SMOG SIGNALS: PREDICTING PM2.5 LEVELS IN MILAN CITY USING SARIMA MODEL

Mohammad Wasim Ashraf

INTRODUCTION

Air pollution has become one of the biggest environmental concerns in many cities around the world, and Milan is no exception. In recent years, Milan has gained attention for its poor air quality, especially in colder months. High levels of PM2.5 and other pollutants pose serious health risks to the people living in the city. Because of this, it is important to study and understand the behaviour of air pollution in Milan over time.

In this project, we are analysing the average PM2.5 levels in Milan to identify patterns, trends, and seasonal behaviour using time series techniques. The dataset used in this project contains daily air quality and weather data for the city of Milan. The data was collected from Open-Meteo APIs, based on validated reanalysis models from European environmental and meteorological institutions. For the purpose of this analysis, we focused on a time frame from January 1, 2020 to December 31, 2023, covering a total of 1,461 daily observations. The dataset consists of 51 columns, including pollutants like PM2.5, PM10, nitrogen dioxide, sulphur dioxide, ozone, as well as meteorological variables such as temperature, humidity, radiation, wind speed, precipitation, and others. From this dataset, the average daily PM2.5 levels were extracted and later aggregated to monthly values to perform time series analysis. The data is validated and checked for accuracy, consistency, and completeness. For this analysis, we have focused specifically on monthly average PM2.5, which is known to be one of the most harmful pollutants to human health.

To conduct this analysis, we first converted the daily data into monthly format and plotted the time series. We then examined five key features in the data: trend, seasonality, change in variance, change point, and behaviour. Based on these observations, we decided to use the SARIMA model, which is suitable for handling both seasonal and non-seasonal components.

We follow a residual approach to build and validate the SARIMA model. At each step, we fit a model and then analyze the residuals to decide the next course of action. We use tools like seasonal ACF and PACF plots, ADF test, EACF table, and BIC values to select appropriate model parameters (p, d, q) and (P, D, Q). Once we have a set of possible models, we compare them using AIC, BIC, and error measures like RMSE, MAE, and MAPE to choose the best fitting model. Diagnostic checks are also performed to ensure that the residuals are random and the model is valid.

By doing this study, we aim to contribute towards a better understanding of air pollution patterns in Milan in the next 10 months. This can help raise awareness and support evidence-based decisions to improve the air quality in the city in the future.

```
rm(list=ls())
library(TSA)
library(tseries)
library(forecast)
library(lmtest)
library(stats)
library(FinTS)
```

```
# Helper function ------------------------------------------------------------

helper <- function(class = c("acf", "pacf"), ...) {

  # Capture additional arguments
  params <- match.call(expand.dots = TRUE)
  params <- as.list(params)[-1]

  # Calculate ACF/PACF values
  if (class == "acf") {
    acf_values <- do.call(acf, c(params, list(plot = FALSE)))
  } else if (class == "pacf") {
    acf_values <- do.call(pacf, c(params, list(plot = FALSE)))
  }

  # Extract values and lags
  acf_data <- data.frame(
    Lag = as.numeric(acf_values$lag),
    ACF = as.numeric(acf_values$acf)
  )

  # Identify seasonal lags to be highlighted
  seasonal_lags <- acf_data$Lag %% 1 == 0

  # Plot ACF/PACF values
  if (class == "acf") {
    do.call(acf, c(params, list(plot = TRUE)))
  } else if (class == "pacf") {
    do.call(pacf, c(params, list(plot = TRUE)))
  }

  # Add colored segments for seasonal lags
  for (i in which(seasonal_lags)) {
    segments(x0 = acf_data$Lag[i], y0 = 0, x1 = acf_data$Lag[i], y1 = acf_data$ACF[i], col = "re
d")
  }
}


# seasonal_acf ---------------------------------------------------------------

seasonal_acf <- function(...) {
  helper(class = "acf", ...)
}


# seasonal_pacf --------------------------------------------------------------

seasonal_pacf <- function(...) {
  helper(class = "pacf", ...)
}
```

```
# Data for Testing ----------------------------------------------------------

# Generate time series data
set.seed(123)
n <- 120 # Number of observations
time <- seq(1, n) # Time index
seasonal_pattern <- sin(2 * pi * time / 12) # Seasonal pattern with a period of 12 months
trend <- 0.1 * time # Linear trend
noise <- rnorm(n, mean = 0, sd = 0.5) # Random noise

# Create a time series object
ts_data <- ts(seasonal_pattern + trend + noise, frequency = 12, start = c(2024, 1))



# seasonal_acf Testing ------------------------------------------------------------

# Minimal arguments
acf(ts_data)
```

```
seasonal_acf(ts_data)
```

```
# Extended arguments
acf(ts_data, lag.max=24, main = "ACF of the Time Series Data", demean = TRUE)
```

```
seasonal_acf(ts_data, lag.max=24, main = "ACF of the Time Series Data", demean = TRUE)
```

```
# seasonal_pacf Testing ------------------------------------------------------------

# Minimal arguments
pacf(ts_data)
```

```
seasonal_pacf(ts_data)
```

```
# Extended arguments
pacf(ts_data, lag.max=120, main = "PACF of the Time Series Data")
```

```
seasonal_pacf(ts_data, lag.max=120, main = "PACF of the Time Series Data")
```

DESCRIPTIVE ANALYSIS

The data in analysis for this report measures the air quality of Milan city in Italy. The dataset contains various air quality measures for example amount of SOx, NOx and other parameters present in the air. The feature of our interest for this analysis is the average PM 2.5 levels in the air. We can rightfully say that, in this analysis our target feature will be average PM2.5 level.

The dataset captures daily average PM 2.5 levels in air in Milan, Italy. For the sake of analysis, we have aggregated the daily data into the monthly data.

The R codes used to convert our daily data into monthly data. First, we extracted the month and year from the “date” column in the dataset. After extracting the month and year we used aggregate() function to group the data by month and year and by taking the mean of average PM2.5 within each group. After aggregation, we sorted the aggregated monthly data in a chronological order.

After, converting the daily data to monthly data, we move forward to plot the time series plot for the monthly data, to achieve this, we will convert the data into a “ts” object with frequency as 12 (since, we converted the data into monthly).

```
AQ <- read.csv("weatheraqDataset.csv", header=TRUE)
AQ$Month <- format(as.Date(AQ$date), "%m")
AQ$Year <- format(as.Date(AQ$date), "%Y")

monthlyPM25 <- aggregate(avg_pm2_5 ~ Month + Year, AQ, mean)

monthlyPM25 <- monthlyPM25[order(monthlyPM25$Year, monthlyPM25$Month), ]
summary(monthlyPM25)
```

```
##     Month               Year             avg_pm2_5
##  Length:48          Length:48          Min.   : 9.121
##  Class :character   Class :character   1st Qu.:12.484
##  Mode  :character   Mode  :character   Median :19.825
##                                        Mean   :22.729
##                                        3rd Qu.:30.176
##                                        Max.   :52.486
```

```
monthlyPM25TS <- ts(monthlyPM25$avg_pm2_5,
                    start = c(2020, 1),
                    frequency = 12)
```

TIME SERIES PLOT

The time series plot of the monthly average PM2.5 level. Before moving forward, lets discuss the 5 bullet points.

1. Trend If we look closely, we observe that there is a linear downward trend in our series.
2. Seasonality Seasonality means repeating pattern in the time series plot. We observe seasonality in our time series plot. Every year in winter the PM2.5 level increases. During winters we see a peak in PM2.5 level every year. Thus, we conclude that we have seasonality in our time series plot.
3. Change in Variance We observe that there is change in variance from winters in 2020 to summer in 2021. The variance changes again from winter 2021 to summer 2022. We see a spike in winter 2022 in PM2.5 and it reduces in summer 2024. If we just look at the seasons i.e., winters in our time series plot, we can see change in variance there as well. There is an increase in variance from winters in 2020 to winters in 2021. The variance changes again from winters in 2021 to winters in 2022. The same pattern is seen from winters in 2022 to winters in 2023.

4. Change point/Intervention We can see a change point in the winters of 2021. There is a sudden increase in the PM2.5 levels in the city. This may be because of some industrial action took place in the city or some hazardous gas leakage happened around that time that increased the PM2.5 level in air.

5. Behaviour In our time series plot as shown in figure – 3, we mostly have succeeding time points. There are few fluctuations that we observe in our time series plot. Succeeding time points indicate Autoregressive behaviour and fluctuations indicate moving average behaviour. Therefore, the time series plot has a mix of both Autoregressive and moving average behaviour.

Since, we have clear seasonality in the series, SARIMA model, will be the best model to fit in our time series plot.

To fit a SARIMA model in our time series, we will use residual approach. We will analyse residuals after fitting the model.

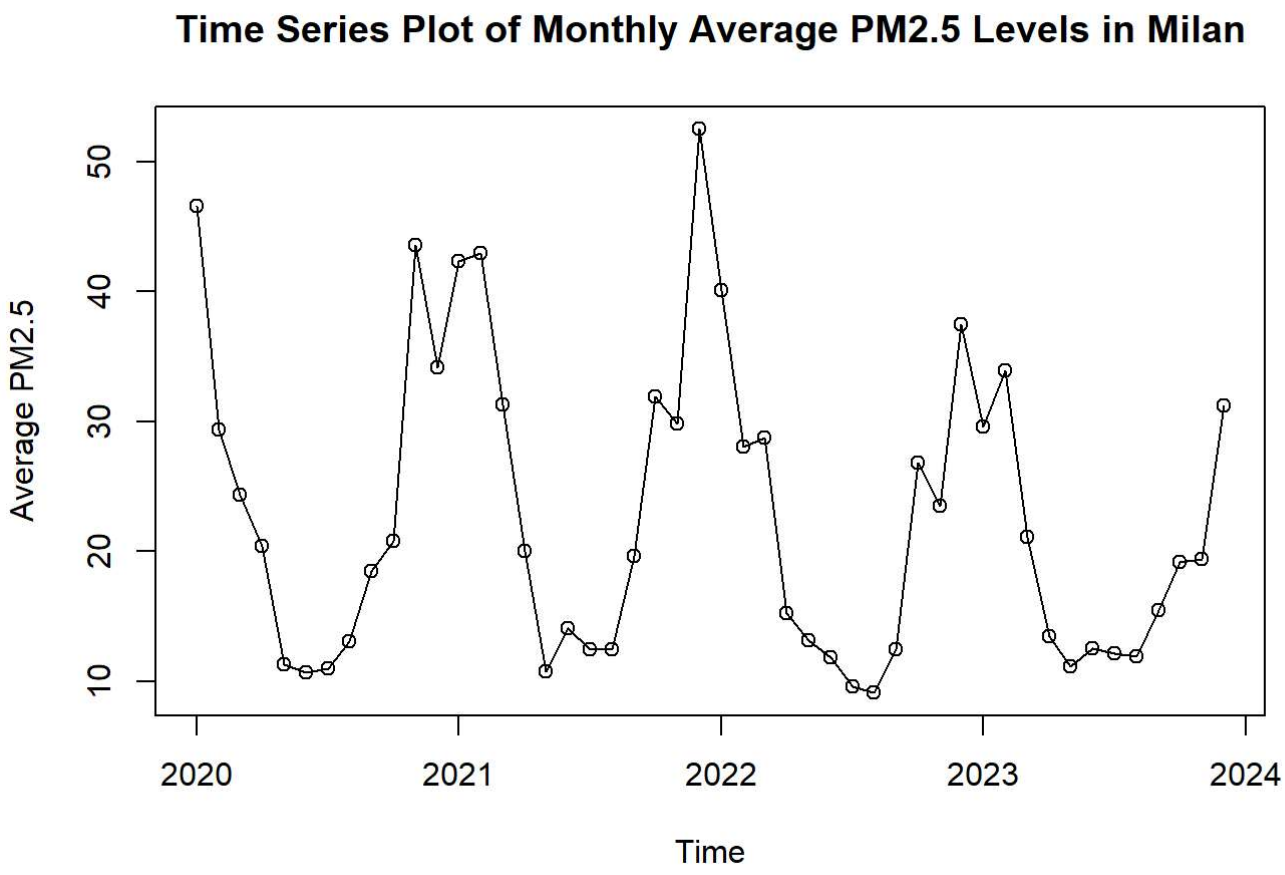
SARIMA model is defined as SARIMA (p, d, q) x (P, D, Q)s

Where,

p, d, q are the ordinary orders P, D, Q are the seasonal orders s is the period or frequency of the series

In SARIMA modelling we will plot seasonal ACF and seasonal PACF. Lets have a look at seasonal ACF and seasonal PACF.

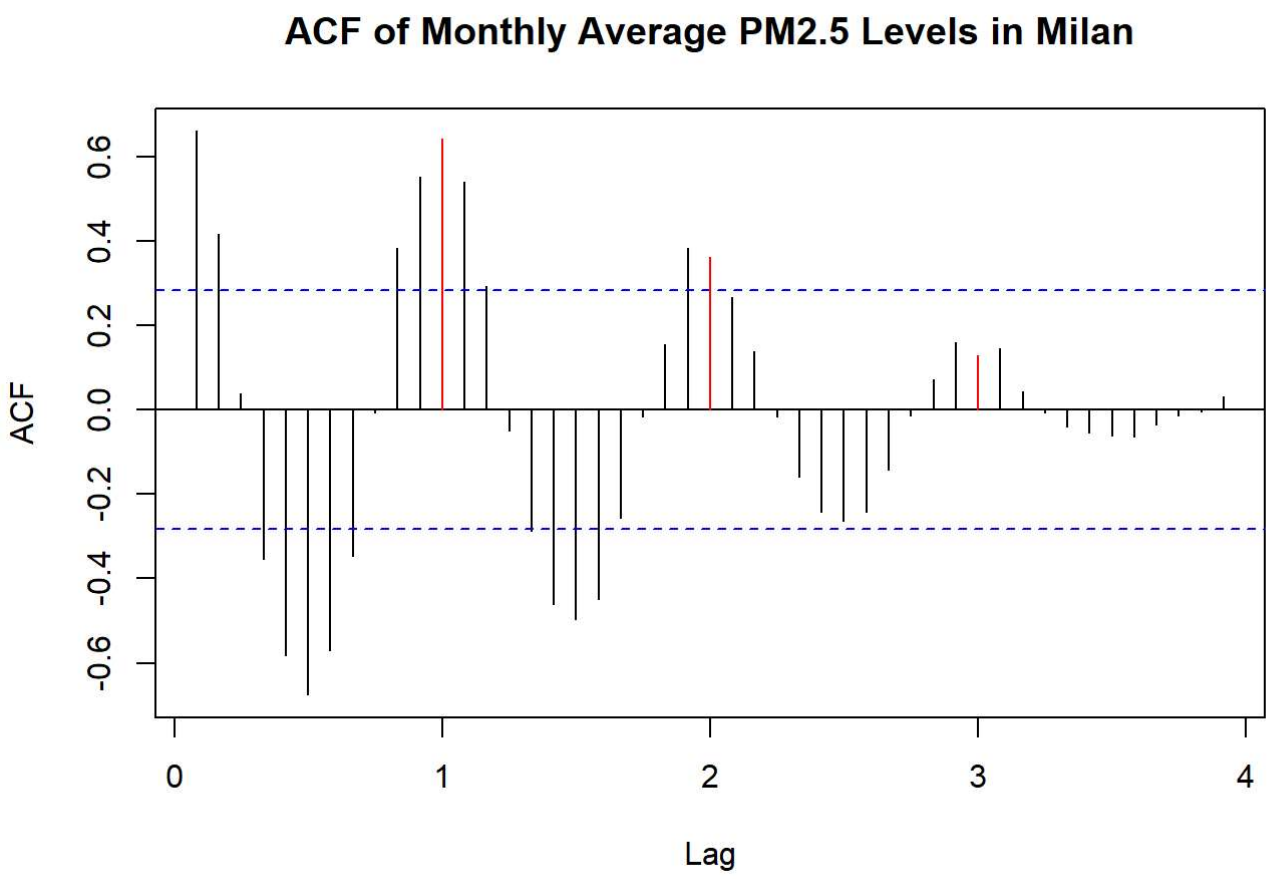
```
plot(monthlyPM25TS,
     ylab = "Average PM2.5",
     xlab = "Time",
     type = "o",
     main = "Time Series Plot of Monthly Average PM2.5 Levels in Milan")
```



SEASONAL ACF

The red bars in the ACF plot denotes the seasons. The red bars makes easier to look at what is going on at the seasons. In ACF plot, we see a wave pattern. If we just look at the red bars (seasons: where the pattern repeats), we see a slowly decaying pattern, red bars are getting smaller and smaller slowly that means seasons are getting smaller and smaller slowly. This slowly decaying pattern shows non-stationarity of the series and we have seasonal trend in the series.

```
seasonal_acf(monthlyPM25TS, lag.max = 100, main = "ACF of Monthly Average PM2.5 Levels in Mila
n")
```

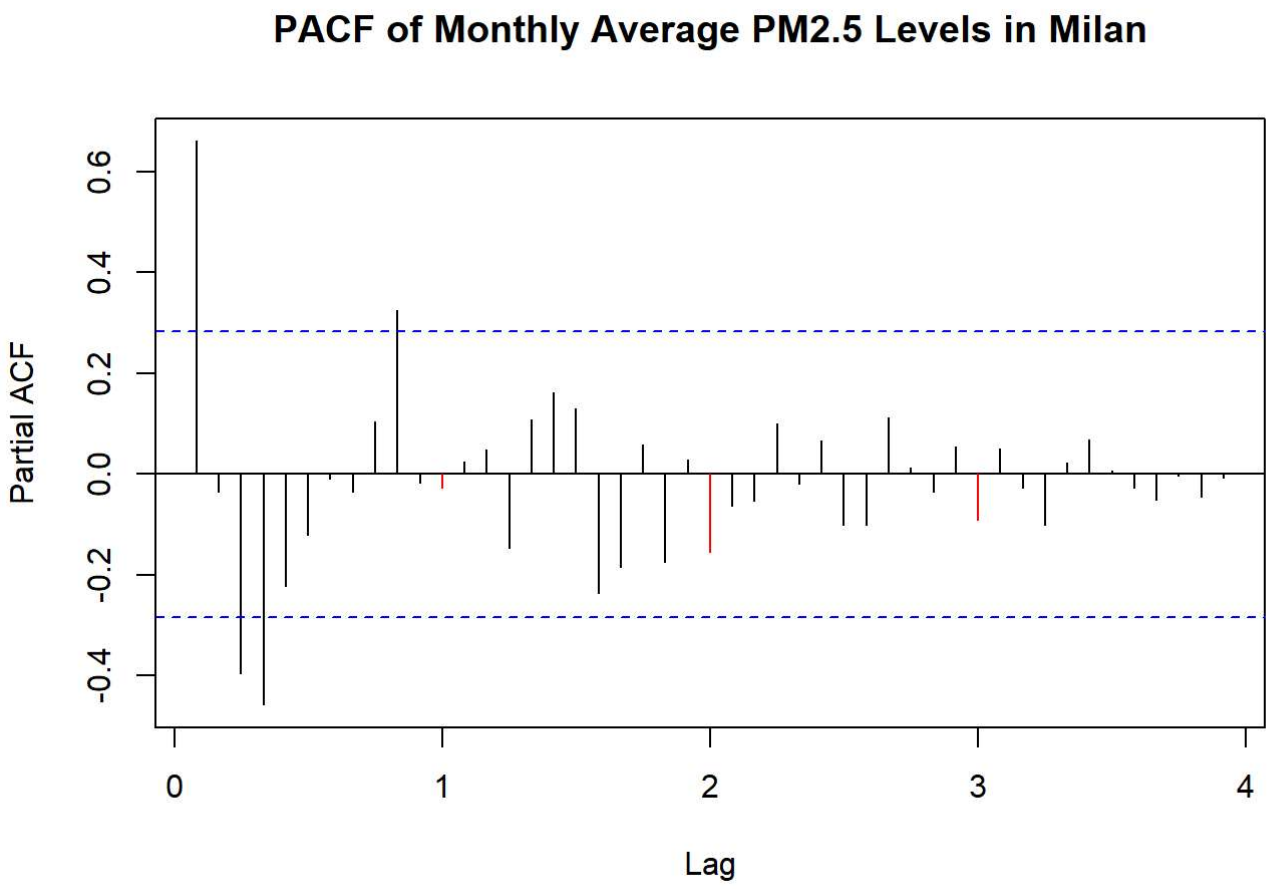


SEASONAL PACF

We observe high first lag, which is around 0.63. High first lag means the series is non-stationary.

