STAT 6555 Final TH

Bheeni Garg

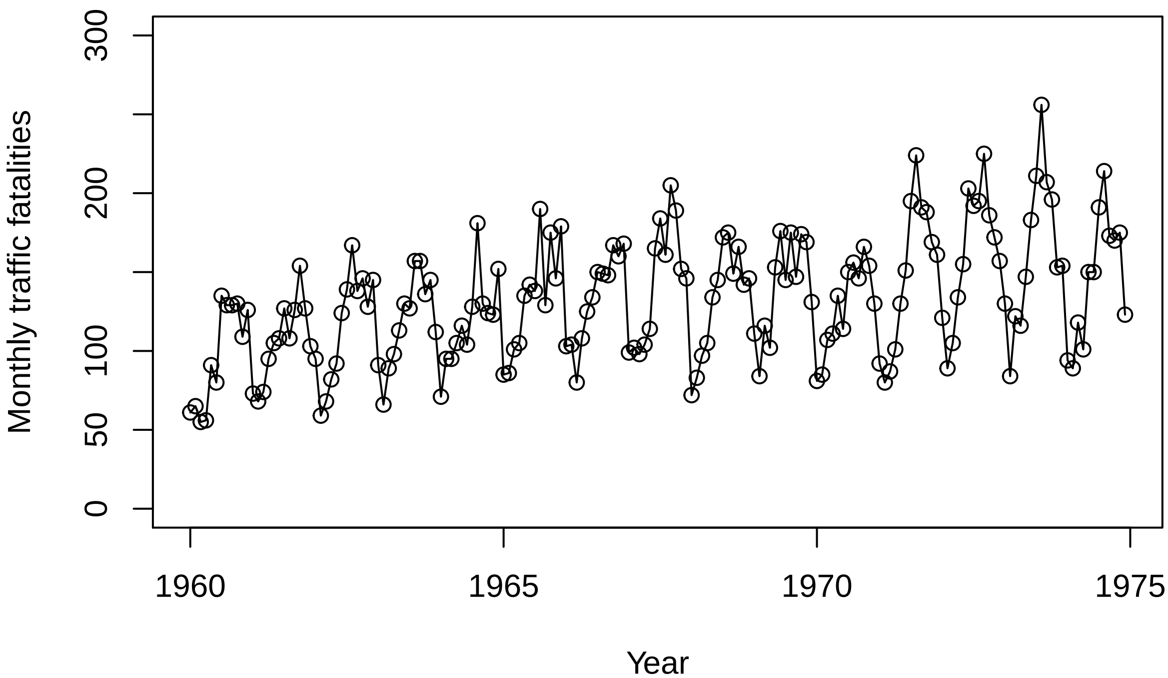
6/7/2017

Time series modeling of Monthly Traffic Fatalities in Ontario from 1960-1974 using SARIMA

Trend assessment and forecasting traffic accident fatalities by using time-series models can provide useful information that can be used by policy-makers and managers for planning and implementing special interventions to prevent and limit future accidental deaths.

