Time Series Forecasting of Atmospheric CO_2 Concentration Using Seasonal ARIMA Models

SOHINI MANDAL

2025-12-26

Abstract

Atmospheric carbon dioxide (CO_2) concentration exhibits a persistent upward trend along with strong seasonal variation. This study aims to model and forecast monthly mean atmospheric CO_2 concentrations measured at the Mauna Loa Observatory using Seasonal Autoregressive Integrated Moving Average (SARIMA) models. Several candidate SARIMA models were fitted to the training data and evaluated using Akaike Information Criterion (AIC), root mean square error (RMSE), and residual diagnostics via the Ljung–Box test. Based on model parsimony, goodness-of-fit, and forecasting accuracy, a final SARIMA model was selected. The results demonstrate that SARIMA models are effective for capturing both trend and seasonal dynamics in long-term environmental time series.

Introduction

Atmospheric carbon dioxide concentration is a key indicator of climate change and has shown a steady increase over the past several decades, primarily due to anthropogenic activities. Accurate modeling and forecasting of CO_2 concentration are important for understanding long-term climate dynamics and supporting environmental policy formulation.

The Mauna Loa Observatory provides one of the longest continuous records of atmospheric CO_2 concentration, making it a standard benchmark dataset for time series analysis. The CO_2 series is inherently non-stationary, exhibiting a strong upward trend and regular annual seasonality as observed from the time series plot.

Traditional regression techniques are inadequate for such data as they fail to capture temporal dependence and seasonal variation present in the observations. Seasonal Autoregressive Integrated Moving Average (SARIMA) models are therefore suitable, as they explicitly account for both non-seasonal and seasonal components.

The objective of this study is to identify an appropriate SARIMA model for monthly CO_2 concentration data and to evaluate its adequacy using diagnostic checks and out-of-sample forecasting performance.

Data Description

The dataset consists of monthly mean atmospheric CO_2 concentration values (in parts per million, ppm) recorded at the Mauna Loa Observatory, Hawaii, covering the period from 1958 to 2023. The observations are equally spaced with a seasonal frequency of 12 months.

To assess predictive performance, the dataset was divided into training and testing subsets. Approximately 80% of the observations were used for model estimation, while the remaining 20% were reserved for forecast evaluation.

Methodology

Preliminary Analysis

The time series plot of the training data shows a clear upward trend along with regular seasonal fluctuations. No cyclical movement is observed, though irregular variations are present. Since the seasonal amplitude remains approximately constant over time, the series exhibits additive trend–seasonal behavior. Therefore, an additive SARIMA framework is appropriate for modeling the data.

Model Framework

Seasonal ARIMA models of the form

$$ARIMA(p, d, q)(P, D, Q)_{12}$$

were considered, where p, d, q denote the non-seasonal autoregressive, differencing, and moving average orders, and P, D, Q represent the corresponding seasonal orders with a periodicity of 12 months.

From the ACF of the original series, the spikes decay very slowly, indicating the presence of a trend component. Additionally, significant spikes appear at regular intervals of 12 months, which suggests seasonal non-stationarity.

To remove the trend, first-order non-seasonal differencing (d=1) is applied. After this differencing, the slow decay pattern disappears, but significant seasonal spikes remain. Therefore, first-order seasonal differencing (D=1) is applied to remove seasonality.

Parameter Estimation

Model parameters are estimated using the maximum likelihood method. Based on the ACF and PACF behavior of the final differenced series, several candidate SARIMA models are fitted to the training data for comparison.

Model Evaluation Criteria

The fitted models were compared using the following criteria:

- **Ljung–Box test** was applied to the residuals of each fitted SARIMA model to assess whether any significant autocorrelation remained after model estimation. The test evaluates the joint null hypothesis that residual autocorrelations up to a specified lag are zero.

Null Hypothesis (H_0): Residuals are uncorrelated (white noise).
Alternative Hypothesis (H_1): Residuals exhibit autocorrelation at one or more lags.

At the 5% significance level**,** a p-value greater than 0.05 indicates failure to reject the null hypothesis, implying that the model adequately captures the temporal dependence in the data.

- **Akaike Information Criterion (AIC)** to balance goodness-of-fit and model complexity

- **Root Mean Square Error (RMSE)** computed on the test dataset to evaluate forecast accuracy

Results and Model Comparison

Comparison of Candidate SARIMA Models

Model	Ljung–Box p-value	AIC	RMSE
ARIMA(1,1,1)(0,1,1)[12]	0.5089 *(Random residuals)*	299.09	3.073
ARIMA(3,1,0)(0,1,1)[12]	0.3454 *(Random residuals)*	302.08	2.977
ARIMA(2,1,0)(0,1,1)[12]	0.1961 *(Random residuals)*	305.59	2.936
ARIMA(0,1,1)(1,1,1)[12]	0.3475 *(Random residuals)*	301.13	2.999
ARIMA(0,1,1)(0,1,1)[12]	0.4044 *(Random residuals)*	299.32	2.957

All models yielded Ljung–Box test p-values greater than 0.05, indicating that the residuals can be regarded as uncorrelated and approximately white noise.

Final Model Selection

The selected model for forecasting is:

$$\boxed{ARIMA(1,1,1)(0,1,1)_{12}}$$

Rationale for Selection

Although certain models achieved marginally lower RMSE values, the improvement was small and not practically significant. The selected model offers the lowest AIC among the candidate models, indicating superior parsimony while maintaining competitive forecast accuracy. Residual diagnostics further confirm the adequacy of the model.

Final Fitted Model

Estimated Parameters

Parameter	Estimate	Standard Error
AR(1)	0.1791	0.1143
MA(1)	−0.5264	0.1005
Seasonal MA(1)	−0.8728	0.0204

- **Ljung–Box p-value:** 0.5089
- **AIC:** 299.09
- **RMSE (test set):** 3.073

Conclusion

The analysis shows that the monthly CO_2 concentration series exhibits both a long-term upward trend and strong annual seasonality. Seasonal ARIMA modeling effectively captures these characteristics. Among the candidate models examined, the ARIMA(1,1,1)(0,1,1)[12] model was selected based on its favorable balance between model simplicity, goodness-of-fit, and forecasting performance. The residual diagnostics indicate that the model assumptions are reasonably satisfied, supporting its suitability for forecasting monthly CO_2 concentrations.

Future work may incorporate exogenous variables through SARIMAX models or apply rolling-origin validation to further assess forecasting performance.

