POLLUTANT LEVELS IN SAN DIEGO

TIME SERIES ANALYSIS

Sneha Thanasekaran
Anuj Mathur

DATASET

Source: data.world

Air Pollution in the U.S. since 2000-2011.

Includes four major pollutants

Nitrogen Dioxide

Sulphur Dioxide

Carbon Monoxide

and Ozone)

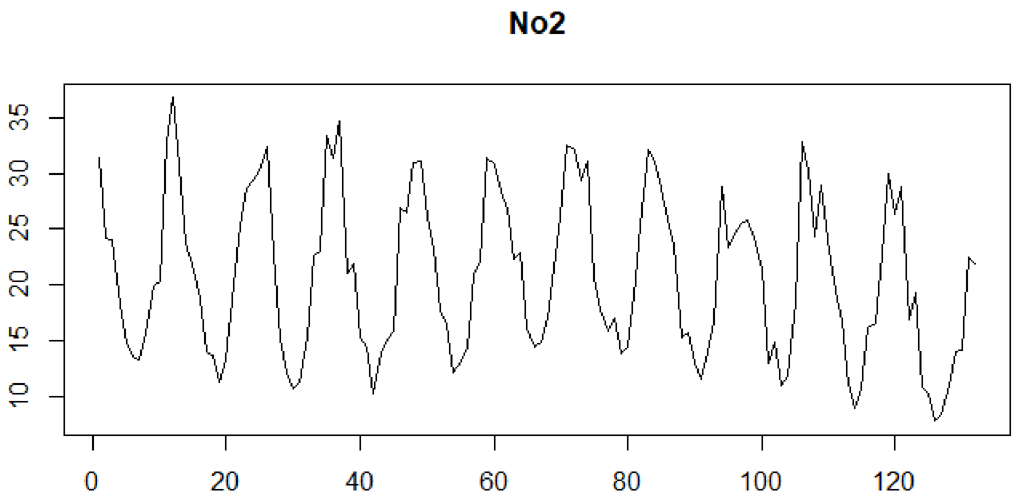
Our Focus is on:

- **City: San Diego**
- **Mean : The arithmetic mean of concentration of NO2, O3, SO2 and O3 within a given Month of the year**

OUR MODELS

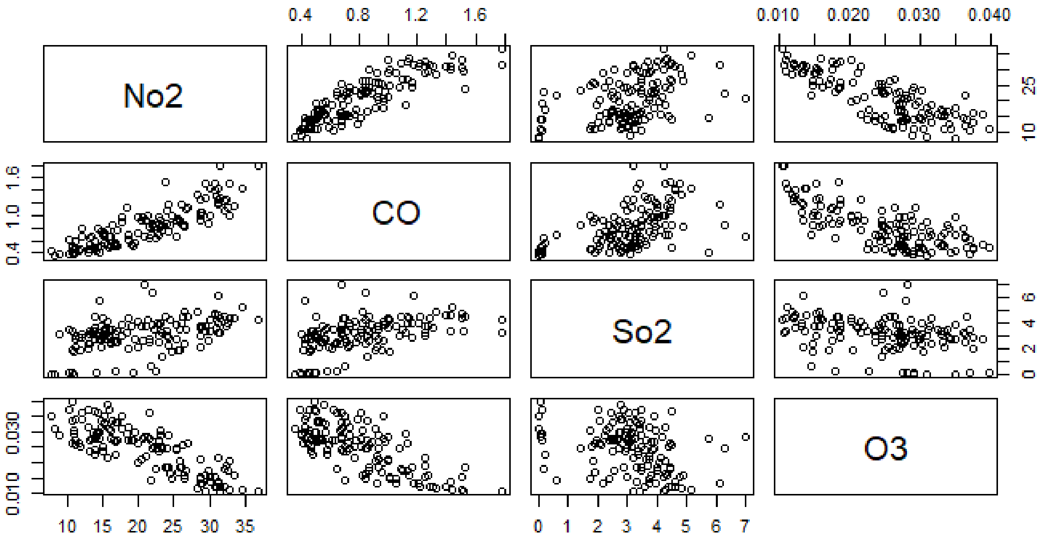
UNIVARIATE

- The arithmetic mean of concentration of NO2within a given Month of the year



MULTIVARIATE

- The arithmetic mean of concentration of NO2, O3, SO2 and O3 within a given Month of the year



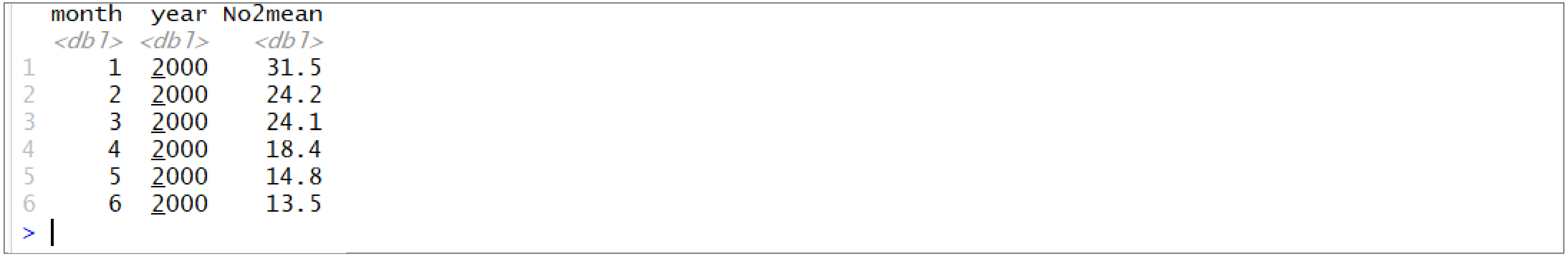
PREPARING THE DATASET

```
data <- read.csv('uspollution_pollution_us_2000_2011.csv’)
dataSD <- data[data$City == "San Diego",]
df <- data.frame(date = dataSD$Date.Local,
                year = year(dataSD$Date.Local),
                month = month(dataSD$Date.Local))
datadate <- cbind(df,dataSD)
testdatadate <- datadate[datadate$year == 2011,]
datadate <- datadate[datadate$year != 2011,]
unique(datadate$year)
unique(testdatadate$year)
```

```
[1] 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
[1] 2011
```

PREPARING THE DATASET

```
df_no2 <- datadate %>%
  mutate(norm = mean(NO2.Mean)) %>%
  group_by(month,year) %>%
  dplyr::summarize(No2mean =mean(NO2.Mean)) %>%
  arrange(year, month)
head(df_no2)
```



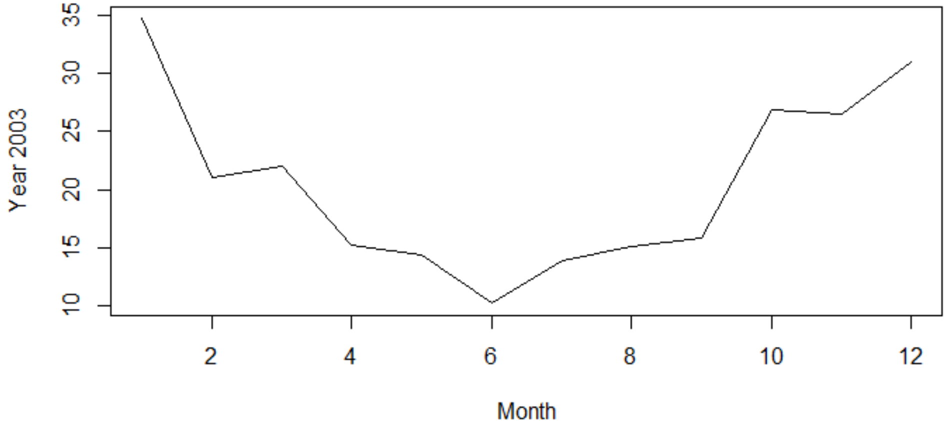
```
  month  year No2mean
  <dbl> <dbl>   <dbl>
1     1  2000    31.5
2     2  2000    24.2
3     3  2000    24.1
4     4  2000    18.4
5     5  2000    14.8
6     6  2000    13.5
> 
```

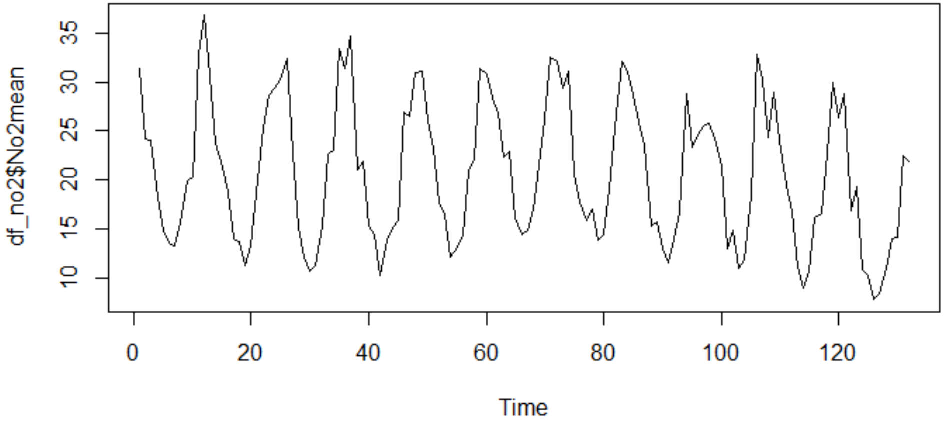
STATIONARITY AND SEASONALITY

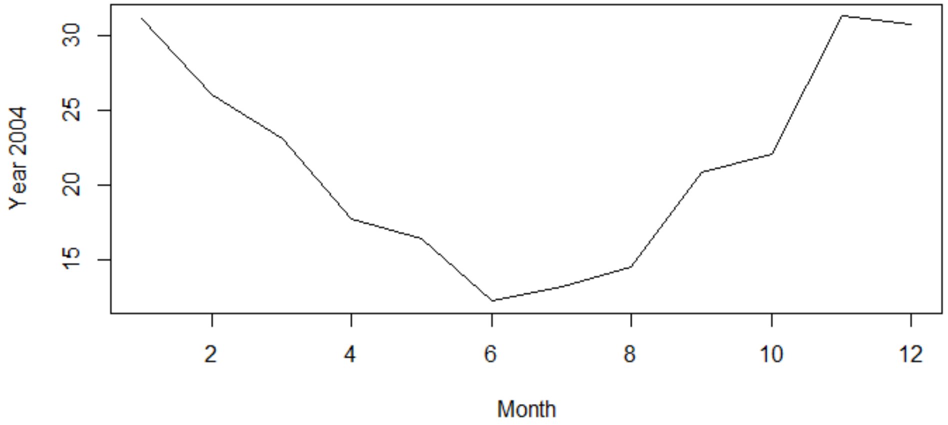
No2 Mean – Time Series Plot

Factor for pattern:

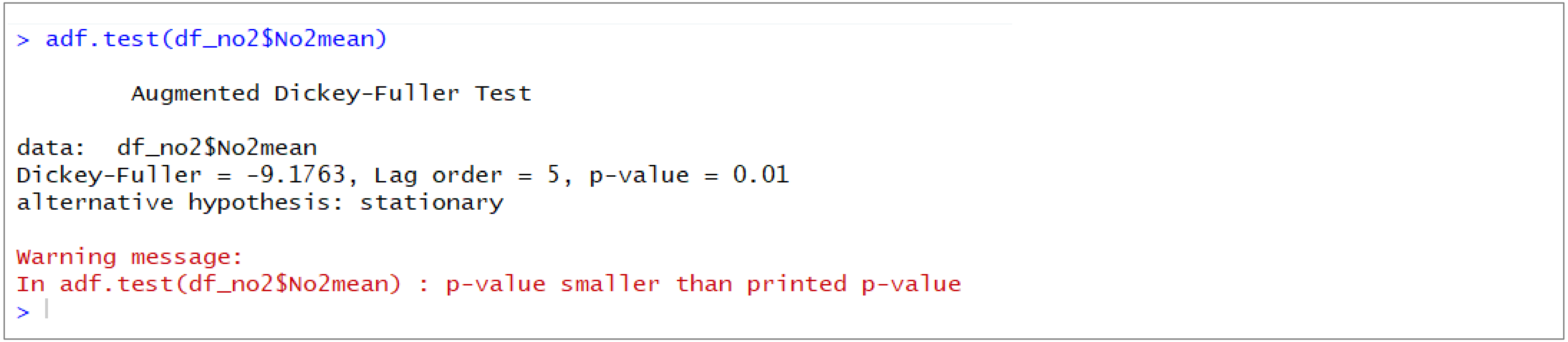
- Weather
- Traffic





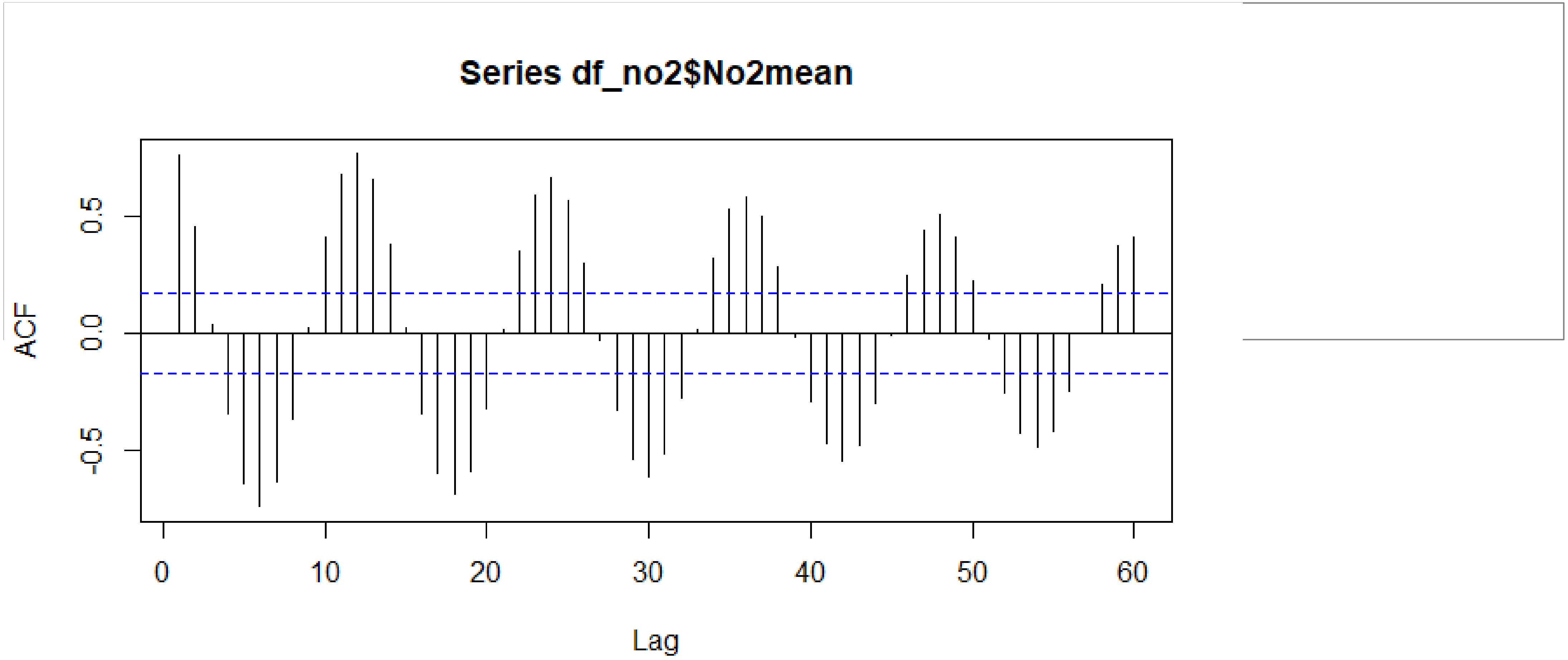


TEST FOR STATIONARITY



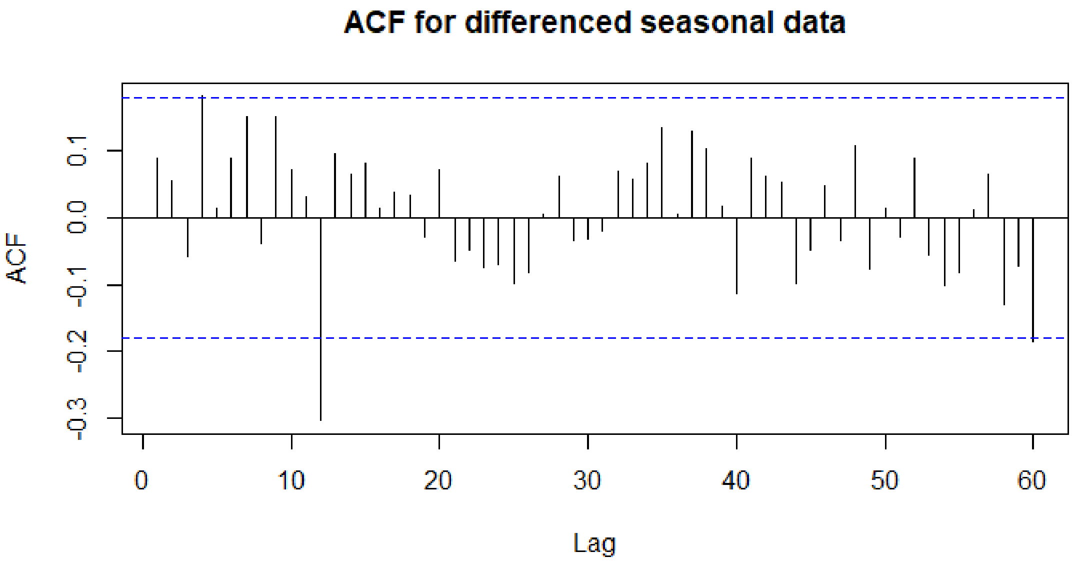
```
> adf.test(df_no2$No2mean)

	Augmented Dickey-Fuller Test

data:  df_no2$No2mean
Dickey-Fuller = -9.1763, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(df_no2$No2mean) : p-value smaller than printed p-value
> 
```

