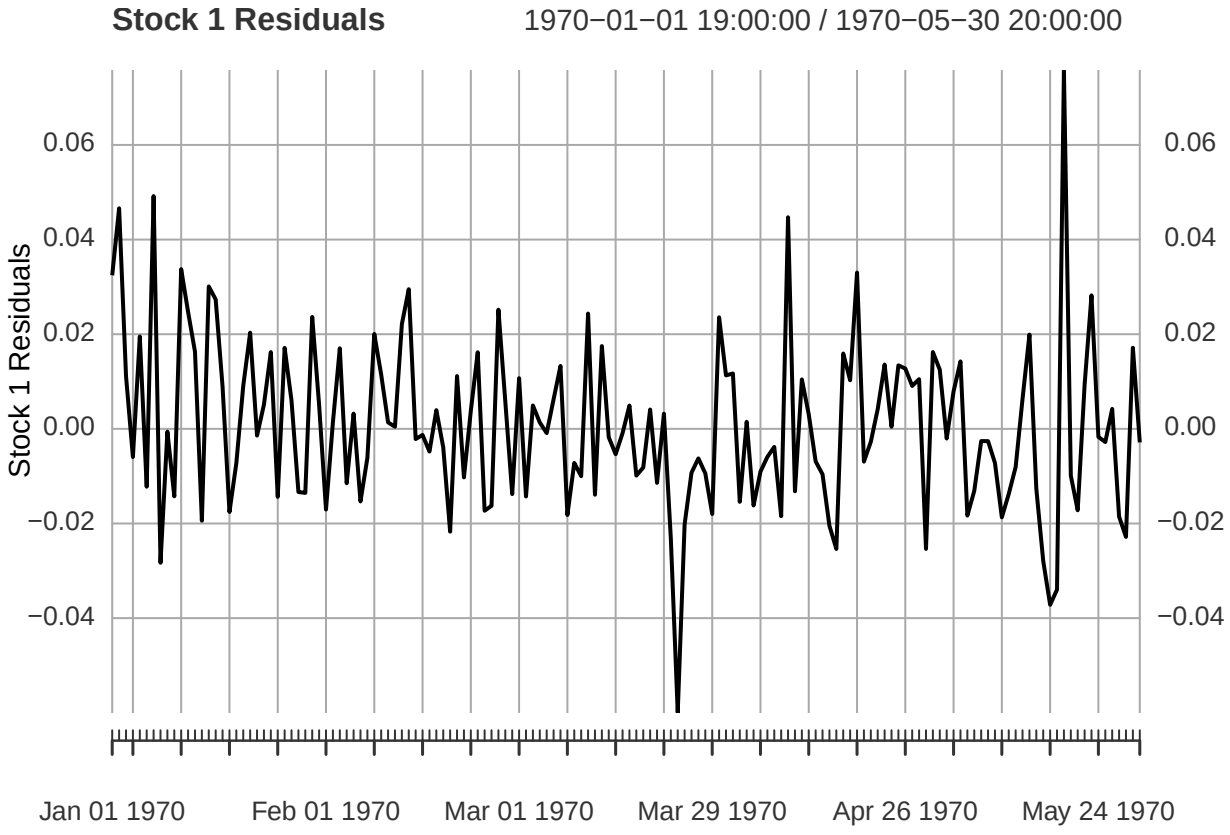
```
## decision
##  1  2  3
##  5 27  8
```

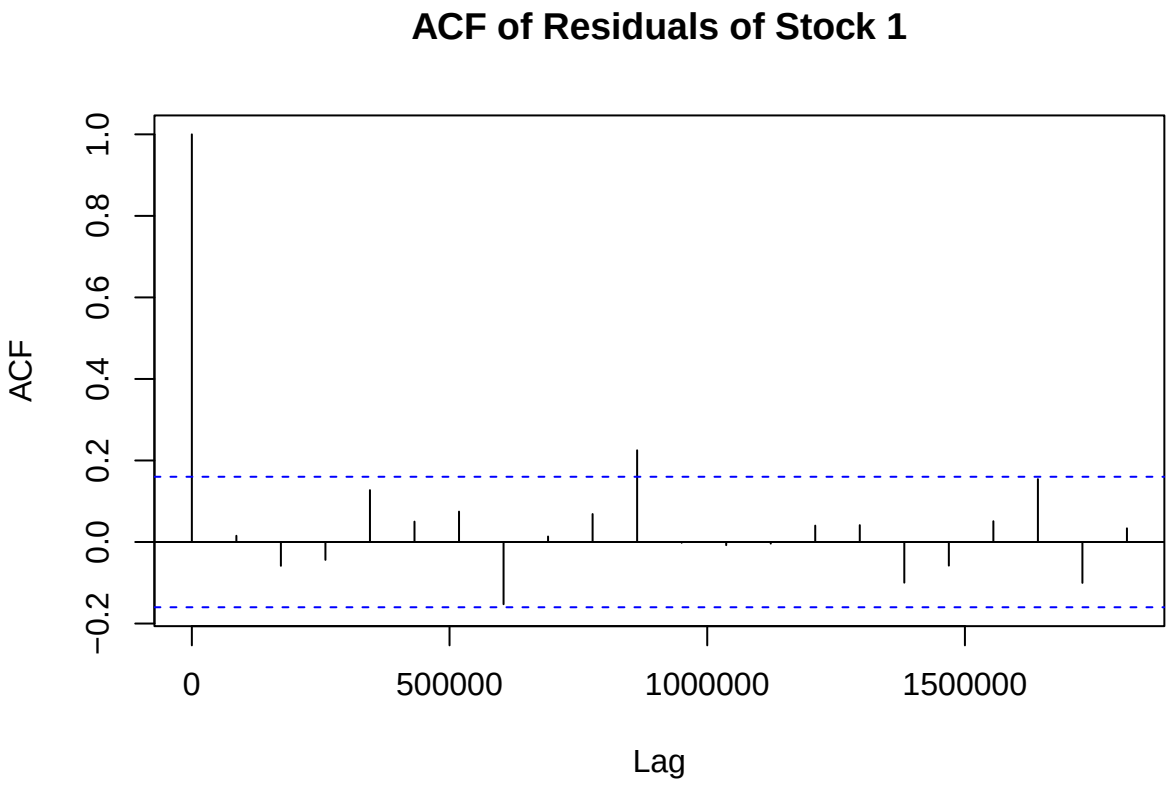
Note that most of the models have the best model of eGARCH among the three predominantly, followed by the GJR-GARCH.

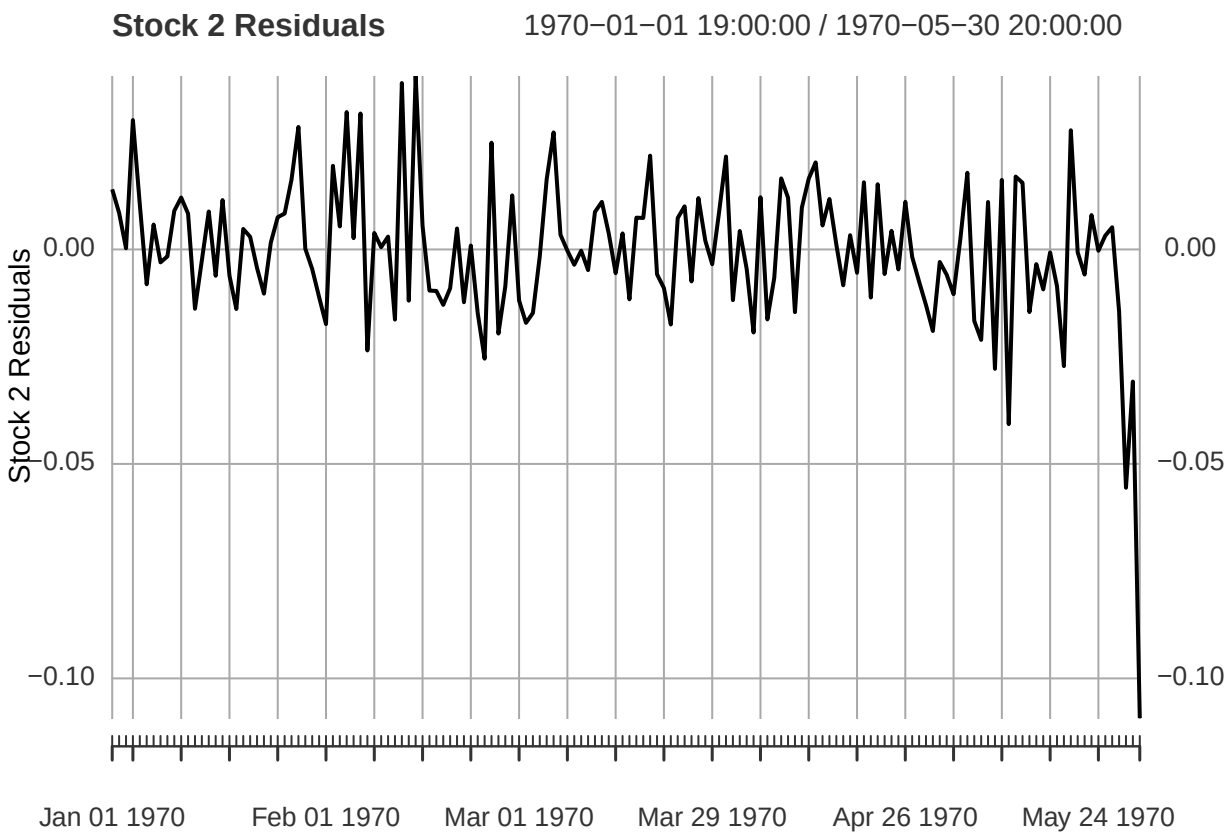
The reason why we only use (1,1) as parameters because we discussed in class how the best models do not provide a significantly better forecast than the GARCH(1,1) model. This result was established by the tests of for superior predictive ability of White (2000) and Hensen (2001).

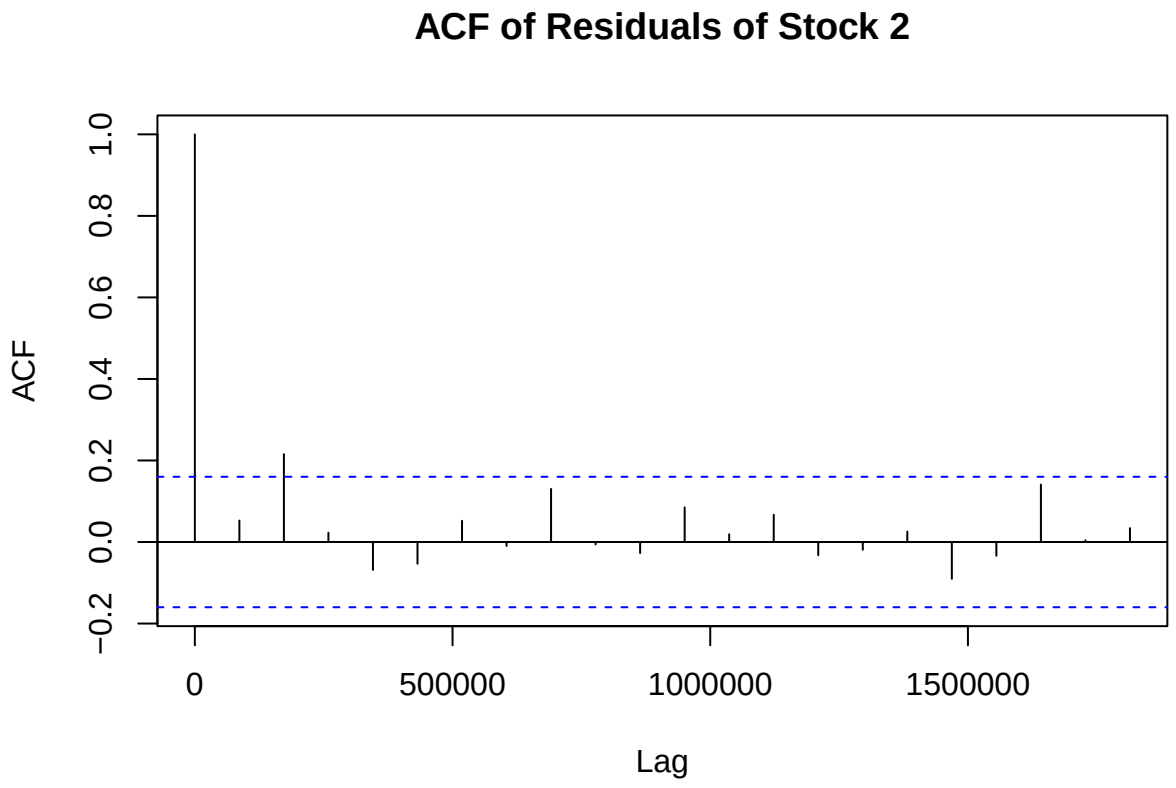
We then fit the best models for each stock (using the decision vector to match each stock to their best models) and find the value at risk levels for each stock.

We also want to confirm the residual diagnostics for these stocks using these models. We present it for the first two stocks since that is what we will be presenting for our forecast plot. The same process can be achieved for the other 38 stocks.