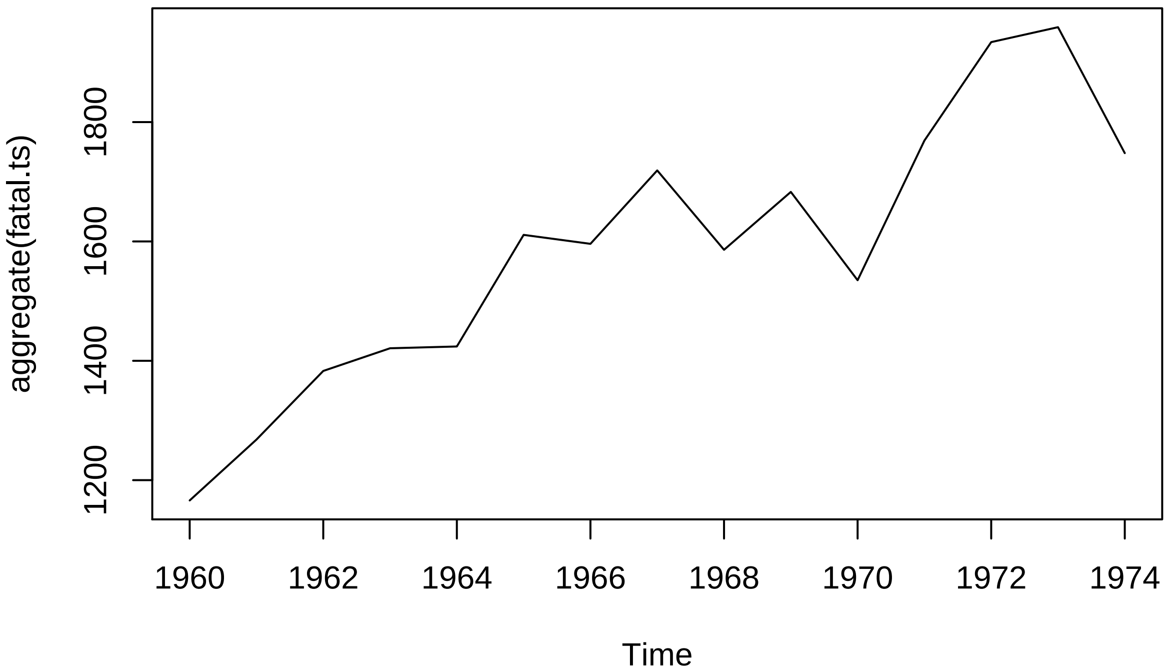
The objective of this report is use trend assessment to develop a forecasting model for traffic accident fatalities for January 1975 to December 1976 in Ontario, considering the seasonal Box-Jenkins modeling approach.

The data for the monthly traffic fatalities in Ontario from January 1960 to December 1974 are loaded into R and the time-series is plotted,

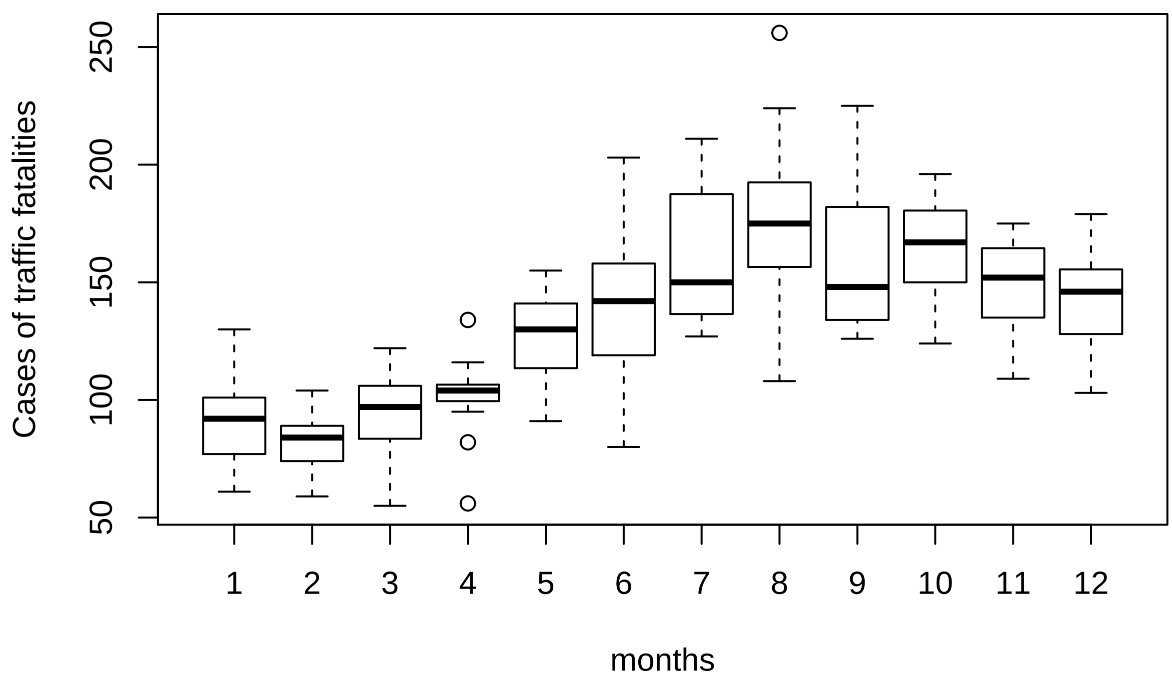


The series is non-stationary with a slight increasing trend and strong seasonal effect. An increase in variance can be noted in the series from January 1971 to December 1973.