R Code Implementation:

```
# Data Description:
# Monthly mean carbon dioxide (in ppm) measured at Mauna Loa Observatory, Haw
aii from 1958 to 2023.
rm(list=ls())
library(astsa)
library(forecast)

## Warning: package 'forecast' was built under R version 4.2.3

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

##
## Attaching package: 'forecast'

## The following object is masked from 'package:astsa':
##
##     gas

cardox

##         Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct
## 1958               315.70 317.45 317.51 317.24 315.86 314.93 313.20 312.43
## 1959 315.58 316.48 316.65 317.72 318.29 318.15 316.54 314.80 313.84 313.33
## 1960 316.43 316.98 317.58 319.03 320.04 319.59 318.18 315.90 314.17 313.83
## 1961 316.89 317.70 318.54 319.48 320.58 319.77 318.57 316.79 314.99 315.31
## 1962 317.94 318.55 319.68 320.57 321.02 320.62 319.61 317.40 316.25 315.42
## 1963 318.74 319.07 319.86 321.38 322.25 321.48 319.74 317.77 316.21 315.99
## 1964 319.57 320.01 320.74 321.84 322.26 321.89 320.44 318.69 316.70 316.87
## 1965 319.44 320.44 320.89 322.14 322.17 321.87 321.21 318.87 317.81 317.30
## 1966 320.62 321.60 322.39 323.70 324.08 323.75 322.38 320.36 318.64 318.10
## 1967 322.33 322.50 323.04 324.42 325.00 324.09 322.54 320.92 319.25 319.39
## 1968 322.57 323.15 323.89 325.02 325.57 325.36 324.14 322.11 320.33 320.25
## 1969 324.00 324.42 325.63 326.66 327.38 326.71 325.88 323.66 322.38 321.78
## 1970 325.06 325.98 326.93 328.13 328.08 327.67 326.34 324.69 323.10 323.06
## 1971 326.17 326.68 327.17 327.79 328.93 328.57 327.36 325.43 323.36 323.56
## 1972 326.77 327.63 327.75 329.72 330.07 329.09 328.04 326.32 324.84 325.20
## 1973 328.55 329.56 330.30 331.50 332.48 332.07 330.87 329.31 327.51 327.18
## 1974 329.35 330.71 331.48 332.65 333.18 332.20 331.07 329.15 327.33 327.28
## 1975 330.73 331.46 331.94 333.11 333.95 333.42 331.97 329.95 328.50 328.36
## 1976 331.56 332.74 333.36 334.74 334.72 333.98 333.08 330.68 328.96 328.72
## 1977 332.68 333.17 334.96 336.14 336.93 336.17 334.89 332.56 331.29 331.28
## 1978 334.94 335.26 336.66 337.69 338.02 338.01 336.50 334.42 332.36 332.45
## 1979 336.14 336.69 338.27 338.82 339.24 339.26 337.54 335.72 333.97 334.24
## 1980 337.90 338.34 340.07 340.93 341.45 341.36 339.45 337.67 336.25 336.14
## 1981 339.29 340.55 341.63 342.60 343.04 342.54 340.82 338.48 336.95 337.05
## 1982 340.93 341.76 342.77 343.96 344.77 343.88 342.42 340.24 338.38 338.41
## 1983 341.57 342.79 343.37 345.40 346.14 345.76 344.32 342.51 340.46 340.53
```

```
## 1984 344.21 344.92 345.68 347.37 347.78 347.16 345.79 343.74 341.59 341.86
## 1985 345.48 346.41 347.91 348.66 349.28 348.65 346.90 345.26 343.47 343.35
## 1986 346.78 347.48 348.25 349.86 350.52 349.98 348.25 346.17 345.48 344.82
## 1987 348.73 348.92 349.81 351.40 352.15 351.58 350.21 348.20 346.66 346.72
## 1988 350.51 351.70 352.50 353.67 354.35 353.88 352.80 350.49 348.97 349.37
## 1989 353.07 353.43 354.08 355.72 355.95 355.44 354.05 351.84 350.09 350.33
## 1990 353.86 355.10 355.75 356.38 357.38 356.39 354.89 353.06 351.38 351.69
## 1991 354.93 355.82 357.33 358.77 359.23 358.23 356.30 353.97 352.34 352.43
## 1992 356.34 357.21 357.97 359.22 359.71 359.44 357.15 354.99 353.01 353.41
## 1993 357.10 357.42 358.59 359.39 360.30 359.64 357.46 355.76 354.14 354.23
## 1994 358.36 359.04 360.11 361.36 361.78 360.94 359.51 357.59 355.86 356.21
## 1995 360.04 361.00 361.98 363.44 363.83 363.33 361.78 359.33 358.32 358.14
## 1996 362.20 363.36 364.28 364.69 365.25 365.06 363.69 361.55 359.69 359.72
## 1997 363.24 364.21 364.65 366.48 366.77 365.73 364.46 362.40 360.44 360.98
## 1998 365.39 366.10 367.36 368.79 369.56 369.13 367.98 366.10 364.16 364.54
## 1999 368.35 369.28 369.84 371.15 371.12 370.46 369.61 367.06 364.95 365.52
## 2000 369.45 369.71 370.75 371.98 371.75 371.87 370.02 368.27 367.15 367.18
## 2001 370.76 371.69 372.63 373.55 374.03 373.40 371.68 369.78 368.34 368.61
## 2002 372.70 373.37 374.30 375.19 375.93 375.69 374.16 372.03 370.92 370.73
## 2003 375.07 375.82 376.64 377.92 378.78 378.46 376.88 374.57 373.34 373.31
## 2004 377.17 378.05 379.06 380.54 380.80 379.87 377.65 376.17 374.43 374.63
## 2005 378.63 379.91 380.95 382.48 382.64 382.40 380.93 378.93 376.89 377.19
## 2006 381.58 382.40 382.86 384.80 385.22 384.24 382.65 380.60 379.04 379.33
## 2007 383.10 384.12 384.81 386.73 386.78 386.33 384.73 382.24 381.20 381.37
## 2008 385.78 386.06 386.28 387.34 388.78 387.99 386.61 384.32 383.41 383.21
## 2009 387.17 387.70 389.04 389.76 390.36 389.70 388.24 386.29 384.95 384.64
## 2010 388.91 390.41 391.37 392.67 393.21 392.38 390.41 388.54 387.03 387.43
## 2011 391.50 392.05 392.80 393.44 394.41 393.95 392.72 390.33 389.28 389.19
## 2012 393.31 394.04 394.59 396.38 396.93 395.91 394.56 392.59 391.32 391.27
## 2013 395.78 397.03 397.66 398.64 400.02 398.81 397.51 395.39 393.72 393.90
## 2014 398.04 398.27 399.91 401.51 401.96 401.43 399.27 397.18 395.54 396.16
## 2015 400.18 400.55 401.74 403.34 404.15 402.97 401.46 399.11 397.82 398.49
## 2016 402.73 404.25 405.06 407.60 407.90 406.99 404.59 402.45 401.23 401.79
## 2017 406.36 406.66 407.54 409.22 409.89 409.08 407.33 405.32 403.57 403.82
## 2018 408.15 408.52 409.59 410.45 411.44 410.99 408.90 407.16 405.71 406.19
## 2019 411.03 411.96 412.18 413.54 414.86 414.16 411.97 410.18 408.76 408.75
## 2020 413.61 414.34 414.74 416.45 417.31 416.60 414.62 412.78 411.52 411.51
## 2021 415.52 416.75 417.64 419.05 419.13 418.94 416.96 414.47 413.30 413.93
## 2022 418.19 419.28 418.81 420.23 420.99 420.99 418.90 417.19 415.95 415.78
## 2023 419.47 420.41 421.00
##         Nov    Dec
## 1958 313.33 314.67
## 1959 314.81 315.58
## 1960 315.00 316.19
## 1961 316.10 317.01
## 1962 316.69 317.70
## 1963 317.07 318.35
## 1964 317.68 318.71
## 1965 318.87 319.42
## 1966 319.78 321.03
```

```
## 1967 320.73 321.96
## 1968 321.32 322.89
## 1969 322.86 324.12
## 1970 324.01 325.13
## 1971 324.80 326.01
## 1972 326.50 327.55
## 1973 328.16 328.64
## 1974 328.31 329.58
## 1975 329.38 330.78
## 1976 330.16 331.62
## 1977 332.46 333.60
## 1978 333.76 334.91
## 1979 335.32 336.81
## 1980 337.30 338.29
## 1981 338.58 339.91
## 1982 339.44 340.78
## 1983 341.79 343.20
## 1984 343.31 345.00
## 1985 344.73 346.12
## 1986 346.22 347.49
## 1987 348.08 349.28
## 1988 350.42 351.62
## 1989 351.55 352.91
## 1990 353.14 354.41
## 1991 353.89 355.21
## 1992 354.42 355.68
## 1993 355.53 357.03
## 1994 357.65 359.10
## 1995 359.61 360.82
## 1996 361.04 362.39
## 1997 362.65 364.51
## 1998 365.67 367.30
## 1999 366.88 368.26
## 2000 368.53 369.83
## 2001 369.94 371.42
## 2002 372.43 373.98
## 2003 374.84 376.17
## 2004 376.33 377.68
## 2005 378.54 380.31
## 2006 380.35 382.02
## 2007 382.70 384.19
## 2008 384.41 385.79
## 2009 386.23 387.63
## 2010 388.87 389.99
## 2011 390.48 392.06
## 2012 393.20 394.57
## 2013 395.36 397.03
## 2014 397.40 399.08
## 2015 400.27 402.06
## 2016 403.72 404.64
```

```
## 2017 405.31 407.00
## 2018 408.21 409.27
## 2019 410.48 411.98
## 2020 413.12 414.26
## 2021 415.01 416.71
## 2022 417.51 418.95
## 2023
```

```
length(cardox)
```

```
## [1] 781
```