Seasonal ACF and PACF plot confirms that our series is non-stationary, we will need some tools like differencing to make the series stationary.

```
seasonal_pacf(monthlyPM25TS, lag.max = 100, main = "PACF of Monthly Average PM2.5 Levels in Mila
n")
```



RESIDUAL APPROACH

In residual approach, we fit a model and then look at its residuals. We observe the left-overs. We repeat this procedure and deal with one characterstic at a time.

As stated earlier SARIMA model is defined as SARIMA (p, d, q) x (P, D, Q)s.

Where,

p, d, q are the ordinary orders P, D, Q are the seasonal orders s is the period or frequency of the series (in our case, s = 12)

In residual approach, we will first look for the values of seasonal orders. Therefore, first step will be to find the seasonal orders and then move forward to find the ordinary orders. In seasonal orders, we will start with finding the value of “D” (Seasonal Differencing), because In ACF plot, we observed a slowly decaying pattern of the seasons, that means, we have seasonal trend.

We will set D = 1 (first seasonal differencing) and set every other order as zero and observe what happens in the residuals.

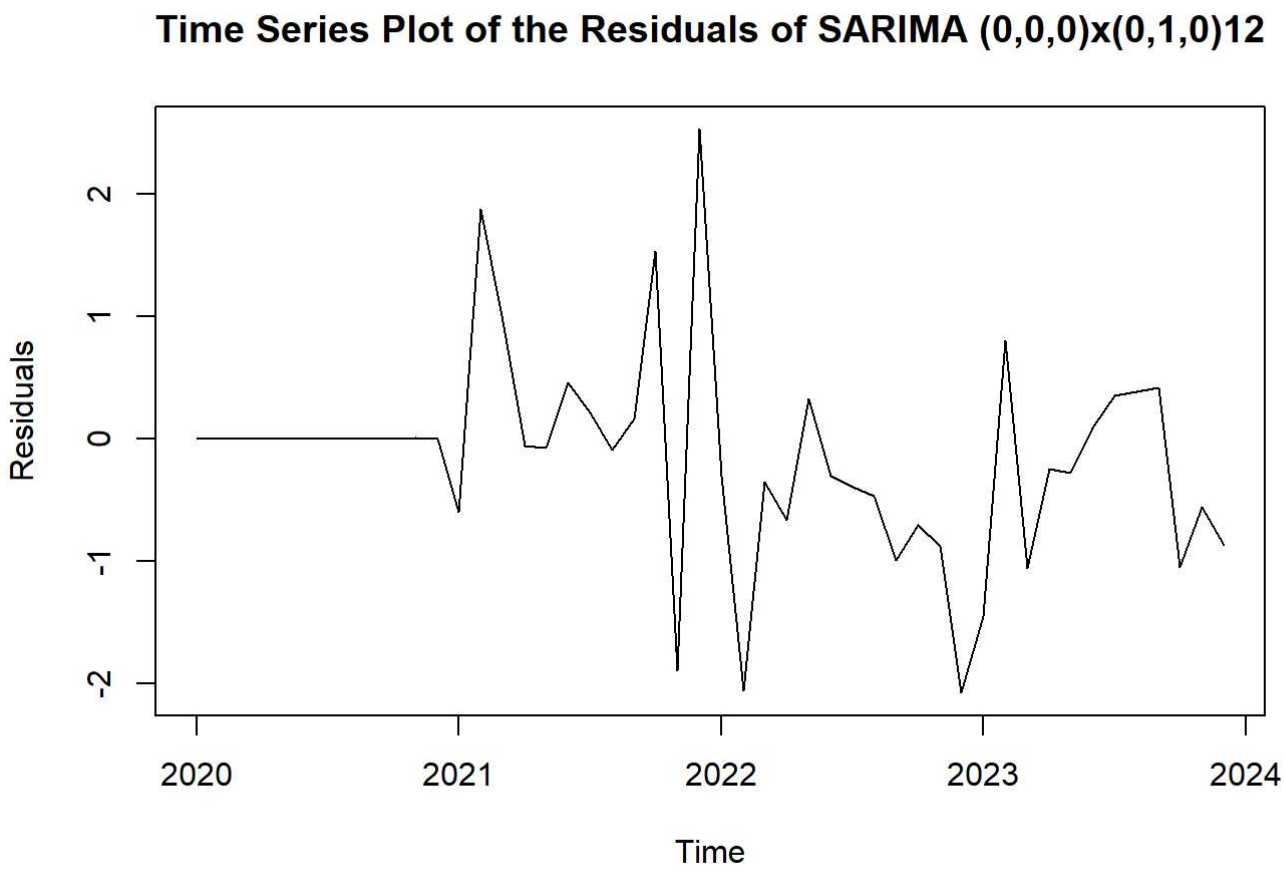
Therefore, the model will be SARIMA (0,0,0)x(0,1,0)12

Now, let’s fit this model and check the residuals.

The time series plot of residuals of SARIMA (0,0,0)x(0,1,0)12. With first seasonal differencing, we lost the first seasonal data points. it goes flat. The heavy repeating patterns somewhat disappear, that means, first seasonal differencing (D = 1) captures seasonal trend.

1. Trend The residual plot does not exhibit any clear upward or downward trend. The early part of the series appears flat due to the loss of the first seasonal data point following seasonal differencing (D = 1). This absence of trend is because the model has removed the linear component from the original series.
2. Seasonality The strong seasonal peaks are seen in the original series have been substantially reduced. The seasonal differencing step (D = 1) has successfully accounted for the repeating seasonal structure in the original data. The residuals now appear more irregular and do not follow an annual seasonal pattern as before, but the seasonality is still noticeable.
3. Change in variance There is a little bit of change in variance across the time period. In particular, sharp spikes occur between late 2021 and early 2022, it is a localized increases in residual volatility. After this point, the residual variance becomes somewhat more stable but still fluctuates, suggesting that some heteroskedasticity may remain in the series.
4. Change point/Intervention There is no clear change point or structural break visible in the residual series. While some spikes in residuals are observed around early 2022, they appear to be random and not indicative of a sustained shift in the series level or variance. This suggests that the SARIMA(0,0,0)×(0,1,0)[12] model has not missed any major interventions or abrupt changes in the underlying process.
5. Behaviour While the model has eliminated seasonal components, the residuals still exhibit clusters of high and low values, implying autocorrelation remains. This points to the presence of underlying autoregressive or moving average behaviour not yet addressed by the current model.

```
#Residual Approach SARIMA (0,0,0)x(0,1,0)12

m1.AQ = Arima(monthlyPM25TS, order = c(0,0,0), seasonal = list(order = c(0,1,0),
                                                               period = 12))
res.m1 = rstandard(m1.AQ)
plot(res.m1, xlab = "Time", ylab = "Residuals", main = "Time Series Plot of the Residuals of SAR
IMA (0,0,0)x(0,1,0)12")
```

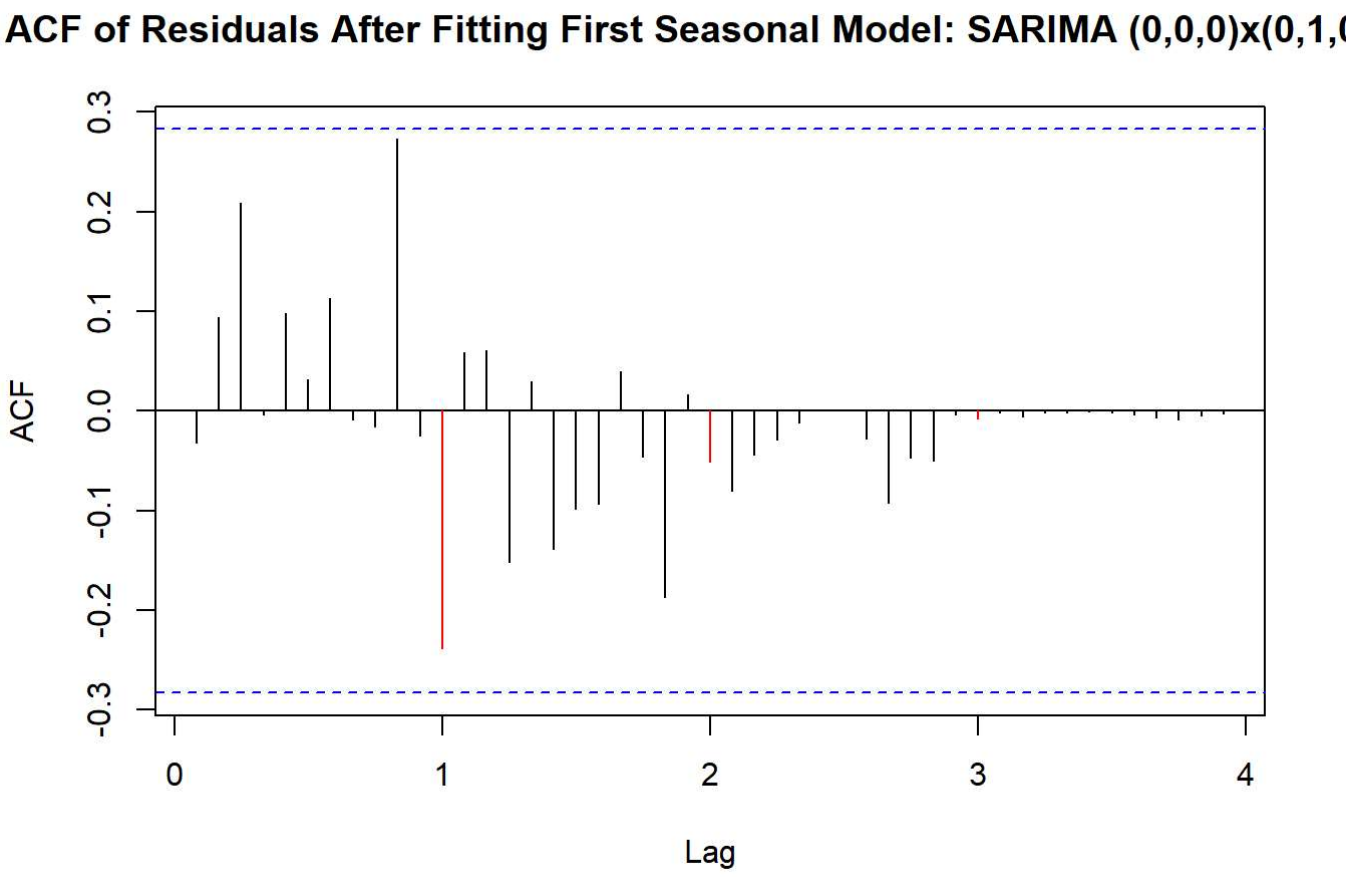


SEASONAL ACF PLOT OF RESIDUALS FOR SARIMA (0,0,0)x(0,1,0)12

We observe that we don't have the wave pattern in our ACF plot of residuals. Therefore, D = 1 captures the seasonal trend. Now, what left is a bunch of significant autocorrelations. No red bar is crossing the blue line (confidence interval), therefore, no seasons are significant.

From this seasonal ACF plot of SARIMA (0,0,0)x(0,1,0)12, we can evaluate "Q". Since there are no significant seasons, therefore, Q = 0.

```
seasonal_acf(res.m1, lag.max = 100, main = "ACF of Residuals After Fitting First Seasonal Model:
SARIMA (0,0,0)x(0,1,0)12")
```



SEASONAL PACF PLOT OF RESIDUALS OF SARIMA (0,0,0)x(0,1,0)12

In PACF plot of residuals, we have the first season as significant as it crosses the confidence interval. The rest of the seasons are insignificant.

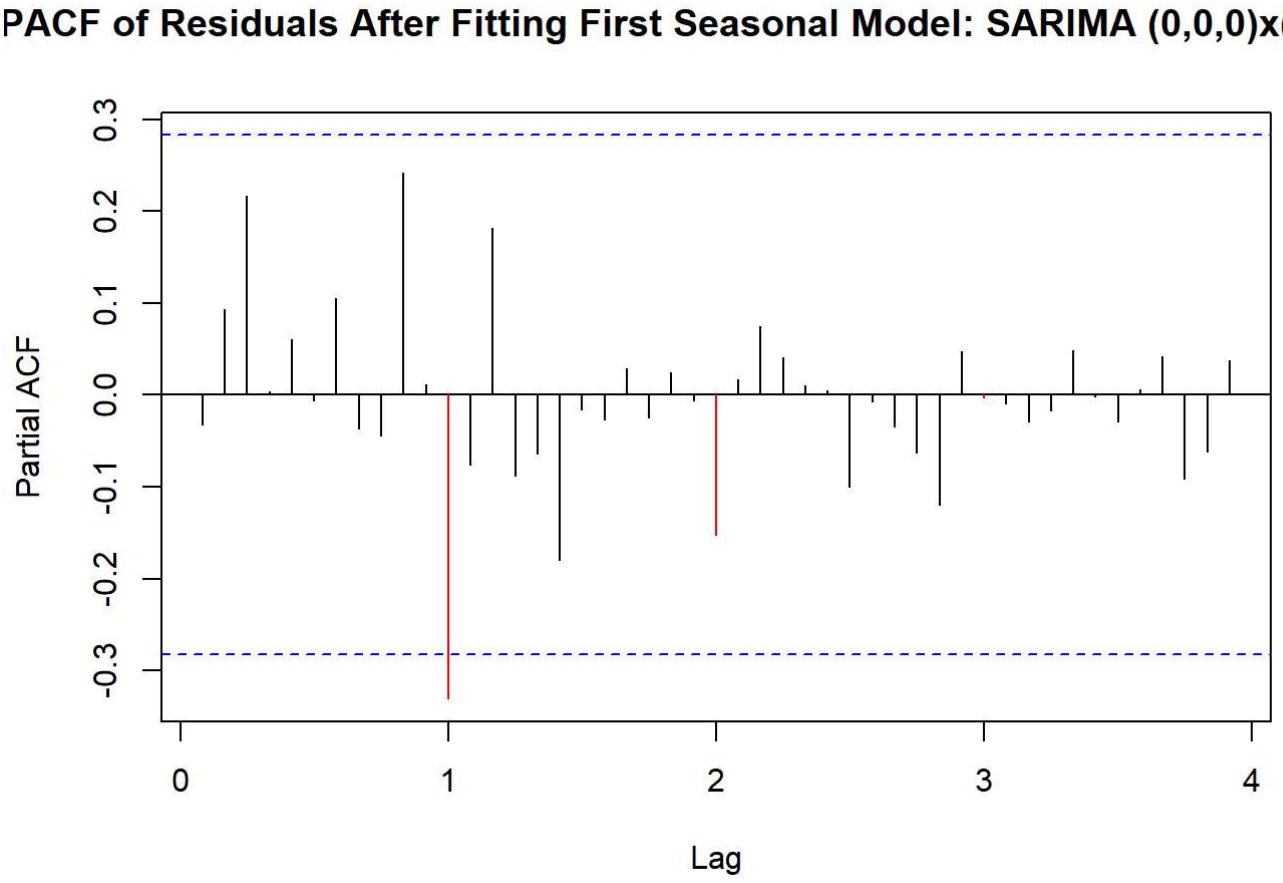
From this seasonal PACF plot of SARIMA (0,0,0)x(0,1,0)12, we can evaluate “P”. Since there is one significant season (first season), therefore, P = 1.

From Seasonal ACF and PACF, we have the values of the seasonal orders P and Q as 1 and 0 respectively. Now we set P = 1 and Q = 0 and fit the model.

Therefore, the model will be SARIMA(0,0,0)x(1,1,0).

Now, lets fit this model and check the residuals.

```
seasonal_pacf(res.m1, lag.max = 100, main = "PACF of Residuals After Fitting First Seasonal Mode
l: SARIMA (0,0,0)x(0,1,0)12")
```



FITTING AND CHECKING THE RESIDUALS OF SARIMA(0,0,0)x(1,1,0)12

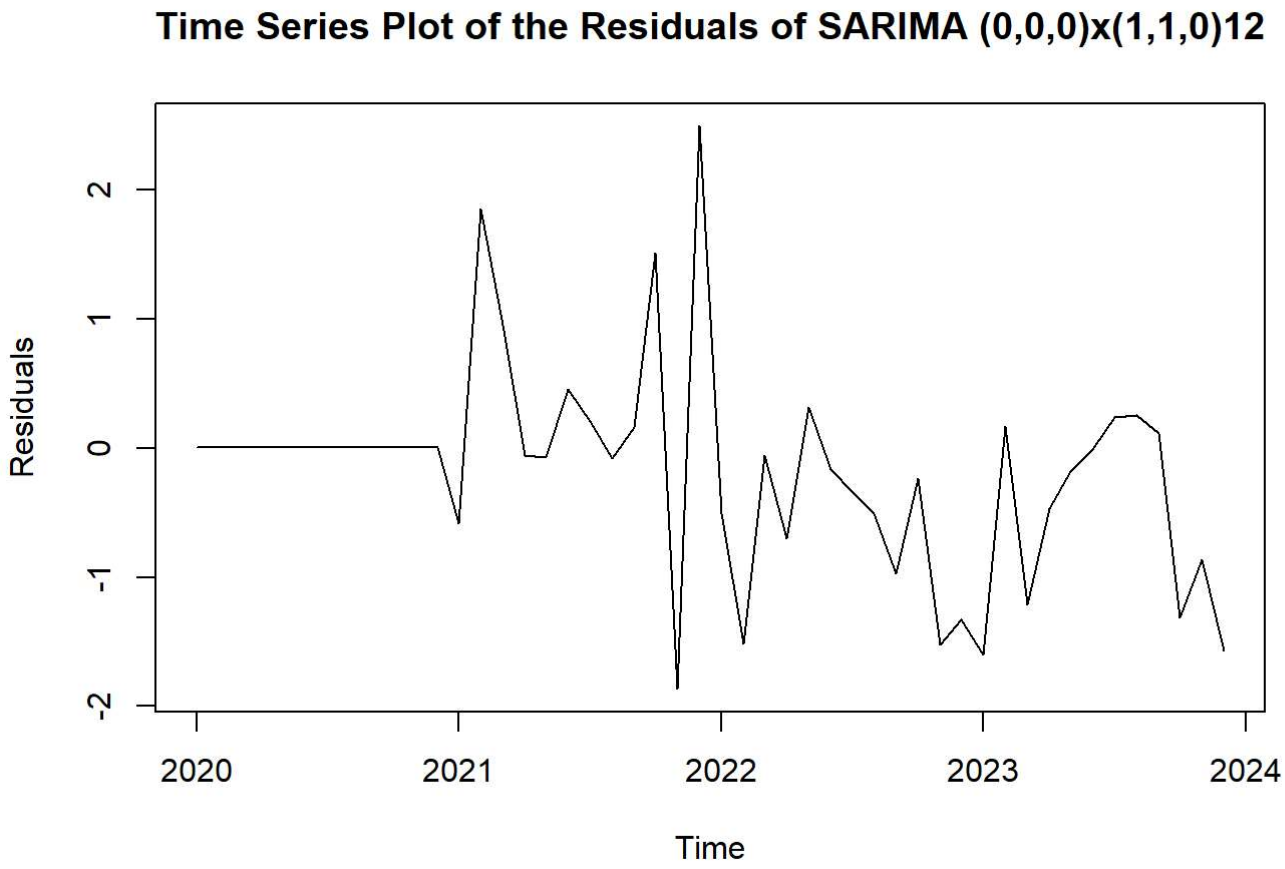
Lets now compare this plot with the time series plot of residuals of SARIMA (0,0,0)x(0,1,0)12 model. Both the residual plots are same, which was expected as we have only one change and that is P = 1 for SARIMA (0,0,0)x(1,1,0)12. Since, P = 1 is not that high enough to capture some of the information.