To get a clearer view of the trend, the seasonal effect can be removed by aggregating the data to the annual level, which can be achieved in R using the aggregate function.
The plot confirms the overall increasing trend with a slight dip around January 1970.



A summary of the values for each season can be viewed using a boxplot.

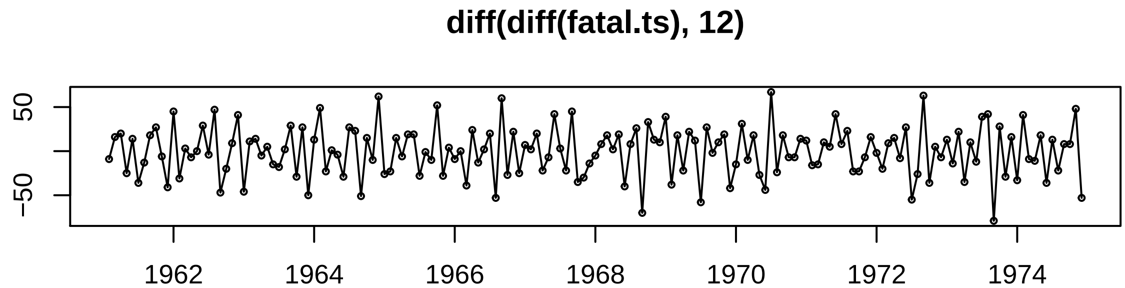


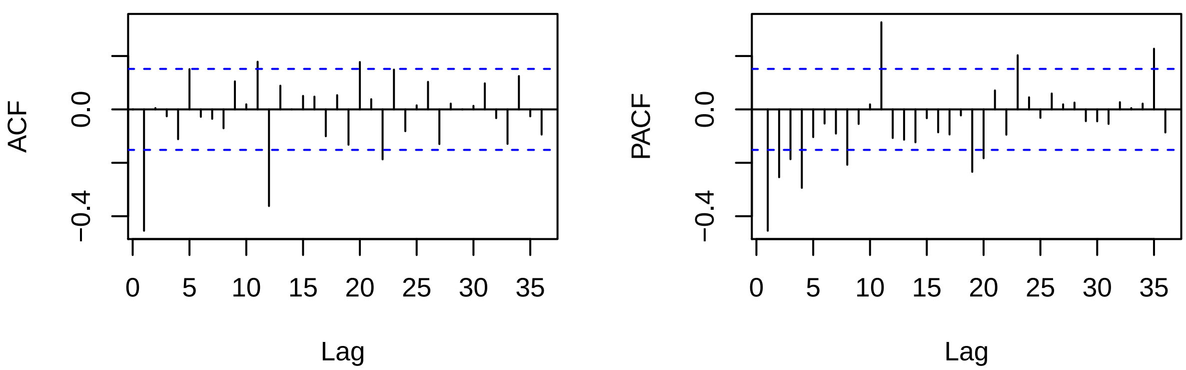
The boxplot suggests a clear monthly variation. There is an increase in traffic fatalities from February to August followed by a small decrease from September to December. The cases of fatalities are highest in August and lowest in February. Calculating the precise percentages in R, we get, **on average, the traffic fatalities are 38% lower in February and 33% higher in August**.

The time series modeling is accomplished using the following steps:

1) Apply a suitable transformation to control non-constant variance.
2) Apply differencing to remove any trends (non-seasonal or seasonal) to make the series stationary.
3) Identify the dependence orders of seasonal ARIMA and coefficient estimation.
4) Run diagnostic checks to choose the best model.
5) Forecast using the chosen model.