TEST FOR SEASONALITY

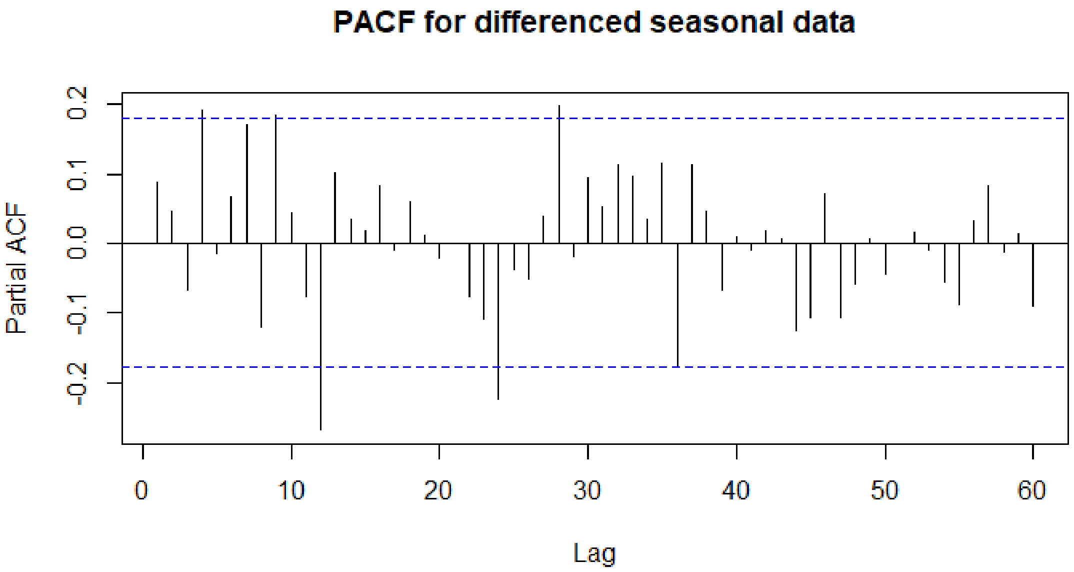


TAKING DIFFERENCE FOR SEASONALITY

```
diffseasonal = diff(df_no2$No2mean,12)

par(mfrow=c(1,2))

acf(diffseasonal,main='ACF for differenced seasonal data',lag.max=60) #MA1

pacf(diffseasonal,main='PACF for differenced seasonal data', lag.max=60) #AR3
```



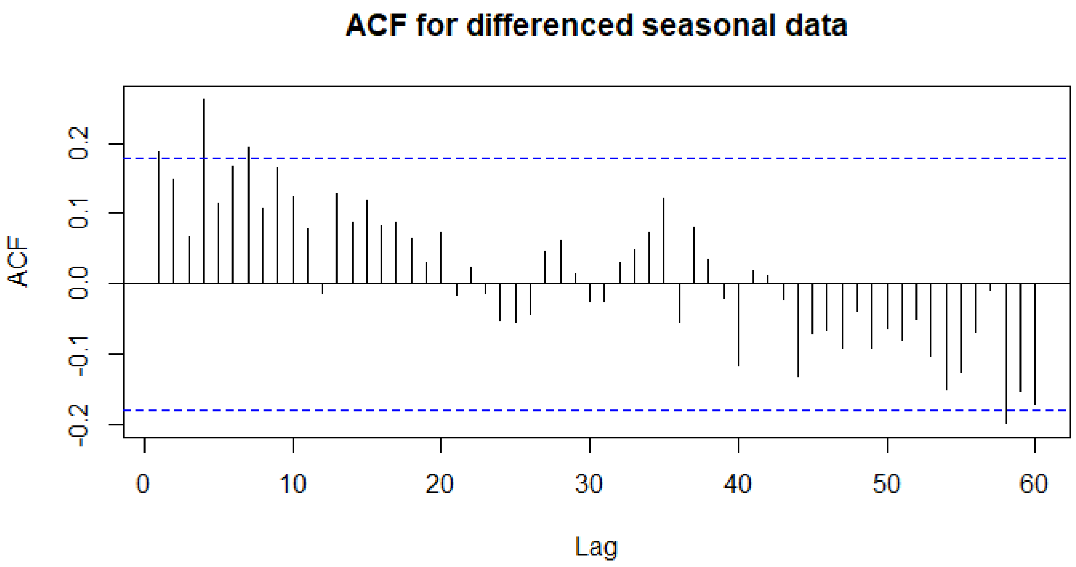
ACF for differenced seasonal data

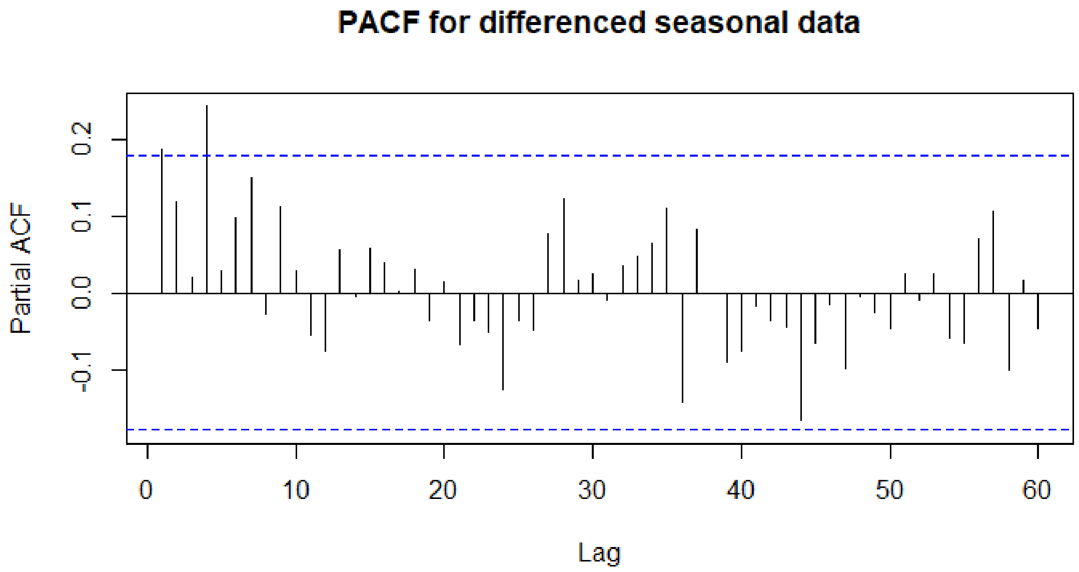


PACF for differenced seasonal data

MODELLING THE SEASONALITY

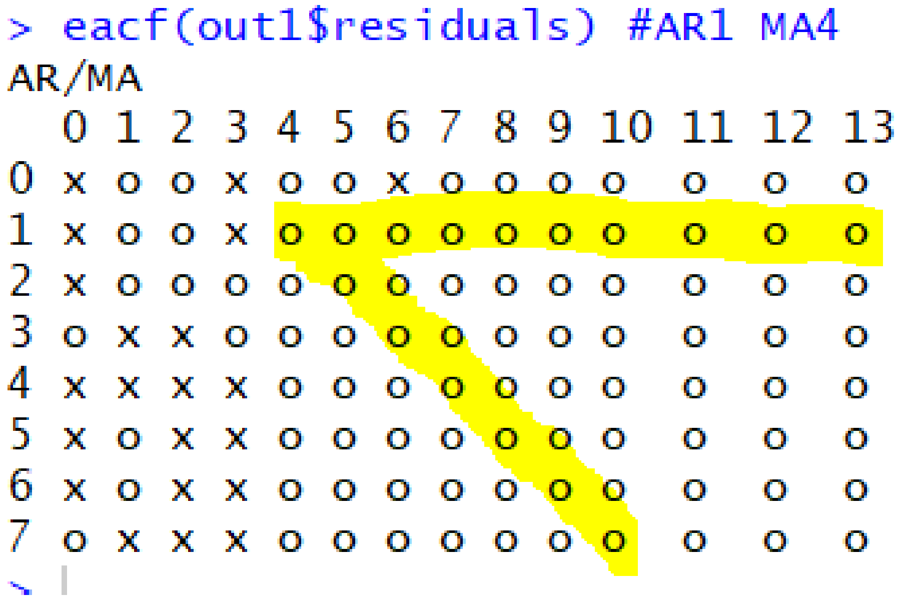
```
out1=arima(diffseasonal,order=c(0,0,0),seasonal=list(order=c(3,0,0),period=12))

par(mfrow=c(1,2))

acf(out1$residuals,main='ACF for differenced seasonal data',lag.max=60) #MA4

pacf(out1$residuals,main='PACF for differenced seasonal data', lag.max=60) #AR4
```





TAKING DIFFERENCE FOR SEASONALITY

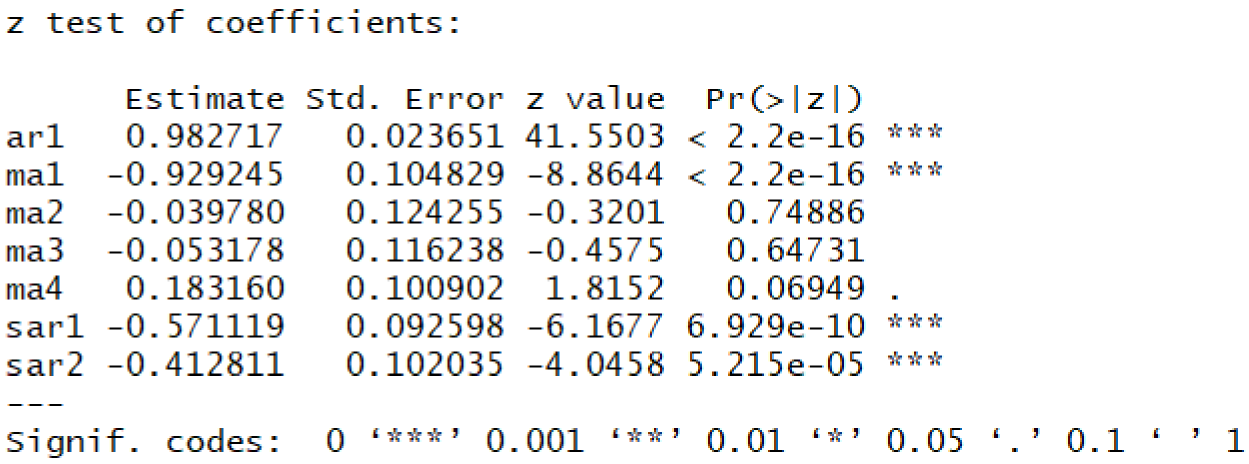
```
eacf(out1$residuals) #AR1 MA4
```



```
> eacf(out1$residuals) #AR1 MA4
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o x o o x o o o o  o  o  o 
1 x o o x o o o o o o o  o  o  o 
2 x o o o o o o o o o o  o  o  o 
3 o x x o o o o o o o o  o  o  o 
4 x x x x o o o o o o o  o  o  o 
5 x o x x o o o o o o o  o  o  o 
6 x o x x o o o o o o o  o  o  o 
7 o x x x o o o o o o o  o  o  o 
> 
```