1. Trend The residual plot does not display any observable trend; the seasonal and non-seasonal trends in the original series have been effectively removed. The flat nature of the initial section and the random fluctuations afterward is a promising picture of the trend component has been well handled.
2. Seasonality The seasonal peaks seen in the original time series have been further reduced compared to the SARIMA(0,0,0)×(0,1,0)[12] model. This improvement confirms that adding P = 1 (seasonal autoregressive order) has enhanced the model's ability to account for repeating seasonal structure, although seasonal influence still exists to some extent.
3. Change in variance This residual plot shows variation in residual magnitude over time, with larger spikes in the middle portion (around early 2022) and slightly reduced variance in later periods. However, the change in variance is not drastic, and the residuals remain within acceptable bounds overall.
4. Change point/Intervention There is no evidence of a structural change or intervention in this residual series. Although there is a pronounced spike in early 2022, it appears to be an isolated fluctuation rather than a sustained shift in level or variance. The residuals return to typical behaviour thereafter we have no change point present.
5. Behaviour The residuals still display some degree of autocorrelation, particularly through clusters of rising and falling values. Although the addition of a seasonal autoregressive component (P = 1) has slightly improved model performance, the residuals are not entirely random.

SEASONAL ACF PLOT OF RESIDUALS FOR SARIMA (0,0,0)x(1,1,0)12

In SARIMA (0,0,0)x(1,1,0)12 model. We observe that we have no red bars (seasons) that cross the confidence interval. Therefore, no seasons are significant.

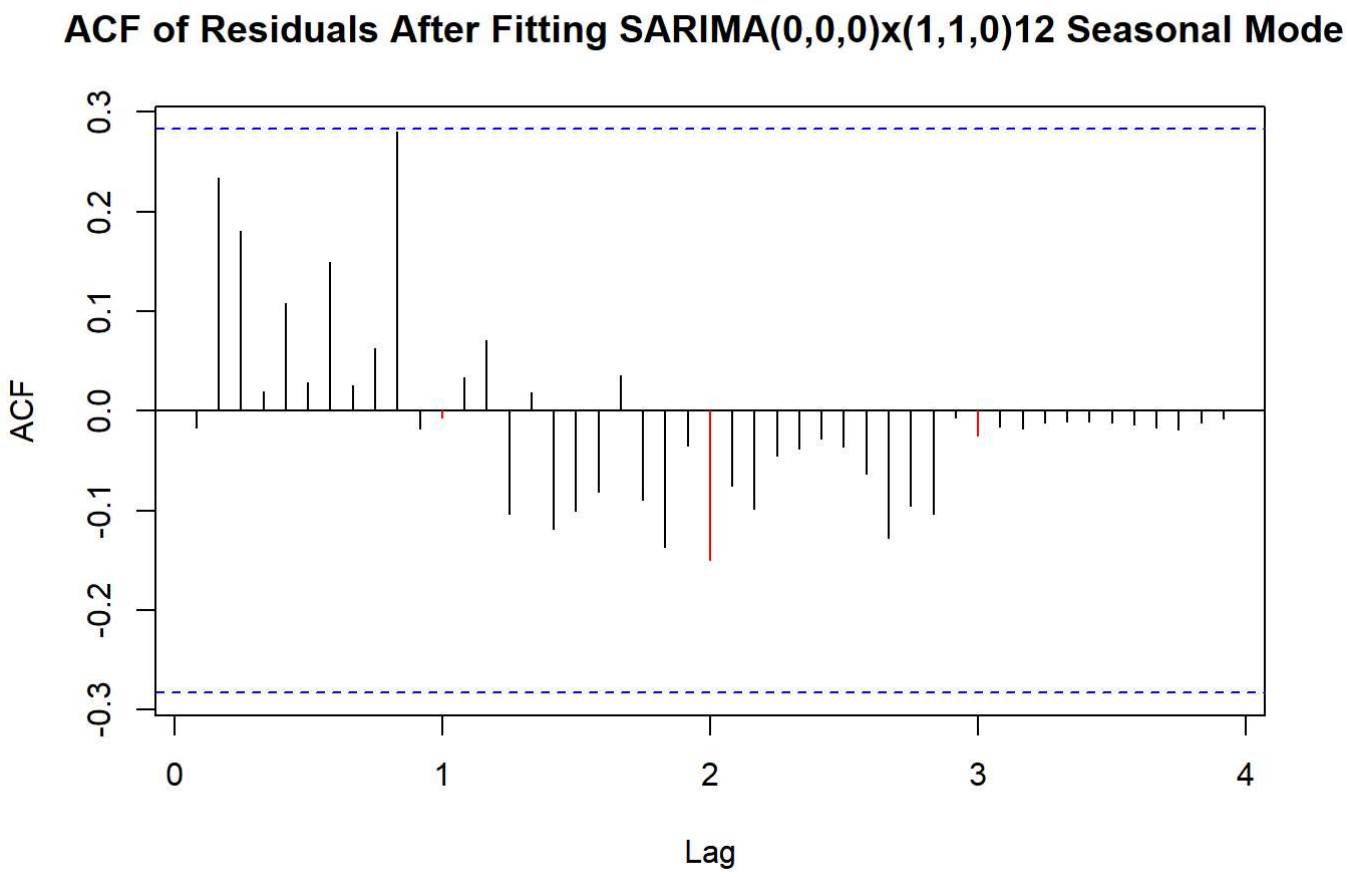
```
#SARIMA(0,0,0)x(1,1,0)12

m2.AQ = Arima(monthlyPM25TS, order = c(0,0,0), seasonal = list(order = c(1,1,0),
                                                               period = 12))
res.m2 = rstandard(m2.AQ)
plot(res.m2, xlab = "Time", ylab = "Residuals", main = "Time Series Plot of the Residuals of SAR
IMA (0,0,0)x(1,1,0)12")
```



SEASONAL ACF PLOT OF RESIDUALS FOR SARIMA (0,0,0)x(1,1,0)12

In SARIMA (0,0,0)x(1,1,0)12 model. We observe that we have no red bars (seasons) that cross the confidence interval. Therefore, no seasons are significant.

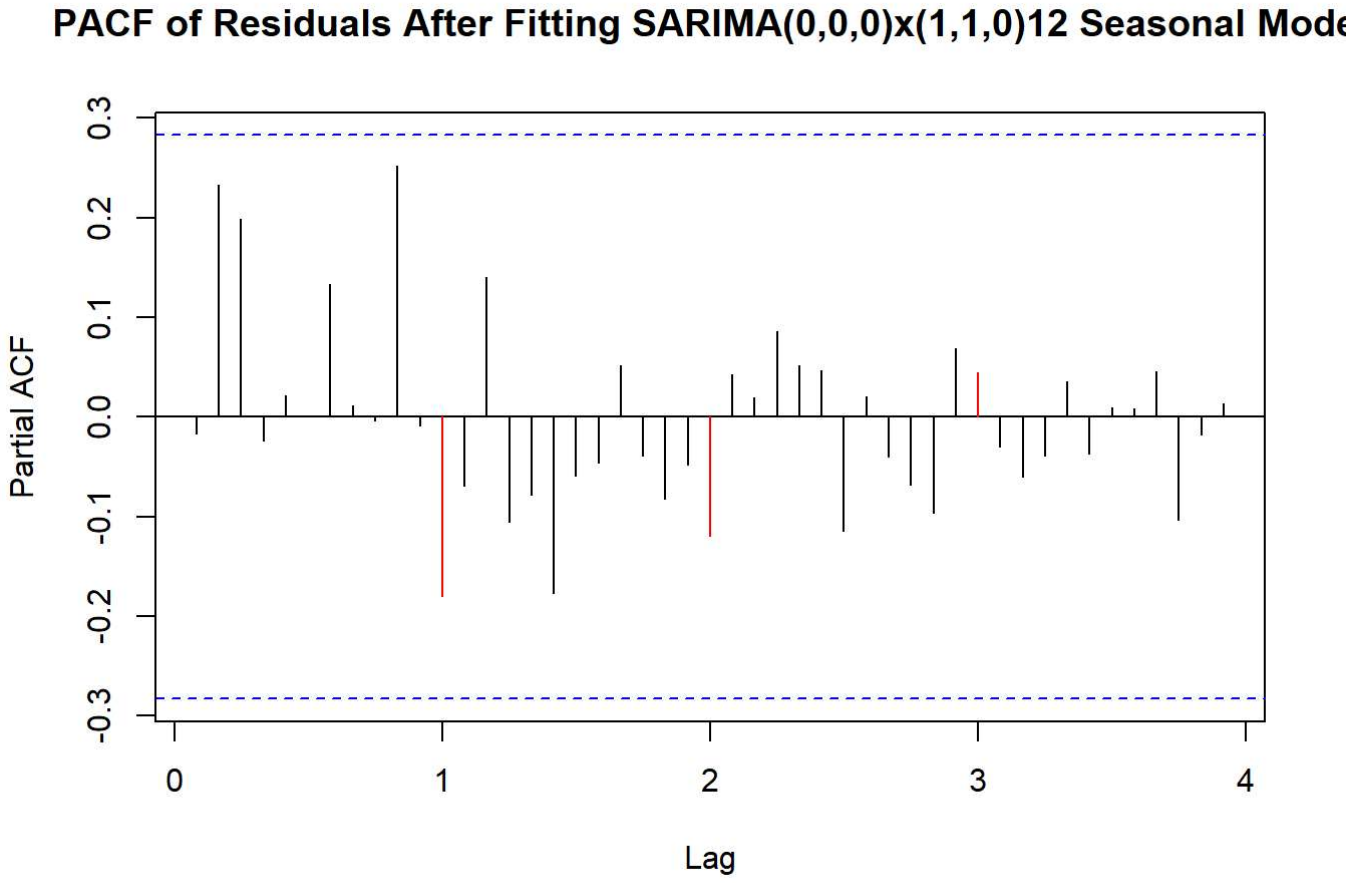
```
seasonal_acf(res.m2, lag.max = 100, main = "ACF of Residuals After Fitting SARIMA(0,0,0)x(1,1,0)
12 Seasonal Model")
```



SEASONAL PACF PLOT OF RESIDUALS FOR SARIMA (0,0,0)x(1,1,0)12

In the seasonal PACF plot of residuals of SARIMA (0,0,0)x(1,1,0) model, We observe that we have no red bars (seasons) that cross the confidence interval. Therefore, no seasons are significant.

```
seasonal_pacf(res.m2, lag.max = 100, main = "PACF of Residuals After Fitting SARIMA(0,0,0)x(1,1,
0)12 Seasonal Model")
```



TEST FOR STATIONARITY

Before moving forward, we need to check whether our series in stationary or not. We will use ADF test to check if our series is stationary. H0 = The process is difference nonstationary (the process is nonstationary but becomes stationary after first differencing). HA = The process is stationary.

the result of ADF test on residuals of SARIMA (0,0,0)x(1,1,0) shows the p-value is 0.2634. since the p-value is greater than the significance level (α = 0.05). Therefore, the residuals are not stationary.

Therefore, we will need to do ordinary differencing as well to make it stationary.

We will repeat the whole procedure again, with d = 1 (first order differencing).

Let's start with fixing d = 1 and D = 1. Model will be SARIMA (0,1,0)x(0,1,0). We will fit the model and check the residuals.

```
adf.test(res.m2)
```

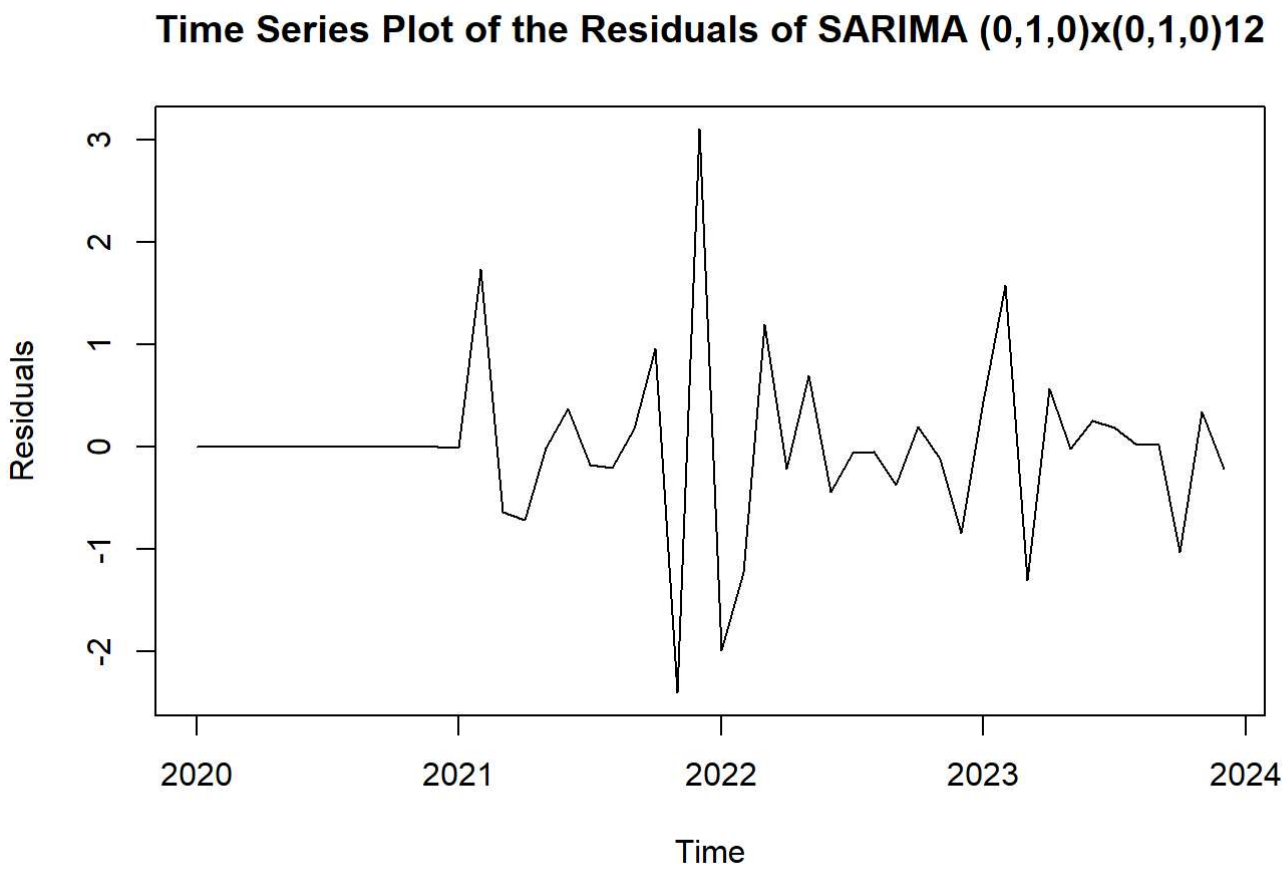
```
##
##  Augmented Dickey-Fuller Test
##
## data:  res.m2
## Dickey-Fuller = -2.778, Lag order = 3, p-value = 0.2634
## alternative hypothesis: stationary
```

FIRST ORDER DIFFERENCING

This chunk captures the time series plot of residuals of SARIMA (0,1,0)x(0,1,0)12. With first order ordinary differencing, the series is now fluctuating at a mean level and there is no trend. The heavy repeationg patterns has also been reduced because of the first order seasonal differencing. This means that first seasonal differencing (D = 1) captures seasonal trend.

1. Trend After applying first order ordinary differencing (d = 1) and seasonal differencing (D = 1), the residuals now fluctuate around a stable mean with no visible trend. Both the long term and seasonal trends in the original PM2.5 series have been effectively removed.
2. Seasonality The plot demonstrates a clear reduction in repeating seasonal patterns. The seasonal differencing (D = 1) has eliminated the seasonal component observed in the original data. The absence of recurring winter peaks in the residuals supports this.
3. Change in Variance The residuals show a little bit variation in magnitude, with slightly larger spikes around early 2022 and early 2023. However, the variance remains stable across time, so no heteroskedasticity. The SARIMA model has reasonably captured the variance structure of the original series.
4. Change Point / Intervention There is no indication of a structural break or change point in the residuals. Although there are short lived spikes, particularly in early 2022, they appear isolated and not part of a sustained shift. The series returns to typical residual behaviour quickly after these fluctuations.
5. Behaviour The residuals display some short-term autocorrelation, as evident in the small clusters of rising or falling values. While the combination of differencing steps has improved the stationarity of the series, no non-seasonal autoregressive or moving average terms have been included yet. As a result, the model may still be underfitted.

```
#Residual Approach SARIMA (0,1,0)x(0,1,0)12

m11.AQ = Arima(monthlyPM25TS, order = c(0,1,0), seasonal = list(order = c(0,1,0),
                                                                period = 12))
res.m11 = rstandard(m11.AQ)
plot(res.m11, xlab = "Time", ylab = "Residuals", main = "Time Series Plot of the Residuals of SA
RIMA (0,1,0)x(0,1,0)12")
```

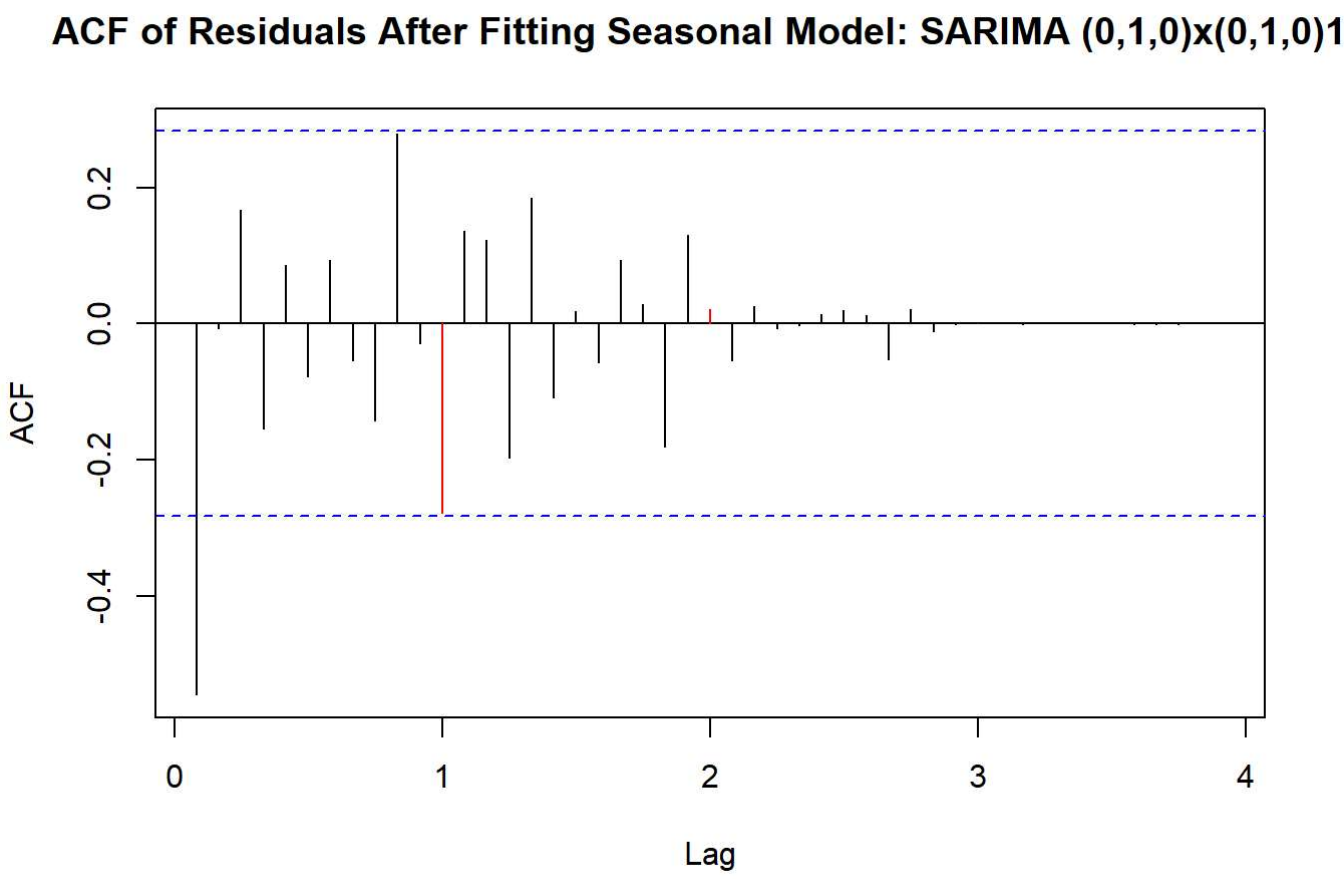


SEASONAL ACF PLOT OF RESIDUALS FOR SARIMA (0,1,0)x(0,1,0)12

This shows the seasonal ACF plot of residuals of SARIMA (0,1,0)x(0,1,0)12. We observe that we don't have the wave pattern in our ACF plot of residuals. Therefore, D = 1 captures the seasonal trend. Now, what left is a bunch of significant autocorrelations. One red bar (first season) is almost touching the confidence interval. We can count it as significant season.