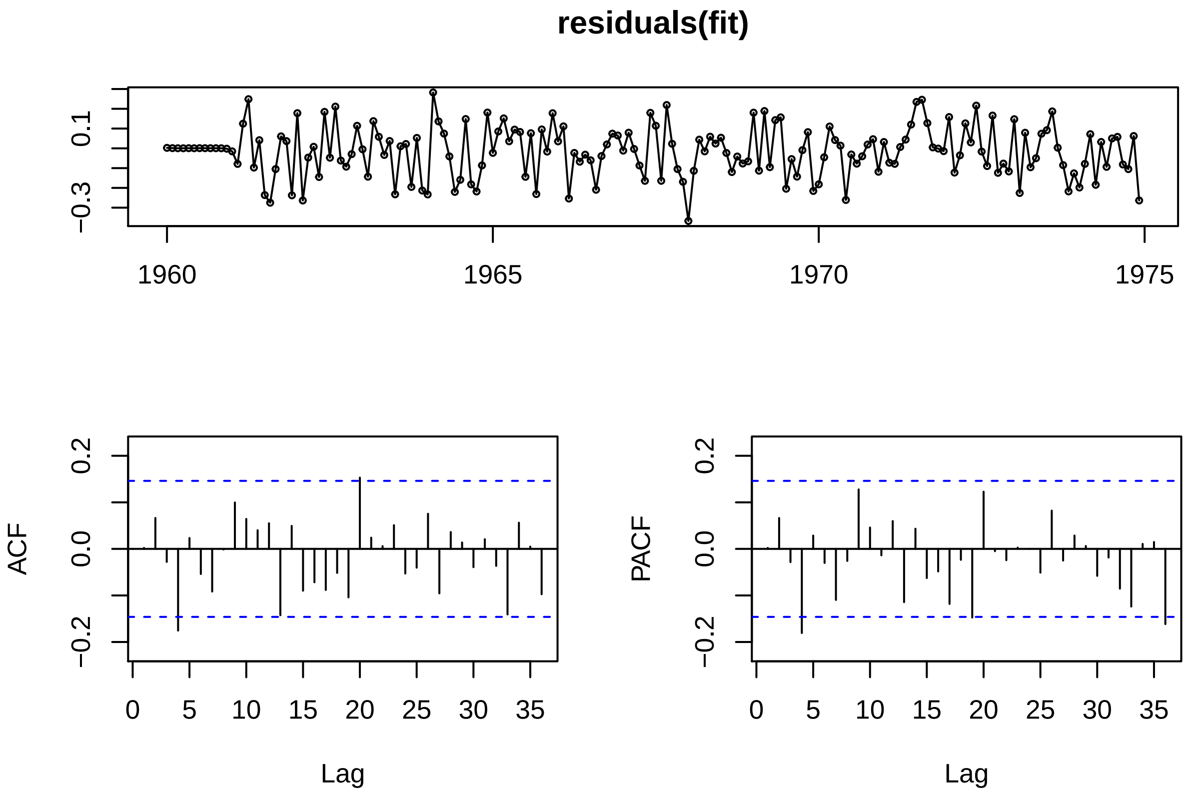
As the first step to start the analysis, we check for constant variance. The series, however, shows a slight linear and strong seasonal trend which hides any effect, so we first difference the original series seasonally and non-seasonally,





The differenced series look stationary with nearly constant variance. Therefore, a transformation is not required.

Our aim now is to find an appropriate seasonal ARIMA model based on the ACF and PACF plots shown above. The significant negative spike at lag 1 in the ACF suggests a non-seasonal MA(1) component and a significant spike at lag 12 in the ACF suggests a seasonal MA(1) component. Consequently, we begin with an ARIMA(0,1,1)(0,1,1)12 model, indicating a first and seasonal difference, and non-seasonal and seasonal MA(1) components.The residuals for the fitted model are as shown below,