ACF of Residuals of Stock 1
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
500000
1000000
1500000
Lag

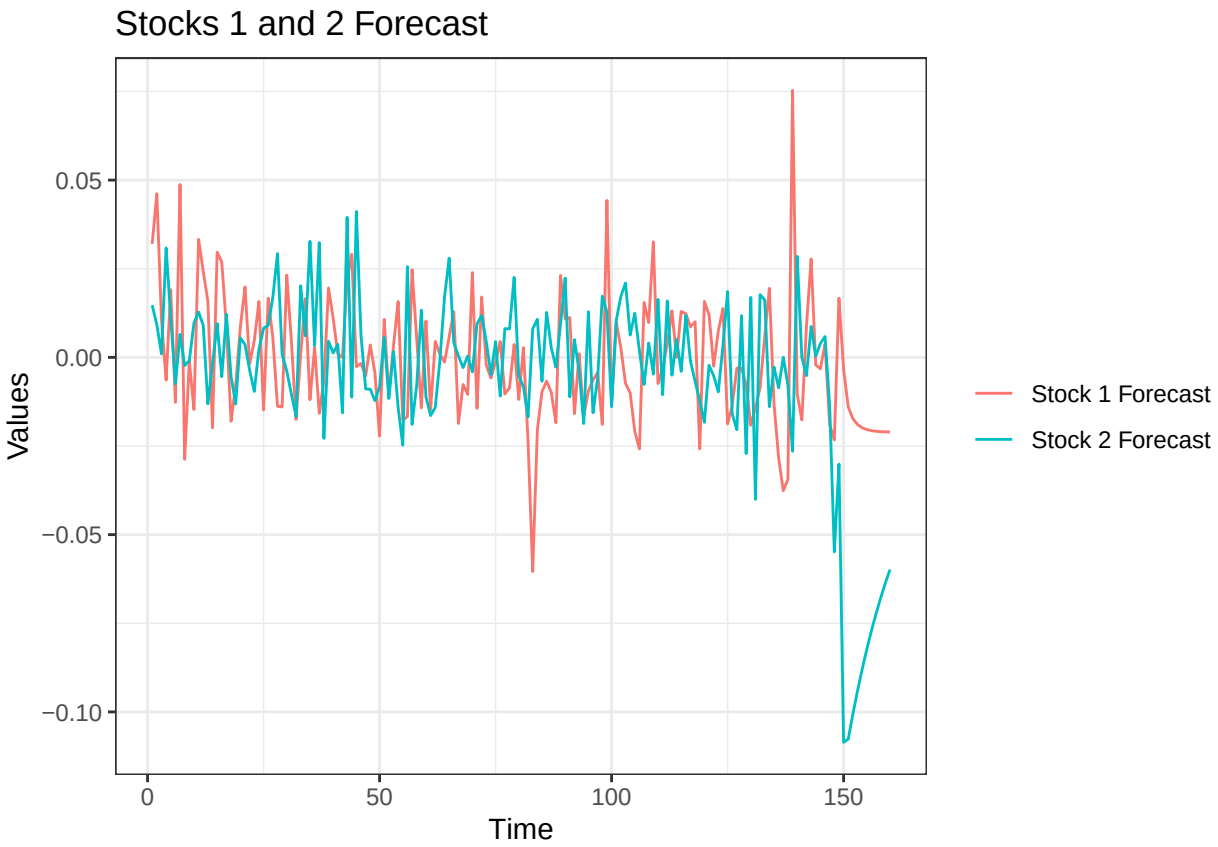

Stock 2 Residuals
1970−01−01 19:00:00 / 1970−05−30 20:00:00
Stock 2 Residuals
0.00
−0.05
−0.10
0.00
−0.05
−0.10
Jan 01 1970
Feb 01 1970
Mar 01 1970
Mar 29 1970
Apr 26 1970
May 24 1970




```
## Ljung-Box test for Stock 1's residuals:

##
##  Box-Ljung test
##
## data:  stock1_residuals
## X-squared = 0.035759, df = 1, p-value = 0.85

## Ljung-Box test for Stock 2's residuals:

##
##  Box-Ljung test
##
## data:  stock2_residuals
## X-squared = 0.42361, df = 1, p-value = 0.5151
```

- Goodness of fit from residuals:
  - Both the residuals from both models look very similar and appear to be randomly/evenly spread out (cloudy pattern)
  - Both ACF's look very similar and appear to be stationary (no major spikes outside interval)
  - This supports the Box-Ljung test for the residuals where our p-values are greater than 0.05

We can now plot the forecast for the first two stocks:



Looking at the forecasts for the first two stocks, the first stock appears to have a decent forecast that captures the volatility of the stock returns well and appears to have low risk value for the stock. Therefore, this stock appears to be safe to invest into.

For Stock 2, the forecast of the Value-At-Risk is super volatile and unpredictable. So the Value-At-Risk is unstable for the first 10 steps ahead for second stock. Therefore, this stock is not safe to invest into.

The same can be performed with all other stocks.

Scenario 3

Number of pages for text + R Output (excluding plots + code): 1.5 pages

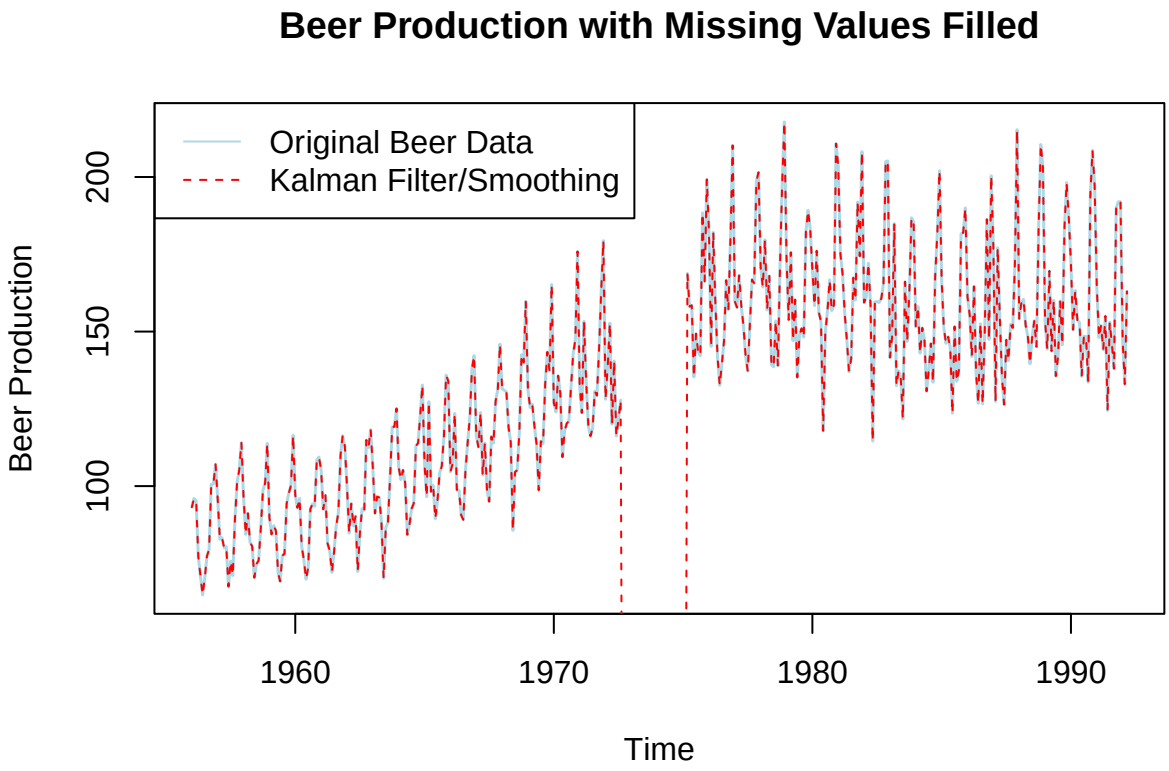
To start, because we have imputed data, we want to use Kalman Filtering and Smoothing to determine the missing values for the beer. We can accomplish this by using the other files on car, steel, gas, electricity, and temperature.

We first combine them into one data frame.

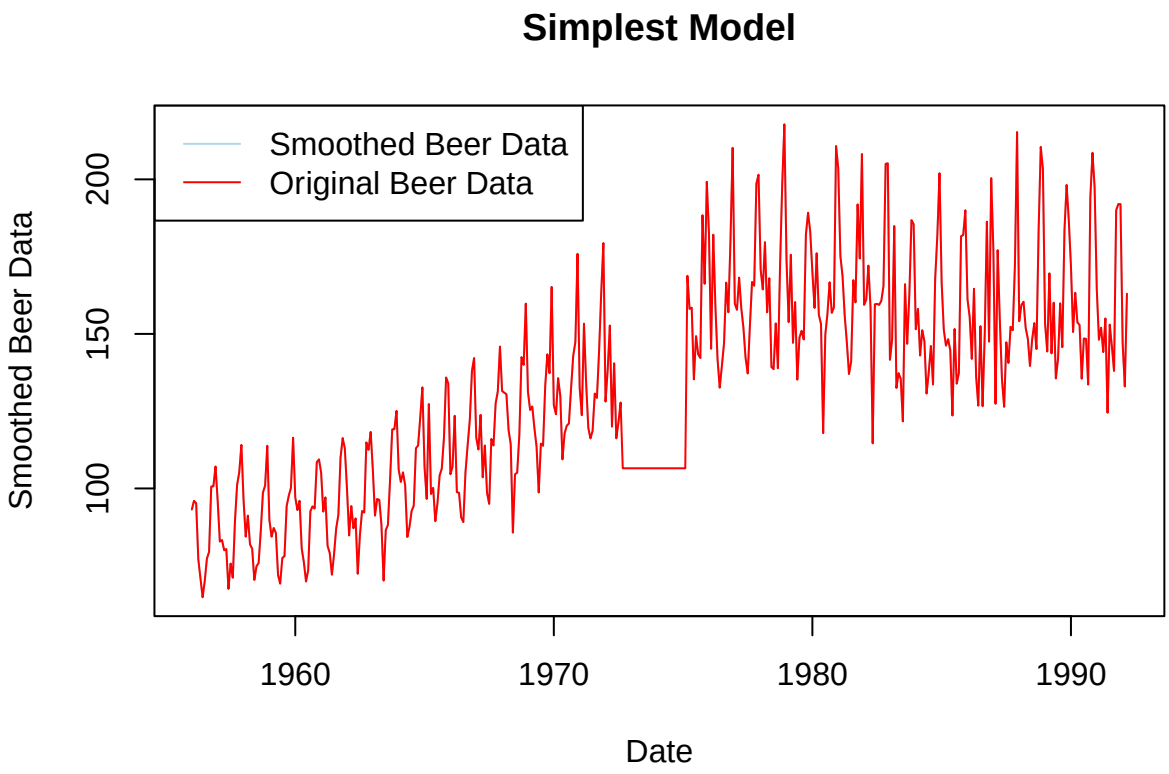
Note that we only used data starting from January 1956 since temperature had an earlier start date of November 1943 (hard to estimate given that no other file has any data).

I tried to fit a Kalman Filtering and Smoothing Model, but the imputed values ended up to be very different from our original data (in the negatives). This is show in the plot next page. After trying different model parameters (including interactions and different relationships), I proceeded to try a different method.

When I tried doing Kalman Filtering, Smoothing, and both, the result ended up having the same parameters.



Even the simplest model with just the intercept value of 1 didn't seem to work very well:

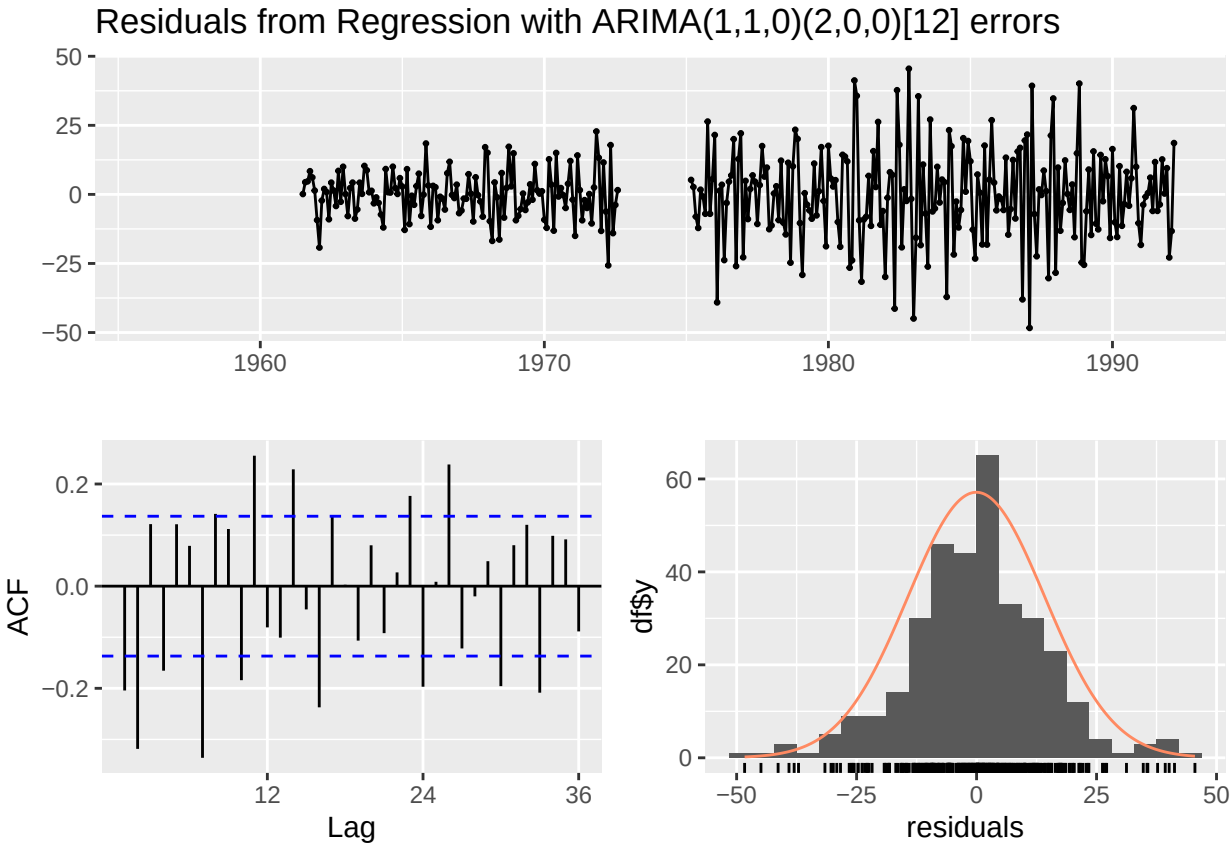


Alternatively, we can also try using the ARIMAX model, since it is excellent at handling missing values by using exogenous variables that are from other time series. We can use normal time series techniques like ACF the most optimal model in this scenario. We also scaled the exogenous variables since they are of different scales.