From this seasonal ACF plot of SARIMA (0,1,0)x(0,1,0)12, we can evaluate "Q". Since there is one significant season, therefore, Q = 1.

```
seasonal_acf(res.m11, lag.max = 100, main = "ACF of Residuals After Fitting Seasonal Model: SARI
MA (0,1,0)x(0,1,0)12")
```



SEASONAL PACF PLOT OF RESIDUALS FOR SARIMA (0,1,0)x(0,1,0)12

This shows the seasonal PACF plot of residuals of SARIMA (0,1,0)x(0,1,0)12. In PACF plot of residuals, no seasons crosses the confidence interval. Therefore, from PACF, we get so seasons as significant.

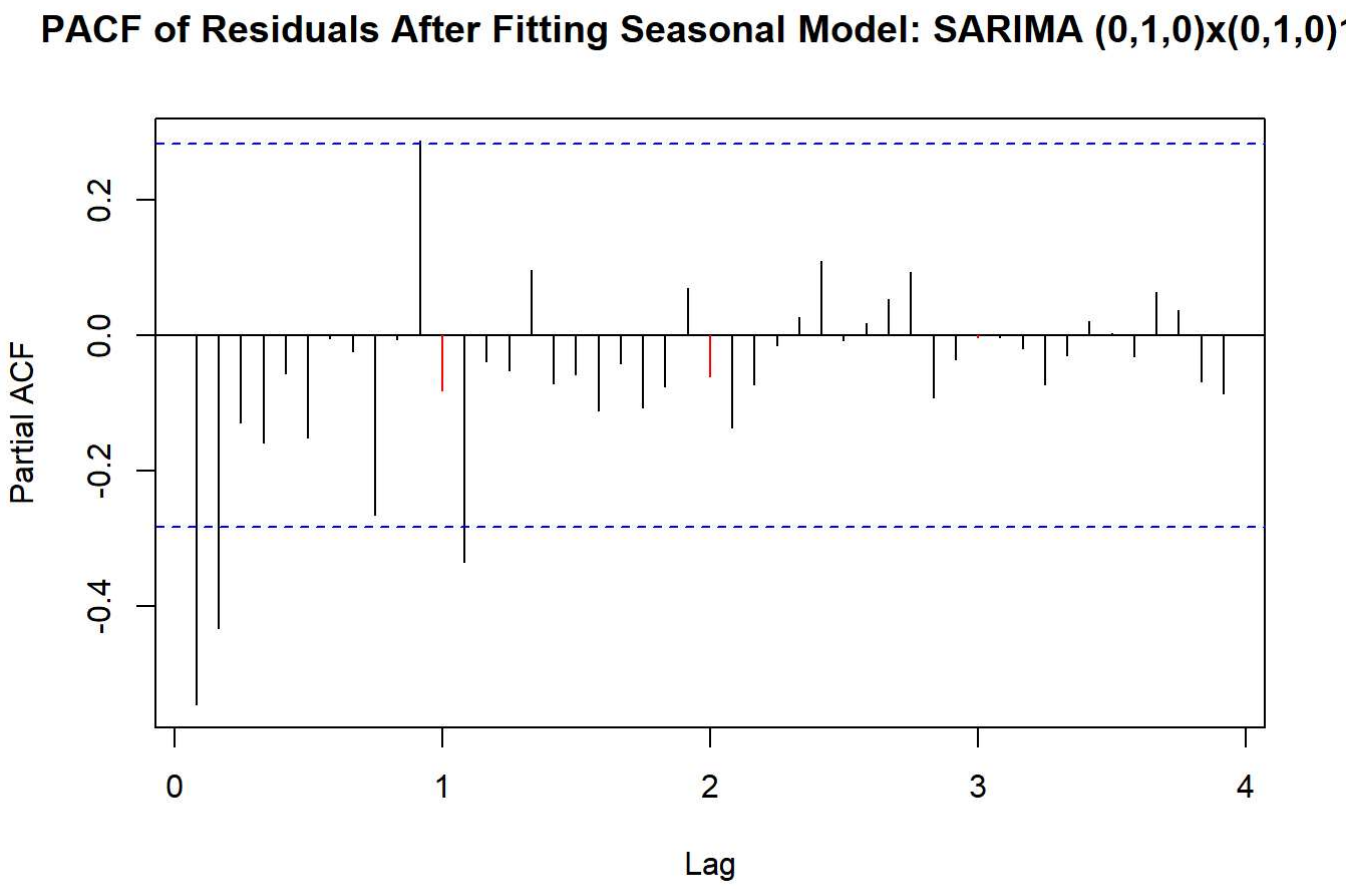
From this seasonal PACF plot of SARIMA (0,1,0)x(0,1,0)12, we can evaluate “P”. Since there are no significant seasons, therefore, P = 0.

From Seasonal ACF and PACF, we have the values of the seasonal orders P and Q as 0 and 1 respectively. Now we set P = 0 and Q = 1 and fit the model.

Therefore, the model will be SARIMA(0,1,0)x(0,1,1).

Now, lets fit this model and check the residuals.

```
seasonal_pacf(res.m11, lag.max = 100, main = "PACF of Residuals After Fitting Seasonal Model: SA
RIMA (0,1,0)x(0,1,0)12")
```



FITTING AND CHECKING THE RESIDUALS OF SARIMA(0,1,0)x(0,1,1)12

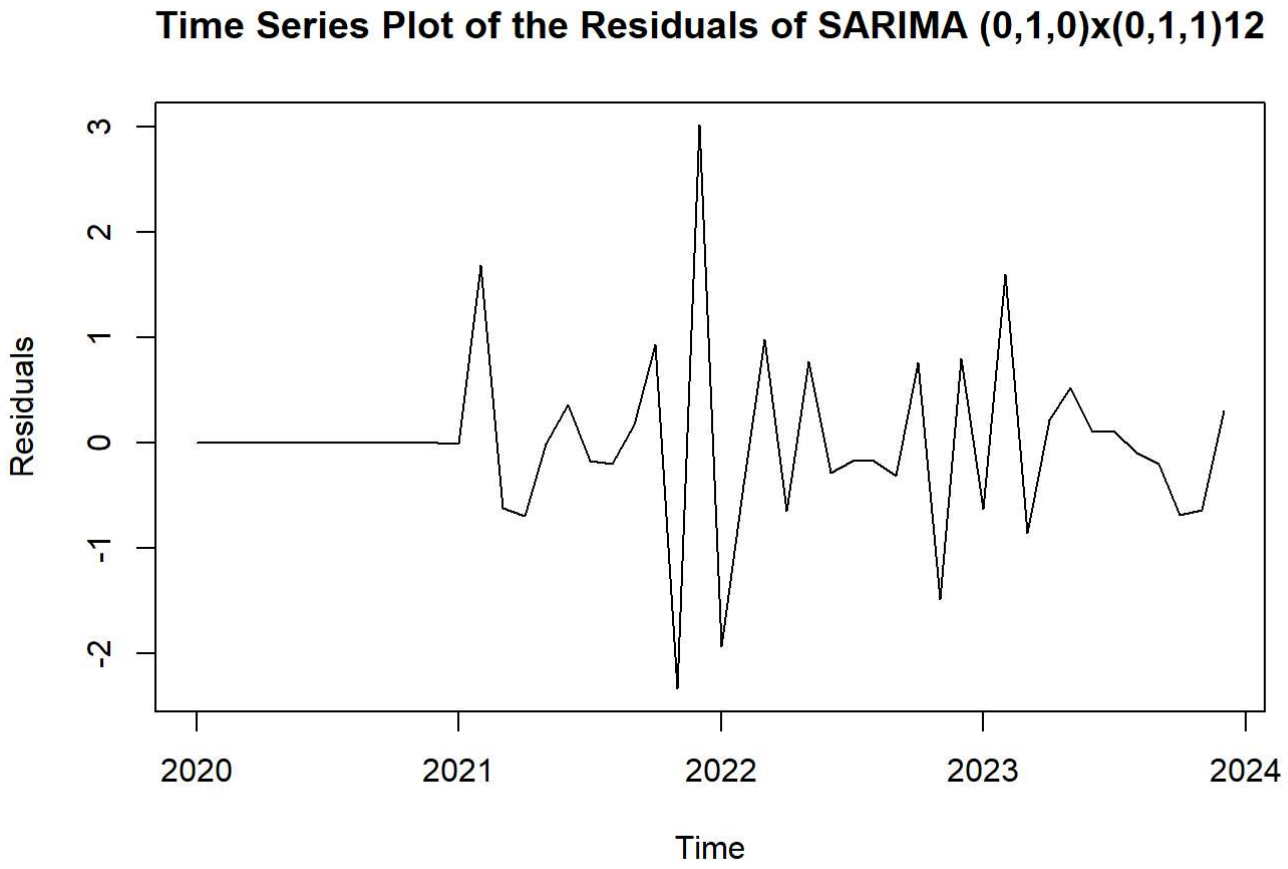
It shows the time series plot of residuals of SARIMA (0,1,0)x(0,1,1)12 model. Lets now compare this plot with the time series plot of residuals of SARIMA (0,1,0)x(0,1,0) model.

Both the residual plots are same, which was expected as we have only one change and that is Q = 1 for SARIMA (0,1,0)x(0,1,1)12. Since, Q = 1 is not that high enough to capture some of the information. However, during the winter season of the year 2022, we can see a change. The residual plot of SARIMA (0,1,0)x(0,1,1)12 is ore random during this stage. This means that our model SARIMA (0,1,0)x(0,1,1)12 captured information for this phase.

1. Trend The residual plot exhibits no apparent trend. Both first-order ordinary differencing (d = 1) and seasonal differencing (D = 1) have eliminated long-term and seasonal trends from the original series. The residuals now fluctuate around a stable mean level.
2. Seasnality Compared to the residuals of SARIMA(0,1,0)×(0,1,0)[12], the addition of the seasonal moving average term (Q = 1) has enhanced the model's ability to capture residual seasonality. The repetitive seasonal pattern from the original time series has been removed, as no visible seasonal cycles persist in the residuals.
3. Change in Variance The variance across time appears to be relatively consistent, though a brief spike is observed during early 2022. Outside this short interval, the magnitude of residuals remains within a controlled range, so it is a generally stable variance structure.
4. Change Point / Intervention There is no clear change point or structural intervention in the residual series. Although some spikes occur, particularly around winter 2022, these are not sustained and return to baseline levels quickly. The model has captured any structural shifts, if present, during that period.

5. Behaviour The residuals appear more random compared to the earlier SARIMA(0,1,0)×(0,1,0)[12] model. The inclusion of the seasonal MA term (Q = 1) has improved the model's ability to address remaining autocorrelation. However, very small clusters of autocorrelations may still exist.

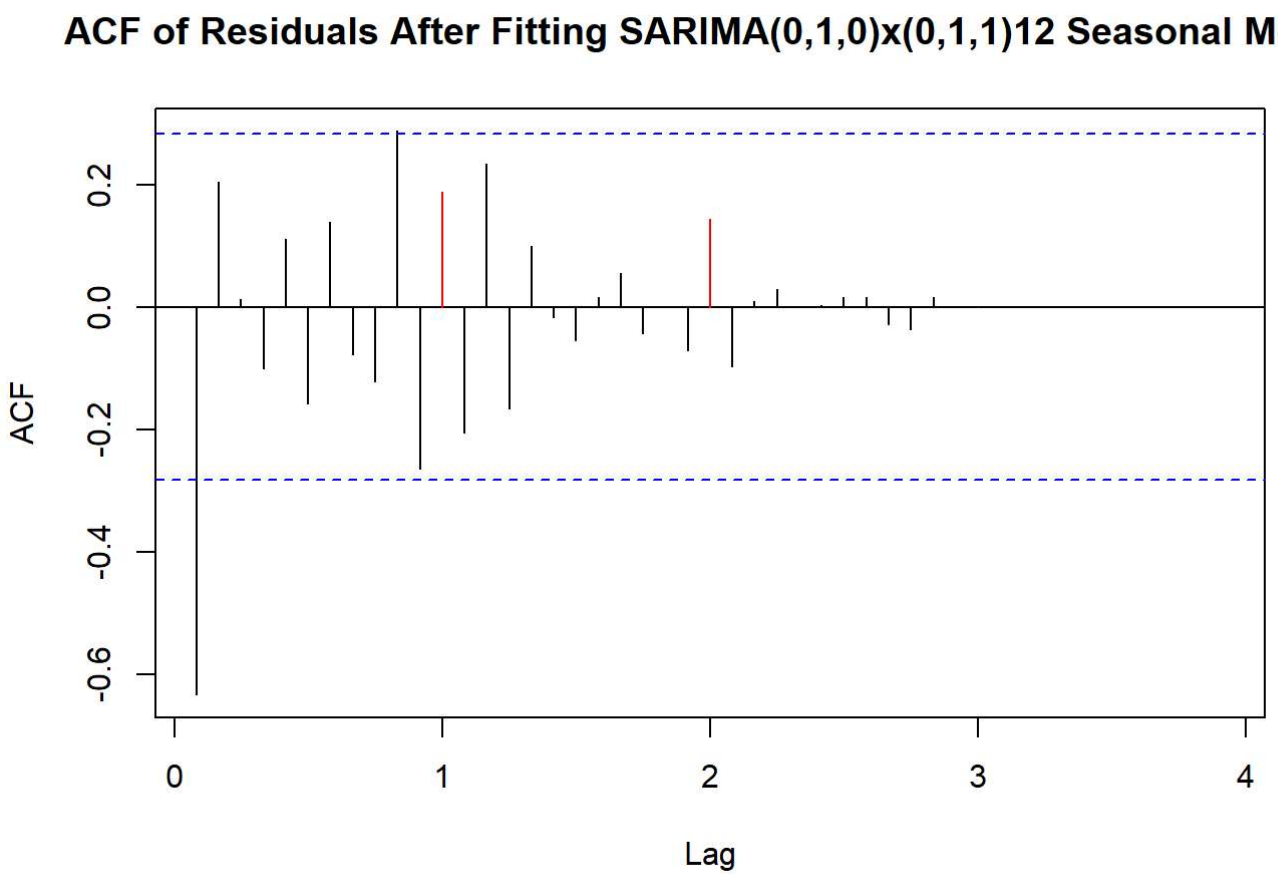
```
#SARIMA(0,1,0)x(0,1,1)12

m21.AQ = Arima(monthlyPM25TS, order = c(0,1,0), seasonal = list(order = c(0,1,1),
                                                                period = 12))
res.m21 = rstandard(m21.AQ)
plot(res.m21, xlab = "Time", ylab = "Residuals", main = "Time Series Plot of the Residuals of SA
RIMA (0,1,0)x(0,1,1)12")
```



SEASONAL ACF PLOT OF RESIDUALS FOR SARIMA (0,1,0)x(0,1,1)12

This illustrates the seasonal ACF plot of residuals of SARIMA (0,1,0)x(0,1,1)12 model. We observe that we have no red bars (seasons) that cross the confidence interval. Therefore, no seasons are significant.

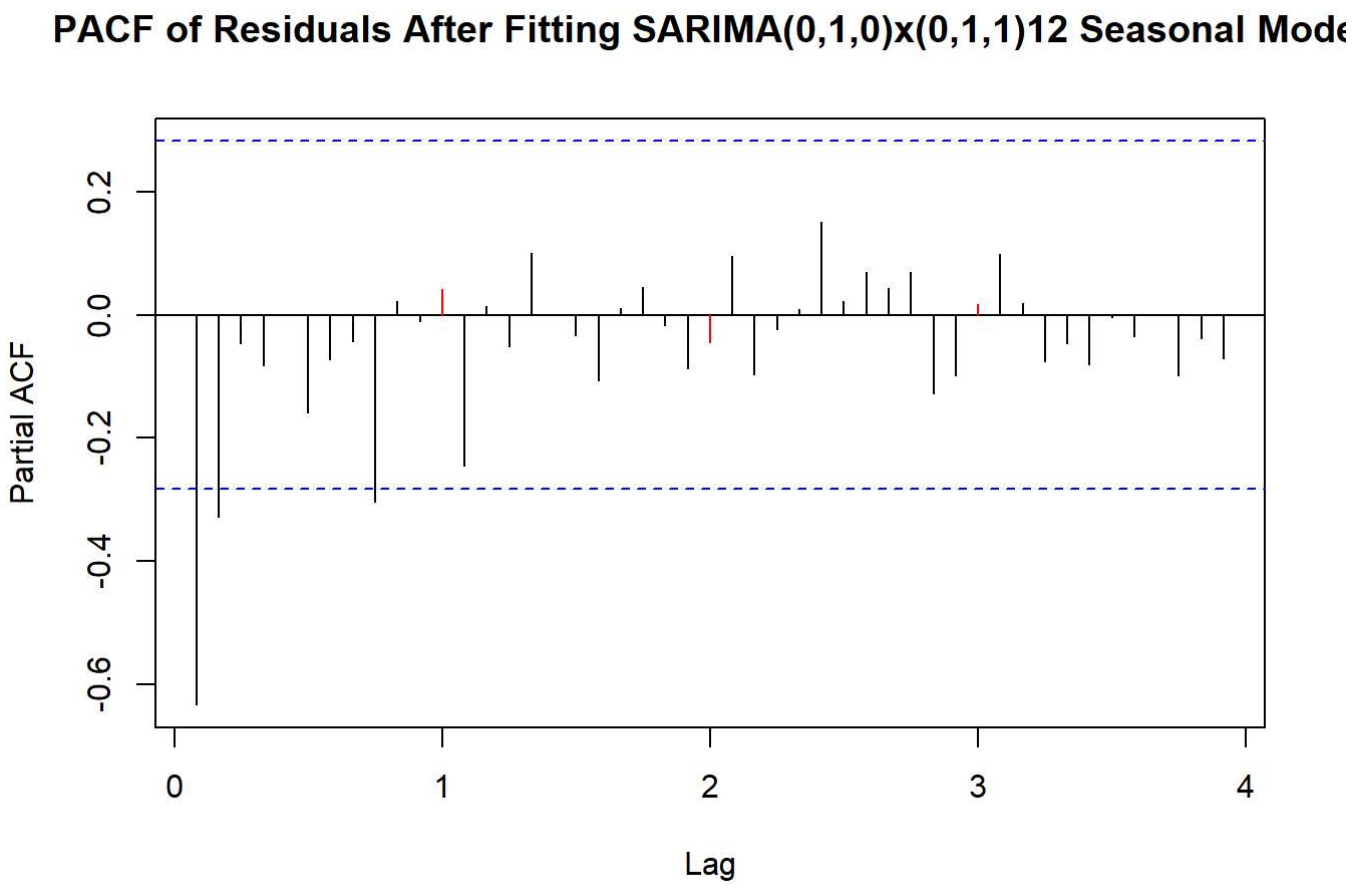
```
seasonal_acf(res.m21, lag.max = 100, main = "ACF of Residuals After Fitting SARIMA(0,1,0)x(0,1,
1)12 Seasonal Model")
```



SEASONAL PACF PLOT OF RESIDUALS FOR SARIMA (0,1,0)x(0,1,1)12

This illustrates the seasonal PACF plot of residuals of SARIMA (0,1,0)x(0,1,1)12 model. We observe that we have no red bars (seasons) that cross the confidence interval. Therefore, no seasons are significant.

```
seasonal_pacf(res.m21, lag.max = 100, main = "PACF of Residuals After Fitting SARIMA(0,1,0)x(0,
1,1)12 Seasonal Model")
```



TEST FOR STATIONARITY

After repeating the whole procedure, let's find out whether our series in stationary or not. We will use ADF test to check if our series is stationary.