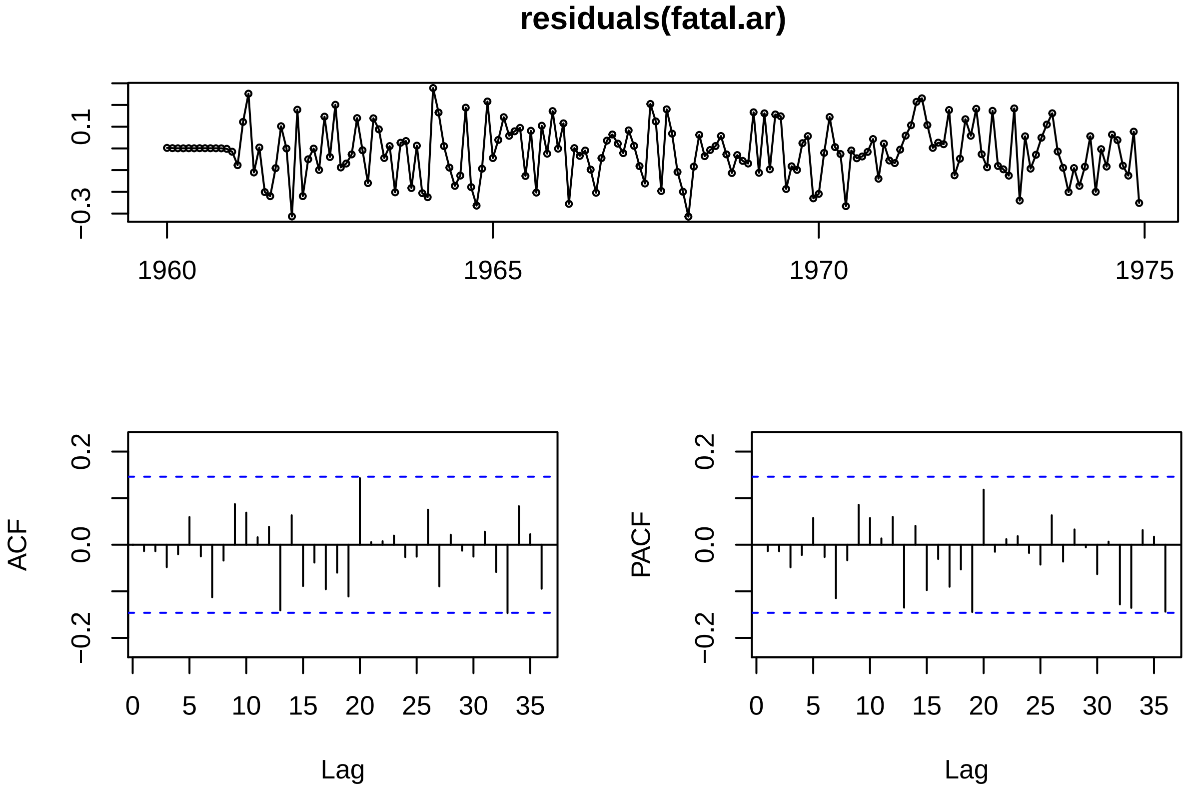


The ACF plot looks fine with a few significant values and the PACF plot has a significant spike at lag 4 which indicates a non-seasonal AR(4) term.

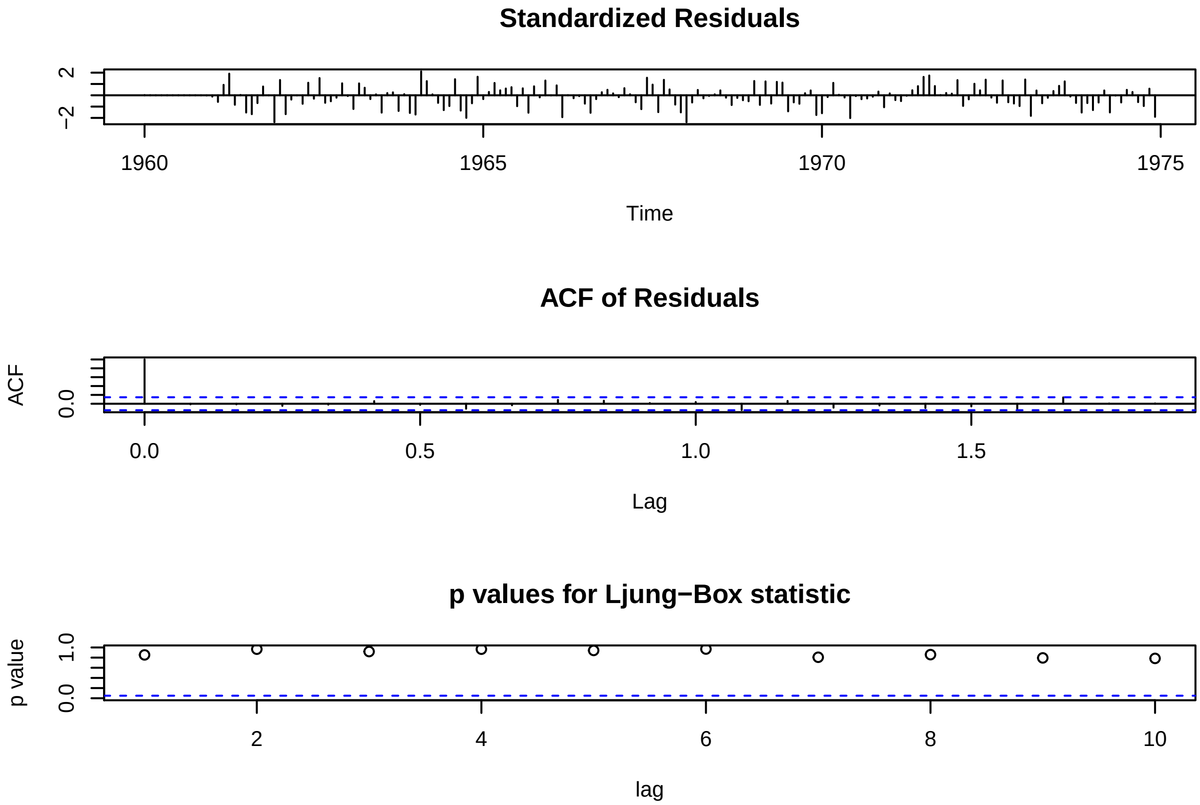
Consequently, this initial analysis suggests that a possible model for these data is an ARIMA(4,1,1)(0,1,1)12. The following table gives the summary of the fitted model,

```
## Series: fatal.ts
## ARIMA(4,1,1)(0,1,1)[12]
## Box Cox transformation: lambda= 0
##
## Coefficients:
##          ar1     ar2      ar3      ar4      ma1     sma1
##       0.0078  0.0898  -0.0096  -0.1733  -0.8083  -0.8844
## s.e.  0.1209  0.1101   0.0938   0.0880   0.1029   0.0979
##
## sigma^2 estimated as 0.01738:  log likelihood=94.77
## AIC=-175.55   AICc=-174.84   BIC=-153.72
```