We can't really see the ACF or the stationarity tests since there are missing data in our beer data.

So we just go ahead and try to fit the best models. We start with the auto.arima model recommended by R:

The default model appears to be the SARIMA(1,1,0,2,0,0) model.




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(1,1,0)(2,0,0)[12] errors
## Q* = 226.81, df = 21, p-value < 2.2e-16
##
## Model df: 3.   Total lags used: 24
```

Note that our ACF values imply that there is some seasonality since they lag peaks occur every 12 values. We try to incorporate this into a SARIMAX model.

I proceeded to test multiple kinds of models, by first resolving differencing by looking at the residual ACF plot and adjusted parameters based on the spikes or patterns presented.

In the end, I decided to go with three SARIMAX models:

- SARIMA(2,1,2,2,1,2)
- SARIMA(3,1,2,2,1,2)
- SARIMA(2,1,3,2,1,1)

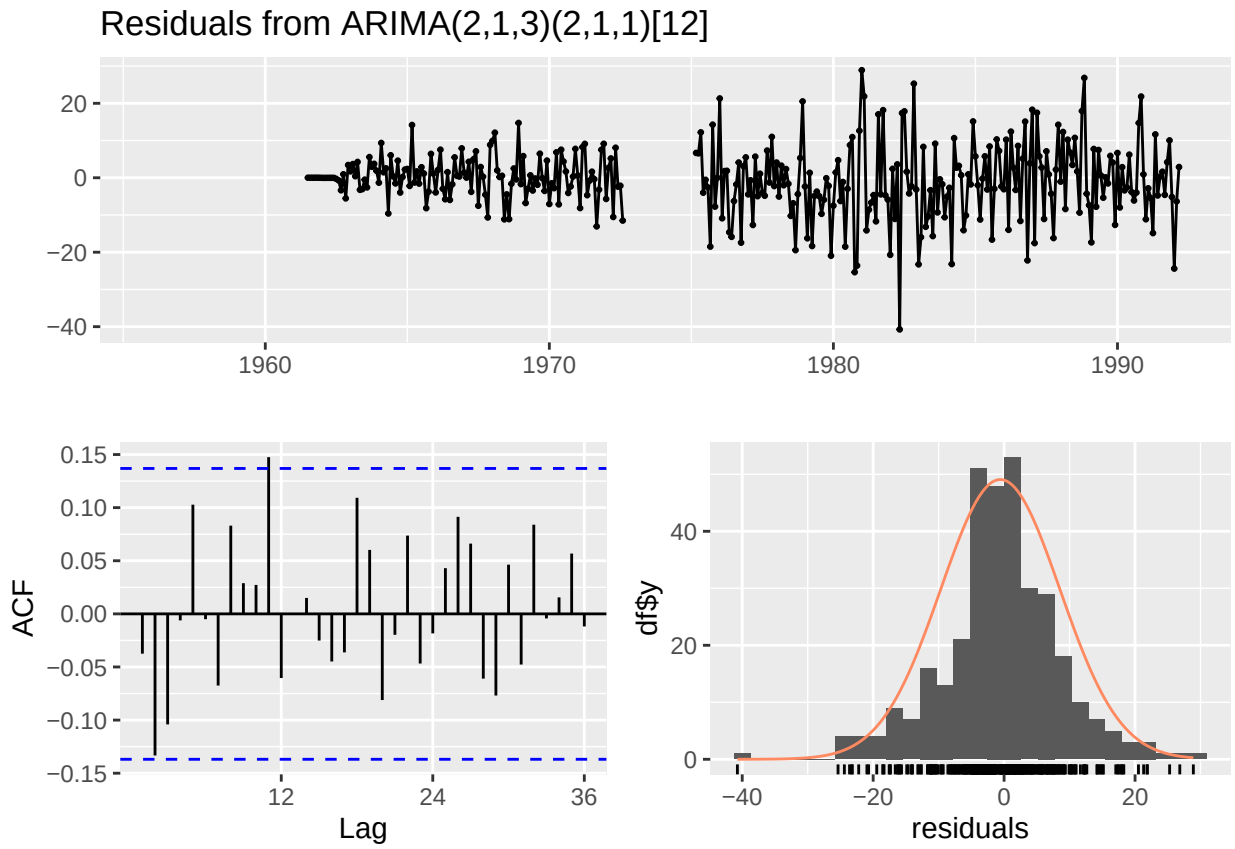
We can now compute and compare AIC values to find the best ARIMAX model:

Table 3: AIC of Different Models

Model	AIC
Auto.Arima	2792.665
SARIMA(2,1,2,2,1,2)	2483.172
SARIMA(3,1,2,2,1,2)	2485.159
SARIMA(2,1,3,2,1,1)	2440.594

So our best model is the SARIMA(2,1,3,2,1,1) model!

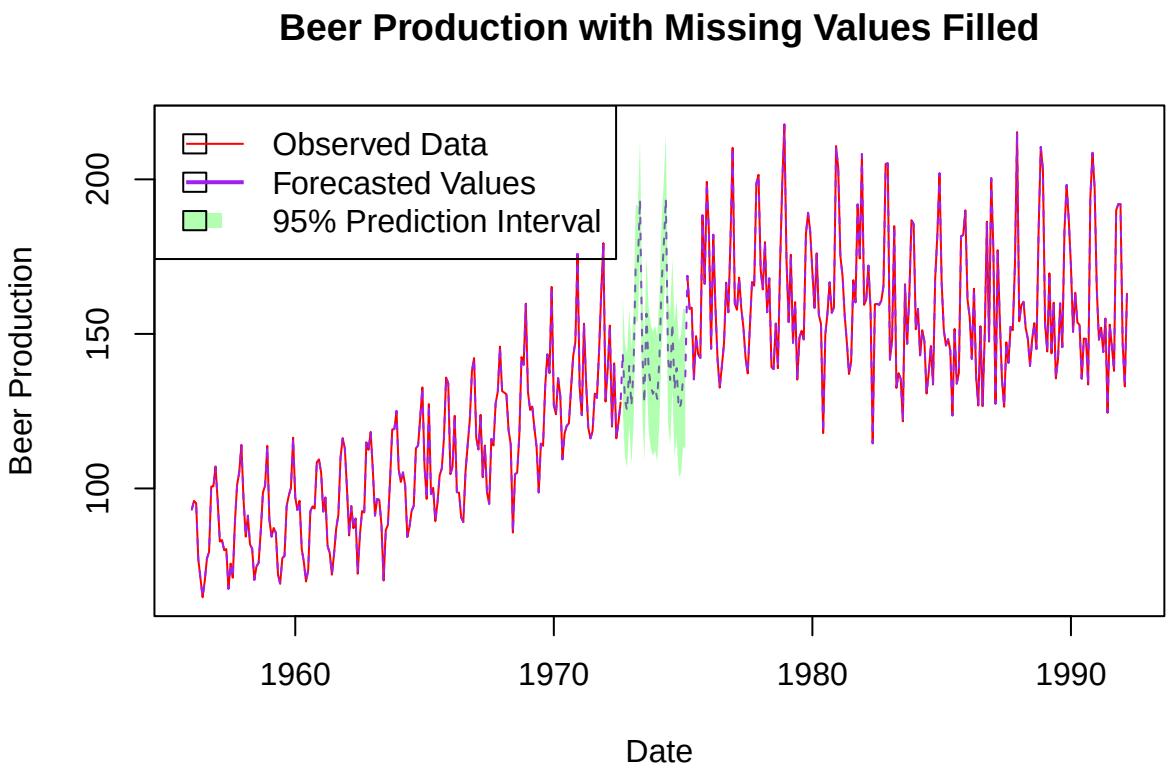
Evaluating the residuals:




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,3)(2,1,1)[12]
## Q* = 32.797, df = 16, p-value = 0.007862
##
## Model df: 8.   Total lags used: 24
```

The ACF plot looks pretty stationary since all the spikes are basically within the line. The residuals look fairly random and normally distributed. Therefore, this model looks appropriate for our beer dataset.

We can now compute the predicted values of the imputed values:



From the plot, the forecast of the imputed values seem to make sense with small prediction intervals as well. So we can conclude scencario 3.

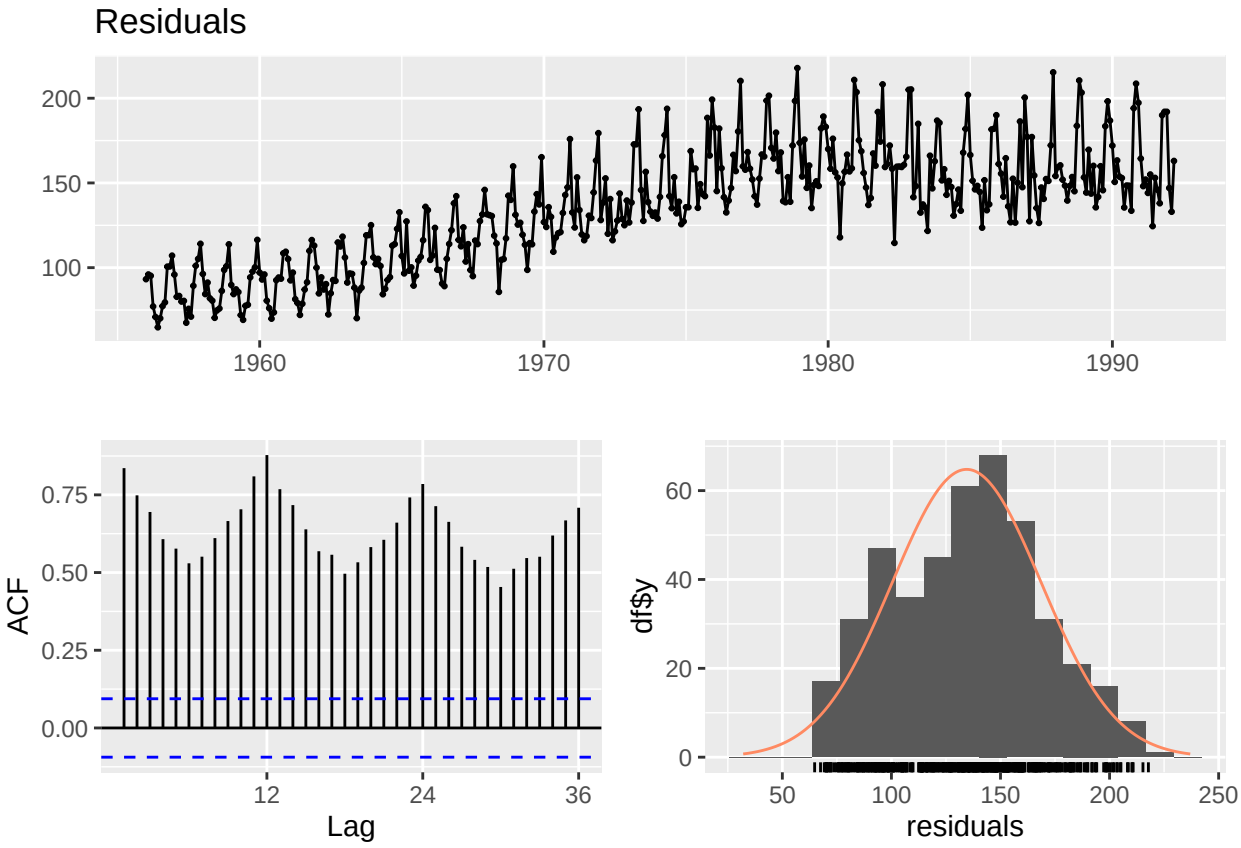
Scenario 4

Number of pages for text + R Output (excluding plots + code): 1.5 pages

Using our imputed data from Scenario 3, we can now refit our data to forecast the 24 steps ahead. We will be re-fitting our models assuming our imputed data is the true value.

Note that we will not be using any exogenous variables for this scenario since none of the other txt files provide any values after the date for the beer data.

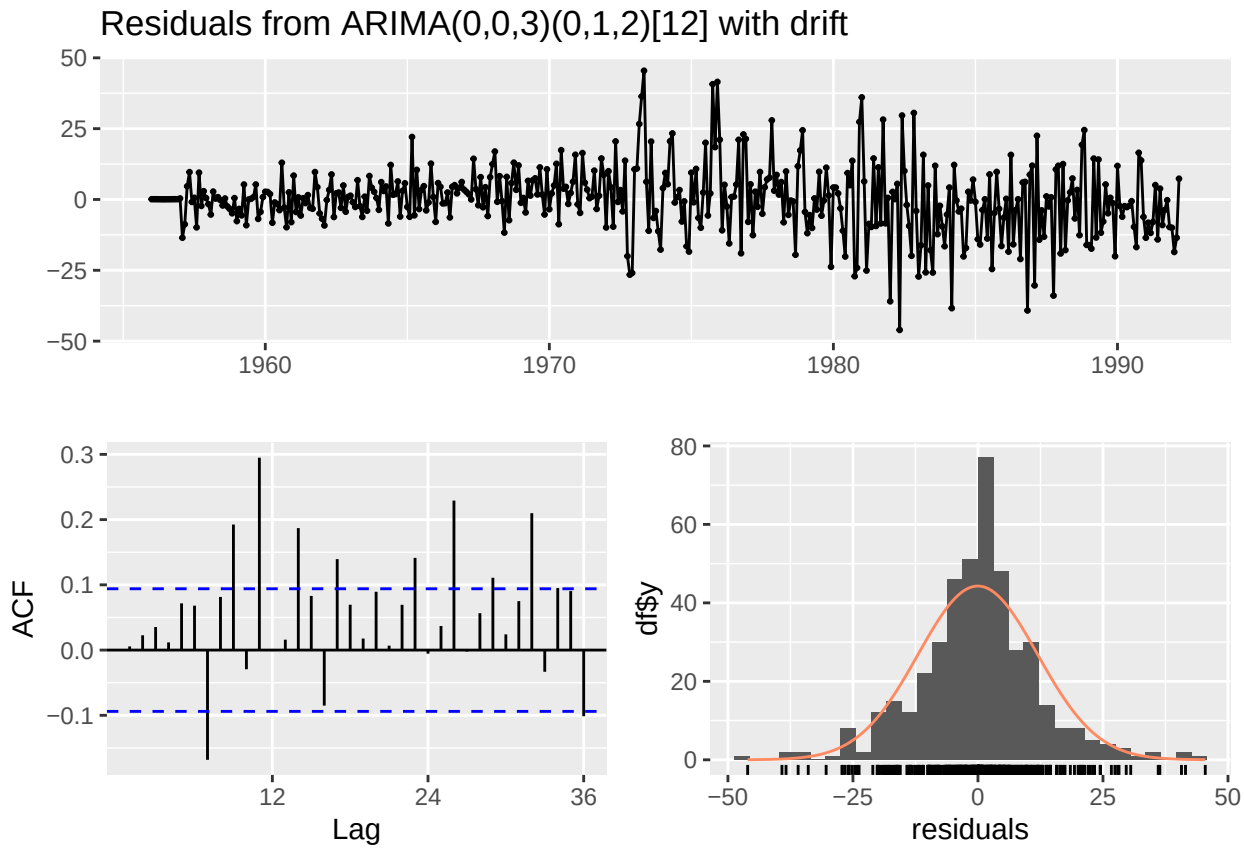
We start with looking at our ACF plot and residual diagnostics.




```
##
##  Ljung-Box test
##
## data:  Residuals
## Q* = 4834.7, df = 24, p-value < 2.2e-16
##
## Model df: 0.   Total lags used: 24
```

Our ACF plot is clearly not stationary, and indicates a strong indication of seasonality every 12 months. The residuals also indicate an upward oscillatory pattern which suggests that more work needs to be done to make our data stationary.