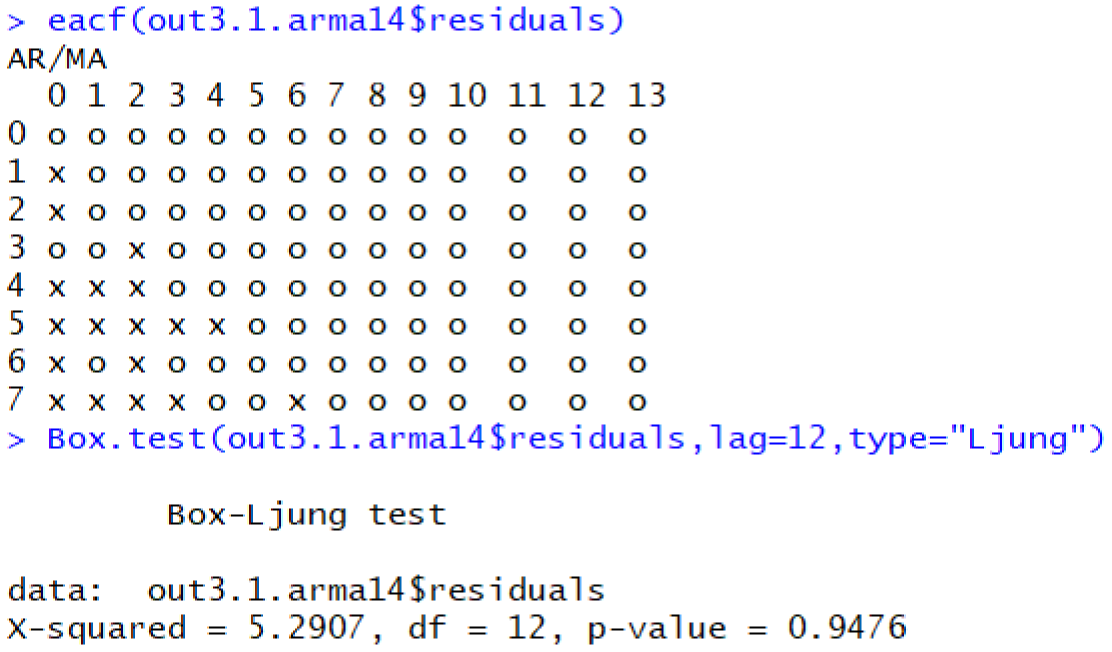
FINALIZING THE MODEL

```
out3.1.arma14=arima(df_no2$No2mean,order=c(1,0,4),seasonal=list(order=c(2,1,0),period=12))

out3.1.arma14

acf(out3.1.arma14$residuals,main='ACF for differenced seasonal differenced data',lag.max=60)

pacf(out3.1.arma14$residuals,main='PACF for differenced seasonal differenced data', lag.max=60)

coeftest(out3.1.arma14)
```



```
z test of coefficients:

      Estimate Std. Error z value  Pr(>|z|)    
ar1   0.982717   0.023651 41.5503 < 2.2e-16 ***
ma1  -0.929245   0.104829 -8.8644 < 2.2e-16 ***
ma2  -0.039780   0.124255 -0.3201   0.74886    
ma3  -0.053178   0.116238 -0.4575   0.64731    
ma4   0.183160   0.100902  1.8152   0.06949 .  
sar1 -0.571119   0.092598 -6.1677 6.929e-10 ***
sar2 -0.412811   0.102035 -4.0458 5.215e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

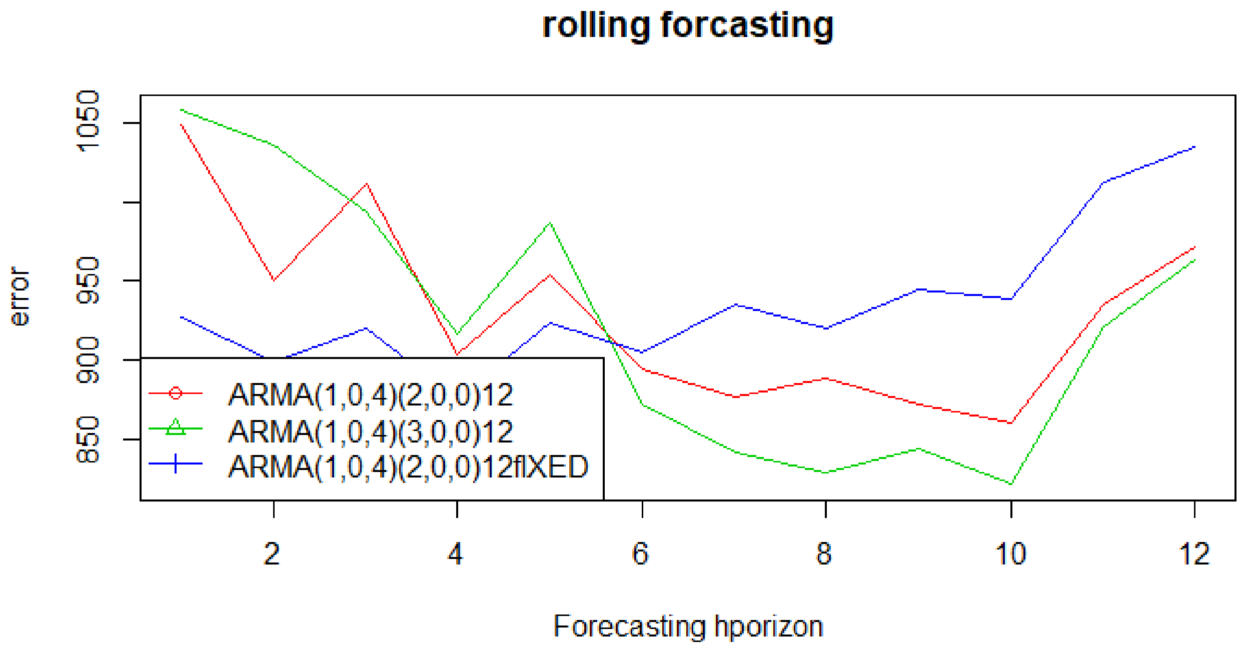
FINALIZING THE MODEL



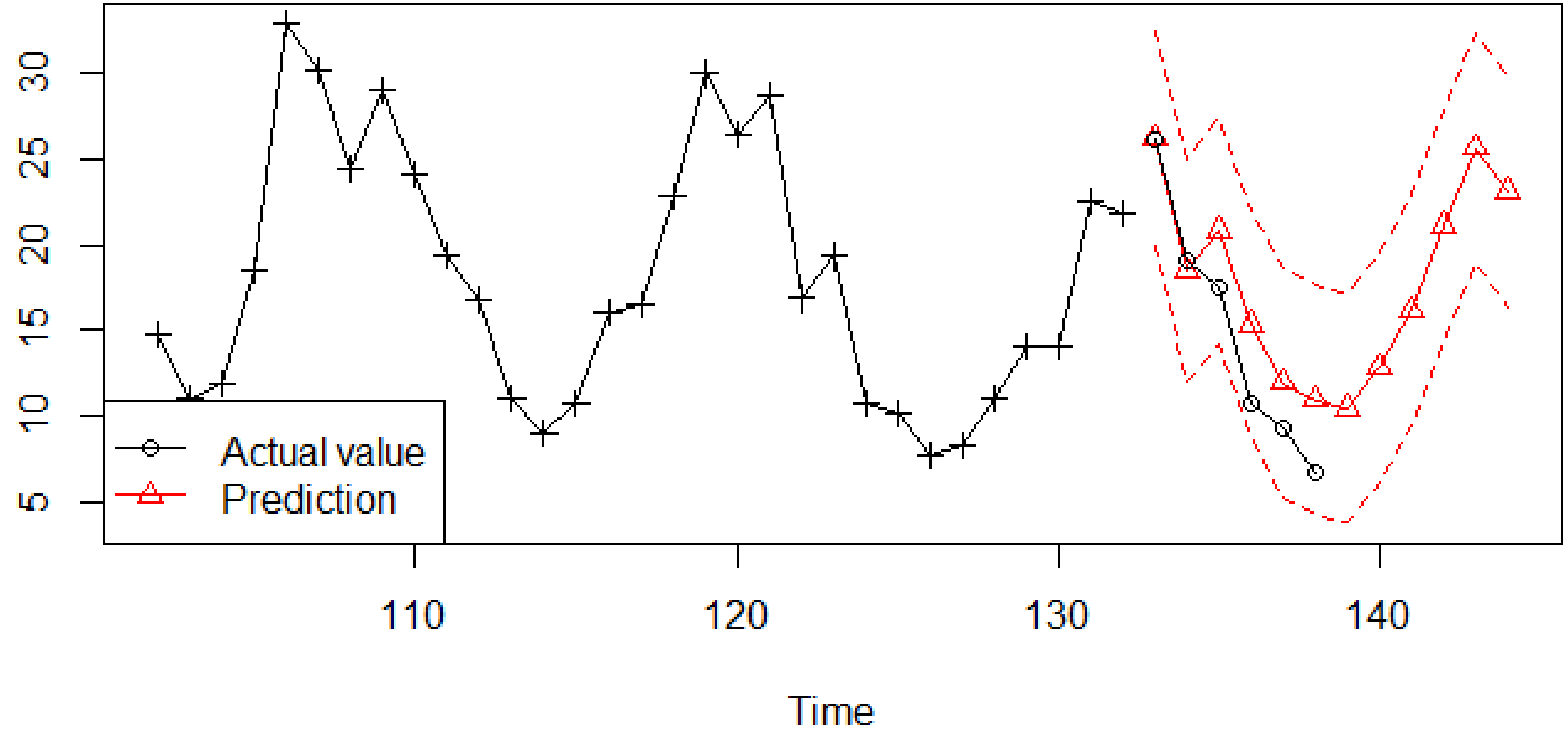
```
> eacf(out3.1.arma14$residuals)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 o o o o o o o o o o o  o  o  o 
1 x o o o o o o o o o o  o  o  o 
2 x o o o o o o o o o o  o  o  o 
3 o o x o o o o o o o o  o  o  o 
4 x x x o o o o o o o o  o  o  o 
5 x x x x x o o o o o o  o  o  o 
6 x o x o o o o o o o o  o  o  o 
7 x x x x o o x o o o o  o  o  o 
> Box.test(out3.1.arma14$residuals,lag=12,type="Ljung")