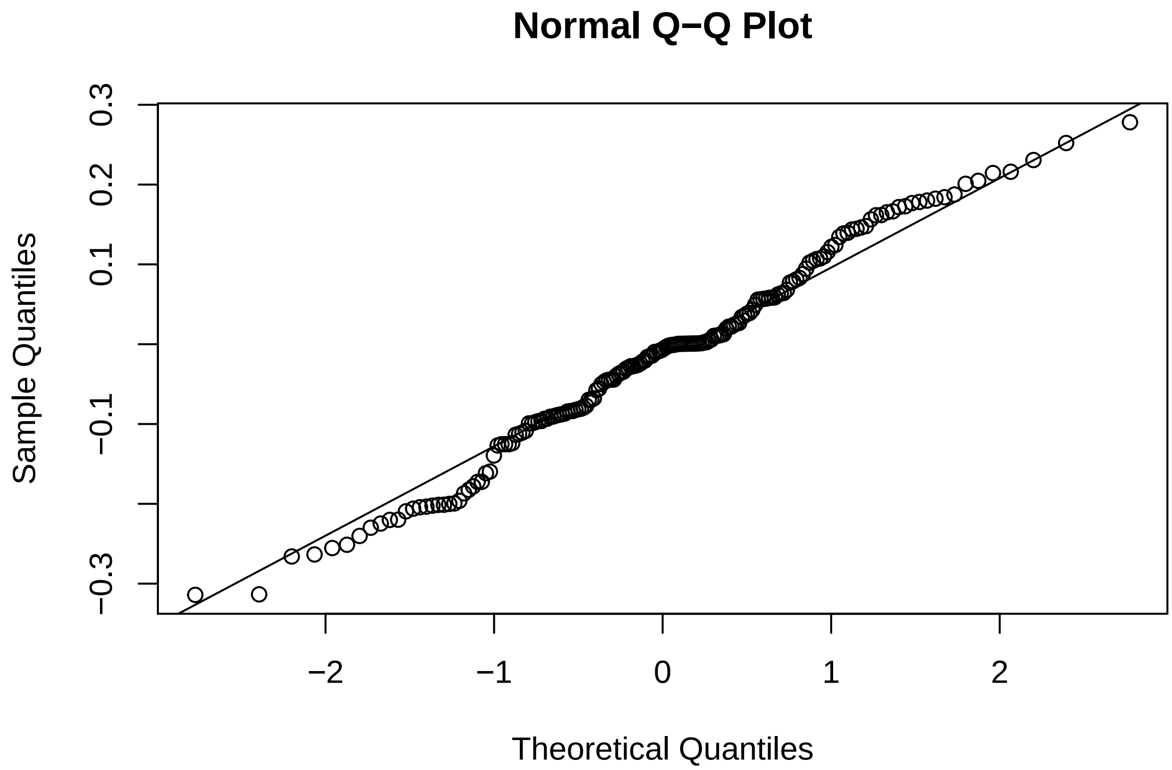
Now, we plot the ACF and PACF plots for the residuals both of which look fine.