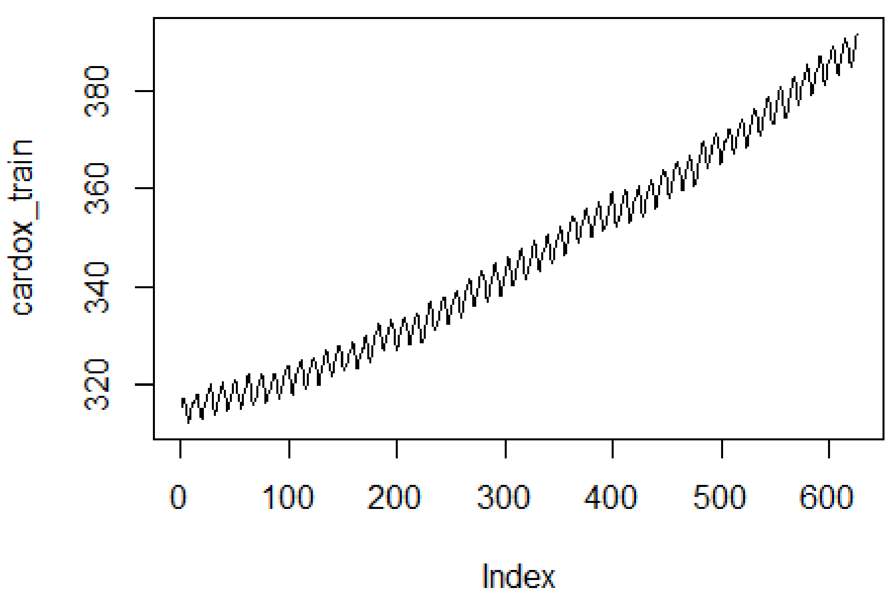
```
str(cardox)
```

```
##  Time-Series [1:781] from 1958 to 2023: 316 317 318 317 316 ...
```

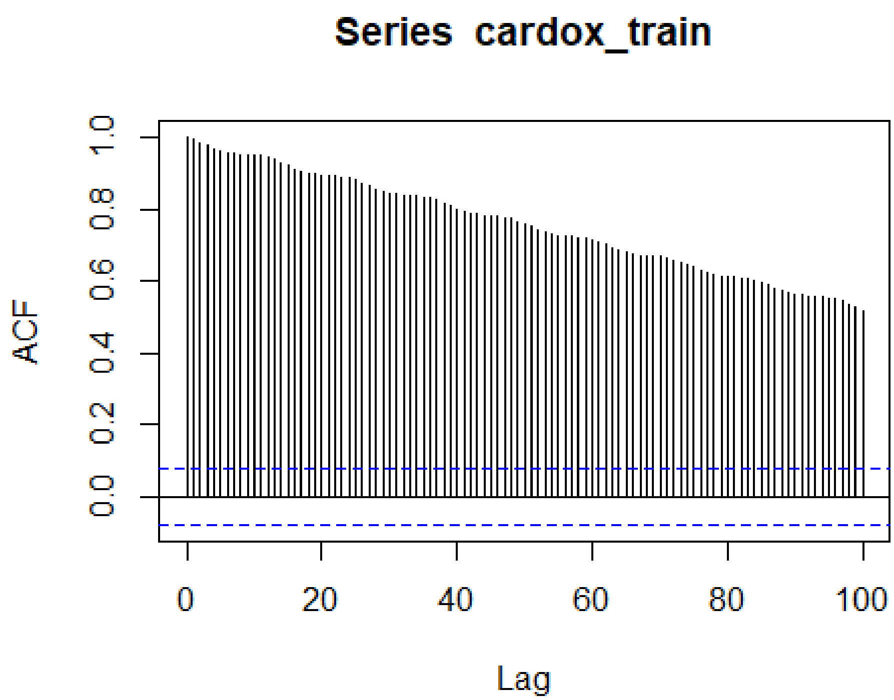
```
frequency(cardox)
```

```
## [1] 12
```

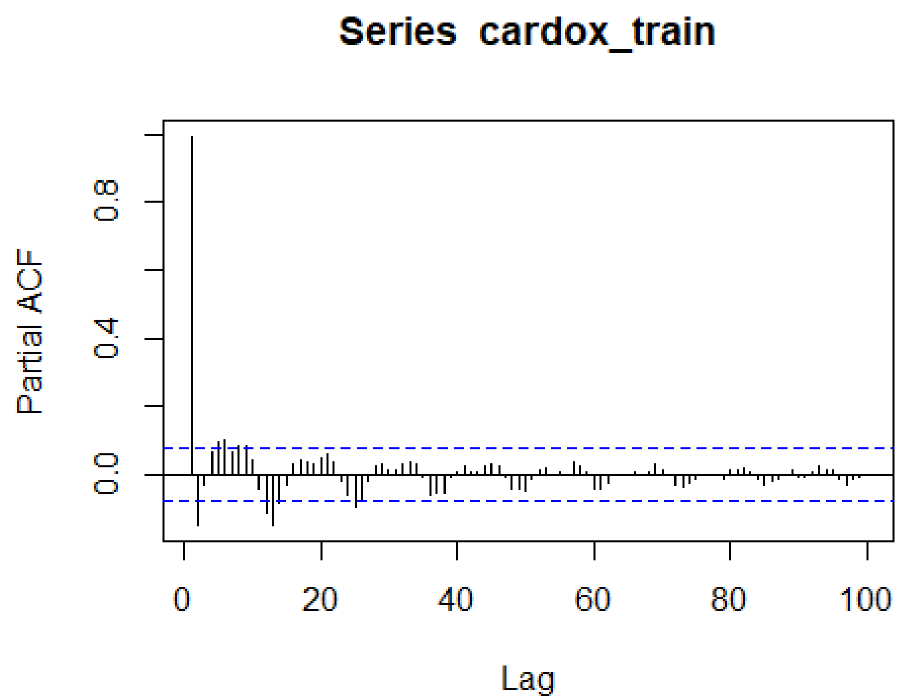
```
# Train-test split
n=length(cardox)
m=ceiling(n*0.8)
cardox_train=cardox[1:m]
cardox_test=cardox[(m+1):n]
plot(cardox_train,type="l")
```