        Box-Ljung test

data:  out3.1.arma14$residuals
X-squared = 5.2907, df = 12, p-value = 0.9476
```

CHOOSING THE RIGHT MODEL

```
➢ print(c(out2.1.arma14$aic, out2.1.arma22$aic, out3.1.arma14$aic, out3.1.arma14.fixed$aic))
➢ [1] 619.1342 621.2042 626.1821 622.5653
```



PREDICTING THE NEXT 12 MONTHS



MULTIVARIATE ANALYSIS

Source: data.world

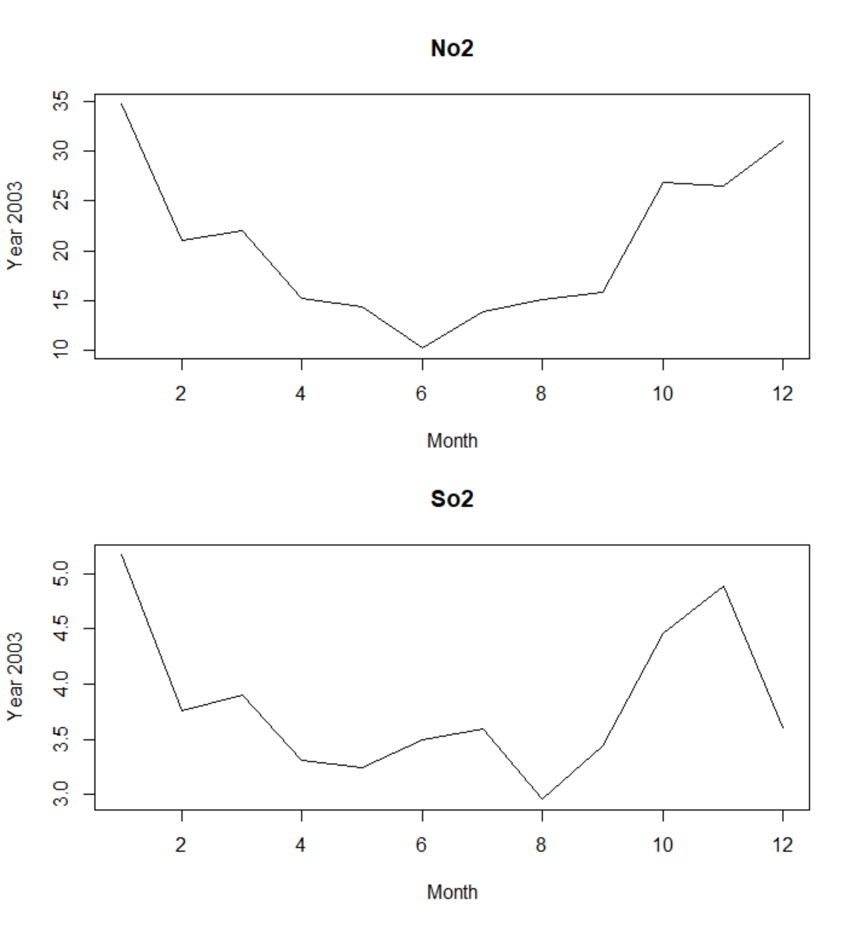
Nitrogen Dioxide

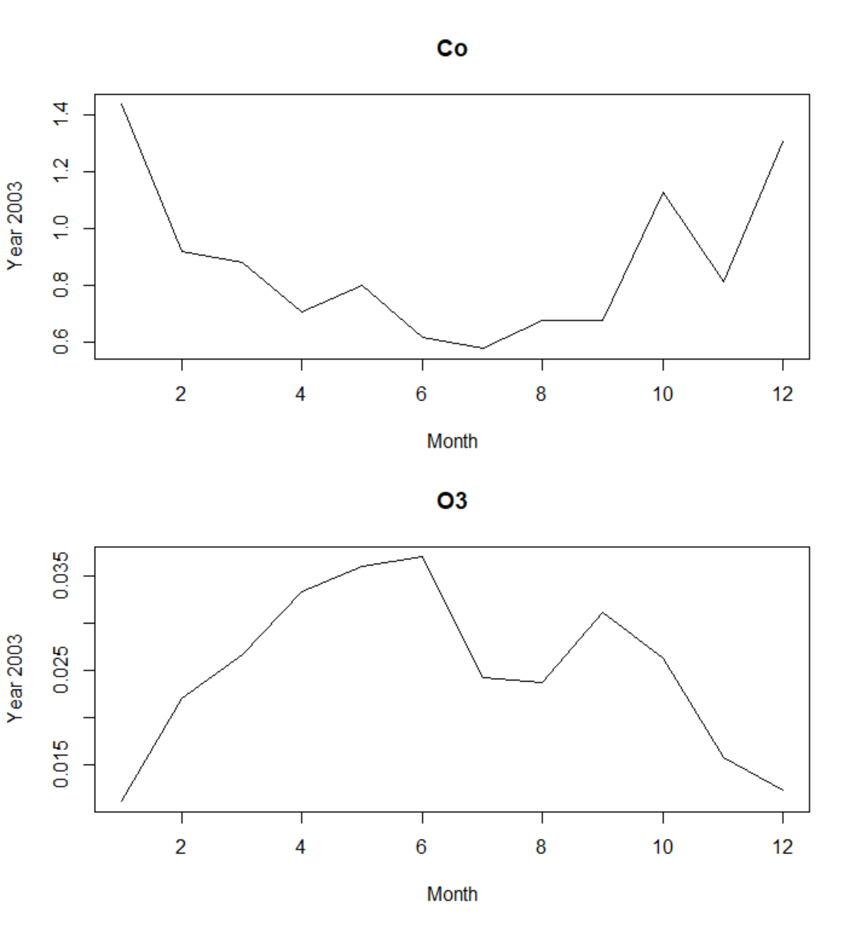
Sulphur Dioxide

Carbon Monoxide

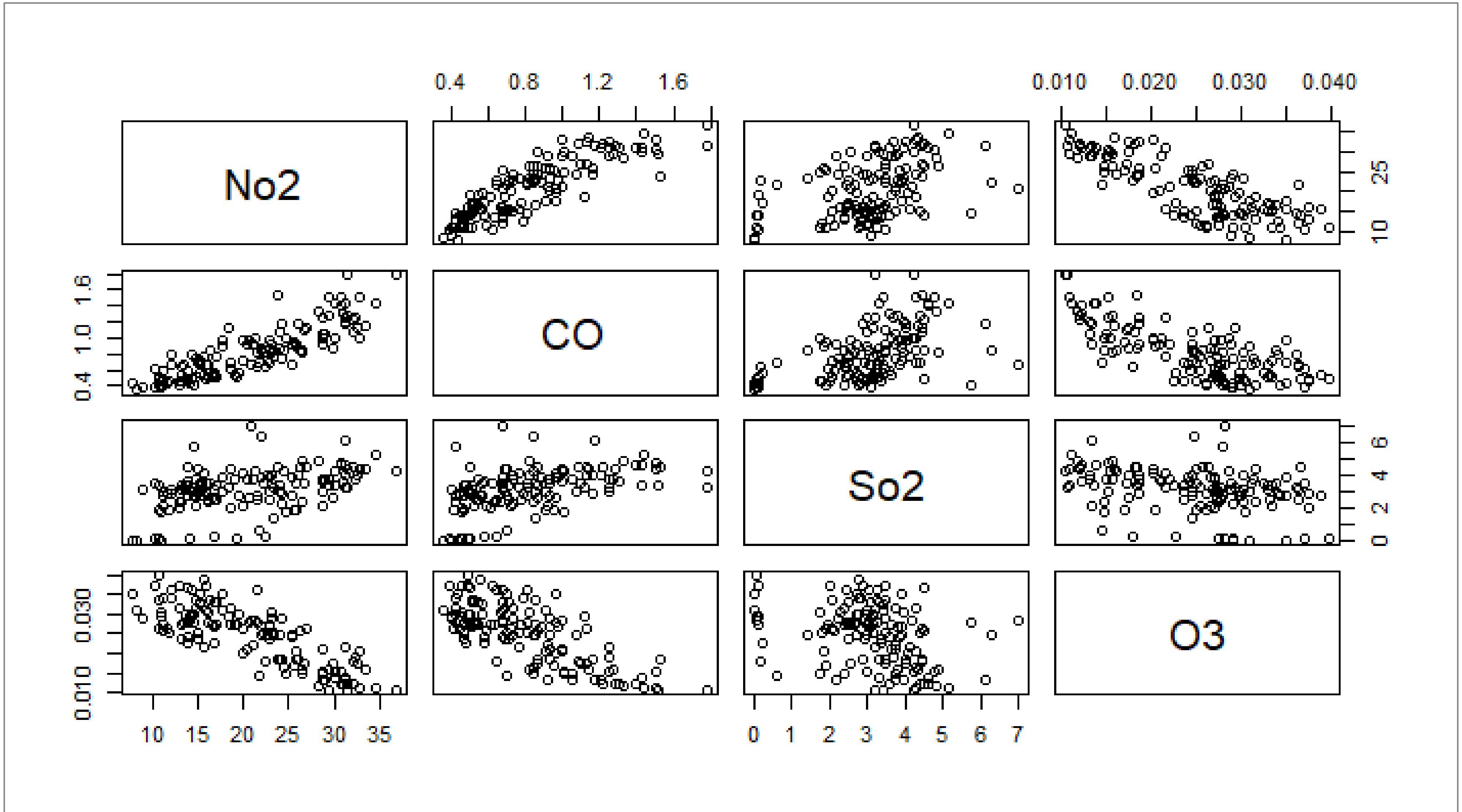
and Ozone