H0 = The process is difference nonstationary (the process is nonstationary but becomes stationary after first differencing). HA = The process is stationary.

The result of ADF test on residuals of SARIMA (0,1,0)x(0,1,1) shows The p-value as 0.01598. since the p-value is less than the significance level (α = 0.05). Therefore, the residuals are stationary.

Now, we have our series stationary, with P = 0, Q = 1, D = 1 and d = 1.

```
adf.test(res.m21)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  res.m21
## Dickey-Fuller = -4.0342, Lag order = 3, p-value = 0.01598
## alternative hypothesis: stationary
```

ORDINARY ORDERS

With this, the seasonal part of the analysis is complete. We now, have to find the possible values of p and to get the possible set of models.

To get the values of p and q, we will use various tools like ACF, PACF, EACF and BIC table.

Let's first use ACF and PACF to get the values of ordinary orders p and q.

From the seasonal ACF plot of residuals of SARIMA (0,1,0)x(0,1,1)12 model, We will only look till the first season, the reason being the seasonality itself. As the seasons repeats itself, so whatever happens in the first season will be repeated in other seasons. Since, we are after the values of ordinary orders p and q, therefore our focus will not be the red bars (seasons).

2 bars go beyond the confidence interval, meaning we have 2 significant lags. Therefore, q = 2.

From the seasonal PACF plot of residuals of (0,1,0)x(0,1,1)12 model, For PACF will only look till the first season.

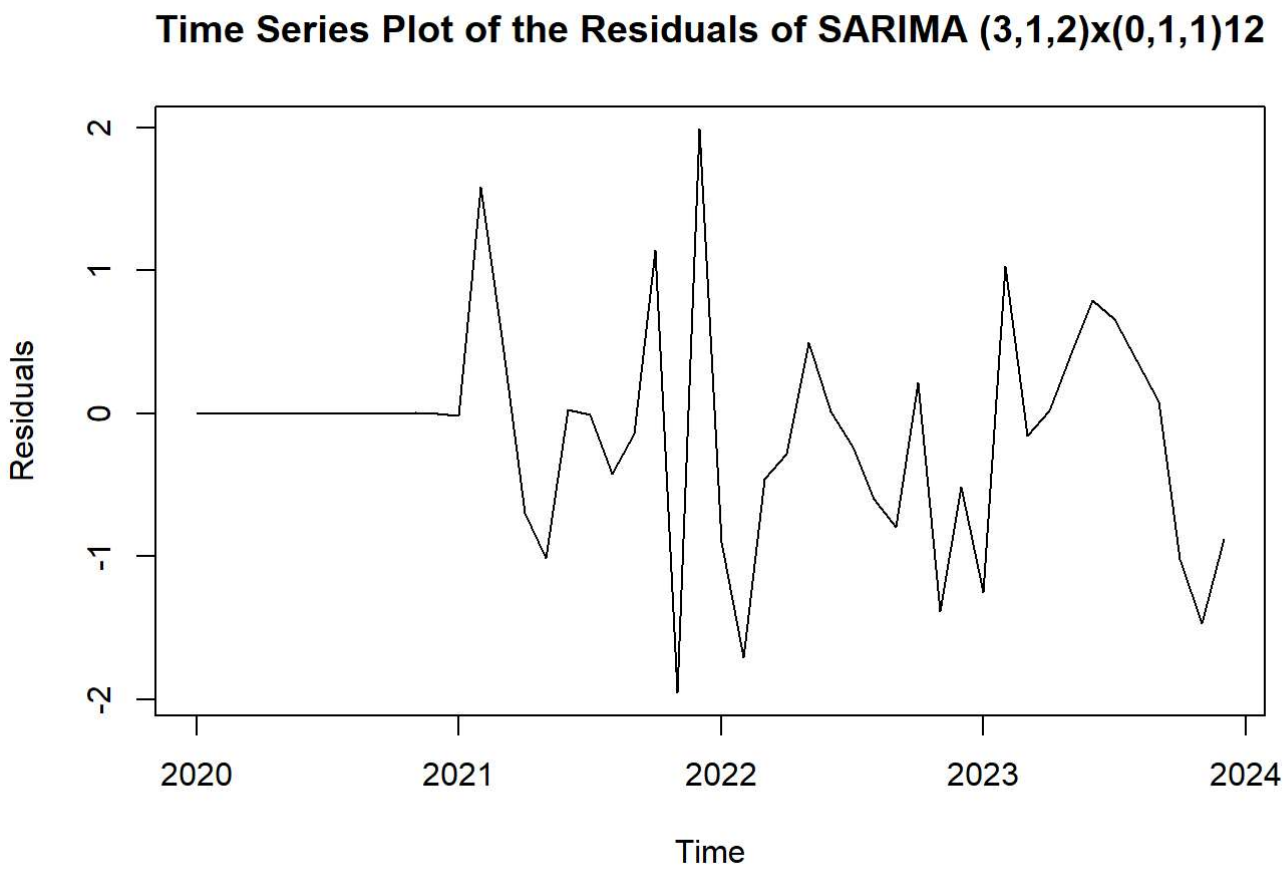
3 bars go beyond the confidence interval, meaning we have 3 significant lags. Therefore, p = 3.

Therefore, from seasonal ACF and PACF plot, we get p = 3 and q = 2. The model will be SARIMA (3,1,2)x(0,1,1)12. Let's fit this model and check the residuals.

1. Trend The residual series displays no apparent trend, Both the first-order ordinary differencing (d = 1) and seasonal differencing (D = 1) have successfully removed the deterministic structure in the original time series. The values fluctuate around a mean level close to zero.
2. Seasonality The seasonal fluctuations visible in earlier models have been eliminated. The inclusion of a seasonal moving average term (Q = 1) appears sufficient to handle seasonal effects, with no clear repeating annual spikes observable in this residual plot.
3. Change in Variance The residual variance remains relatively stable, although a notable spike is still visible around the start of 2022. However, unlike previous models, this version displays less abrupt variance swings, and the residuals appear better balanced throughout the time frame, indicating greater homoscedasticity.
4. Change Point / Intervention No sustained change point or structural break is present in this residual plot. The fluctuation observed around early 2022 appears to be an isolated event rather than an indication of a systematic shift in the process.
5. Behaviour The residuals appear noticeably more random compared to simpler models like SARIMA(0,1,0)×(0,1,1)[12]. This improvement is due to the addition of ordinary autoregressive (p = 3) and moving average (q = 2) components.

Now let's check the seasonal ACF and PACF of residuals of SARIMA (3,1,2)x(0,1,1)12 model.

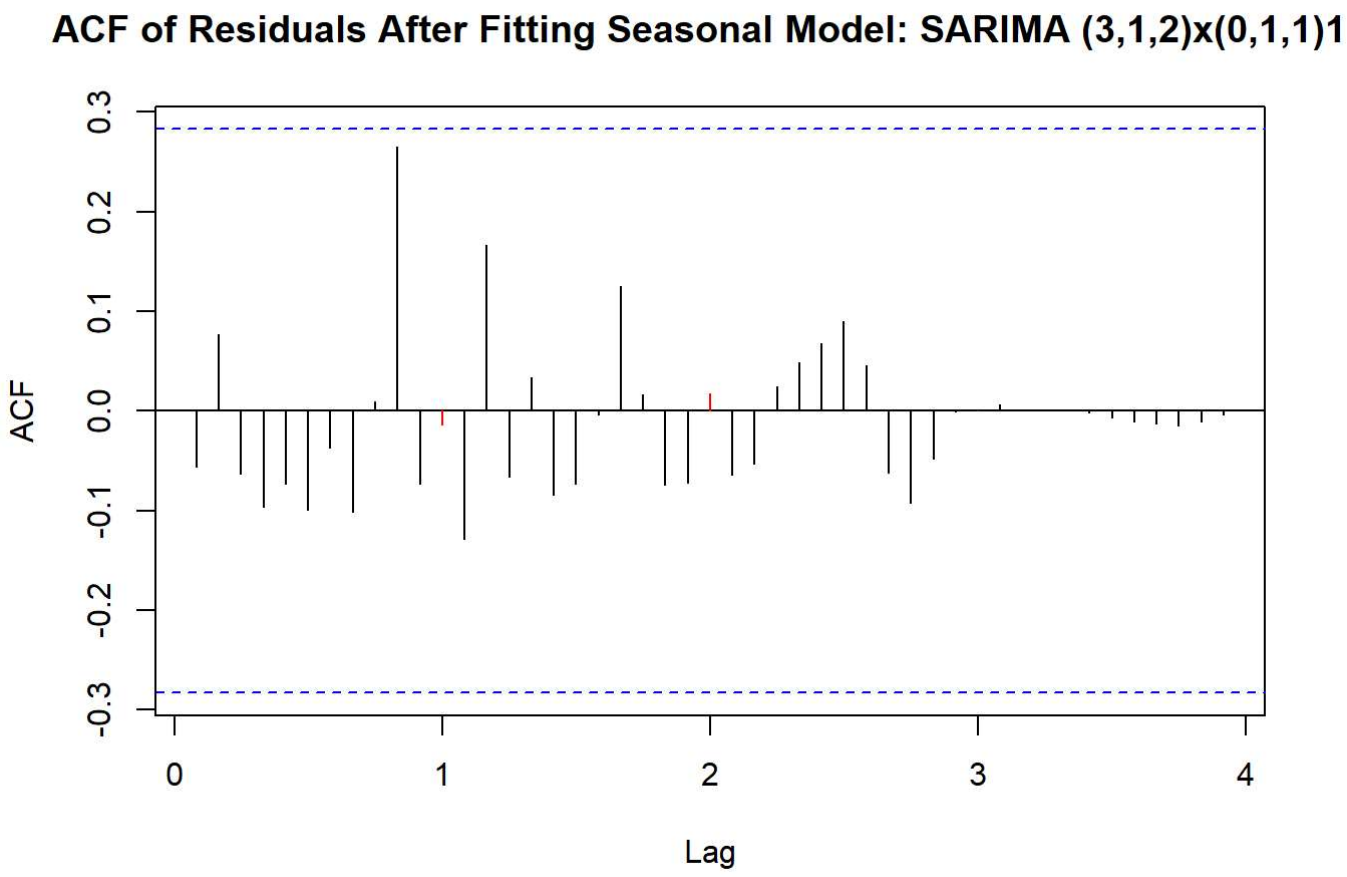
```
m3.AQ = Arima(monthlyPM25TS, order = c(3,1,2), seasonal = list(order = c(0,1,1),
                                                               period = 12))
res.m3 = rstandard(m3.AQ)
plot(res.m3, xlab = "Time", ylab = "Residuals", main = "Time Series Plot of the Residuals of SAR
IMA (3,1,2)x(0,1,1)12")
```



SEASONAL ACF PLOT OF RESIDUALS FOR SARIMA (3,1,2)x(0,1,1)12

This plot shows the seasonal ACF plot of residuals of SARIMA (3,1,2)x(0,1,1)12 model. We observe that no bars go beyond the confidence interval. This means that there is no significant information left in the residuals.

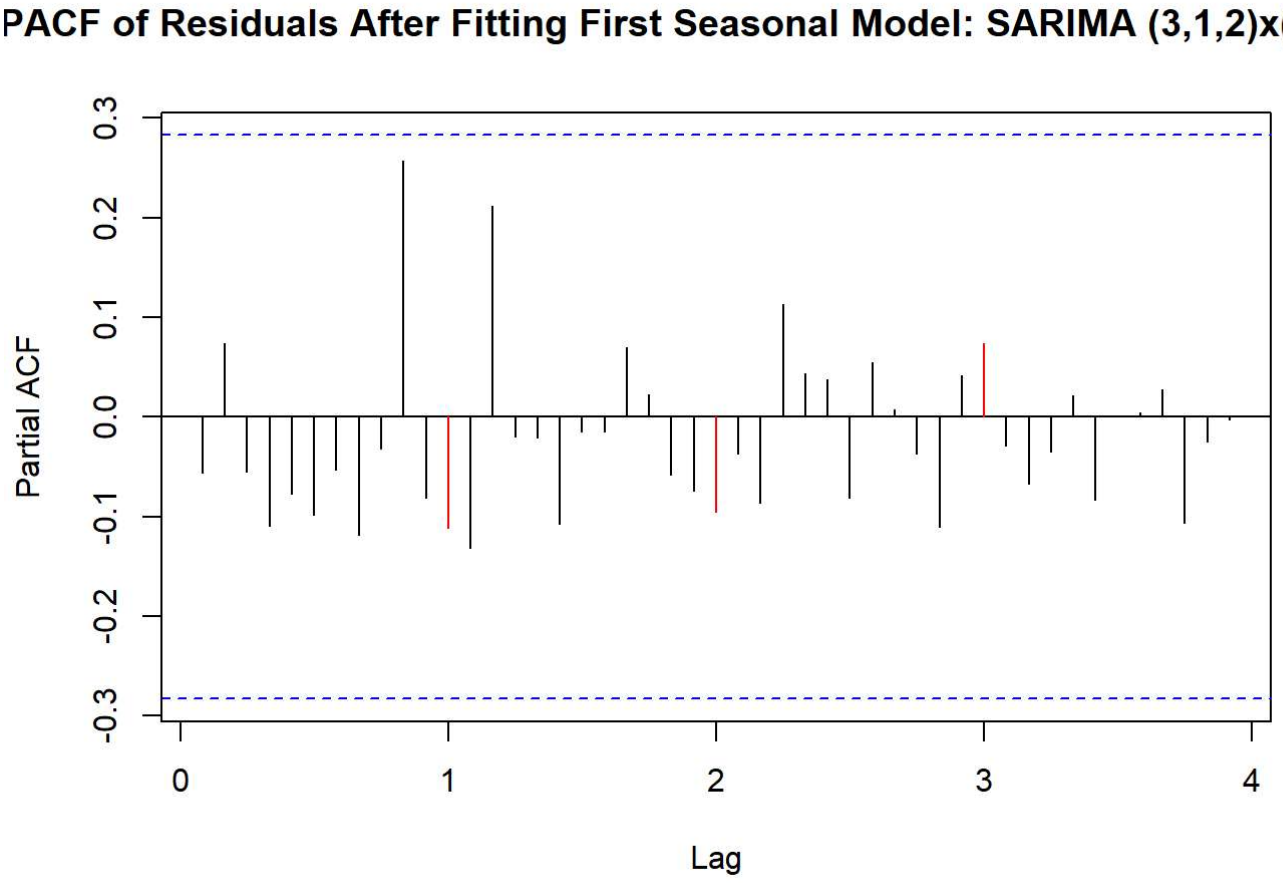
```
seasonal_acf(res.m3, lag.max = 100, main = "ACF of Residuals After Fitting Seasonal Model: SARIM
A (3,1,2)x(0,1,1)12")
```



SEASONAL PACF PLOT OF RESIDUALS FOR SARIMA (3,1,2)x(0,1,1)12

This plot shows the seasonal PACF plot of residuals of SARIMA (3,1,2)x(0,1,1)12 model. We observe that no bars go beyond the confidence interval. This means that there is no significant information left in the residuals.

```
seasonal_pacf(res.m3, lag.max = 100, main = "PACF of Residuals After Fitting First Seasonal Mode
l: SARIMA (3,1,2)x(0,1,1)12")
```



EACF PLOT

To expand the possible set of models, we will now use EACF plot to find possible values of p and q.

We will plot an EACF plot from the residuals of SARIMA (0,1,0)x(0,1,1)12, because it contains more information.

In EACF plot, the first column represents the possible values of p and the first row represents the possible values of q. In an EACF plot, we look for the top-left zero that have consecutive zeros that is not interrupted by x's, then we draw a vertex from that zero and spot the neighbour zeros. Next, we find the corresponding values of p and q for the zeros.

We will first locate the top-left zero. As we can see the top-left zero will be the zero that correspond to the value of AR = 0 and MA = 1. This particular zero must be the top-left zero, where we draw the vertex, because it is not interrupted by any x's. now we will locate the neighbour zeros of the top-left zero.

The first column gives the value of p and the first row gives us the value of q. For the top-left zero, the value of MA(q) is 1 and the value of AR(p) is 0. Therefore, the model we get is SARIMA(0,1,1)x(0,1,1)12. The neighbouring zero located on the right of the top-left zero, gives the value of p and q as 0 and 2 respectively. Therefore, the model we get is SARIMA(0,1,2) x(0,1,1)12. The neighbouring zero located at the bottom of the top-left zero, gives the value of p and q as 1 and 1 respectively. Therefore, the model we get is SARIMA(1,1,1) x(0,1,1)12. The neighbour zero located diagonally to the top-left zero, gives the value of p and q as 1 and 2 respectively. Therefore, the model we get is SARIMA(1,1,2) x(0,1,1)12.