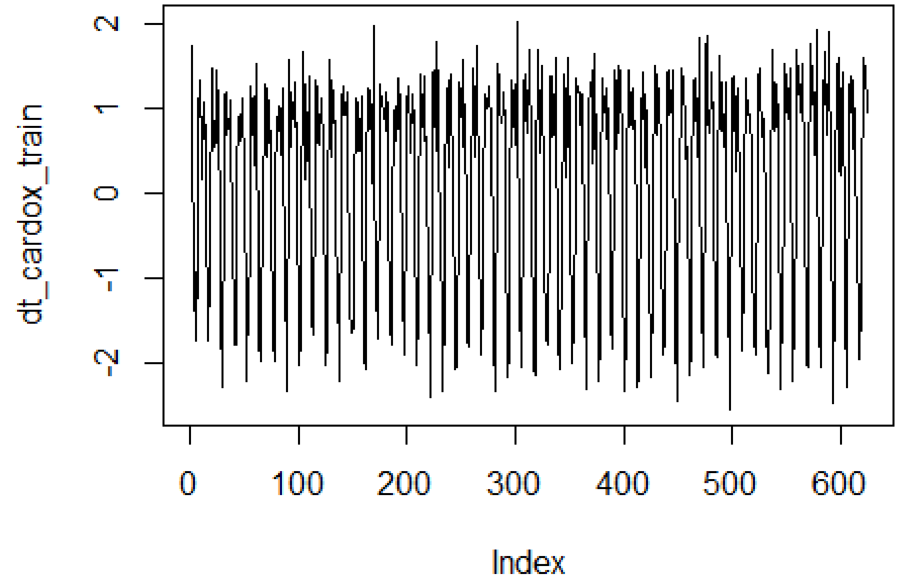
```r
# Upward trend is present. Seasonality is of period 12 months is present. Irr
egularity fluctuations can be seen but no cyclical fluctuations can be observ
ed.
# Also the seasonal variations are more or less constant with time.
# So here we use an additive model for the time series Xt= Tt+St+It.

acf(cardox_train,lag.max=100)
```



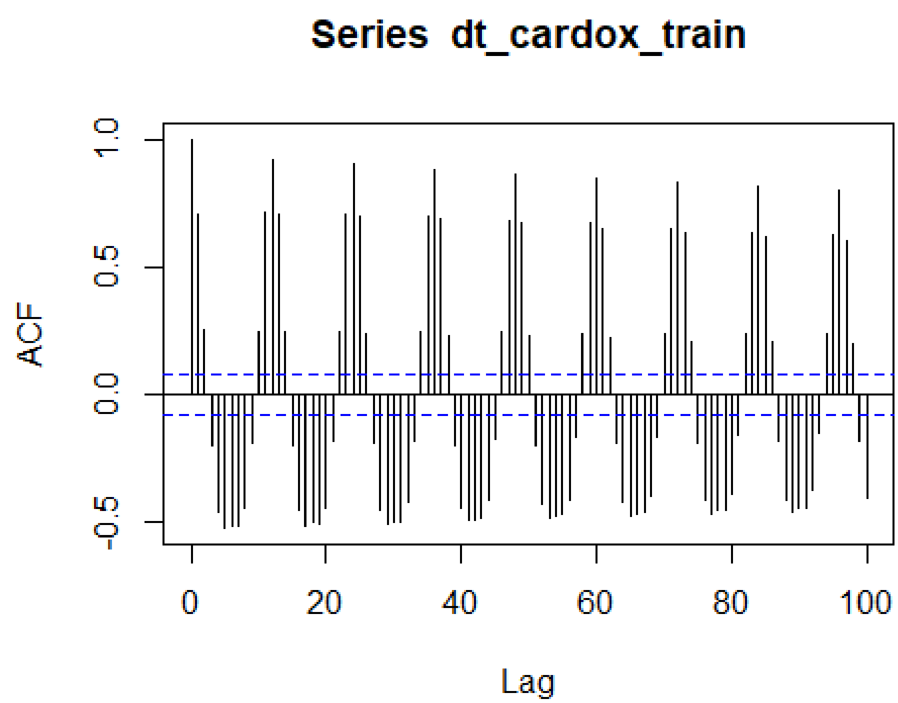

```r
# The spikes are decaying very slowly which means a trend component is presen
t.
# The spikes are repeating its pattern after regular intervals where the inte
rval is 12 months which indicate that seasonality with period 12 months is pr
esent in the data.
pacf(cardox_train,lag.max=100)
```



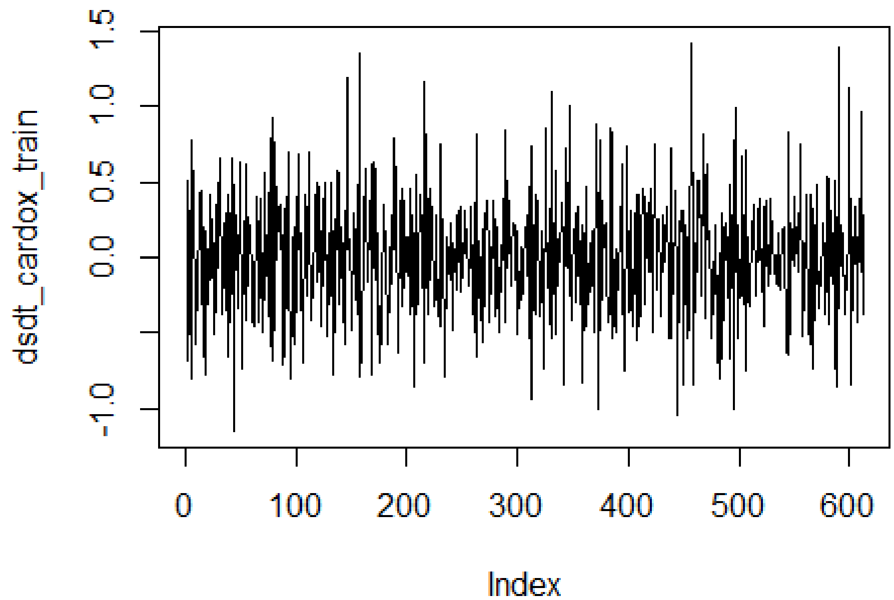

```
# Highly significant spike can be seen at lag 1 and after that the spikes are more or less insignificant.

# Now we will perform differencing to remove trend and seasonality from the data.
dt_cardox_train=diff(cardox_train)
plot(dt_cardox_train,type="l")
```