From Weather reports, it is observed for CO, SO2 and NO2 with the maximum concentrations in the winter and the minimum in the summer, while O3 exhibited an opposite trend



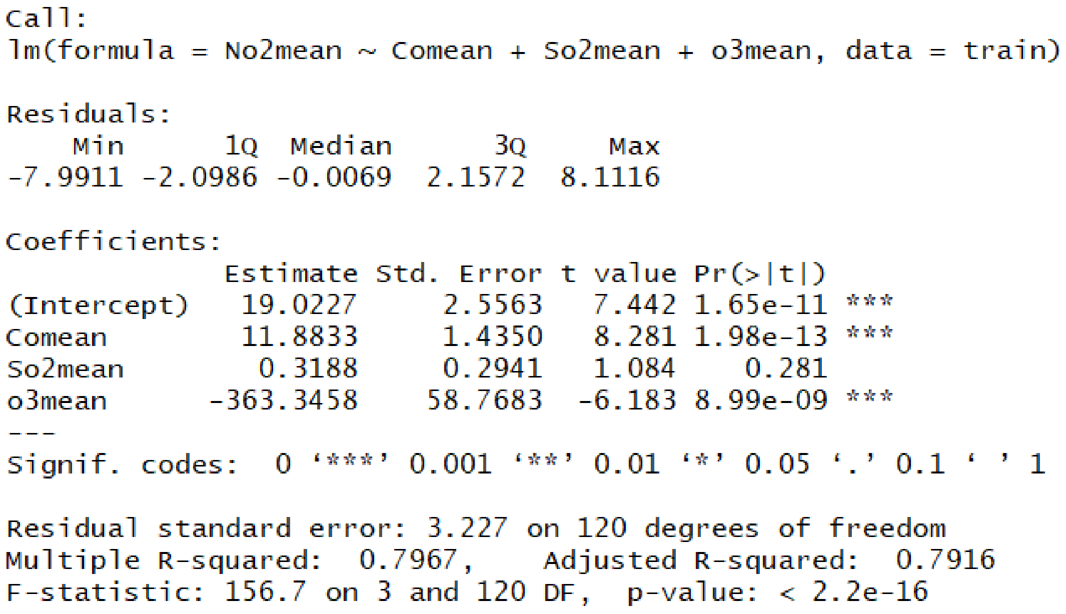


MULTIVARIATE ANALYSIS



LINEAR REGRESSION

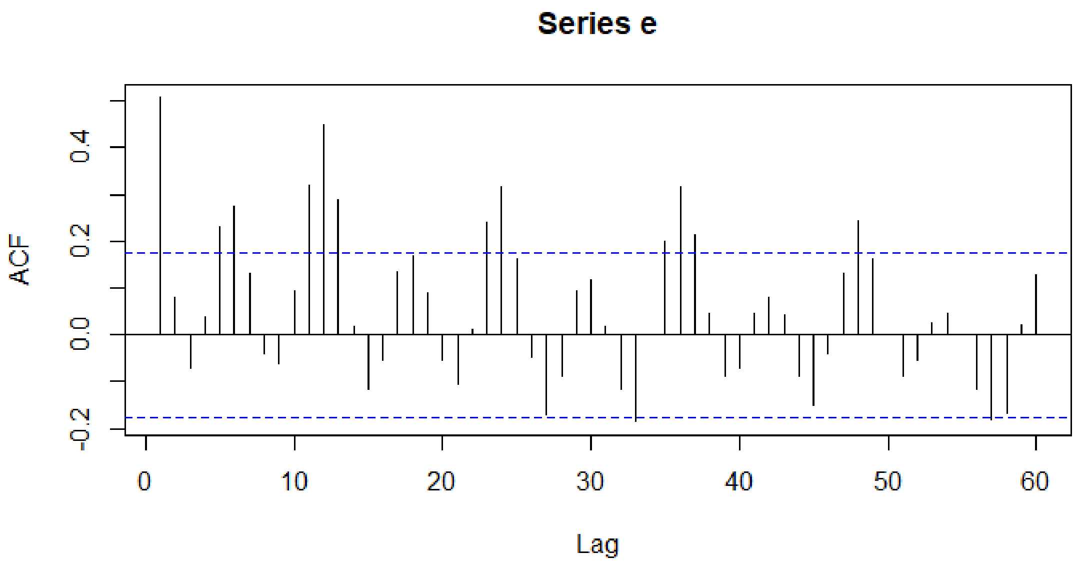
```
data=cbind(df_no2[3], df_co[3], df_So2[3], df_o3[3])
train=as.data.frame(data[1:124,])
new.data=data[125:n,2:4]
lm=lm(No2mean~Comean + So2mean + o3mean,data=train)
summary(lm)
```

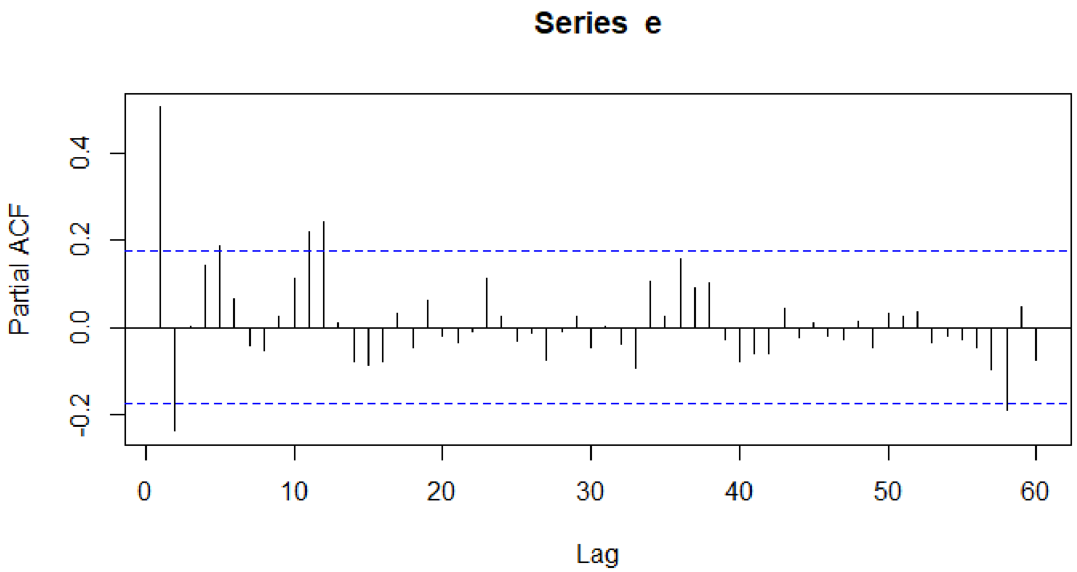