After estimating the parameters of this model, we assess their adequacy by analyzing their residuals. Hence, running the diagnostic check on seasonal ARIMA(4,1,1)(0,1,1)12 gives the following plots,



The standardized residuals plot suggests the standardized residuals estimated from this model should behave as an independent and identically distributed sequence with a mean of zero and a constant variance. The ACF plot suggests that the residuals did not deviate significantly from a zero mean white noise process. The last panel in the plot shows p-values for the Ljung-Box statistic. Given the high p-values associated with the statistics, we cannot reject the null hypothesis of independence in this residual series. Thus, we can say that the ARIMA (4,1,1)(0,1,1)12 model fits the data well.

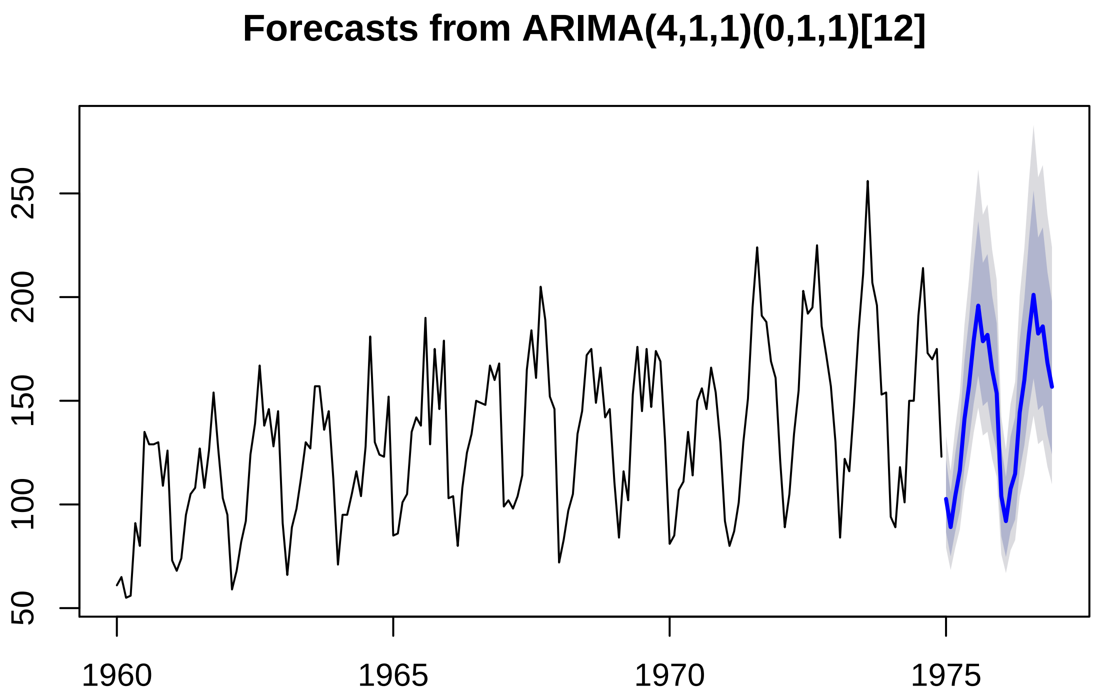


The Q-Q plot suggests that the residuals are approximately normally distributed. In addition, the Shapiro-Wilk test gives no reason to reject the assumption that the distribution of residuals is normal (p-value 0.1247). However, the residuals deviate from normality as we move towards either extremities.