We first try with the auto.arima model again. It appears that it is the SARIMA(0,0,3,0,1,2) with drift.




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,3)(0,1,2)[12] with drift
## Q* = 125.22, df = 19, p-value < 2.2e-16
##
## Model df: 5.   Total lags used: 24
```

The ACF plot still indicate minor seasonality levels with peaks every 12 lags.

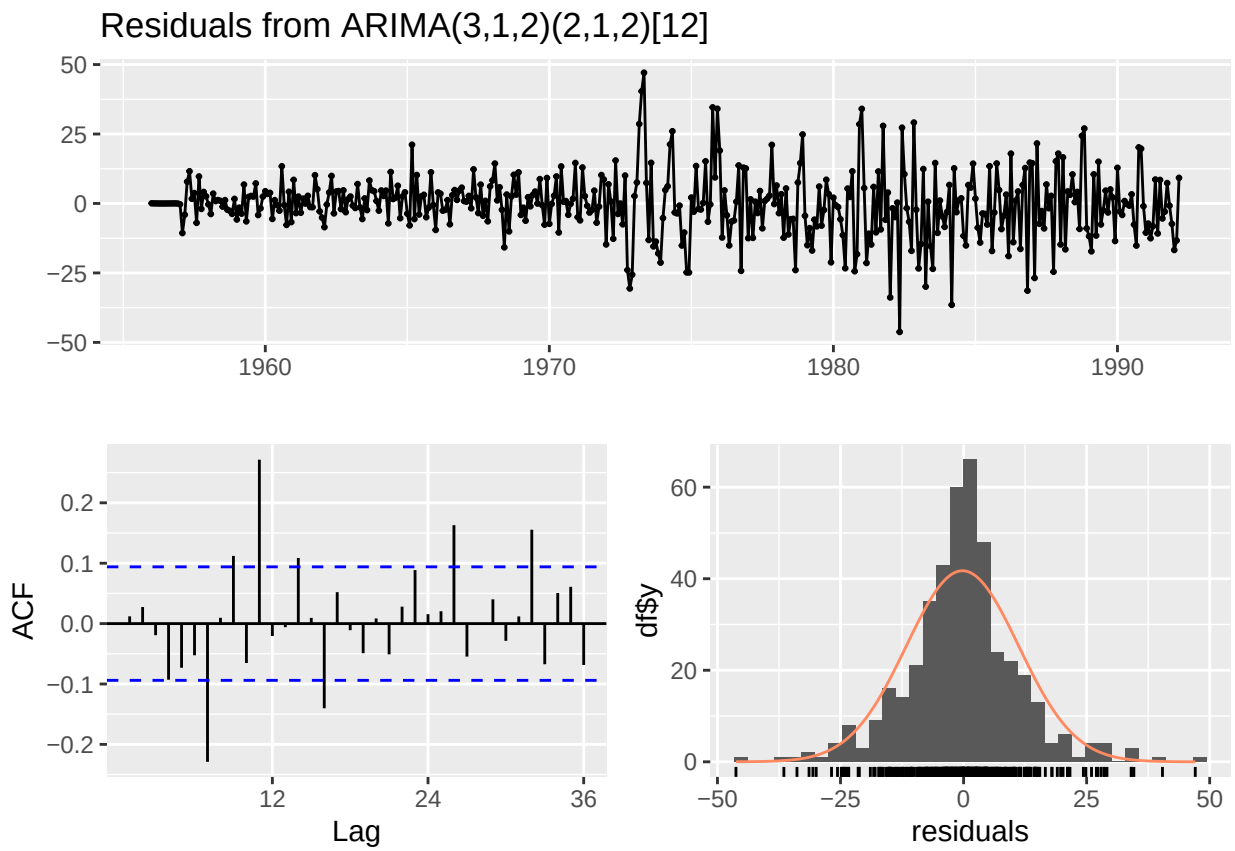
We first try our previous models from Scenario 3 to test if our best model is the same as the one for the imputed values.

Table 4: AIC of Different Models

Model	AIC
Auto.Arima	3330.585
SARIMA(2,1,2,2,1,2)	3307.056
SARIMA(3,1,2,2,1,2)	3300.255
SARIMA(2,1,3,2,1,1)	3303.513

This time, our best model is the SARIMA(3,1,2,2,1,2) model! It barely beat the initial model in Scenario 3

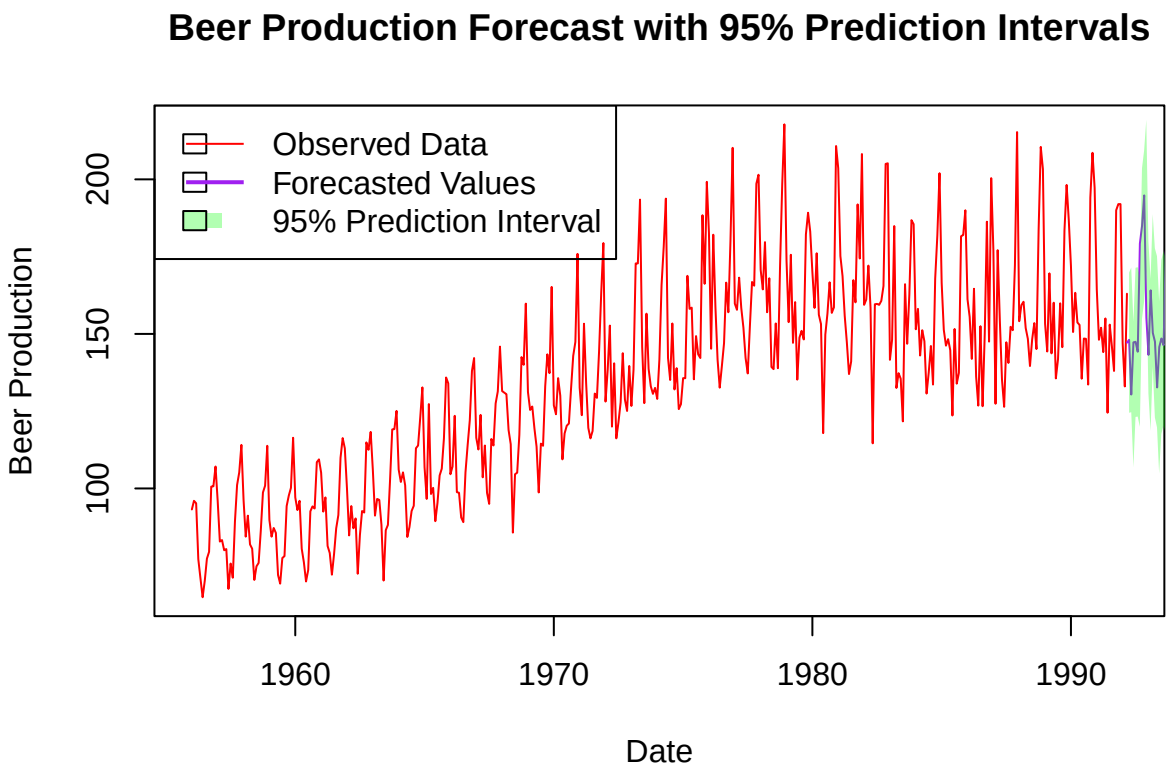
Evaluating the residuals:




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,1,2)(2,1,2)[12]
## Q* = 93.983, df = 15, p-value = 1.782e-13
##
## Model df: 9.   Total lags used: 24
```

The residual ACF plot doesn't look too bad, with only a few spikes outside the interval. In addition, our residual plots appear normal and random. Therefore, this models looks appropriate for our data.

Now we plot our forecast for 24 steps ahead:

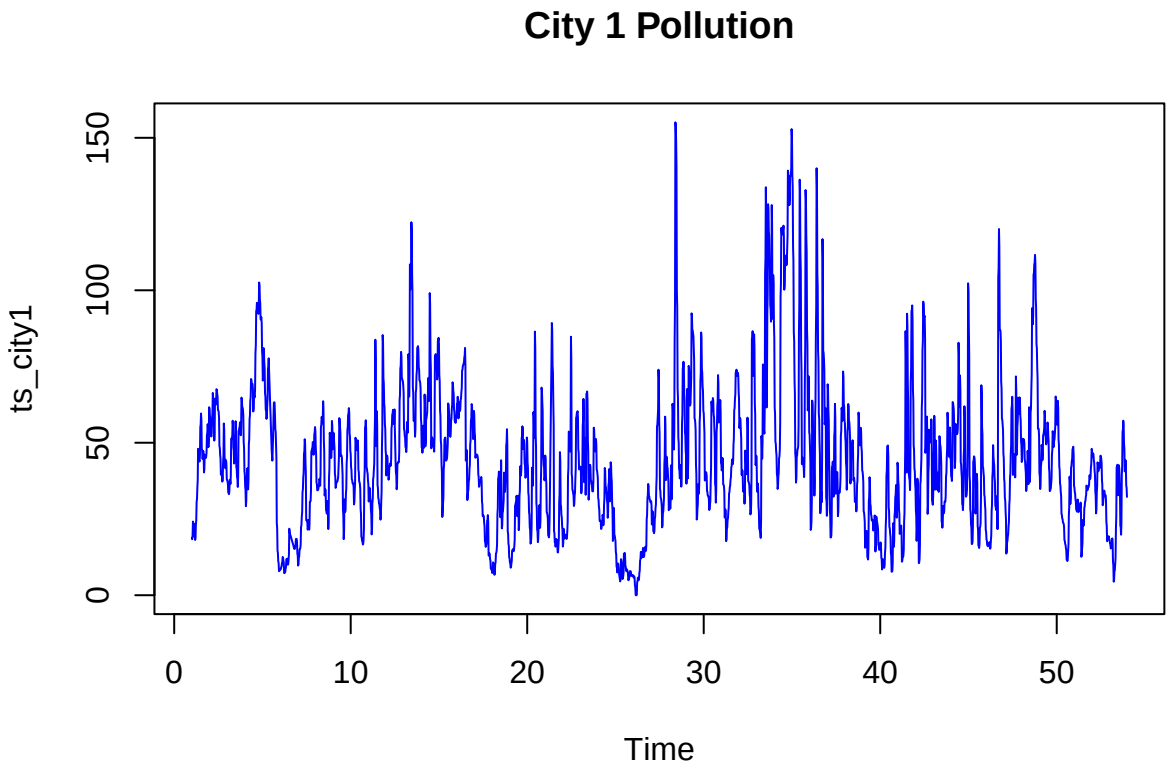


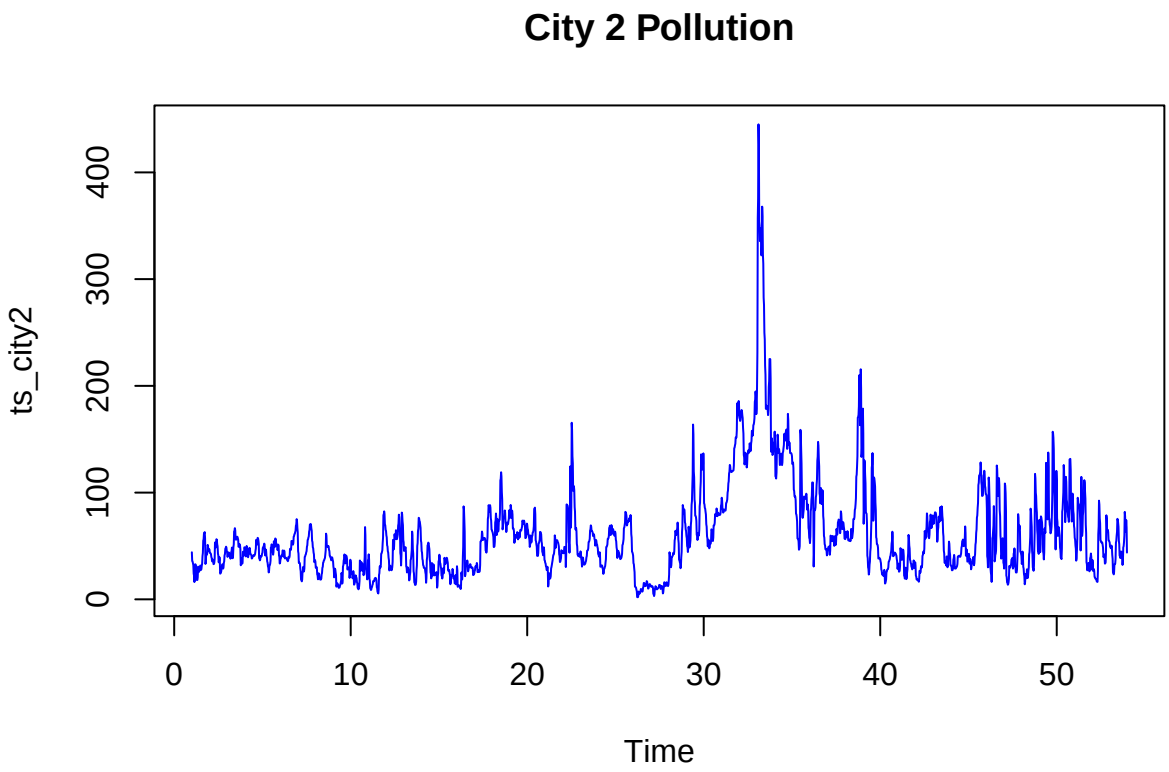
The forecast for 24 steps ahead seems to make sense and the confidence interval doesn't seem too large. Therefore, this seems to be an appropriate model for our new beer dataset.