```
Call:
lm(formula = No2mean ~ Comean + So2mean + o3mean, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.9911 -2.0986 -0.0069  2.1572  8.1116 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   19.0227     2.5563   7.442 1.65e-11 ***
Comean        11.8833     1.4350   8.281 1.98e-13 ***
So2mean        0.3188     0.2941   1.084    0.281    
o3mean      -363.3458    58.7683  -6.183 8.99e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.227 on 120 degrees of freedom
Multiple R-squared:  0.7967,	Adjusted R-squared:  0.7916 
F-statistic: 156.7 on 3 and 120 DF,  p-value: < 2.2e-16
```

MODELING MULTIVARIATE ANALYSIS

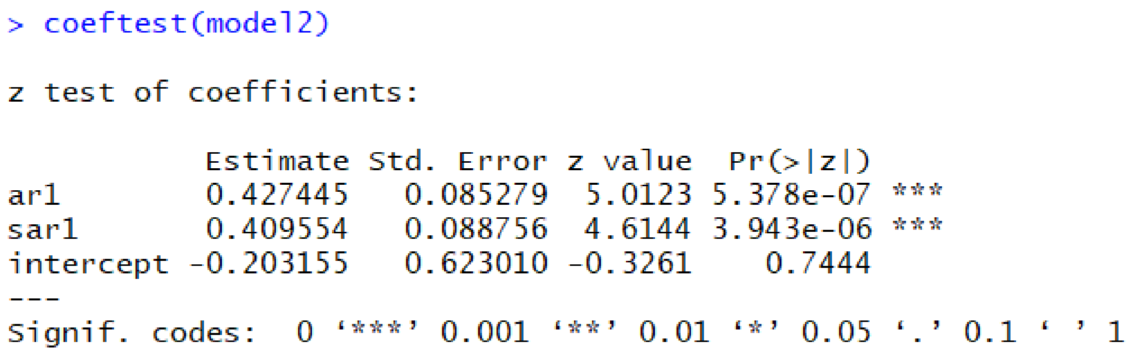
```
e=lm$residuals
plot(e,type='l’)
par(mfrow=c(1,2))
acf(e, lag.max = 60) #MA4 seasonality
pacf(e, lag.max = 60) #AR1
```



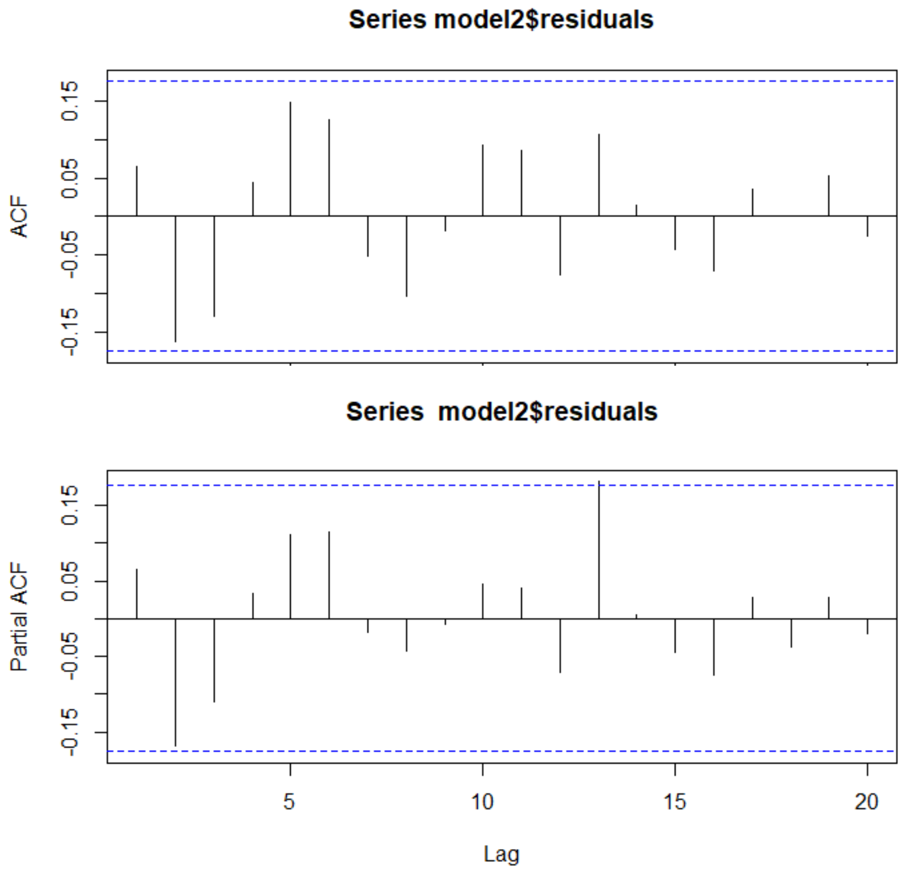


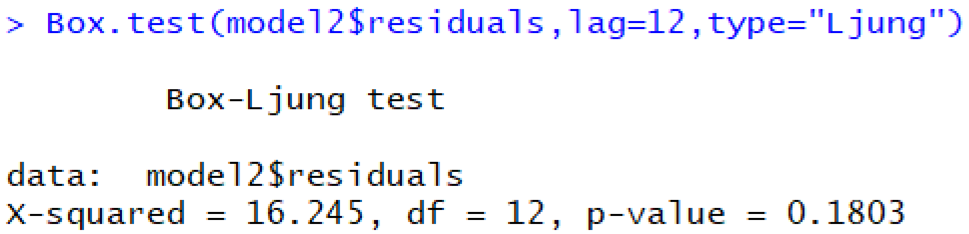
MODELING MULTIVARIATE ANALYSIS

```
model1=arima(e,order=c(0,0,0),seasonal=list(order=c(1,0,0),period=12))

model2=arima(e,order=c(1,0,0),seasonal=list(order=c(1,0,0),period=12))
```



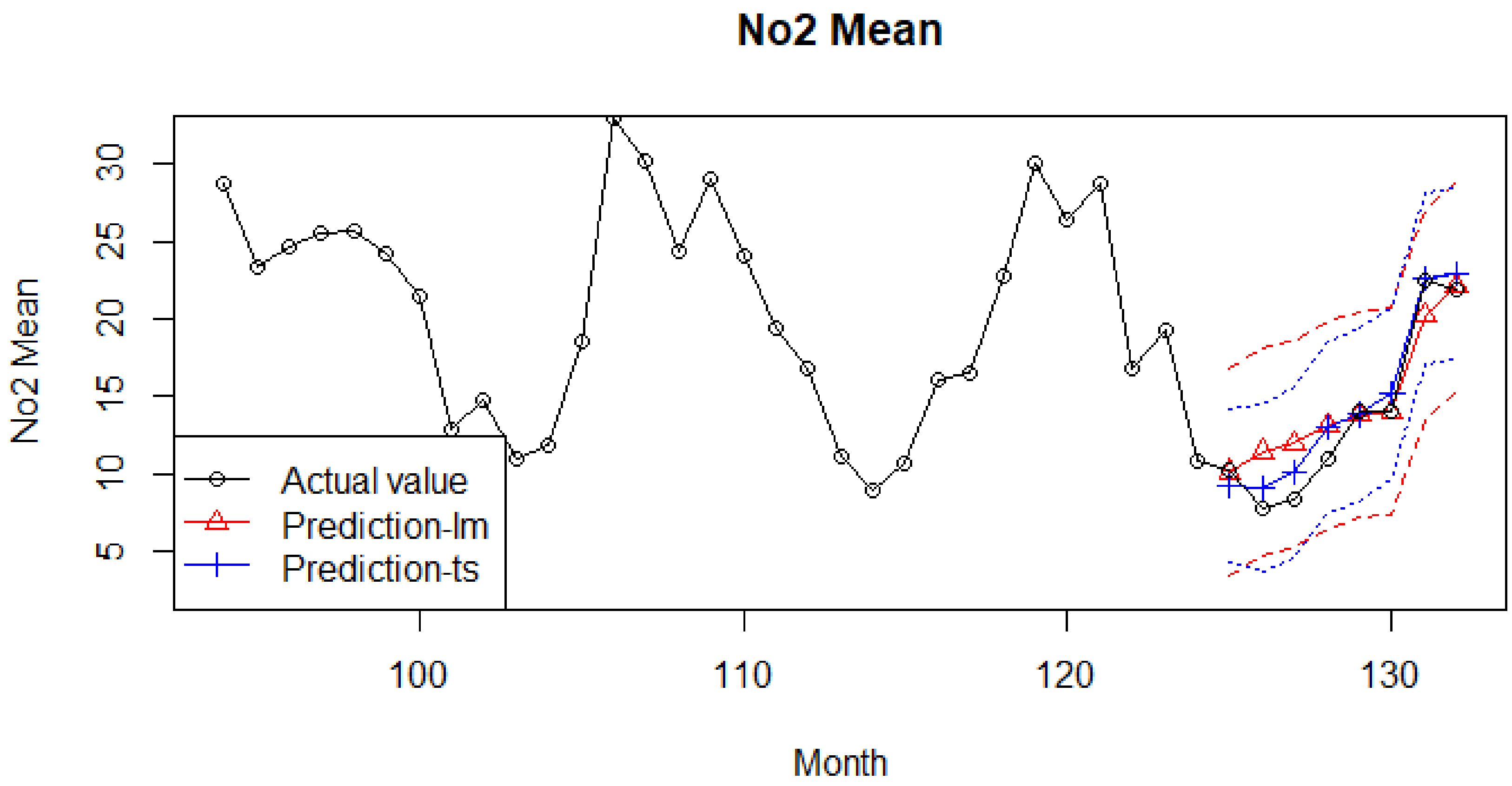
```
> coeftest(model2)

z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)    
ar1        0.427445   0.085279  5.0123 5.378e-07 ***
sar1       0.409554   0.088756  4.6144 3.943e-06 ***
intercept -0.203155   0.623010 -0.3261    0.7444    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```