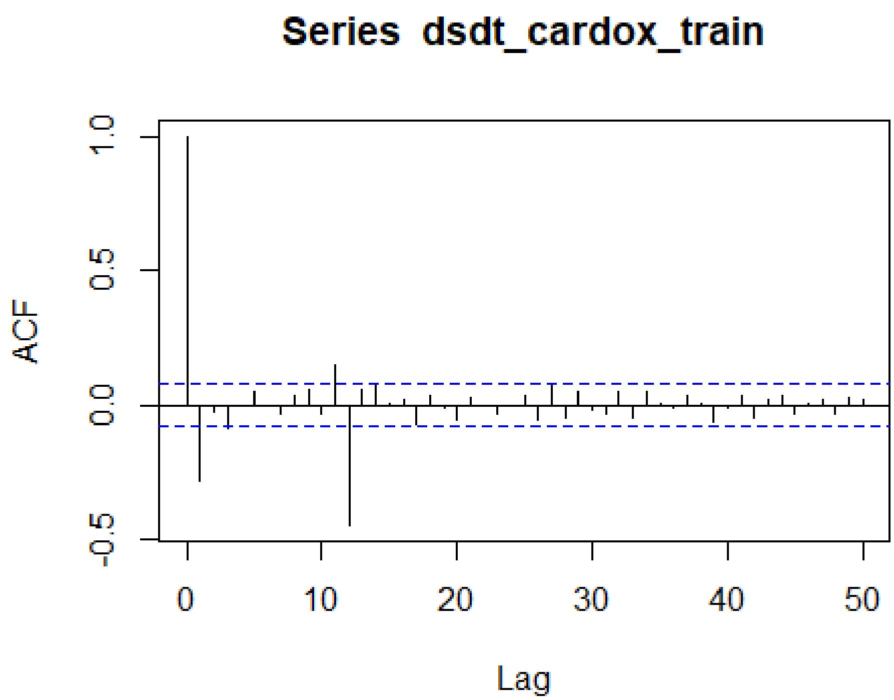

```
acf(dt_cardox_train,lag.max=100)
```



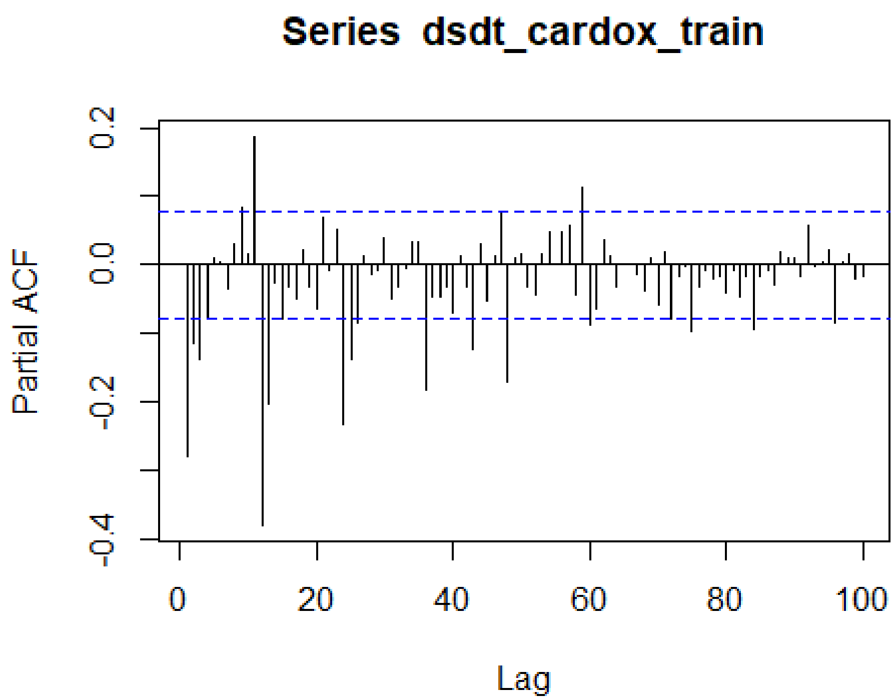
```
# The slow decaying pattern of the spikes has been eliminated after 1st diffe
rencing of the series at lag 1.
# But seasonal non-stationarity is there as significantly high spikes can be
seen at seasonal lags which are decaying very slowly.
dsdt_cardox_train=diff(dt_cardox_train,lag=12)
plot(dsdt_cardox_train,type="l")
```




```
acf(dsdt_cardox_train,lag.max=50)
```

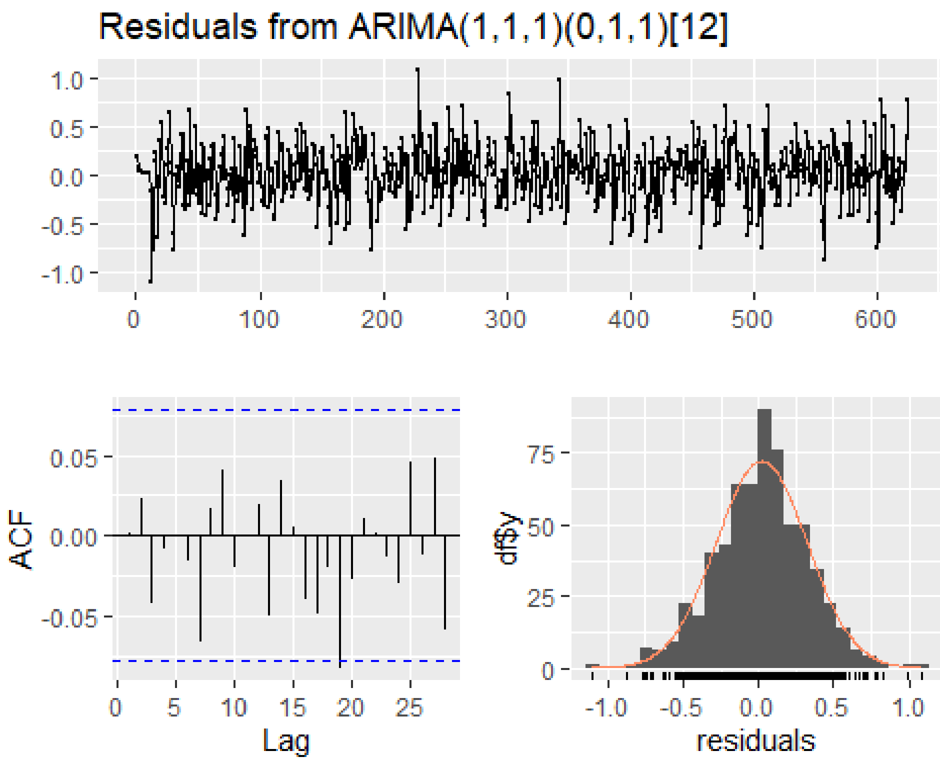



```
# After differencing at lag 12 on the detrended data, seasonality is now remo
ved.
# So no more differencing is now needed
# So trial value for d is 1 and also for D is 1.
# Significantly high spike can be seen at lag 1 so trial value of q=1.
# A very high significant spike can be seen at seasonal lag 1(lag 12), so tri
al value of Q=1.
pacf(dsdt_cardox_train,lag.max=100)
```