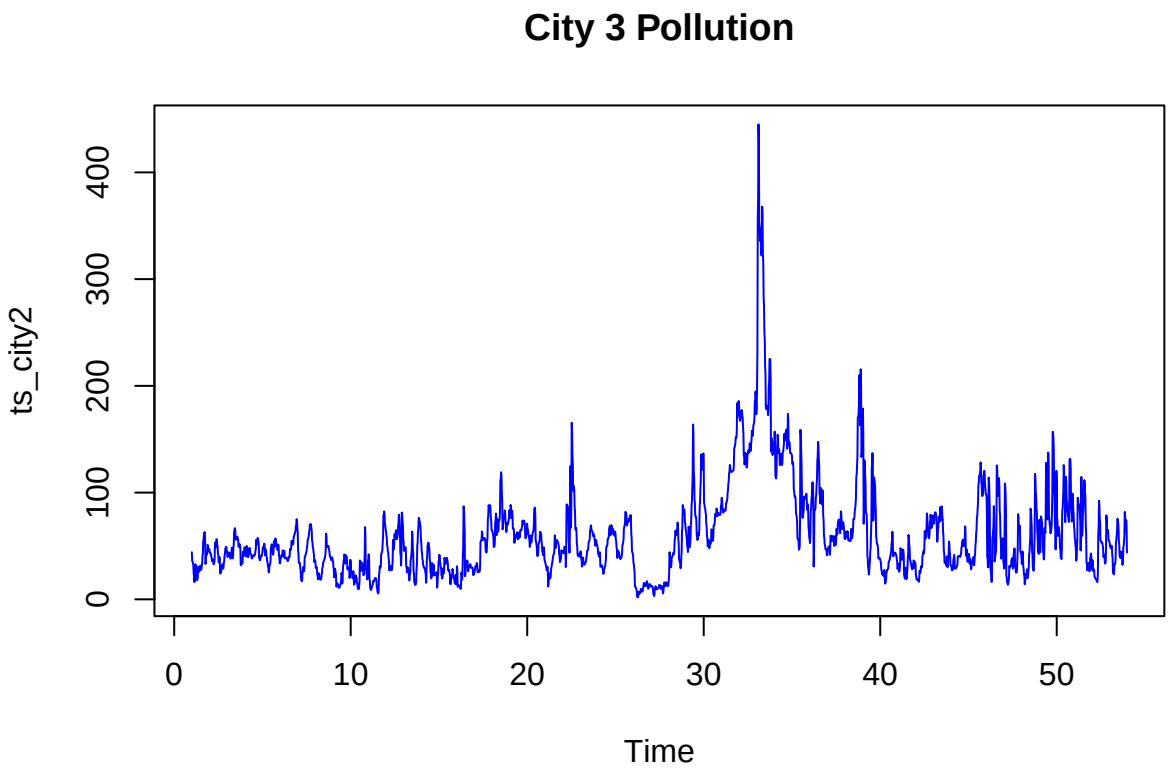
Scenario 5

Number of pages for text + R Output (excluding plots + code): 1.75 pages

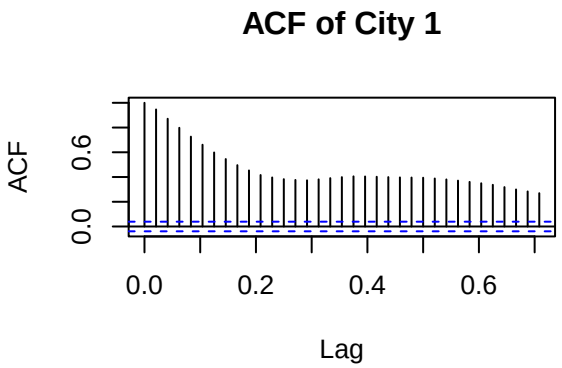
We first start by plotting our three series.

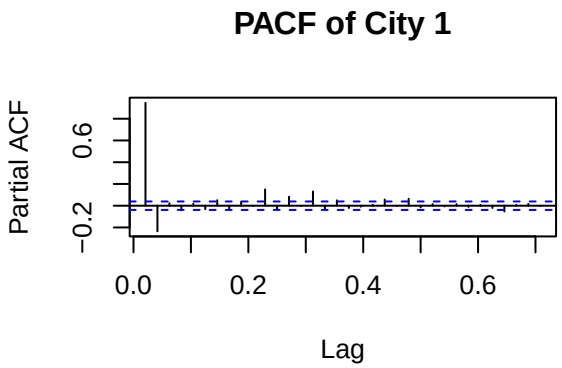


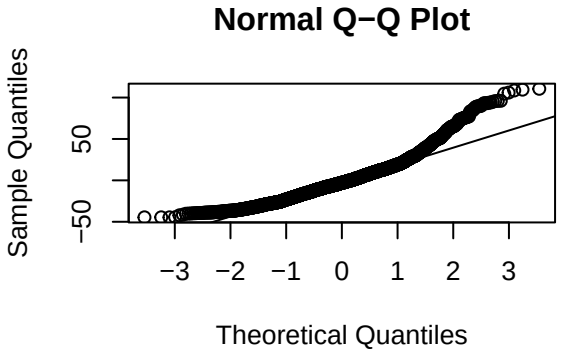

City 2 Pollution
ts_city2
0
100
200
300
400
0
10
20
30
40
50
Time

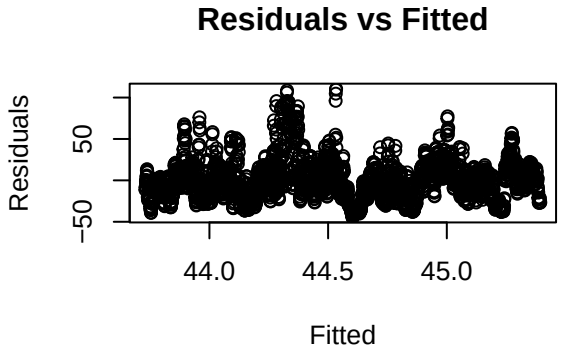


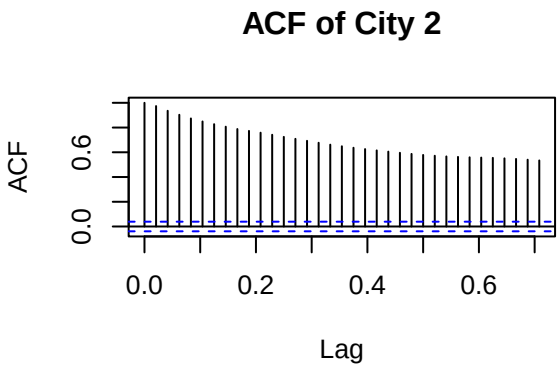
We will first diagnose stationarity by plotting the ACF/PACF plots of the original data with residual diagnostics:

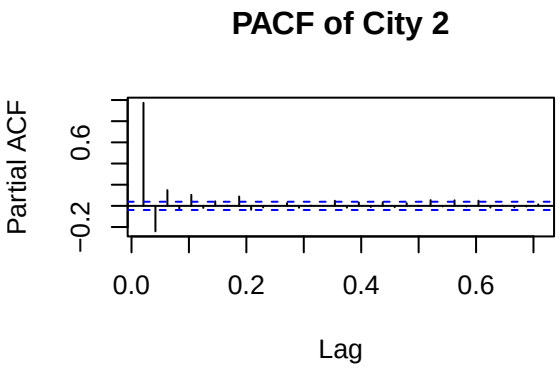

ACF of City 1
ACF
0.0
0.6
0.0
0.2
0.4
0.6
Lag

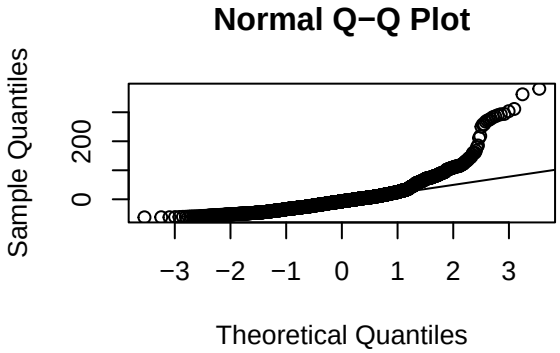

PACF of City 1
Partial ACF
−0.2
0.6
0.0
0.2
0.4
0.6
Lag

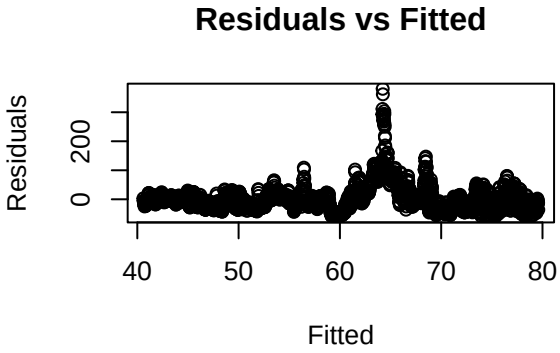

Normal Q−Q Plot
Sample Quantiles
−50
50
−3
−2
−1
0
1
2
3
Theoretical Quantiles

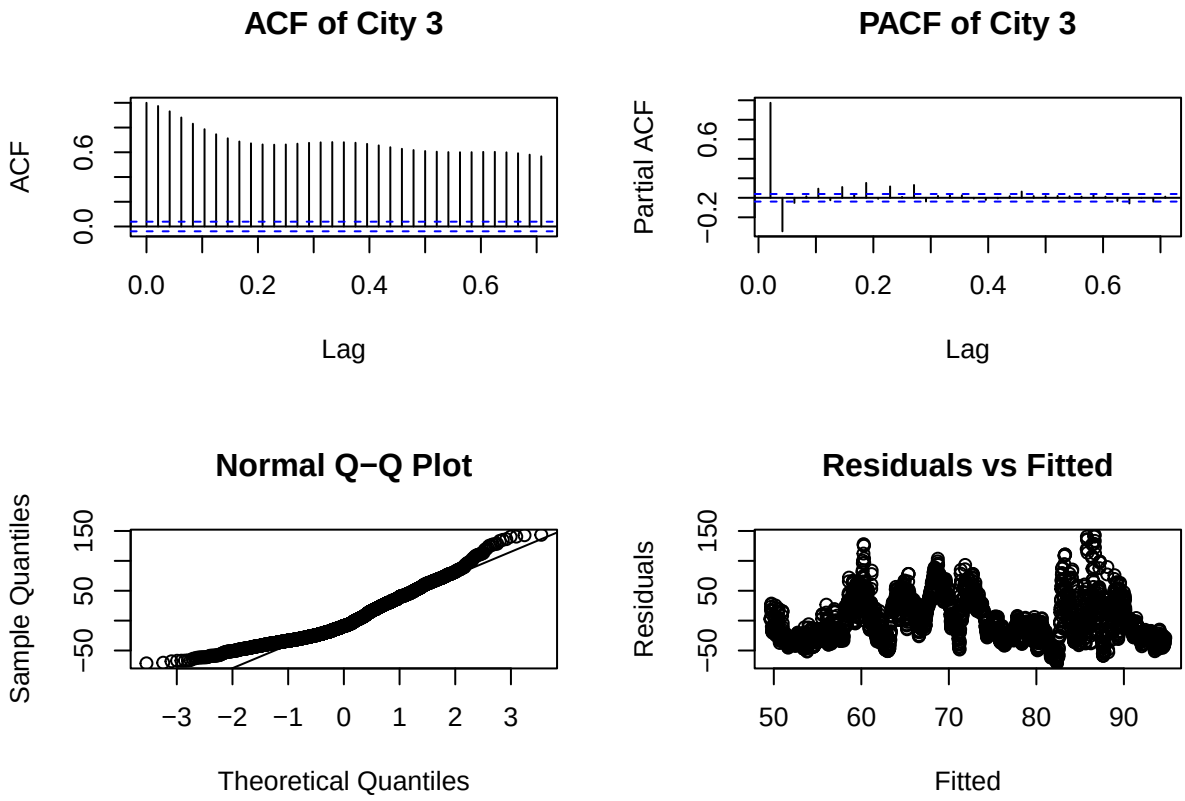

Residuals vs Fitted
Residuals
−50
50
44.0
44.5
45.0
Fitted


ACF of City 2
ACF
0.0
0.6
0.0
0.2
0.4
0.6
Lag


PACF of City 2
Partial ACF
−0.2
0.6
0.0
0.2
0.4
0.6
Lag

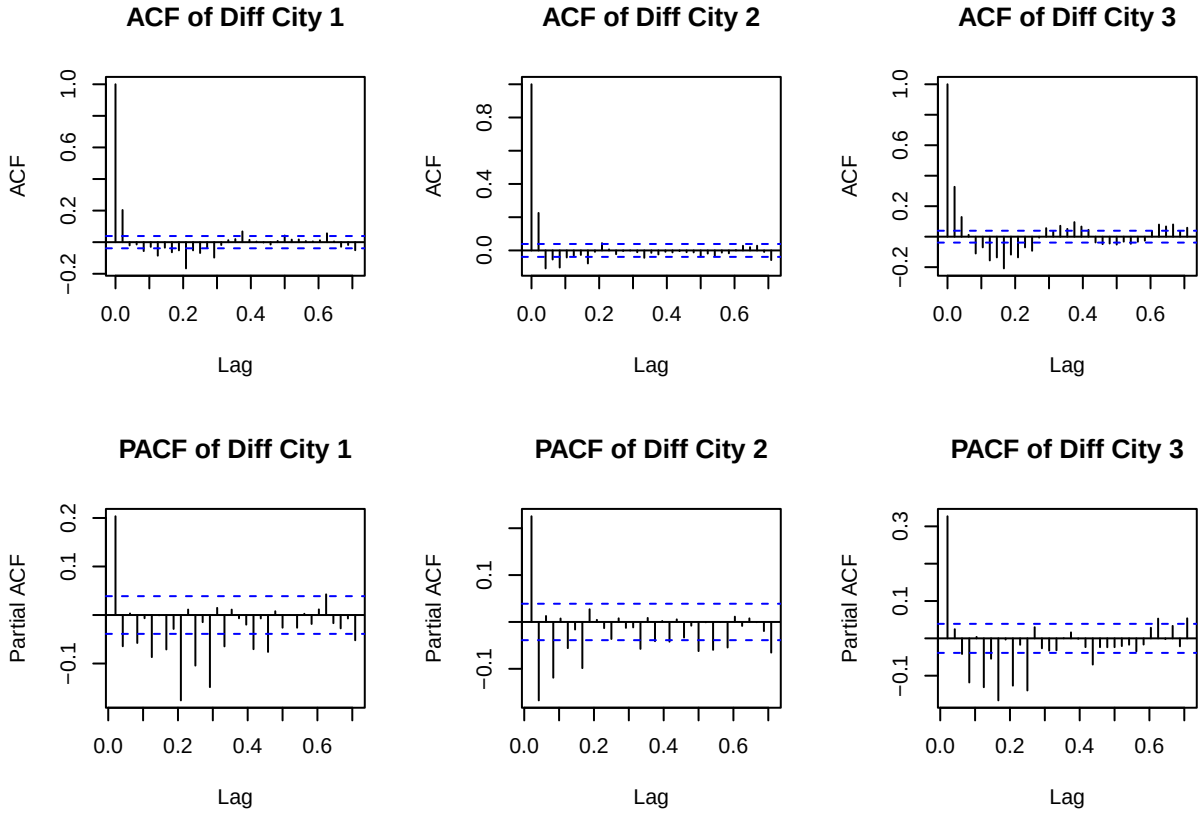

Normal Q−Q Plot
Sample Quantiles
0
200
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Residuals vs Fitted
Residuals
0
200
40
50
60
70
80
Fitted



For all the ACF's for the 3 cities, they are clearly not stationary because there is no clear cutoff and a steady decrease in the spikes of the ACF. All the PACF's also look like they exceed the cutoff periodically. All the normal QQ-plots violate normality since the QQ-points are off the QQ-line and the residuals vs fitted plot all indicate a somewhat oscillatory pattern.

We first take the first differences of the time series and plot their ACF's and PACF's.



For City 1, our ACF cuts off at $q = 1$ with mild spikes every other 0.2 lags (maybe seasonality). While investigating the correct parameters, residual diagnostics indicated sharp spikes periodically and so seasonal variation was accounted for as well.

For City 2, the ACF cuts off abruptly, indicating that we have a stationary model. This also implies that our q parameter is 1 from the first spike and p parameter is 3 from the first 3 spikes being past the line.

For City 3, the ACF appears to be following a somewhat sinusodial pattern, indicating some seasonality in our model. Perhaps a first differencing in the seasonal component may remove that pattern from the ACF.

This shows that first differencing was effective, and second differencing is no longer needed.

We test out some ARMA parameters that matches these plots:

We also have our auto.arima models as well to use as reference.

We determine which the best models are by comparing the AIC values against each other.

Table 5: AIC for City 1

Model	AIC
Auto.Arima	17350.33
SARIMA(2,1,1,1,0,2)	17318.96
SARIMA(1,1,2,1,0,2)	17289.80
SARIMA(2,1,0,1,0,2)	17380.36

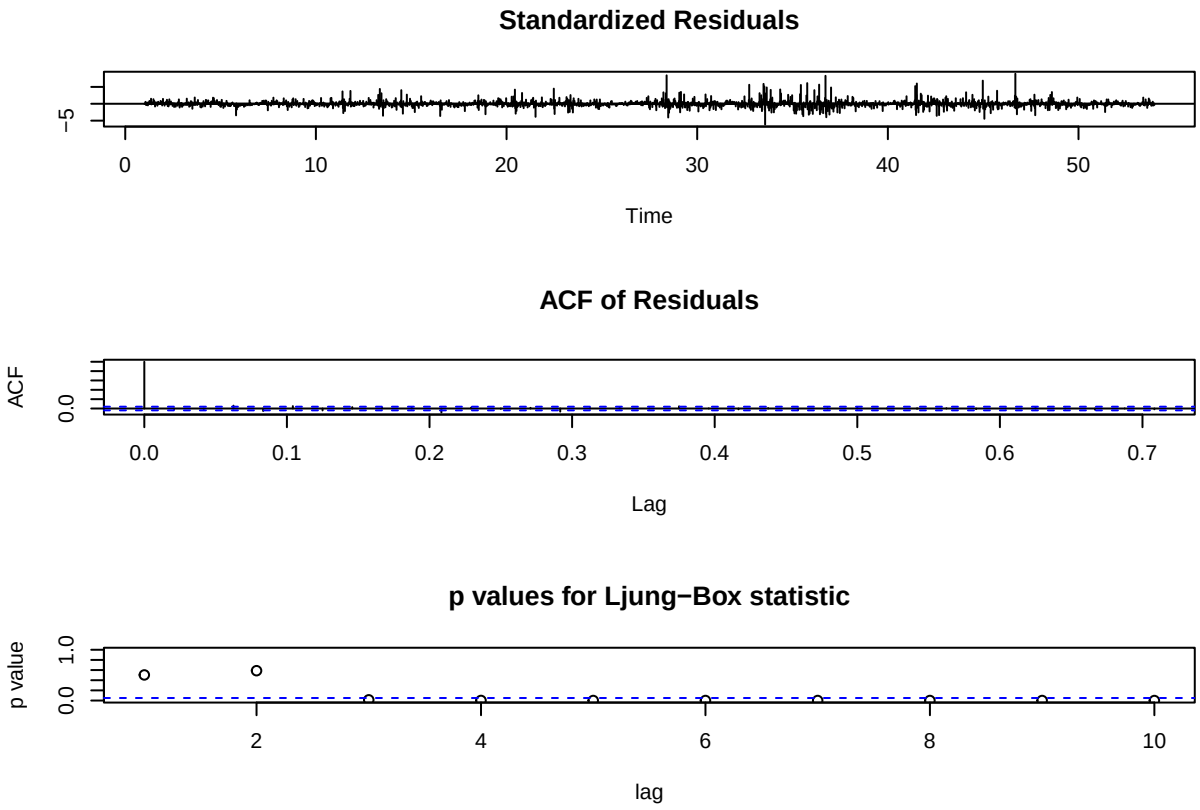
Table 6: AIC for City 2

Model	AIC
Auto.Arima	18745.20
SARIMA(0,1,0,0,1,1)	18830.32
SARIMA(3,1,2,2,1,2)	18783.02
SARIMA(2,1,3,2,1,1)	18745.30

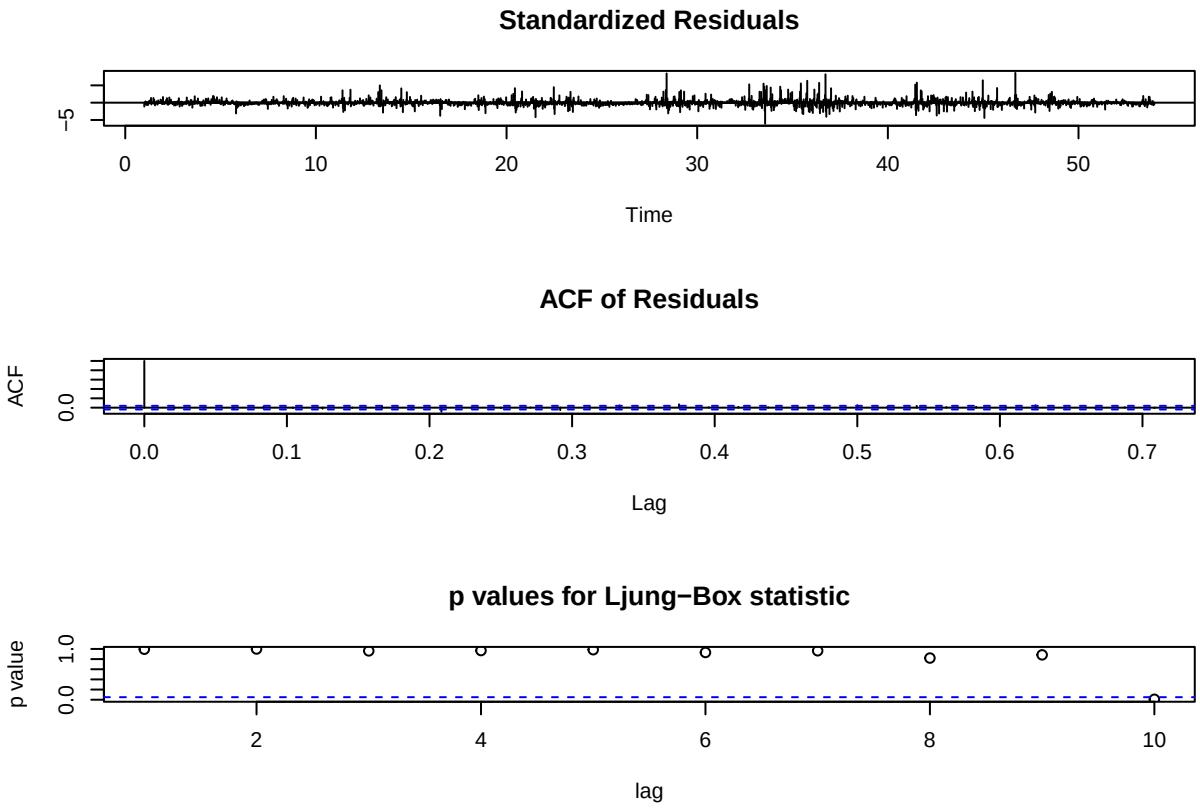
Table 7: AIC for City 3

Model	AIC
Auto.Arima	17887.39
SARIMA(2,1,2,2,1,2)	17867.64
SARIMA(1,1,0,1,1,2)	17622.77
SARIMA(1,1,0,0,0,2)	17885.39

For City 1, although the best model with the lowest AIC value is SARIMA(1,1,2,1,0,2). The residual diagnostics doesn't imply stationarity from the Box-Ljung Test:



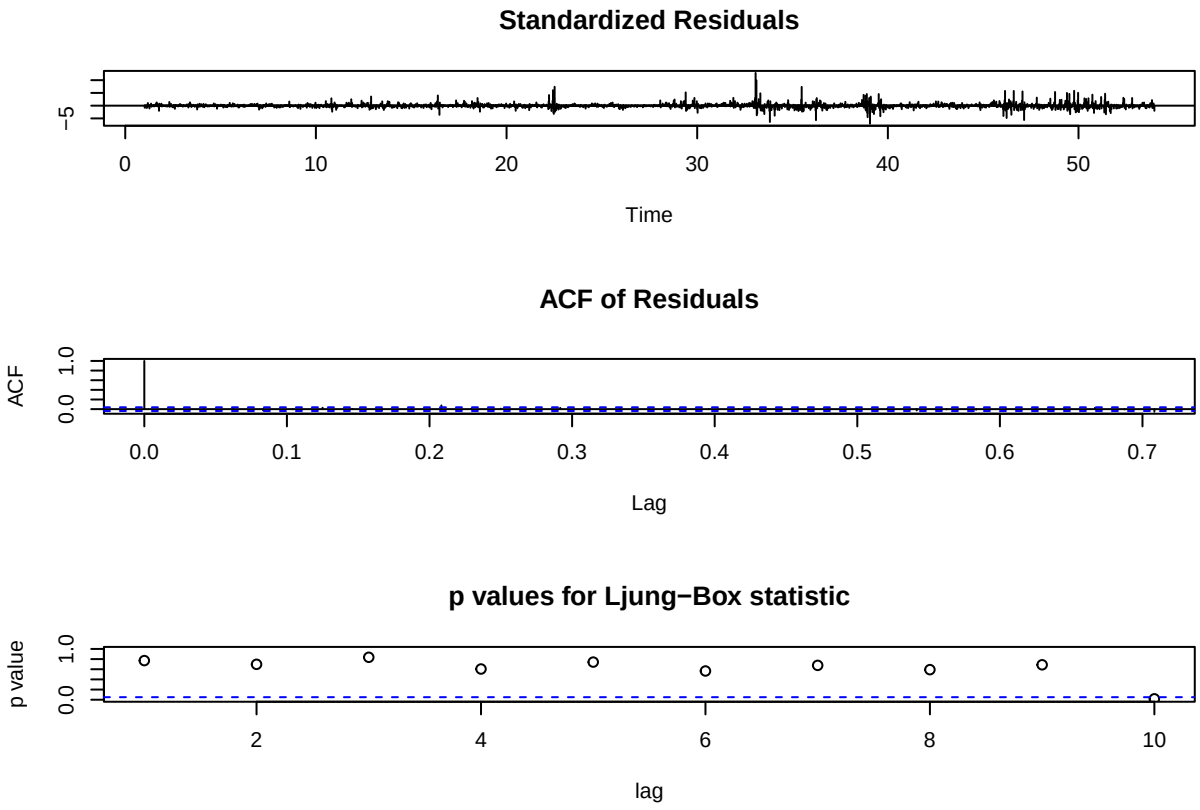
Instead, we will continue to use the auto.arima model which is the SARIMA(3,0,1,0,0,2) that has p-values for the Box-Ljung Pierce test to be > 0.05. The residuals look randomly spread and the ACF looks stationary for the auto.arima model for City 1.