Therefore, from EACF, we have the following possible set of models: 1. SARIMA(0,1,1)x(0,1,1)12 2. SARIMA(0,1,2)x(0,1,1)12 3. SARIMA(1,1,1)x(0,1,1)12 4. SARIMA(1,1,2)x(0,1,1)12

```
eacf(res.m21)
```

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x o o o o o o o o o o  o  o  o
## 1 x o o o o o o o o o o  o  o  o
## 2 o o o o o o o o o o o  o  o  o
## 3 x o o o o o o o o o o  o  o  o
## 4 o x o o o o o o o o o  o  o  o
## 5 o x o o o o o o o o o  o  o  o
## 6 x o x o o o o o o o o  o  o  o
## 7 x o o o o o o o o o o  o  o  o
```

BIC TABLE

This plot shows the BIC table for the residuals of SARIMA (0,1,0)x(0,1,1)12. In BIC table, the models are arranged as best to worst from top to bottom. The p-lag values give us the value of p and error lag gives us the values of q. The shaded part tells us the value of p and q in these models. We will look at the top three best models to get the possible set of models.

The best model located at the top, has a BIC value of -15. The best model has p-lag1 shaded and is supported by other models as well. The best model has p-lag2 shaded as well. However, it is not supported by the other two best models (second best and third best) but supported by rest of the models. Therefore, we will not consider it. We want to consider this as well to increase the possible the possible set of models. In the top-best model no error-lag is shaded. Therefore, from the top-best model we have p = 1 and q = 0.

The second-best model has a BIC value of -14. The shaded region in p-lag is p-lag1, which is supported by other models as well. In the second-best model no error-lag is shaded. Therefore, from second best model we have value of p as 1 and q as 0.

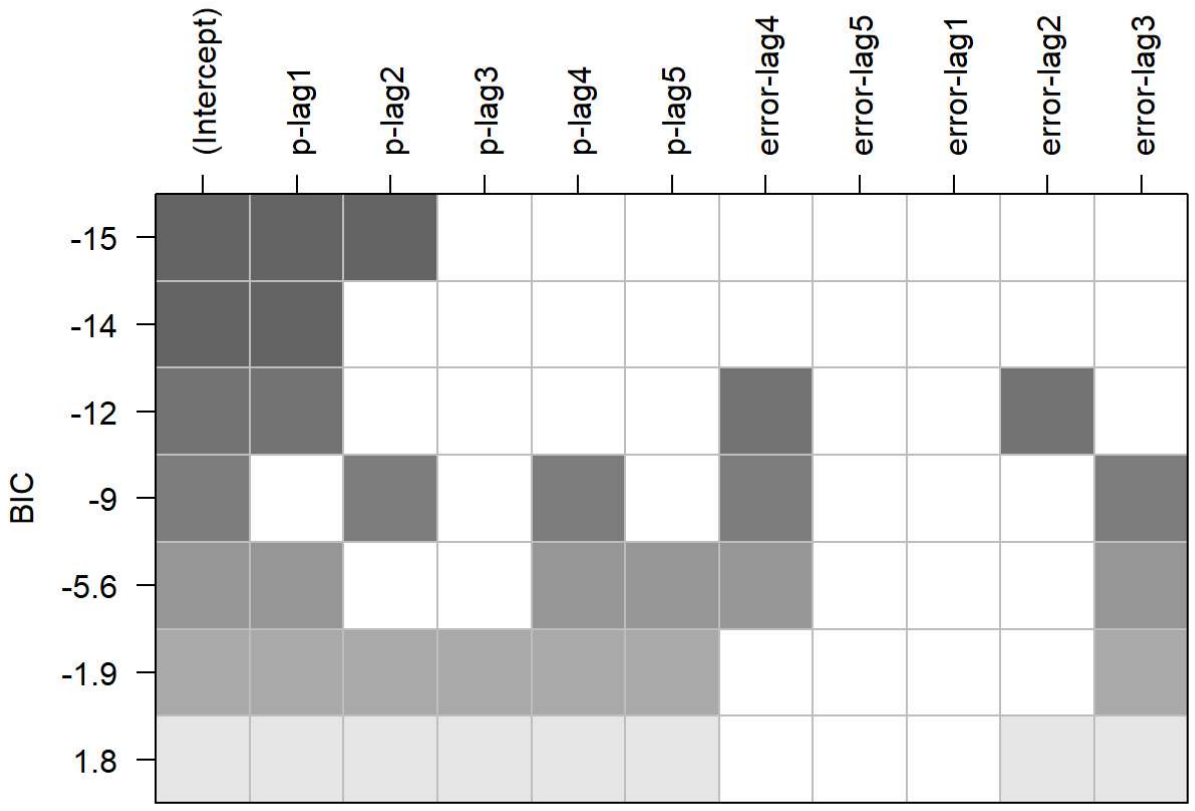
The third-best model has a BIC value of -12. The shaded region in p-lag is p-lag1, which is supported by other models as well. In the third best model error-lag2 and error-lag4 are shaded. Error-lag4 is supported by other models but error-lag2 is not supported by other models. Therefore, we will consider error-lag4 but not error-lag2. Therefore, from third best model we have value of p as 1 and q as 4.

From BIC table we have the following possible set of models:

1. SARIMA(1,1,0)x(0,1,1)
2. SARIMA(1,1,4)x(0,1,1)

```
plot(armasubsets(y=res.m21, nar = 5, nma = 5, y.name = "p", ar.method = "ols"))
```

```
## Reordering variables and trying again:
```



From ACF, PACF, EACF and BIC table, we have the following possible models:

1. SARIMA (3,1,2)x(0,1,1)12 (from ACF and PACF)
2. SARIMA (0,1,1)x(0,1,1)12 (from EACF)
3. SARIMA (0,1,2)x(0,1,1)12 (from EACF)
4. SARIMA (1,1,1)x(0,1,1)12 (from EACF)
5. SARIMA (1,1,2)x(0,1,1)12 (from EACF)
6. SARIMA (1,1,0)x(0,1,1)12 (from BIC Table)
7. SARIMA (1,1,4)x(0,1,1)12 (from BIC Table)

We, usually expect some models to be consistent, meaning same models coming from two different methods. Unfortunately, we don't see this to happen in our case. We have no consistant models. Now, we will fit these 7 models.

MODEL FITTING

To fit SARIMA models, we have 2 methods:

1. Maximum Likelihood Estimation - ML Estimation
2. Least Squares (Conditional SS - CSS

To find the model which best fit our series, we must find the model which have maximum number of estimate coefficients as significant. We will use both Maximum Likelihood Estimation - ML Estimation and Least Squares (Conditional SS - CSS) model fitting methods to check the significance of the estimate coefficients.

1. SARIMA (3,1,2)x(0,1,1)12 - ML FITTED

Here we have 3 ordinary AR coefficients, 2 ordinary MA coefficients and 1 seasonal MA coefficient. The p-value of every coefficient is greater than the significance level (α=0.05). therefore, no coefficients are significant.

```
m312.AQ = Arima(monthlyPM25TS, order = c(3,1,2), seasonal = list(order = c(0,1,1),
                                                                 period = 12), method = "ML")
coeftest(m312.AQ)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ar1  -0.21720    0.85744 -0.2533   0.8000
## ar2   0.16394    0.28214  0.5810   0.5612
## ar3   0.21783    0.21405  1.0176   0.3088
## ma1  -0.76830    0.87566 -0.8774   0.3803
## ma2  -0.10832    0.71865 -0.1507   0.8802
## sma1 -0.64495    0.40879 -1.5777   0.1146
```

2. SARIMA (3,1,2)x(0,1,1)12 - CSS FITTED

Here we have 3 ordinary AR coefficients, 2 ordinary MA coefficients and 1 seasonal MA coefficient. The p-value of all the AR coefficients are less than the significance level (α=0.05). Therefore, all AR coefficients are significant. The coefficient ma1 is significant as well as its p-value is less than the significance level (α=0.05). The seasonal coefficient is insignificant.

```
m312.AQCSS = Arima(monthlyPM25TS, order = c(3,1,2), seasonal = list(order = c(0,1,1),
                                                                   period = 12), method = "CSS")
coeftest(m312.AQCSS)
```

```
##
## z test of coefficients:
##
##        Estimate Std. Error  z value  Pr(>|z|)
## ar1  -0.1787468  0.0421840  -4.2373 2.262e-05 ***
## ar2   0.0824269  0.0073722  11.1808 < 2.2e-16 ***
## ar3   0.1035620  0.0358123   2.8918   0.00383 **
## ma1  -1.2208050  0.0982655 -12.4235 < 2.2e-16 ***
## ma2  -0.0420805  0.1167663  -0.3604   0.71856
## sma1 -0.3737872  0.1946228  -1.9206   0.05479 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. SARIMA (3,1,2)x(0,1,1)12 - CSS-ML FITTED

Here we have 3 ordinary AR coefficients, 2 ordinary MA coefficients and 1 seasonal MA coefficient. The p-value of every coefficient is greater than the significance level (α=0.05). therefore, no coefficients are significant.

```
m312.AQCSSML = Arima(monthlyPM25TS, order = c(3,1,2), seasonal = list(order = c(0,1,1),
                                                                      period = 12), method = "CSS-
ML")
coeftest(m312.AQCSSML)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ar1  -0.21720    0.85744 -0.2533   0.8000
## ar2   0.16394    0.28214  0.5810   0.5612
## ar3   0.21783    0.21405  1.0176   0.3088
## ma1  -0.76830    0.87566 -0.8774   0.3803
## ma2  -0.10832    0.71865 -0.1507   0.8802
## sma1 -0.64495    0.40879 -1.5777   0.1146
```

4. SARIMA (0,1,1)x(0,1,1)12 - ML FITTED

Here we got 1 MA coefficient as significant as its p-value is less than the significance level (α=0.05). the seasonal coefficient is insignificant.

```
m011.AQ = Arima(monthlyPM25TS, order = c(0,1,1), seasonal = list(order = c(0,1,1),
                                                                 period = 12), method = "ML")
coeftest(m011.AQ)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ma1  -0.86030    0.11114 -7.7406 9.892e-15 ***
## sma1 -0.62385    0.38634 -1.6148    0.1064
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. SARIMA (0,1,1)x(0,1,1)12 - CSS FITTED

Here we got 1 MA and 1 sma coefficient as significant as their p-value is less than the significance level (α=0.05).

```
m011.AQCSS = Arima(monthlyPM25TS, order = c(0,1,1), seasonal = list(order = c(0,1,1),
                                                                   period = 12), method = "CSS")
coeftest(m011.AQCSS)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value Pr(>|z|)
## ma1  -0.891705   0.095885 -9.2997  < 2e-16 ***
## sma1 -0.420017   0.170782 -2.4594  0.01392 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. SARIMA (0,1,2)x(0,1,1)12 - ML FITTED

Here we got 1 MA coefficient as significant as its p-value is less than the significance level (α=0.05).

```
m012.AQ = Arima(monthlyPM25TS, order = c(0,1,2), seasonal = list(order = c(0,1,1),
                                                                 period = 12), method = "ML")
coeftest(m012.AQ)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ma1  -0.95923    0.16289 -5.8889 3.887e-09 ***
## ma2   0.15210    0.16330  0.9314   0.35165
## sma1 -0.58694    0.35023 -1.6759   0.09376 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7. SARIMA (0,1,2)x(0,1,1)12 - CSS FITTED

Here we got 1 MA and 1 sma coefficient as significant as their p-value is less than the significance level (α=0.05).

```
m012.AQCSS = Arima(monthlyPM25TS, order = c(0,1,2), seasonal = list(order = c(0,1,1),
                                                                   period = 12), method = "CSS")
coeftest(m012.AQCSS)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ma1  -0.97236    0.15361 -6.3300 2.452e-10 ***
## ma2   0.12804    0.16357  0.7828   0.43377
## sma1 -0.44763    0.17406 -2.5718   0.01012 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8. SARIMA (1,1,1)x(0,1,1)12 - ML FITTED

Here we got 1 MA coefficient as significant as its p-value is less than the significance level (α=0.05).

```
m111.AQ = Arima(monthlyPM25TS, order = c(1,1,1), seasonal = list(order = c(0,1,1),
                                                                 period = 12), method = "ML")
coeftest(m111.AQ)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ar1  -0.22233    0.22435 -0.9910   0.32169
## ma1  -0.76017    0.19155 -3.9685 7.231e-05 ***
## sma1 -0.59459    0.34859 -1.7057   0.08806 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9. SARIMA (1,1,1)x(0,1,1)12 - CSS FITTED

Here we got all the coefficient as significant as their p-values are less than the significance level (α=0.05).

```
m111.AQCSS = Arima(monthlyPM25TS, order = c(1,1,1), seasonal = list(order = c(0,1,1),
                                                                   period = 12), method = "CS
S")
coeftest(m111.AQCSS)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ar1  -0.41877    0.16695 -2.5084 0.012129 *
## ma1  -0.53683    0.16853 -3.1855 0.001445 **
## sma1 -0.41306    0.19270 -2.1435 0.032071 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10. SARIMA (1,1,2)x(0,1,1)12 - ML FITTED

Here we got no coefficients as significant as their p-values are greater than the significance level (α=0.05).

```
m112.AQ = Arima(monthlyPM25TS, order = c(1,1,2), seasonal = list(order = c(0,1,1),
                                                                 period = 12), method = "ML")
coeftest(m112.AQ)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value Pr(>|z|)
## ar1  -0.260773   0.544974 -0.4785  0.63229
## ma1  -0.721094   0.545387 -1.3222  0.18611
## ma2  -0.035909   0.476429 -0.0754  0.93992
## sma1 -0.597097   0.350925 -1.7015  0.08885 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

11. SARIMA (1,1,2)x(0,1,1)12 - CSS FITTED

Here we got ar1, ma1 and sma1 coefficients as significant as their p-values are less than the significance level (α=0.05).

```
m112.AQCSS = Arima(monthlyPM25TS, order = c(1,1,2), seasonal = list(order = c(0,1,1),
                                                                    period = 12), method = "CSS")
coeftest(m112.AQCSS)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value Pr(>|z|)
## ar1  -0.420494   0.214244 -1.9627  0.04968 *
## ma1  -0.534903   0.226256 -2.3642  0.01807 *
## ma2  -0.003767   0.293892 -0.0128  0.98977
## sma1 -0.412194   0.204321 -2.0174  0.04366 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12. SARIMA (1,1,2)x(0,1,1)12 - CSS-ML FITTED

Here we got no coefficients as significant as their p-values are greater than the significance level (α=0.05).

```
m112.AQCSSML = Arima(monthlyPM25TS, order = c(1,1,2), seasonal = list(order = c(0,1,1),
                                                                      period = 12), method = "CSS-M
L")
coeftest(m112.AQCSSML)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value Pr(>|z|)
## ar1  -0.258590   0.545670 -0.4739  0.63558
## ma1  -0.723166   0.545540 -1.3256  0.18497
## ma2  -0.034098   0.475931 -0.0716  0.94288
## sma1 -0.596869   0.350757 -1.7017  0.08882 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

13. SARIMA (1,1,0)x(0,1,1)12 - ML FITTED

Here we got ar1 coefficients as significant as it's p-value is less than the significance level (α=0.05).

```
m110.AQ = Arima(monthlyPM25TS, order = c(1,1,0), seasonal = list(order = c(0,1,1),
                                                                 period = 12), method = "ML")
coeftest(m110.AQ)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ar1  -0.61439    0.13833 -4.4414 8.939e-06 ***
## sma1 -0.99230    1.65440 -0.5998    0.5486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

14. SARIMA (1,1,0)x(0,1,1)12 - CSS FITTED

Here we got ar1 and sma1 coefficients as significant as their p-values are less than the significance level (α=0.05).

```
m110.AQCSS = Arima(monthlyPM25TS, order = c(1,1,0), seasonal = list(order = c(0,1,1),
                                                                 period = 12), method = "CSS")
coeftest(m110.AQCSS)
```

```
##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ar1  -0.58211    0.12976 -4.4859 7.261e-06 ***
## sma1 -0.42106    0.16538 -2.5461   0.01089 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

15. SARIMA (1,1,4)x(0,1,1)12 - ML FITTED

Here we got no coefficients as significant as their p-values are greater than the significance level (α=0.05).

```
m114.AQ = Arima(monthlyPM25TS, order = c(1,1,4), seasonal = list(order = c(0,1,1),
                                                                 period = 12), method = "ML")
coeftest(m114.AQ)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value Pr(>|z|)
## ar1  -0.308179   1.336420 -0.2306   0.8176
## ma1  -0.679379   1.335909 -0.5086   0.6111
## ma2  -0.016814   1.296708 -0.0130   0.9897
## ma3   0.111068   0.382711  0.2902   0.7717
## ma4  -0.217812   0.204361 -1.0658   0.2865
## sma1 -0.704345   0.515968 -1.3651   0.1722
```

16. SARIMA (1,1,4)x(0,1,1)12 - CSS FITTED

Here we got ar1, ma2, ma3, and ma4 coefficients as significant as their p-values are less than the significance level (α=0.05).

```
m114.AQCSS = Arima(monthlyPM25TS, order = c(1,1,4), seasonal = list(order = c(0,1,1),
                                                                    period = 12), method = "CSS")
coeftest(m114.AQCSS)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1  -0.502763   0.010486 -47.9453 < 2.2e-16 ***
## ma1  -0.035585   0.123957  -0.2871  0.774055
## ma2  -0.664882   0.138663  -4.7950 1.627e-06 ***
## ma3   0.659098   0.148790   4.4297 9.436e-06 ***
## ma4  -0.450999   0.137574  -3.2782  0.001045 **
## sma1 -0.389856   0.211994  -1.8390  0.065915 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

17. SARIMA (1,1,4)x(0,1,1)12 - CSS-ML FITTED

Here we got no coefficients as significant as their p-values are greater than the significance level (α=0.05).

```
m114.AQCSSML = Arima(monthlyPM25TS, order = c(1,1,4), seasonal = list(order = c(0,1,1),
                                                                      period = 12), method = "CSS-
ML")
coeftest(m114.AQCSSML)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value Pr(>|z|)
## ar1  -0.308179   1.336420 -0.2306   0.8176
## ma1  -0.679379   1.335909 -0.5086   0.6111
## ma2  -0.016814   1.296708 -0.0130   0.9897
## ma3   0.111068   0.382711  0.2902   0.7717
## ma4  -0.217812   0.204361 -1.0658   0.2865
## sma1 -0.704345   0.515968 -1.3651   0.1722
```

SORTING MODELS BASED ON AIC AND BIC VALUES

After fitting the all the set of possible models, we must find the best fitting model. To achieve this goal, we will have a look at the AIC and BIC values of these models. The model having the lowest AIC and BIC value will be the best fitting model.

```
#AIC and BIC Sorting

sort.score <- function(x, score = c("bic", "aic")){
  if (score == "aic"){
    x[with(x, order(AIC)),]
  } else if (score == "bic") {
    x[with(x, order(BIC)),]
  } else {
    warning('score = "x" only accepts valid arguments ("aic","bic")')
  }
}
```

AIC VALUES OF FITTED MODELS

The model SARIMA(0,1,1)x(0,1,1)12 has the minimum AIC value (240.4393) among the fitted models. The second best model according to AIC value is SARIMA(1,1,1)x(0,1,1)12 with an AIC value of 241.4536, followed by SARIMA(0,1,2)x(0,1,1) with AIC value of 241.6328. Therefore, based on AIC values the best fitted model is SARIMA(0,1,1)x(0,1,1), followed by SARIMA(1,1,1)x(0,1,1). The Third best model is SARIMA(0,1,2)x(0,1,1).

```
sort.score(AIC(m312.AQ,m011.AQ,m012.AQ,m111.AQ,m112.AQ,
               m110.AQ,m114.AQ), score = "aic")
```

```
##         df      AIC
## m011.AQ  3 240.4393
## m111.AQ  4 241.4536
## m012.AQ  4 241.6328
## m112.AQ  5 243.4478
## m114.AQ  7 245.9689
## m110.AQ  3 246.3765
## m312.AQ  7 246.4054
```

BIC VALUES OF FITTED MODELS

The model SARIMA(0,1,1)x(0,1,1)12 has the minimum BIC value (245.1053) among the fitted models. The second best model according to BIC value is SARIMA(1,1,1)x(0,1,1)12 with a BIC value of 247.6750, followed by SARIMA(0,1,2)x(0,1,1) with a BIC value of 247.8542. Therefore, based on BIC values the best fitted model is SARIMA(0,1,1)x(0,1,1), followed by SARIMA(1,1,1)x(0,1,1). The Third best model is SARIMA(0,1,2)x(0,1,1).

We cannot just choose a best fitted model using 1 tool. We need to verify our results using some other tool. The second tool that we can use to verify our result is Error Measures.

We will check the errors for each fitted model and choose the one which have the lowest errors.

```
sort.score(BIC(m312.AQ,m011.AQ,m012.AQ,m111.AQ,m112.AQ,
               m110.AQ,m114.AQ), score = "bic")
```

```
##         df      BIC
## m011.AQ  3 245.1053
## m111.AQ  4 247.6750
## m012.AQ  4 247.8542
## m110.AQ  3 251.0425
## m112.AQ  5 251.2245
## m114.AQ  7 256.8563
## m312.AQ  7 257.2928
```

ERROR MEASURES

. We will check RMSE (Root Mean Square Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error) and MASE (Mean Absolute Standardised Error).