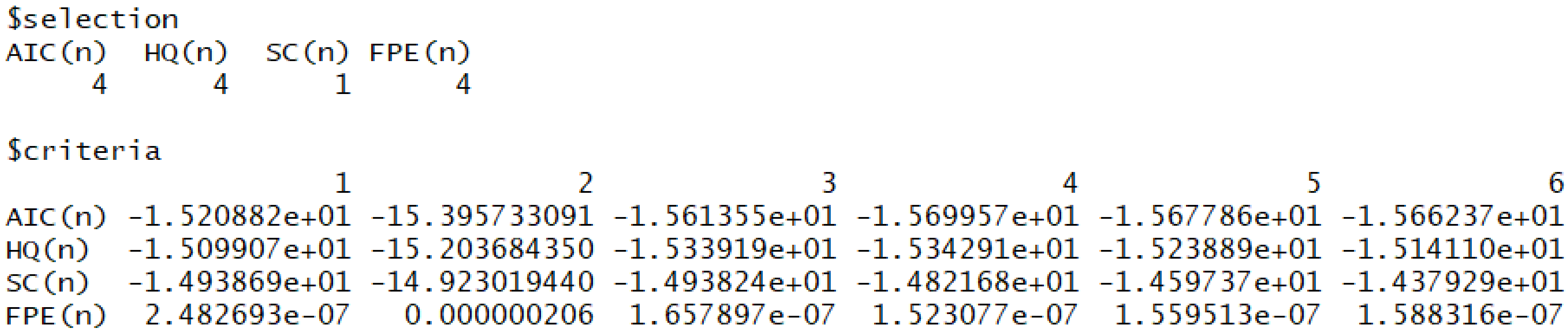
```
> Box.test(model2$residuals,lag=12,type="Ljung")

        Box-Ljung test

data:  model2$residuals
X-squared = 16.245, df = 12, p-value = 0.1803
```

PREDICTING THE NEXT 12 MONTHS



MODELING VAR MODEL

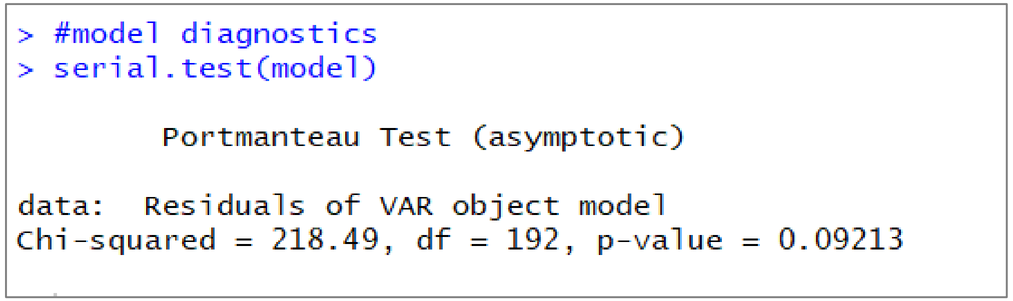
```
#model selection

VARselect(data[,2:4],lag.max=6)
```



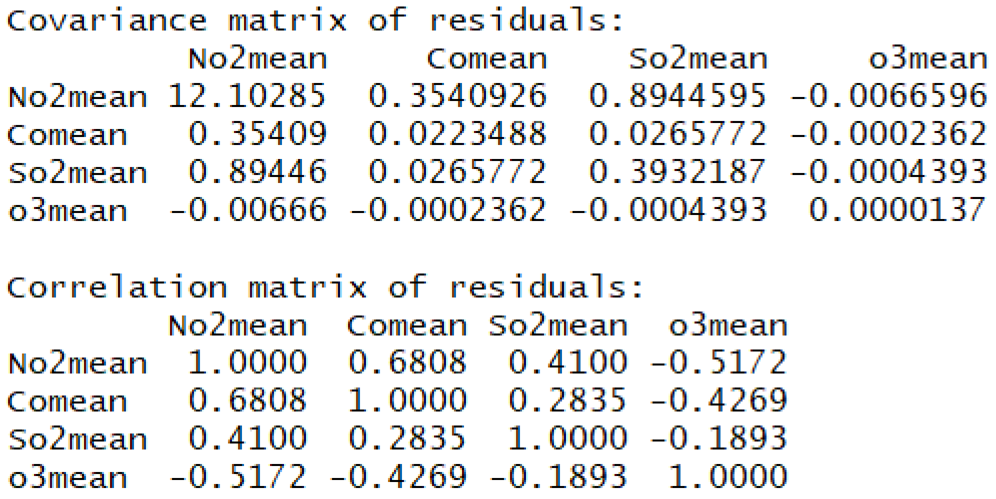
```
$selection
AIC(n)  HQ(n)  SC(n) FPE(n) 
     4      4      1      4 

$criteria
                   1              2             3             4             5             6
AIC(n) -1.520882e+01 -15.395733091 -1.561355e+01 -1.569957e+01 -1.567786e+01 -1.566237e+01
HQ(n)  -1.509907e+01 -15.203684350 -1.533919e+01 -1.534291e+01 -1.523889e+01 -1.514110e+01
SC(n)  -1.493869e+01 -14.923019440 -1.493824e+01 -1.482168e+01 -1.459737e+01 -1.437929e+01
FPE(n)  2.482693e-07   0.000000206  1.657897e-07  1.523077e-07  1.559513e-07  1.588316e-07
```

AUTO CORRELATION

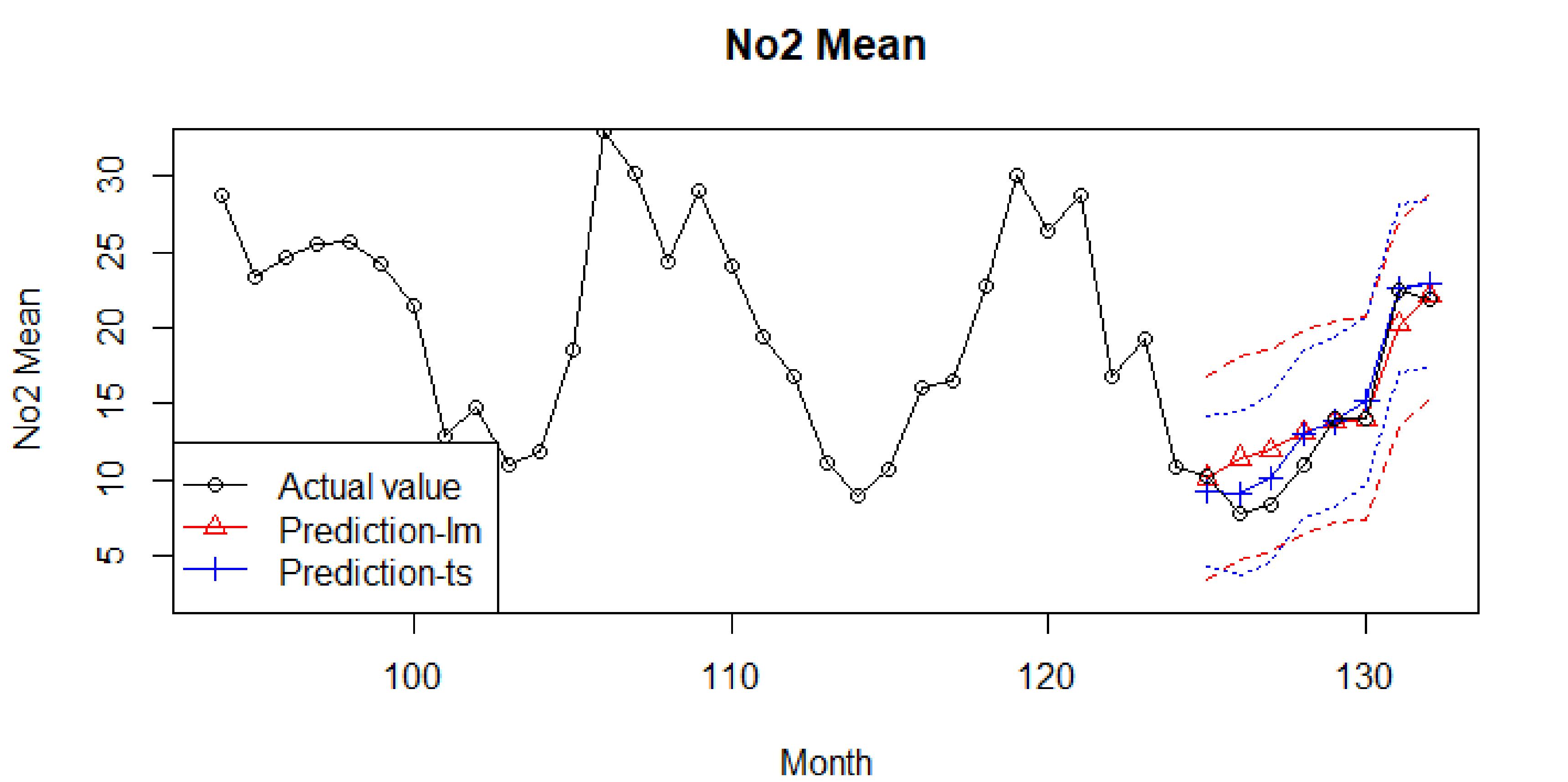
```
model=VAR(data,p=4)

summary(model)

#model diagnostics

serial.test(model)
```



```
> #model diagnostics
> serial.test(model)

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object model
Chi-squared = 218.49, df = 192, p-value = 0.09213
```



```
Covariance matrix of residuals:
         No2mean    Comean    So2mean     o3mean
No2mean 12.10285 0.3540926 0.8944595 -0.0066596
Comean   0.35409 0.0223488 0.0265772 -0.0002362
So2mean  0.89446 0.0265772 0.3932187 -0.0004393
o3mean  -0.00666 -0.0002362 -0.0004393  0.0000137

Correlation matrix of residuals:
        No2mean  Comean So2mean  o3mean
No2mean  1.0000  0.6808  0.4100 -0.5172
Comean   0.6808  1.0000  0.2835 -0.4269
So2mean  0.4100  0.2835  1.0000 -0.1893
o3mean  -0.5172 -0.4269 -0.1893  1.0000
```



THANK YOU

Any questions?