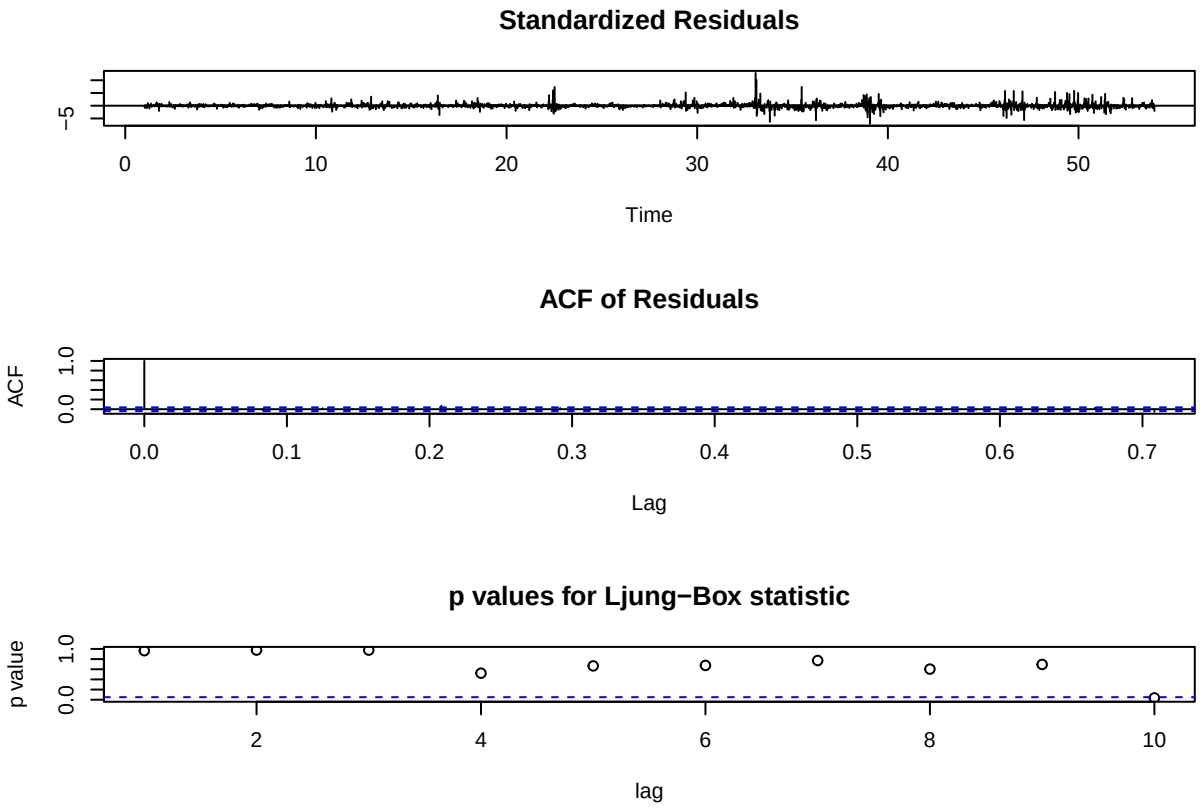
For City 2, we pick the model with the lowest AIC value which is SARIMA(2,1,3,2,1,1). They equivalently have the same diagnostics and AIC value as the auto.arima model which was SARIMA(3,1,3,0,0,2). We just picked the other model arbitrarily but either can work.

For both diagnostics, they have p-values for the Box-Ljung Pierce test to be > 0.05. The residuals look randomly spread and the ACF looks stationary.

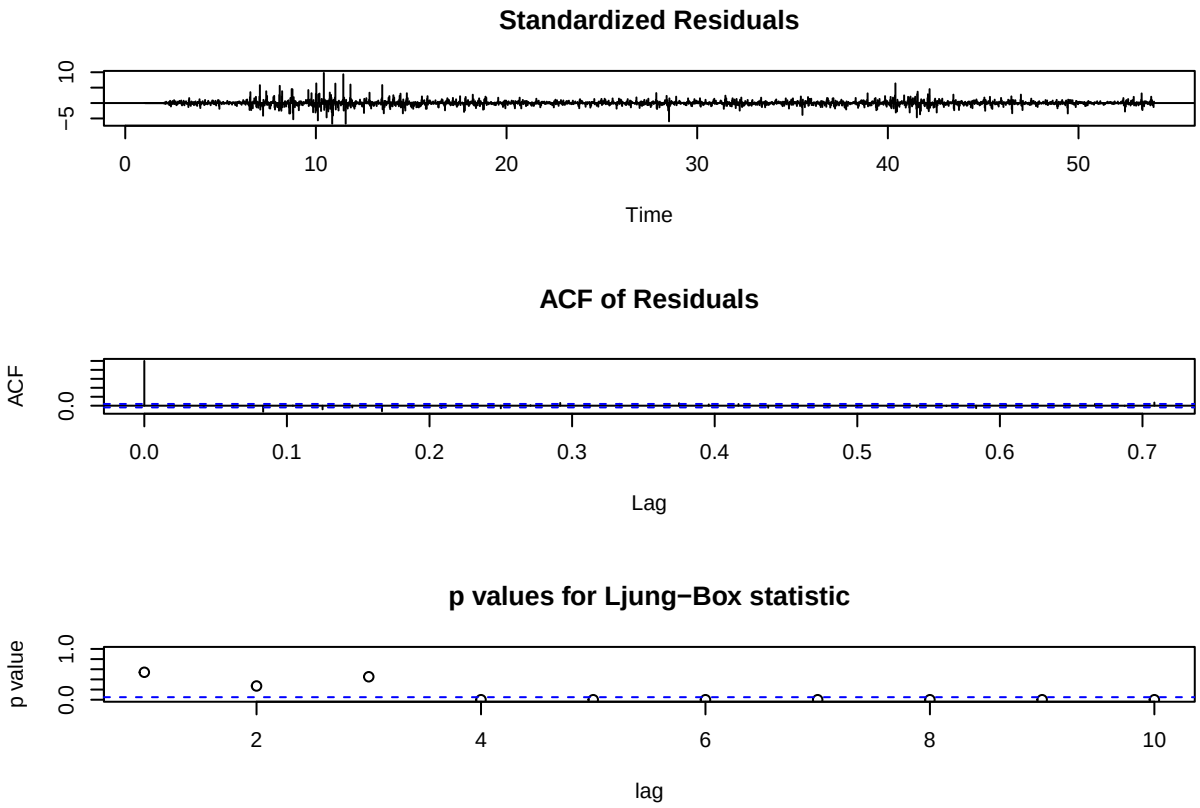
Below is the diagnostics for the final model on SARIMA(2,1,3,2,1,1)



Below are the diagnostics for the auto.arima model:

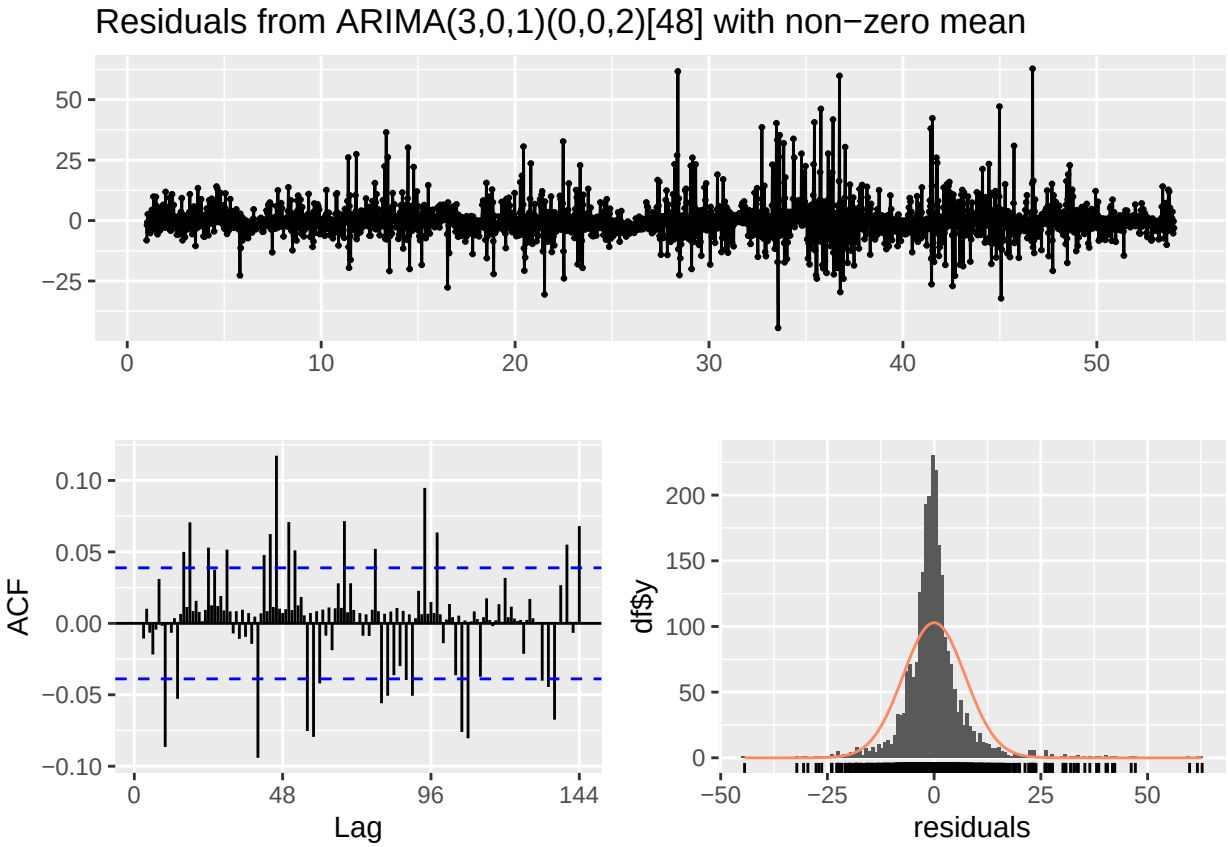


For City 3, the model with the lowest AIC value is SARIMA(1,1,0,1,1,2). This is supported by the Box-Ljung Pierce Test below:

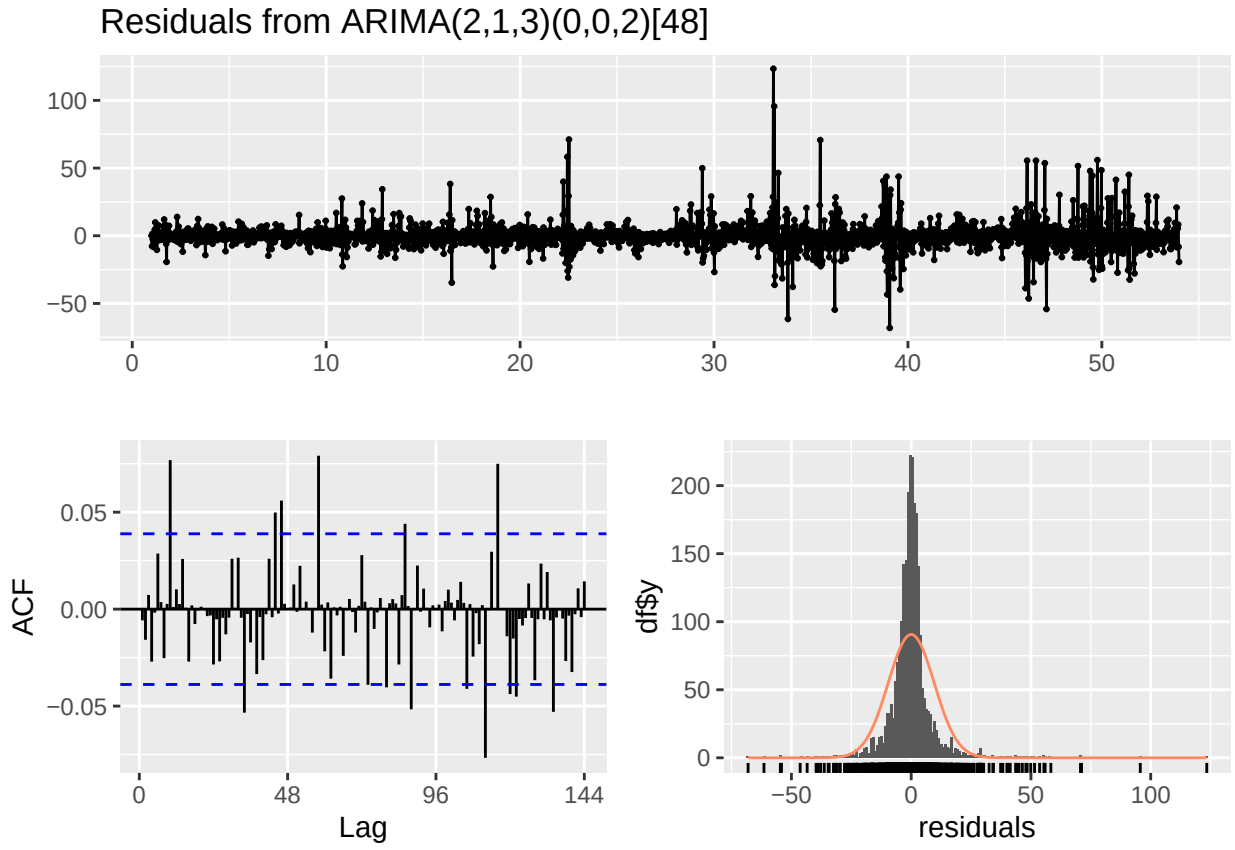


The residuals look randomly spread and the ACF looks stationary.