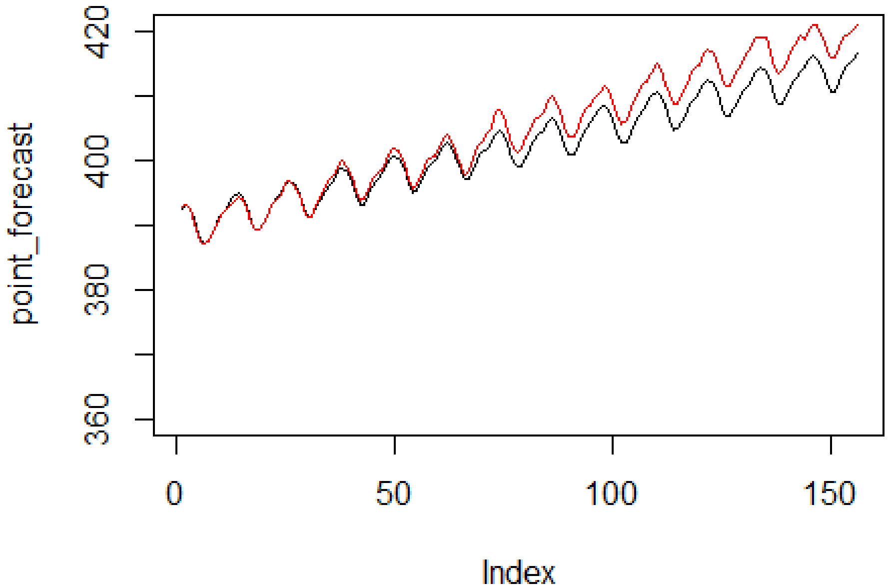



```
# At seasonal lags, as if the PACF is tailing off.
# Highly significant spikes can be seen at seasonal lags 1,2,3.
# So trial values of P=1,2,3.
# Significant spikes can also be seen at lag 1,2,3.
# Also the PACF seems to be tailing off.
# So trial values for p=0,1,2,3.

# Possible choices:
# ARIMA(p=1,d=1,q=1,P=0,D=1,Q=1)
# ARIMA(p=3,d=1,q=0,P=0,D=1,Q=1)
# ARIMA(p=2,d=1,q=0,P=0,D=1,Q=1)
# ARIMA(p=0,d=1,q=1,P=1,D=1,Q=1)
# ARIMA(p=0,d=1,q=1,P=0,D=1,Q=1)
# And S=12

ARIMA=function(p,d,q,P,D,Q)
{
  fit=arima(cardox_train,order=c(p,d,q),
            seasonal=list(order=c(P,D,Q),period=12),
            optim.control=list(maxit=1500),
            method="ML")
  forecast=forecast::forecast(fit,h=length(cardox_test))
  point_forecast=as.numeric(forecast$mean)
  checkresiduals(fit,lag=20)
  plot(point_forecast, type="l", ylim=c(360,420))   # To zoom we used 360 to
420 as cardox_test starts from 392.67 and ends at 421.00
  lines(cardox_test,col="red")
```

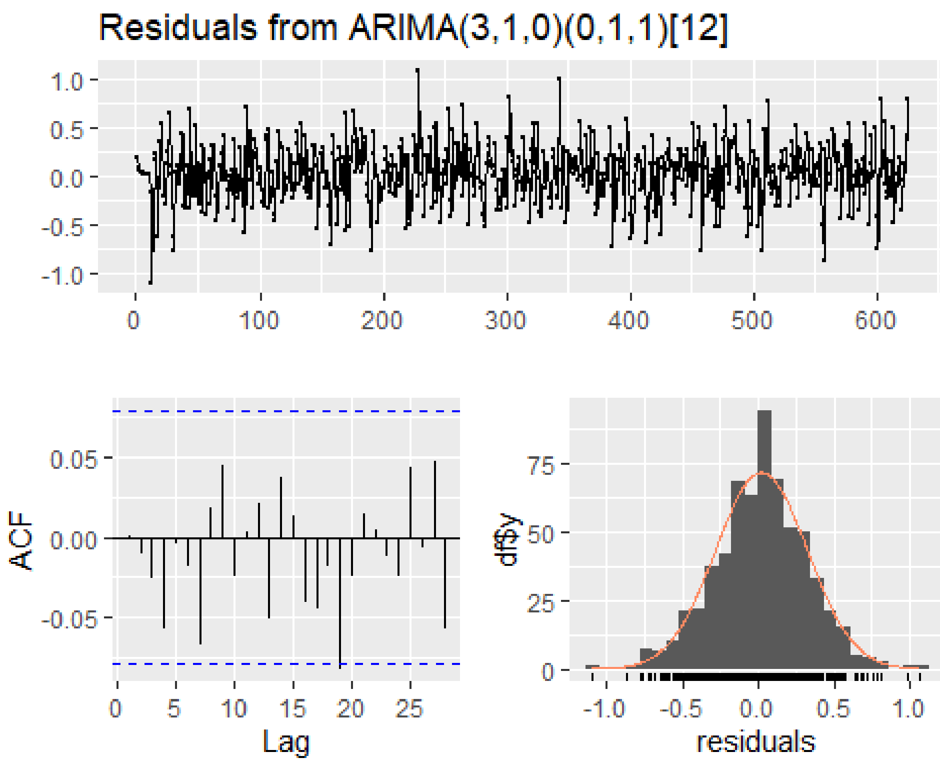
```
  RMSE=sqrt(mean((point_forecast-cardox_test)**2))
  return(list(fit=fit,RMSE=RMSE))
}

ARIMA(p=1,d=1,q=1,P=0,D=1,Q=1)
```