• SARIMA(1,1,4)x(0,1,1)12 has the lowest RMSE value (7.014195), followed by SARIMA(3,1,2)x(0,1,1)12 with RMSE value of 7.348100. • SARIMA(3,1,2)x(0,1,1)12 has the lowest MAE value (5.564214), followed by SARIMA(1,1,4)x(0,1,1)12 with MAE value of 5.605617 • SARIMA(3,1,2)x(0,1,1)12 has the lowest MAPE value (26.24107), followed by SARIMA(1,1,2)x(0,1,1)12 with MAPE value of 26.36976. • SARIMA(3,1,2)x(0,1,1)12 has the lowest MASE value (1.0055323), followed by SARIMA(1,1,4)x(0,1,1)12 with MASE value of 1.012803.

After observing the AIC, BIC amd error measures values, we can conclude that SARIMA(0,1,1)x(0,1,1)12 is the best fitted model. SARIMA(0,1,1)x(0,1,1)12 has the lowest AIC and BIC values. Error measures for the this model are not bad either. We are getting significant coefficients through both ML and CSS methods.

We didn't choose SARIMA(3,1,2)x(0,1,1)12 and SARIMA(1,1,4)x(0,1,1)12 as best model because they have very high AIC and BIC values. We were also not getting significant coefficients through CSS-ML methods.

```
# Error Measures

model.312A <- Arima(monthlyPM25TS, order = c(3,1,2), method = "ML")
model.011A <- Arima(monthlyPM25TS, order = c(0,1,1), method = "ML")
model.012A <- Arima(monthlyPM25TS, order = c(0,1,2), method = "ML")
model.111A <- Arima(monthlyPM25TS, order = c(1,1,1), method = "ML")
model.112A <- Arima(monthlyPM25TS, order = c(1,1,2), method = "ML")
model.110A <- Arima(monthlyPM25TS, order = c(1,1,0), method = "ML")
model.114A <- Arima(monthlyPM25TS, order = c(1,1,4), method = "ML")

Smodel.312A <- accuracy(model.312A)[1:7]
Smodel.011A <- accuracy(model.011A)[1:7]
Smodel.012A <- accuracy(model.012A)[1:7]
Smodel.111A <- accuracy(model.111A)[1:7]
Smodel.112A <- accuracy(model.112A)[1:7]
Smodel.110A <- accuracy(model.110A)[1:7]
Smodel.114A <- accuracy(model.114A)[1:7]


df.Smodels <- data.frame(
  rbind(Smodel.312A, Smodel.011A, Smodel.012A, Smodel.111A, Smodel.112A,
        Smodel.110A, Smodel.114A)
)


colnames(df.Smodels) <- c("ME", "RMSE", "MAE", "MPE", "MAPE", "MASE", "ACF1")


rownames(df.Smodels) <- c("SARIMA(3,1,2)", "SARIMA(0,1,1)", "SARIMA(0,1,2)",
                          "SARIMA(1,1,1)", "SARIMA(1,1,2)", "SARIMA(1,1,0)",
                          "SARIMA(1,1,4)")


df.Smodels
```

```
##                         ME     RMSE      MAE        MPE     MAPE     MASE
## SARIMA(3,1,2) -0.19334783 7.348100 5.564214  -3.023126 26.24107 1.005323
## SARIMA(0,1,1) -0.32944005 8.663613 6.331453  -6.720066 27.42072 1.143944
## SARIMA(0,1,2) -0.17033850 7.910845 5.917502  -2.485524 27.46622 1.069153
## SARIMA(1,1,1) -0.35329567 8.620002 6.296364  -7.093627 27.70452 1.137605
## SARIMA(1,1,2) -0.08083933 7.831457 5.824764  -1.077768 26.36976 1.052398
## SARIMA(1,1,0) -0.33471466 8.662577 6.331897  -6.813523 27.49200 1.144025
## SARIMA(1,1,4) -1.30509237 7.014195 5.605617 -14.156773 28.55941 1.012803
##                        ACF1
## SARIMA(3,1,2) -0.019705753
## SARIMA(0,1,1) -0.006157493
## SARIMA(0,1,2)  0.082631671
## SARIMA(1,1,1)  0.074442396
## SARIMA(1,1,2)  0.010320963
## SARIMA(1,1,0)  0.005698577
## SARIMA(1,1,4) -0.017491994
```

OVERFITTING

The best fitted model according to AIC, BIC and error measures values is SARIMA(0,1,1)x(0,1,1)12. Now we will perform overfitting. In overfitting, we will find two new models by increasing the values of p and q by one unit at a time.

The model we get are:

1. SARIMA(1,1,1)x(0,1,1)12
2. SARIMA(0,1,2)x(0,1,1)12

We already got these models by EACF plot.

DIAGNOSTIC CHECKING

We fill first have a look at the time series plot of residuals of the SARIMA(0,1,1)x(0,1,1)12 model.

TIME SERIES PLOT OF RESIDUALS

• Trend

We cannot see any trend in the series. The data points are fluctuating at a mean level. No trend can be seen.

• Seasonality

Seasonality is repeating pattern in the time series plot. The residual plot of SARIMA(0,1,1)x(0,1,1)12 there are no repeating pattern. The plot is quite random. This means, our model SARIMA(0,1,1)x(0,1,1) captured seasonality very well.

• Change in variance

We can still see change in variance in the residual plot. The range at the end of the year 2021 is higher than the range at the start of the year 2021. This means that our model SARIMA(0,1,1)x(0,1,1)12 failed to capture change in variance.

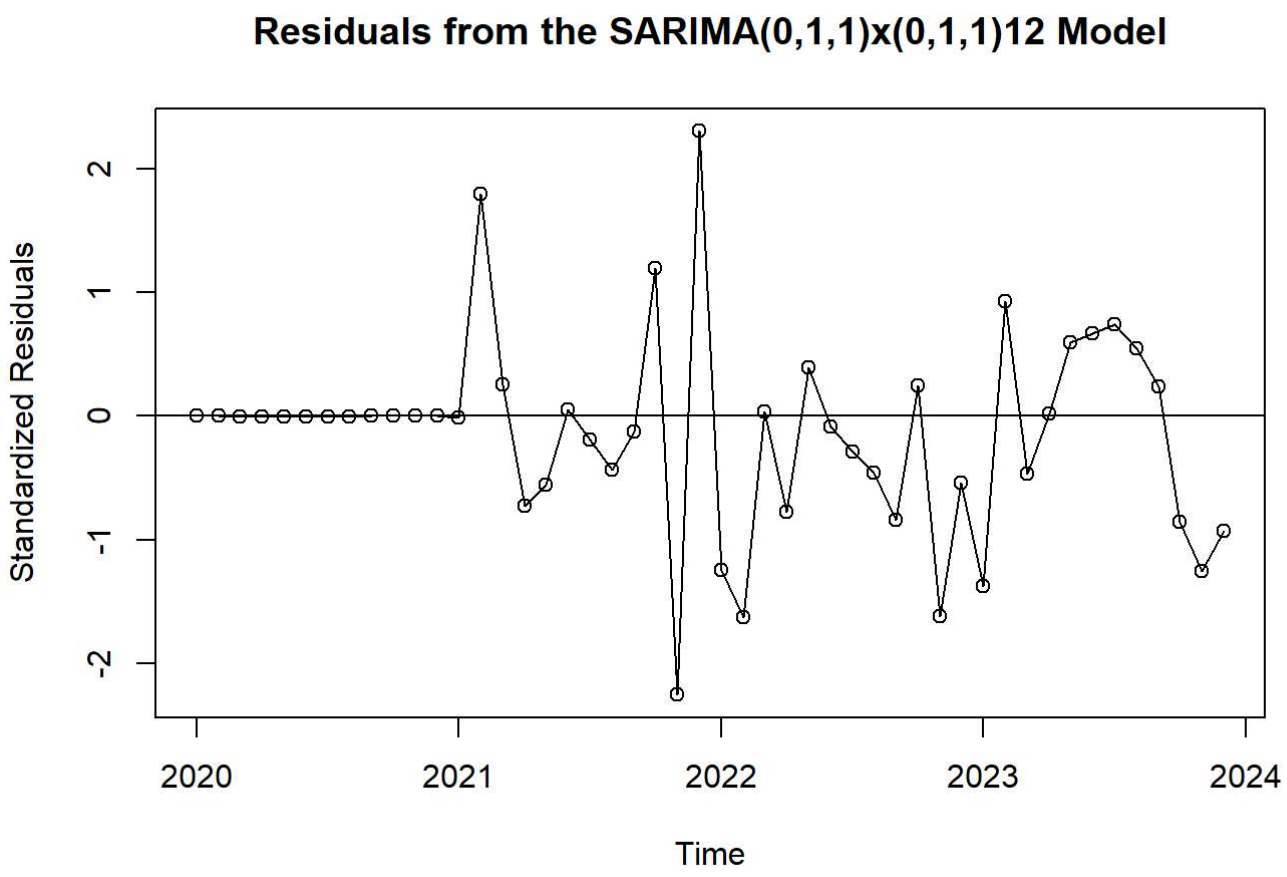
• Change point/Intervention

In our residual plot of SARIMA(0,1,1)x(0,1,1)12, we can see two change points. At the end of the year 2021 we can see a sudden increment and then sudden decrement. This means that our model SARIMA(0,1,1)x(0,1,1)12 failed to capture change point.

• Behaviour

The residual plot has mostly succeeding data points indicating that the behaviour is autoregressive.

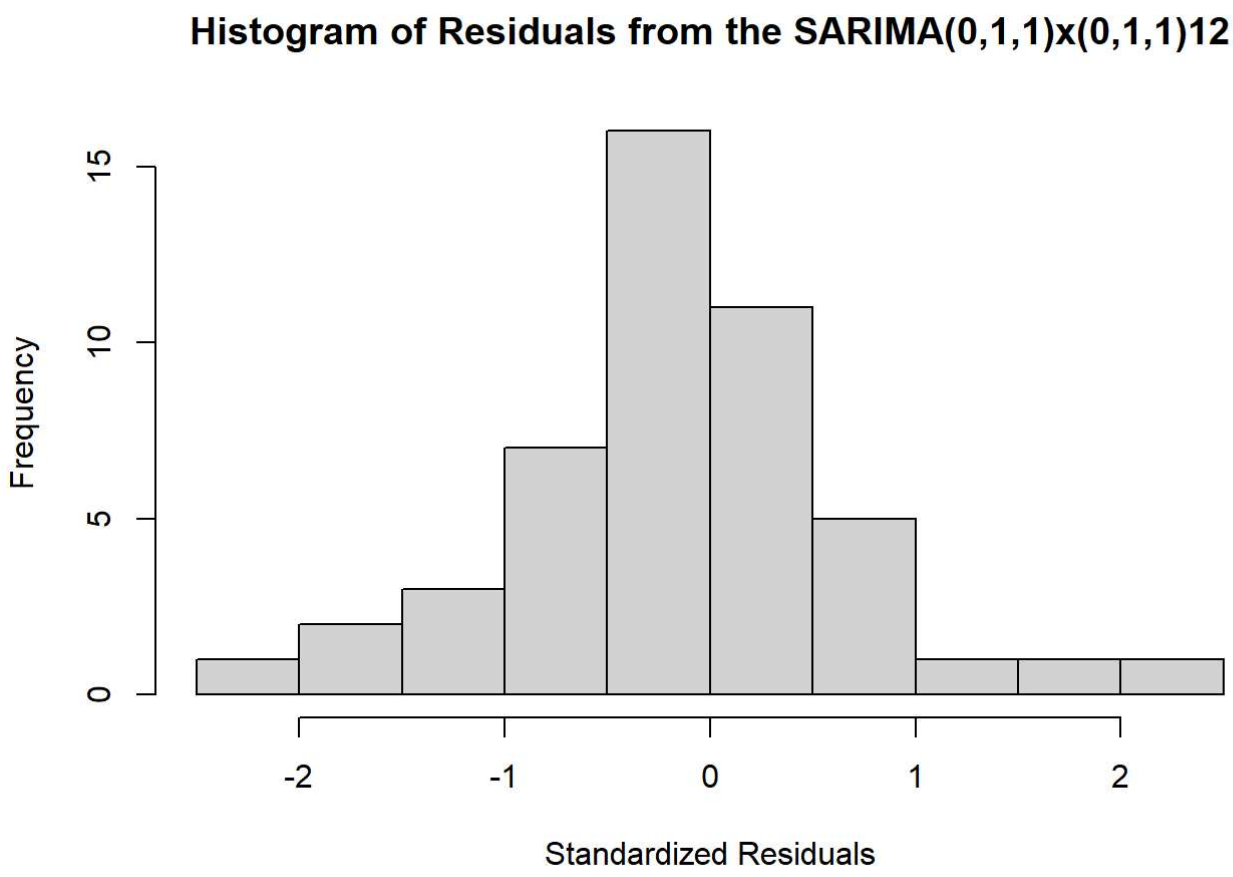
```
plot(rstandard(m011.AQ),
     ylab='Standardized Residuals',type='o',
     main="Residuals from the SARIMA(0,1,1)x(0,1,1)12 Model")
abline(h=0)
```



HISTOGRAM OF RESIDUALS

Histogram gives us an idea about symmetricity of the distribution. Our histogram is distributed between -3 to +3. This means that we have no outliers in the residuals.

```
hist(rstandard(m011.AQ),xlab='Standardized Residuals',main="Histogram of Residuals from the SARI
MA(0,1,1)x(0,1,1)12")
```

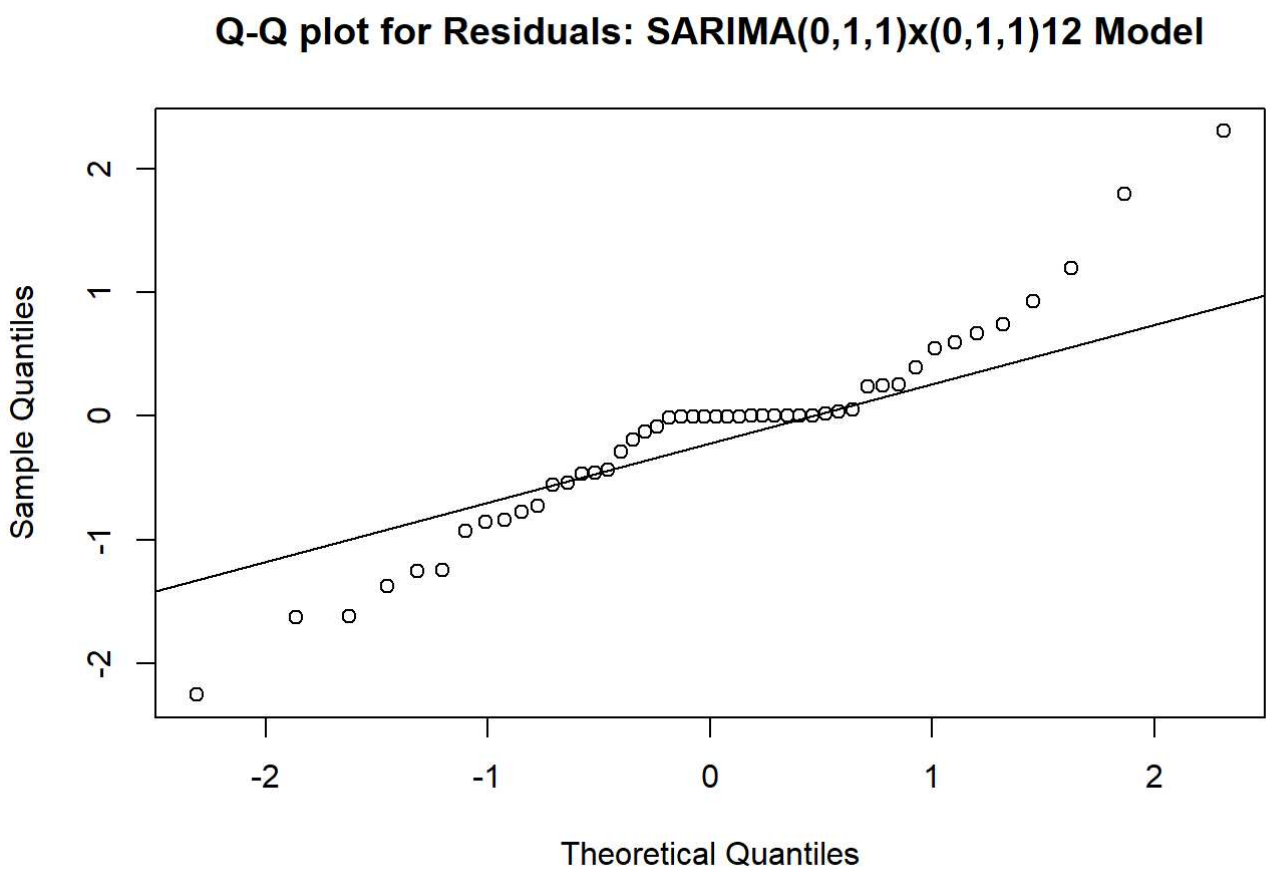


Q-Q PLOT

To check normality of the residuals, we will look at the Q-Q plot. For the residuals to be normal we expect the data points to stick to the reference line. In our case, the data points are deviated from the reference line at the bottom tail. At the top tail, the data points are far away from the reference line. A small amount of data points sticks to the reference line in the middle.

As there is huge deviation of data points from the reference line, we can conclude that the residuals are not normally distributed.

```
qqnorm(rstandard(m011.AQ),main="Q-Q plot for Residuals: SARIMA(0,1,1)x(0,1,1)12 Model")
qqline(rstandard(m011.AQ))
```



SHAPIRO-WILK TEST

To check normality of the residuals, we use another tool called as Shapiro-wilk test.

The p-value of Shapiro-wilk test is 0.03179, which is less than the significance level (α=0.05). Therefore, the residuals are not normally distributed.

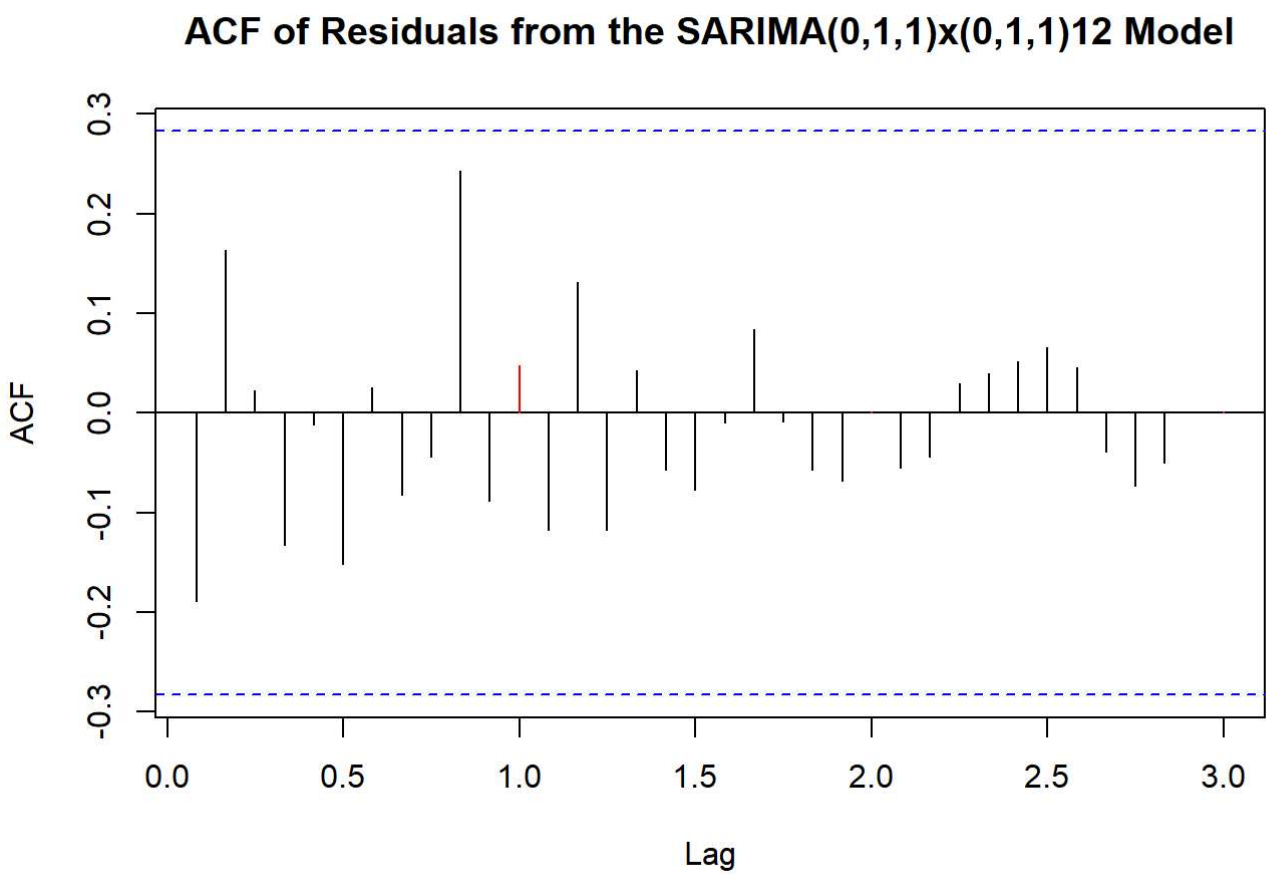
```
shapiro.test(rstandard(m011.AQ))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(m011.AQ)
## W = 0.94749, p-value = 0.03179
```

SEASONAL ACF OF RESIDUALS OF SARIMA(0,1,1)x(0,1,1)12

We observe that there are no significant bars.