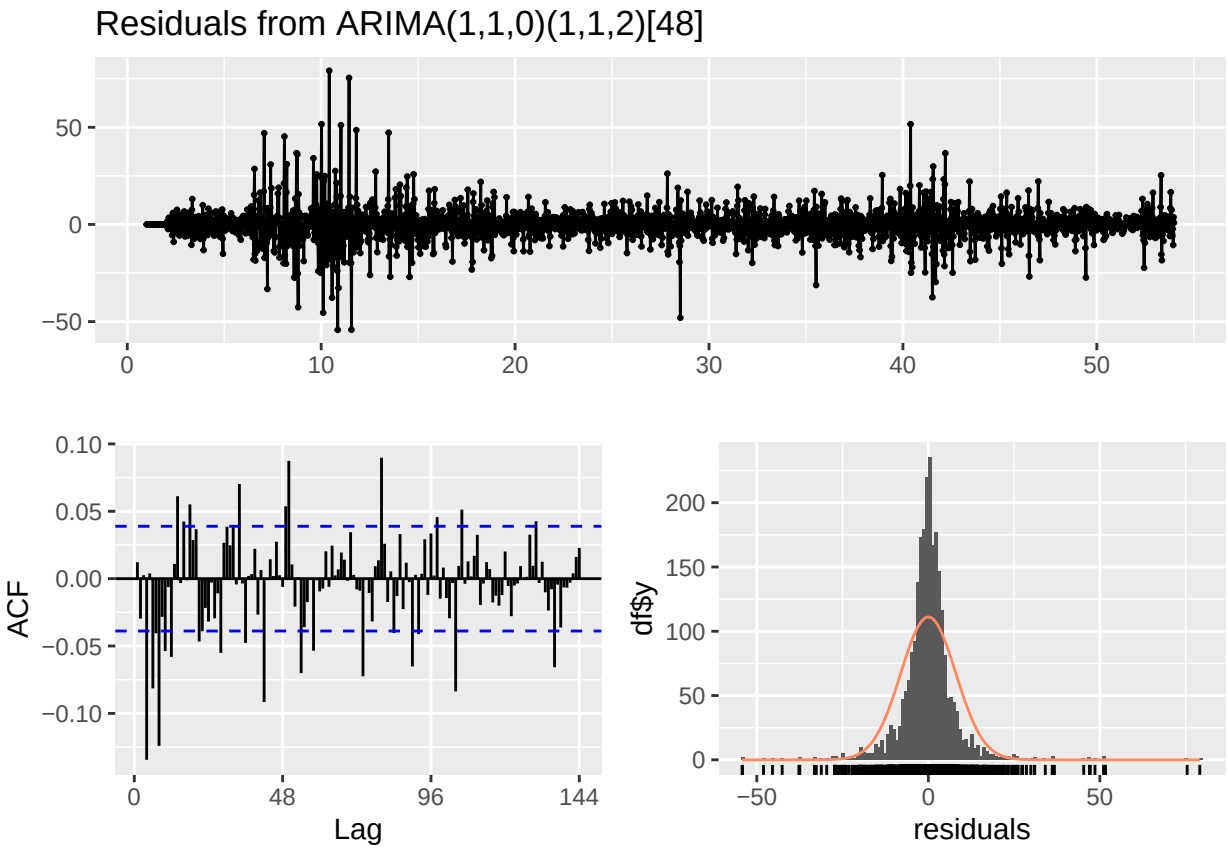
We want to reconfirm this by checking the residual diagnostics of our final models:




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,0,1)(0,0,2)[48] with non-zero mean
## Q* = 293.88, df = 90, p-value < 2.2e-16
##
## Model df: 6.   Total lags used: 96
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,3)(0,0,2)[48]
## Q* = 115.43, df = 89, p-value = 0.03131
##
## Model df: 7.   Total lags used: 96
```

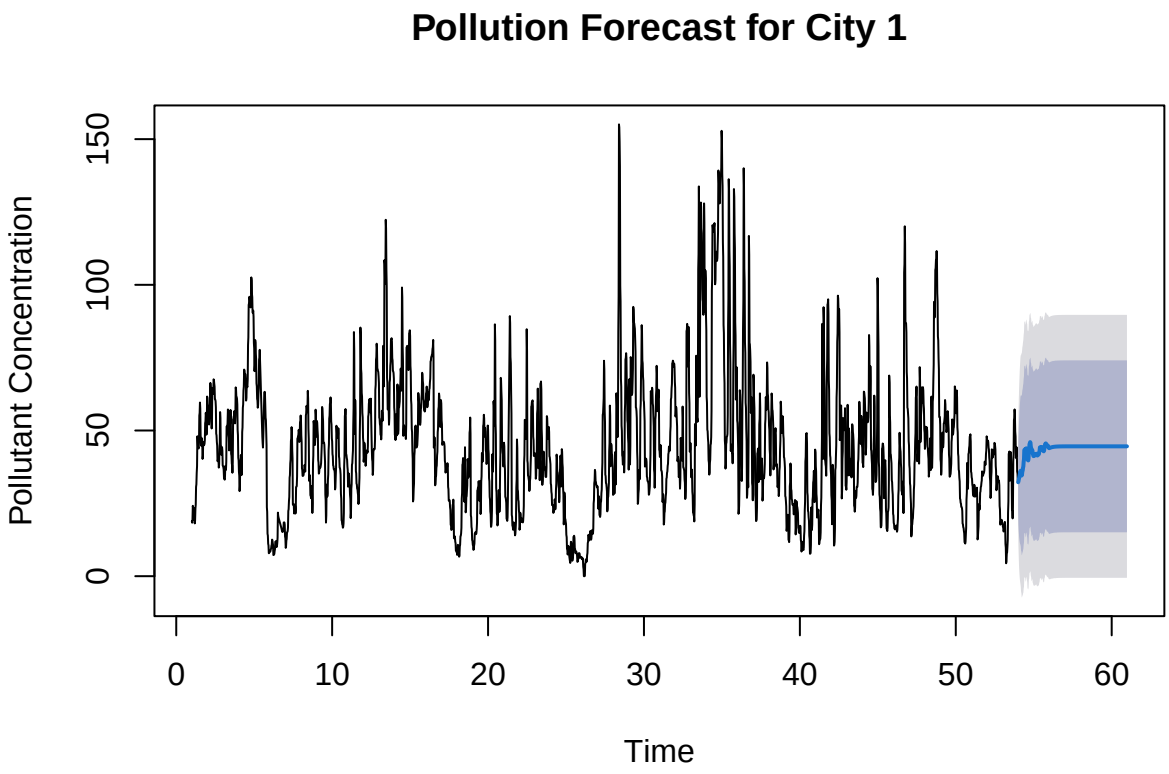


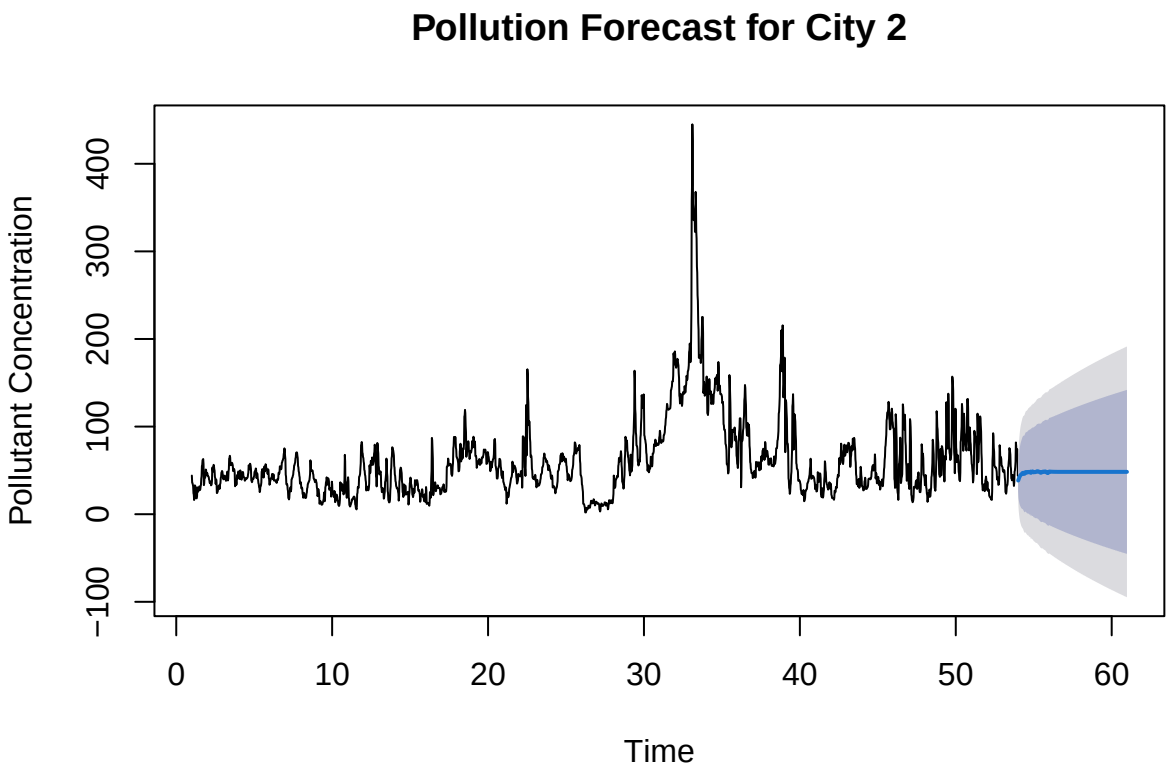

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,0)(1,1,2)[48]
## Q* = 369.47, df = 92, p-value < 2.2e-16
##
## Model df: 4.   Total lags used: 96
```

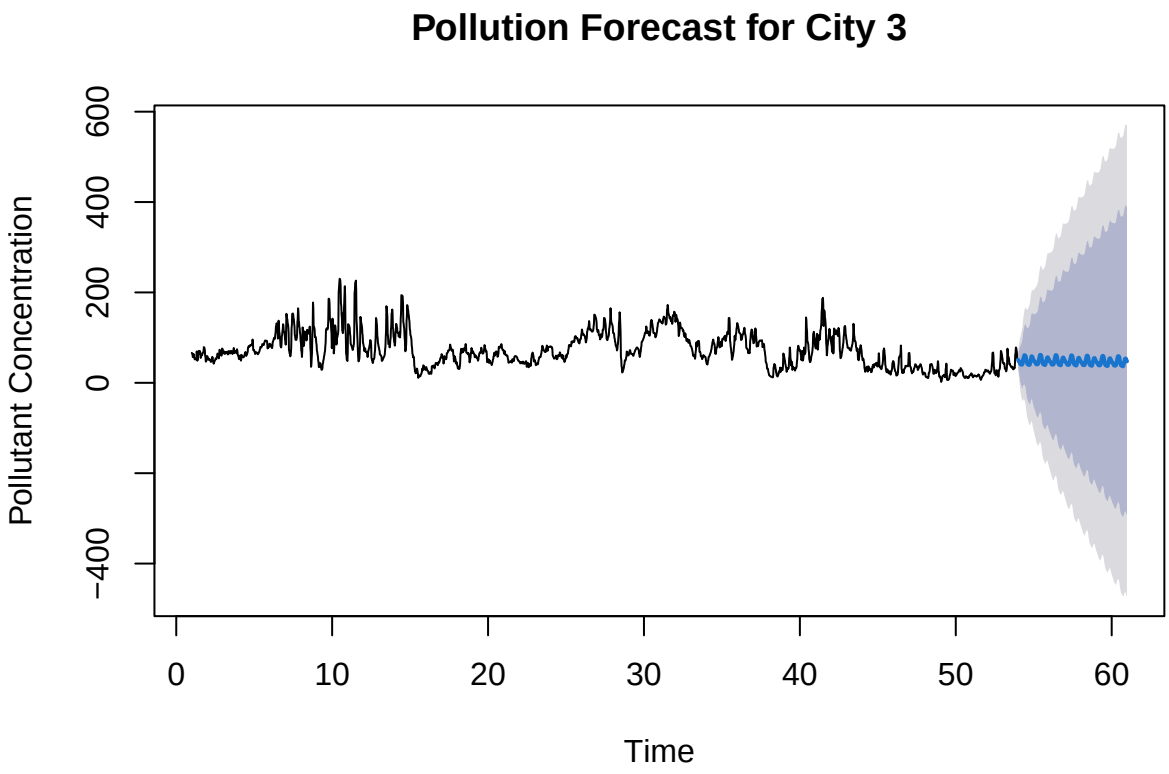
Confirming this with our residual diagnostics:

- The residuals appear random in all plots
- The ACF's only have spikes that exceed the threshold every 48 lags
- The residual bell-curves are all symmetrically normal.

Now we want to forecast the next 336 points


Pollution Forecast for City 1
Pollutant Concentration
0
50
100
150
0
10
20
30
40
50
60
Time


Pollution Forecast for City 2
Pollutant Concentration
−100
0
100
200
300
400
0
10
20
30
40
50
60
Time



For all the forecasts, the dark blue portion refers to the 80% prediction interval and the light grey portion refers to the 95% prediction interval.

For City 1, the forecasts look conservative compared to our data, and then it flattens out eventually. This is indicated by our prediction intervals varying somewhat and then flattening out as well. It appears that the best forecasts in the long run is the mean of the series.

For City 2, the forecasts appear to be varying a little bit around the mean, but the prediction intervals slowly get bigger over time, so our forecasts become more uncertain.

For City 3, the forecasts are varying much more sporadically in a sinusodial pattern with much larger prediction intervals that grow quicker over time.

Therefore, we complete the forecasting report. Thank you for reading!