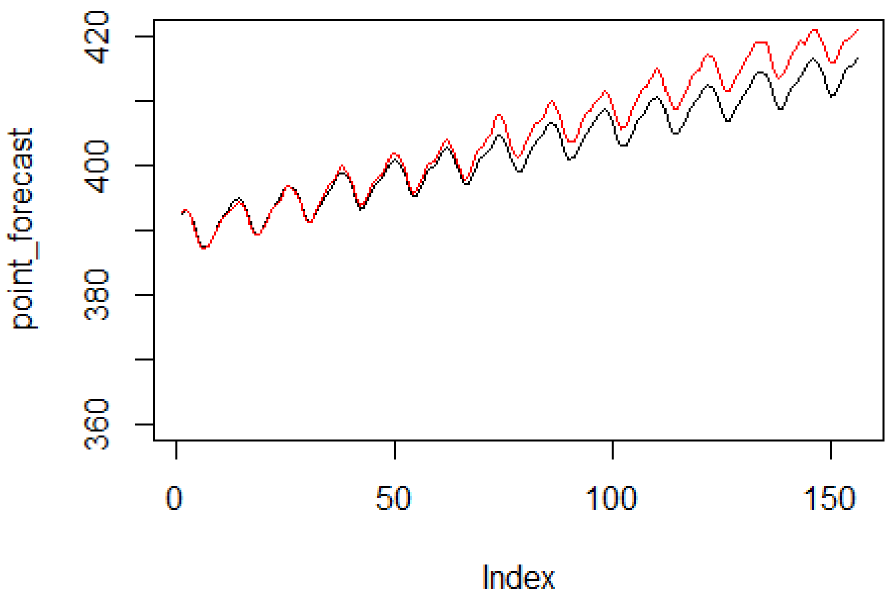

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,1)(0,1,1)[12]
## Q* = 16.211, df = 17, p-value = 0.5089
##
## Model df: 3.   Total lags used: 20
```



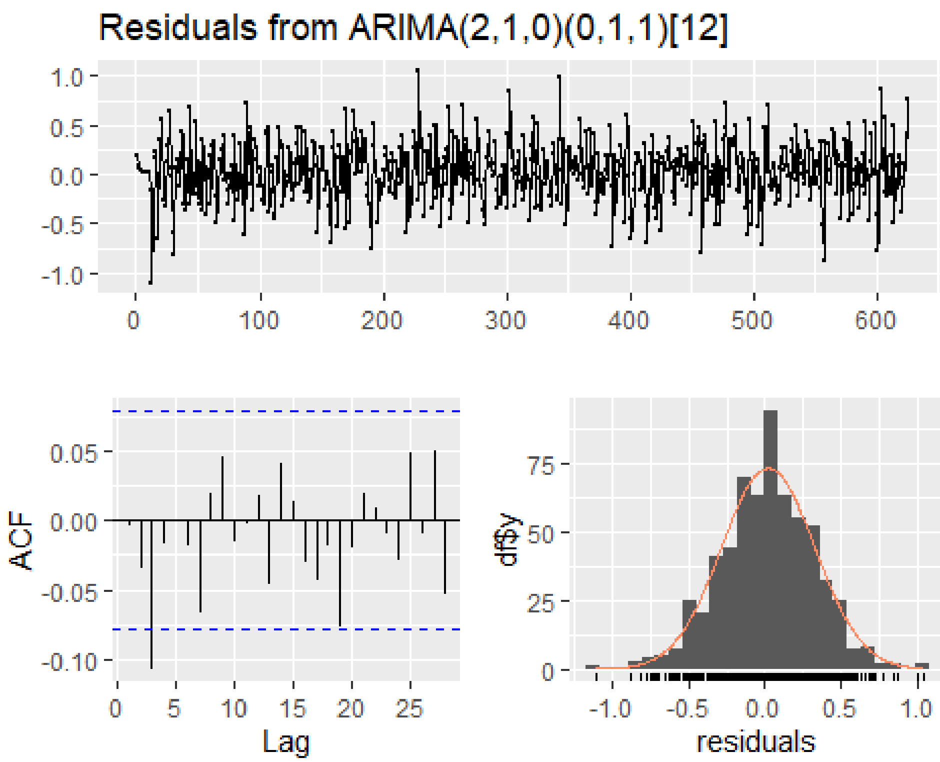

```
## $fit
##
## Call:
## arima(x = cardox_train, order = c(p, d, q), seasonal = list(order = c(P, D
,
##     Q), period = 12), method = "ML", optim.control = list(maxit = 1500))
##
## Coefficients:
##          ar1      ma1     sma1
##       0.1791  -0.5264  -0.8728
## s.e.  0.1143   0.1005   0.0204
##
## sigma^2 estimated as 0.09157:  log likelihood = -145.54,  aic = 299.09
##
## $RMSE
## [1] 3.072857

ARIMA(p=3,d=1,q=0,P=0,D=1,Q=1)
```



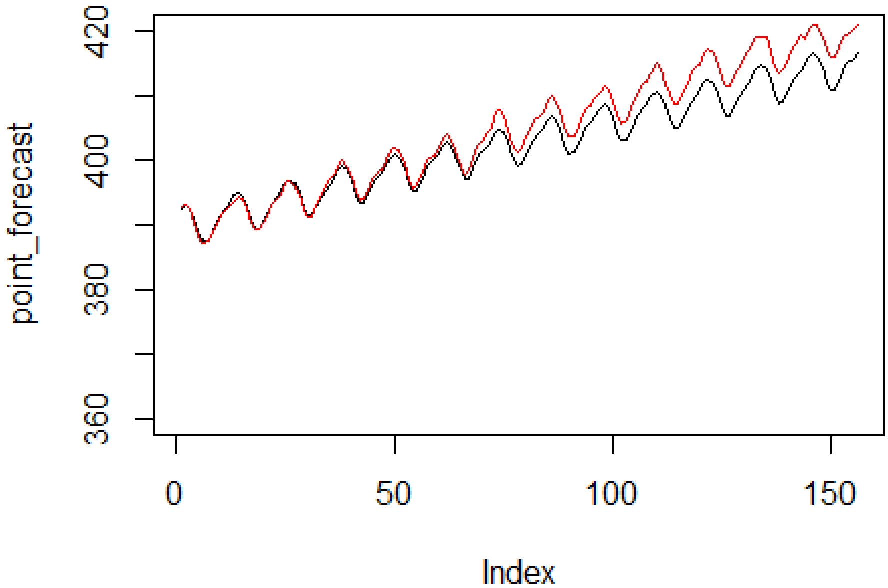

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,1,0)(0,1,1)[12]
## Q* = 17.64, df = 16, p-value = 0.3454
##
## Model df: 4.   Total lags used: 20
```



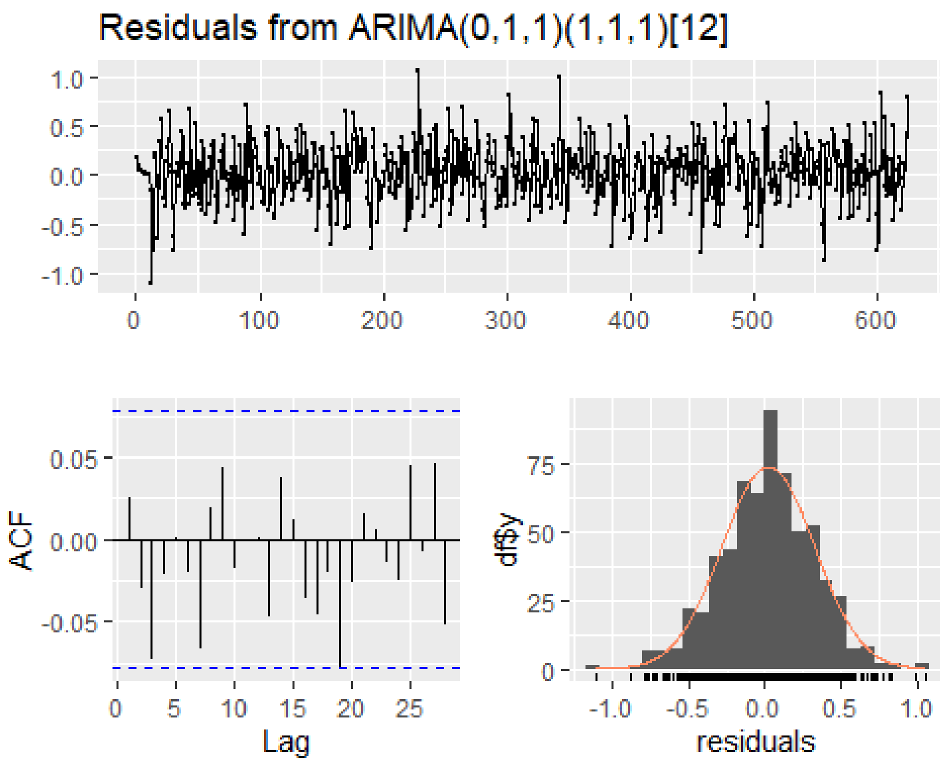

```
## $fit
##
## Call:
## arima(x = cardox_train, order = c(p, d, q), seasonal = list(order = c(P, D
,
##     Q), period = 12), method = "ML", optim.control = list(maxit = 1500))
##
## Coefficients:
##           ar1      ar2      ar3     sma1
##       -0.3432  -0.1446  -0.0959  -0.8733
## s.e.   0.0411   0.0431   0.0408   0.0205
##
## sigma^2 estimated as 0.09172:  log likelihood = -146.04,  aic = 302.08
##
## $RMSE
## [1] 2.976651

ARIMA(p=2,d=1,q=0,P=0,D=1,Q=1)
```