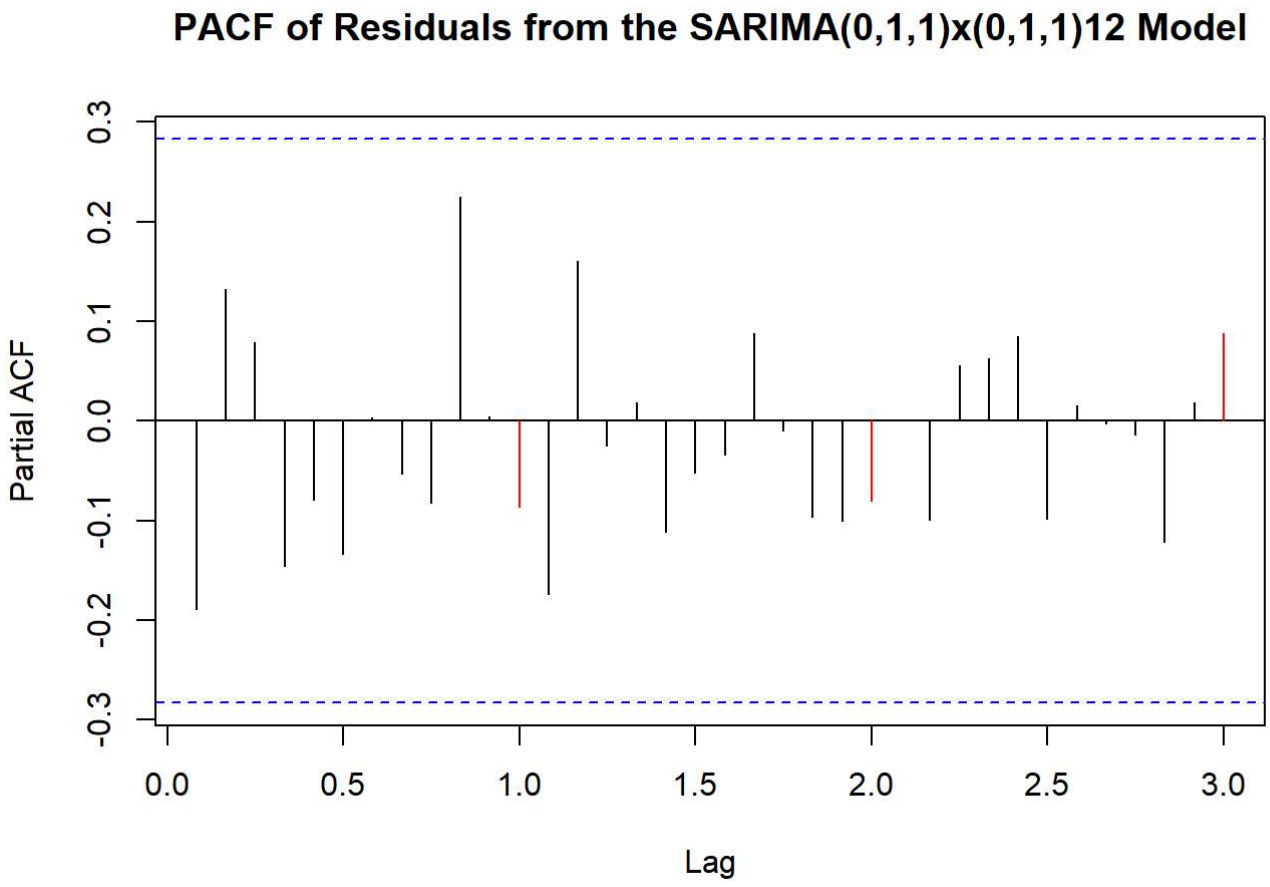
```
seasonal_acf(rstandard(m011.AQ),
             lag.max=36,
             main="ACF of Residuals from the SARIMA(0,1,1)x(0,1,1)12 Model")
```



SEASONAL PACF OF RESIDUALS OF SARIMA(0,1,1)x(0,1,1)12

In PACF plot, we don't see any significant bars.That means there is no significant autocorrelations in the residuals. The model SARIMA(0,1,1)x(0,1,1)12 captured all the significant autocorrelation

```
seasonal_pacf(rstandard(m011.AQ),
              lag.max=36,
              main="PACF of Residuals from the SARIMA(0,1,1)x(0,1,1)12 Model")
```

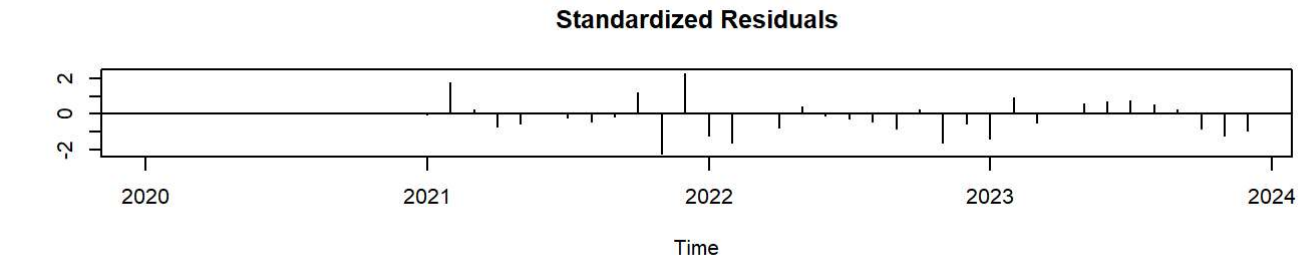


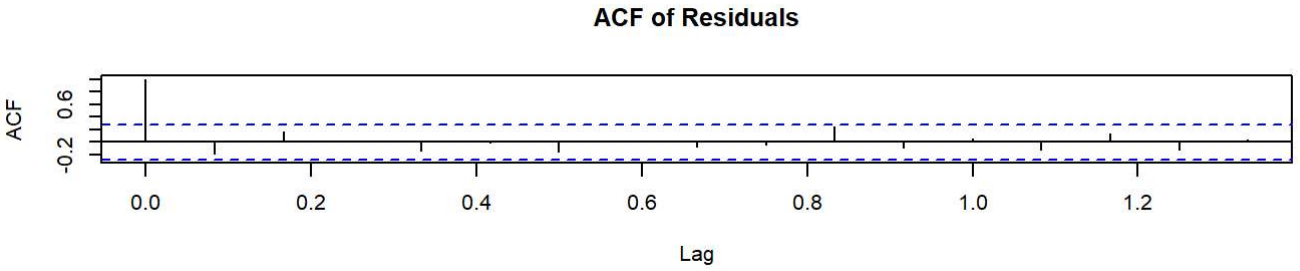
LJUNG-BOX TEST

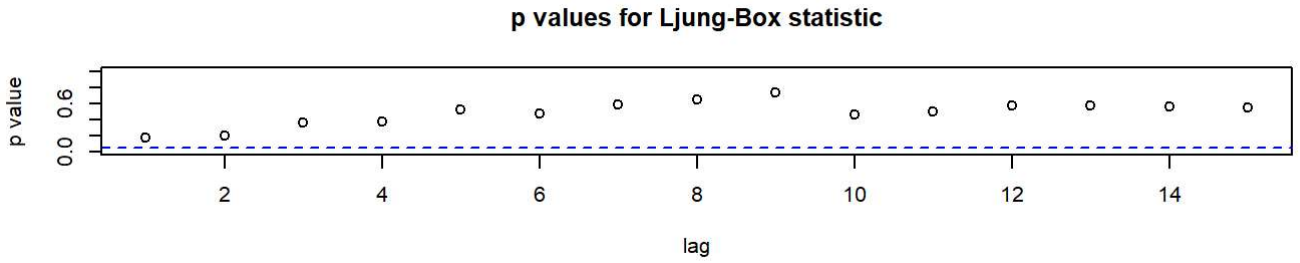
Another tool that helps us to decide whether there is any significant autocorrelation left in the residual is ljung-box test.

The bottom plot shows the p-values for ljung-box test. All the p-values are above the reference line (α = 0.05), this means that there is no left-over autocorrelation in the residuals.

```
tsdiag(m011.AQ,gof=15,omit.initial=F)
```







FORECASTING

We have presented the forecast of PM2.5 levels for the next 10 months using ML and CSS method. We will perform forecasting using the best fitted model which is SARIMA(0,1,1)x(0,1,1)12.

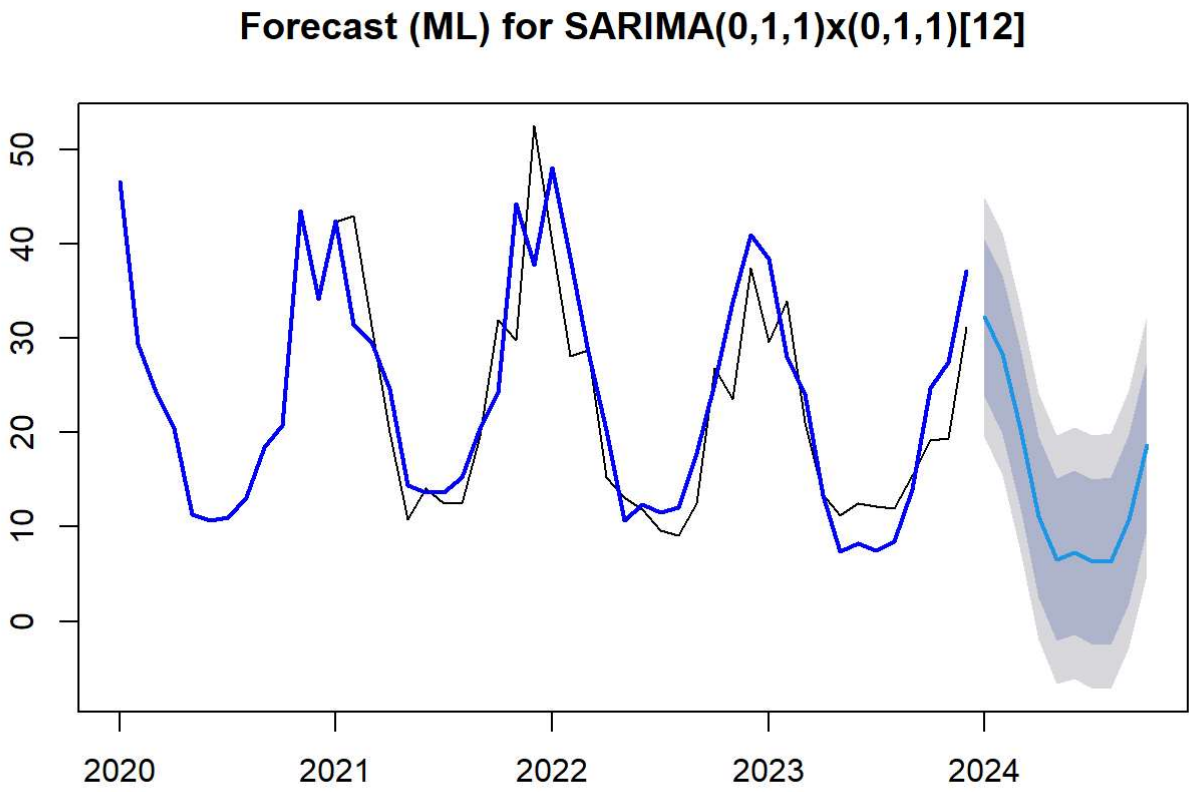
1. Forecast using ML method

As expected, we can see the PM2.5 levels decreasing during the summer (January to August). After that the PM2.5 levels starts increasing as the winter approaches (September and October).

```
frcML  <- forecast::forecast(m011.AQ,    h = 10)
print(frcML)
```

```
##          Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## Jan 2024      32.193965 23.874363 40.51357 19.470231 44.91770
## Feb 2024      28.268224 19.868741 36.66771 15.422323 41.11413
## Mar 2024      20.384507 11.905895 28.86312  7.417589 33.35142
## Apr 2024      11.094103  2.537094 19.65111 -1.992713 24.18092
## May 2024       6.553186 -2.081508 15.18788 -6.652439 19.75881
## Jun 2024       7.257667 -1.454020 15.96935 -6.065708 20.58104
## Jul 2024       6.287645 -2.500360 15.07565 -7.152448 19.72774
## Aug 2024       6.373082 -2.490583 15.23675 -7.182724 19.92889
## Sep 2024      10.831249  1.892563 19.76993 -2.839291 24.50179
## Oct 2024      18.691663  9.678581 27.70474  4.907344 32.47598
```

```
plot(frcML, main = "Forecast (ML) for SARIMA(0,1,1)x(0,1,1)[12]")
lines(fitted(m011.AQ), col = "blue", lwd = 2)
```



2. Forecast using CSS method

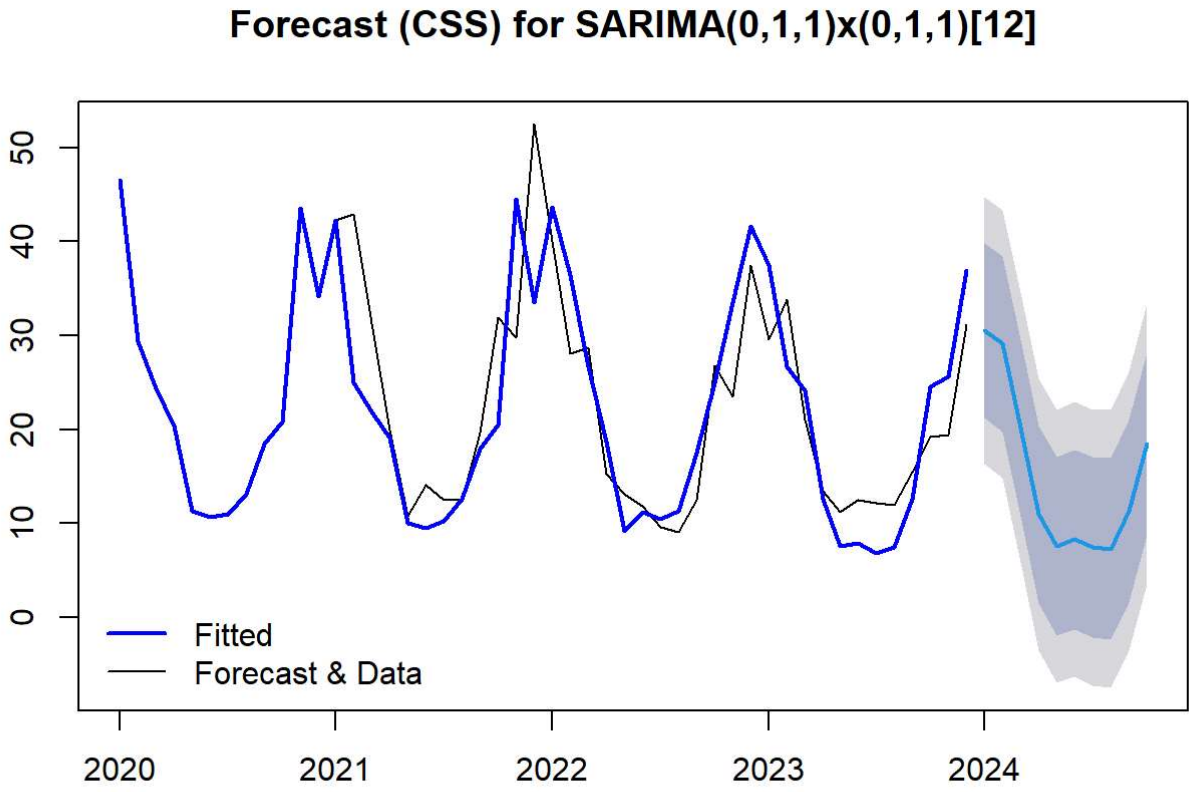
As expected, we can see the PM2.5 levels decreasing during the summer (January to August). After that the PM2.5 levels starts increasing as the winter approaches (September and October).

We are getting consistent results by ML and CSS methods.

```
frcCSS <- forecast::forecast(m011.AQCSS, h = 10)
print(frcCSS)
```

```
##          Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## Jan 2024      30.528475 21.240392 39.81656 16.323577 44.73337
## Feb 2024      29.098933 19.756739 38.44113 14.811280 43.38659
## Mar 2024      20.206273 10.810280 29.60227  5.836341 34.57621
## Apr 2024      10.946376  1.496889 20.39586 -3.505367 25.39812
## May 2024       7.553352 -1.949326 17.05603 -6.979740 22.08644
## Jun 2024       8.326503 -1.229071 17.88208 -6.287486 22.94049
## Jul 2024       7.433013 -2.175165 17.04119 -7.261428 22.12745
## Aug 2024       7.283793 -2.376703 16.94429 -7.490661 22.05825
## Sep 2024      11.309124  1.596592 21.02166 -3.544912 26.16316
## Oct 2024      18.474618  8.710327 28.23891  3.541423 33.40781
```

```
plot(frcCSS, main = "Forecast (CSS) for SARIMA(0,1,1)x(0,1,1)[12]")
lines(fitted(m011.AQCSS), col = "blue", lwd = 2)
legend("bottomleft",
       legend = c("Fitted", "Forecast & Data"),
       col = c("blue", "black"),
       lty = c(1, 1),
       lwd = c(2, 1),
       bty = "n")
```



CONCLUSION

In this report, we tried to analyse the average PM2.5 level in Milan, Italy. In every winter, the air gets polluted and the concentration of PM2.5 increases. We found seasonality and tried to fit SARIMA model. In fitting SARIMA model, we used residual approach. In residual approach, we first fit the model and then check the residuals. We found SARIMA(0,1,1)x(0,1,1)12 as the best fitted model base on AIC, BIC and error measures values.

We then, performed diagnostic checking on residuals. We plotted time series plot of residuals to check if there is trend, seasonality, change in variance, change point and behaviour of the residuals. We plotted histogram to check the symmetricity of the residuals and at last plotted Q-Q plot and performed Shapiro-wilk test to check normality of the residuals. Using the best fitted model we predicted the average PM2.5 concentration for the next 10 months.