So we now have a seasonal ARIMA model that passes the required checks and is ready for forecasting. The seasonal ARIMA(4,1,1)(0,1,1)12 model is used to forecast the traffic fatalities for the next two years (January 1975- December 1976). The predicted values with 80% and 95% prediction bounds are :

```
##          Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## Jan 1975      102.65909  86.65254 121.6224  79.21570 133.0404
## Feb 1975       89.08179  74.94156 105.8900  68.38882 116.0360
## Mar 1975      104.05031  86.96417 124.4934  79.08625 136.8944
## Apr 1975      116.28068  96.87132 139.5789  87.94473 153.7465
## May 1975      141.08682 117.51980 169.3799 106.68228 186.5867
## Jun 1975      157.48994 130.87726 189.5141 118.66129 209.0242
## Jul 1975      179.02730 148.53109 215.7850 134.55029 238.2067
## Aug 1975      195.88078 162.14301 236.6385 146.70351 261.5430
## Sep 1975      178.70216 147.49243 216.5159 133.24214 239.6724
## Oct 1975      181.78421 149.68550 220.7662 135.05590 244.6802
## Nov 1975      164.93798 135.47681 200.8059 122.07520 222.8506
## Dec 1975      153.61591 125.88987 187.4483 113.29985 208.2778
## Jan 1976      103.99031  84.69568 127.6805  75.97624 142.3338
## Feb 1976       91.95371  74.68366 113.2173  66.89605 126.3974
```

```
## Mar 1976      107.46271  87.01855 132.7100  77.82110 148.3946
## Apr 1976      114.79852  92.70569 142.1563  82.78758 159.1869
## May 1976      144.36704 116.32562 179.1681 103.75871 200.8684
## Jun 1976      160.02081 128.60726 199.1074 114.55727 223.5272
## Jul 1976      182.54172 146.34570 227.6902 130.18774 255.9494
## Aug 1976      201.10297 160.82127 251.4742 142.87461 283.0622
## Sep 1976      182.41032 145.50137 228.6819 129.08997 257.7545
## Oct 1976      185.88299 147.90481 233.6130 131.05048 263.6578
## Nov 1976      168.45859 133.70946 212.2385 118.31781 239.8481
## Dec 1976      156.74005 124.10719 197.9534 109.68014 223.9917
```



The plot above shows the forecasted trend in blue and the upper and lower bounds highlighted in grey.

Conclusion: In this study, the ARIMA (4,1,1)(0,1,1)12 model, we showed that the number of traffic fatalities in a given month can be estimated by the number of fatality cases occurring 1 (Q = 1), 4 (p = 4) and 12 (S = 12 and P = 0) months prior, and we found that a moving-average component of order q equal to 1 is adequate for the data.

R Code

```r
# Load the data
fatalities <- read.csv("monthly-traffic-fatalities-in-on.csv")

# convert the data to a time series object
fatal.ts <- ts(fatalities[, 2], start = c(1960, 1), end = c(1974, 12), frequency = 12)

# plot the data
plot(fatal.ts, type = "o", ylab = "Monthly traffic fatalities", xlab = "Year",
    ylim = c(0, 300))

# trend effect
plot(aggregate(fatal.ts))

# seasonal effect
boxplot(fatal.ts ~ cycle(fatal.ts), ylab = "Cases of traffic fatalities", xlab = "months

# percentage fatalities during Feb and Aug
fatal.Feb <- window(fatal.ts, start = c(1960, 2), freq = TRUE)
fatal.Aug <- window(fatal.ts, start = c(1960, 8), freq = TRUE)
Feb.ratio <- mean(fatal.Feb)/mean(fatal.ts)
Feb.ratio
Aug.ratio <- mean(fatal.Aug)/mean(fatal.ts)
Aug.ratio

# ACF, PACF
par(mfrow = c(1, 2))
acf(fatal.ts)
pacf(fatal.ts)

# Differencing
tsdisplay(diff(diff(fatal.ts), 12))

# fitting the seasonal ARIMA model (0,1,1)(0,1,1)12
fit <- Arima(fatal.ts, order = c(0, 1, 1), seasonal = c(0, 1, 1), lambda = 0)
tsdisplay(residuals(fit))

# fitting the seasonal ARIMA model (4,1,1)(0,1,1)12
fatal.ar <- Arima(fatal.ts, order = c(4, 1, 1), seasonal = c(0, 1, 1), lambda = 0)
fatal.ar

# ACF, PACF of residuals
tsdisplay(residuals(fatal.ar))
```

```r
# diagnostic plots
tsdiag(fatal.ar)
qqnorm(fatal.ar$residuals)
qqline(fatal.ar$residuals)
shapiro.test(fatal.ar$residuals)

# 2-year forecasting
forecast(fatal.ar, h = 24)
plot(forecast(fatal.ar, h = 24))
```