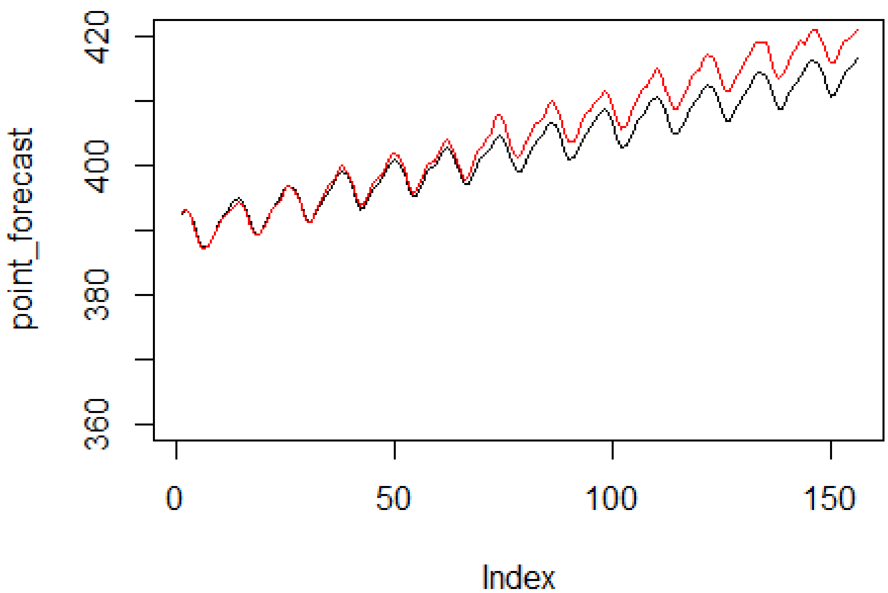

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,0)(0,1,1)[12]
## Q* = 21.71, df = 17, p-value = 0.1961
##
## Model df: 3.   Total lags used: 20
```



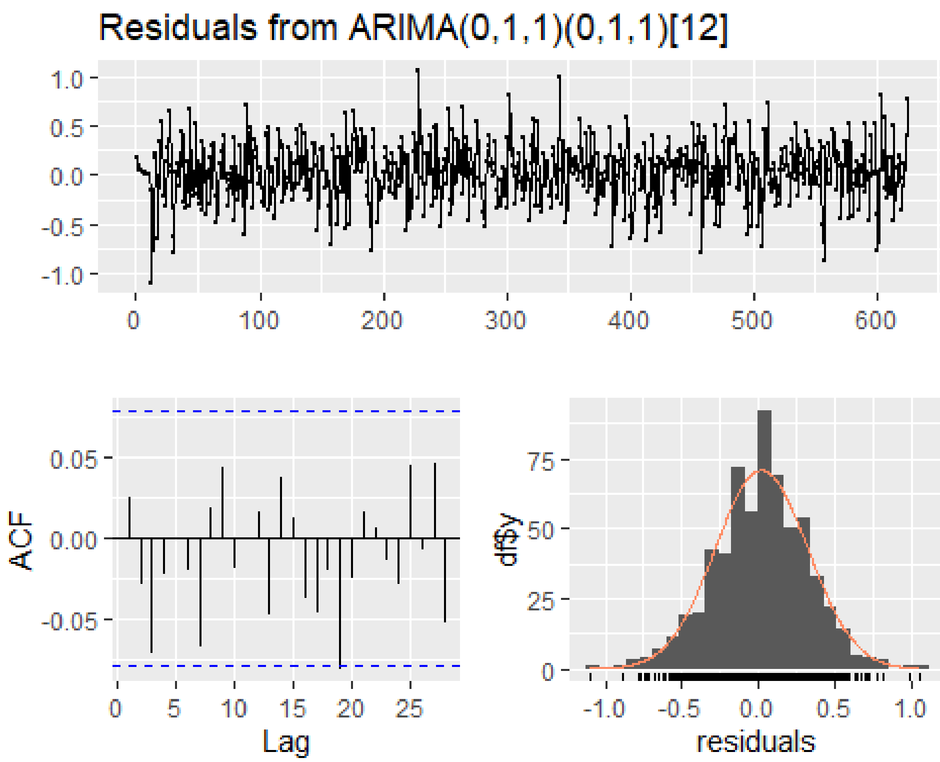

```
## $fit
##
## Call:
## arima(x = cardox_train, order = c(p, d, q), seasonal = list(order = c(P, D
,
##     Q), period = 12), method = "ML", optim.control = list(maxit = 1500))
##
## Coefficients:
##           ar1      ar2     sma1
##       -0.3306  -0.1114  -0.8781
## s.e.   0.0409   0.0408   0.0201
##
## sigma^2 estimated as 0.09248:  log likelihood = -148.79,  aic = 305.59
##
## $RMSE
## [1] 2.935511

ARIMA(p=0,d=1,q=1,P=1,D=1,Q=1)
```



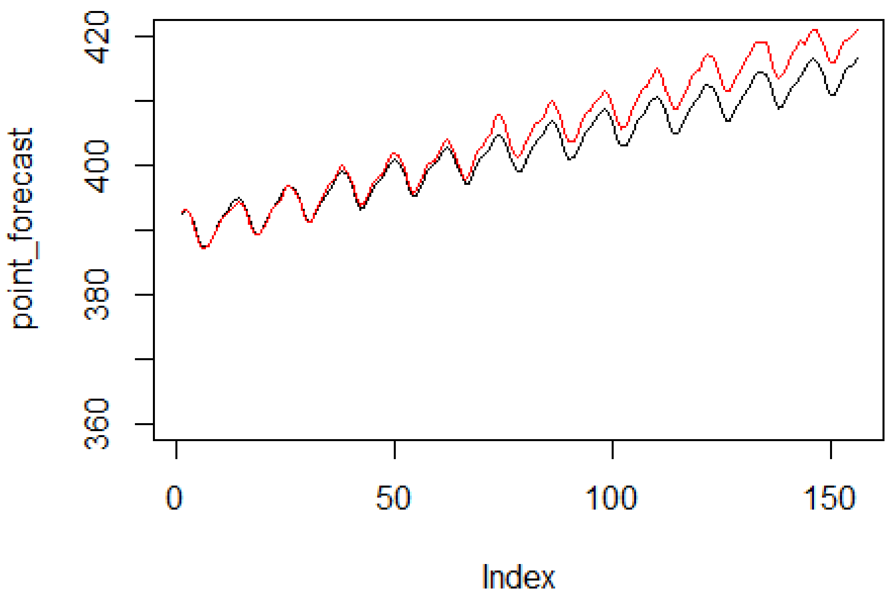

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1)(1,1,1)[12]
## Q* = 18.675, df = 17, p-value = 0.3475
##
## Model df: 3.   Total lags used: 20
```




```
## $fit
##
## Call:
## arima(x = cardox_train, order = c(p, d, q), seasonal = list(order = c(P, D
,
##     Q), period = 12), method = "ML", optim.control = list(maxit = 1500))
##
## Coefficients:
##           ma1    sar1     sma1
##       -0.3642  0.0199  -0.8777
## s.e.   0.0422  0.0460   0.0220
##
## sigma^2 estimated as 0.09188:  log likelihood = -146.56,  aic = 301.13
##
## $RMSE
## [1] 2.999335

ARIMA(p=0,d=1,q=1,P=0,D=1,Q=1)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1)(0,1,1)[12]
## Q* = 18.797, df = 18, p-value = 0.4044
##
## Model df: 2.   Total lags used: 20
```




```
## $fit
##
## Call:
## arima(x = cardox_train, order = c(p, d, q), seasonal = list(order = c(P, D
,
##     Q), period = 12), method = "ML", optim.control = list(maxit = 1500))
##
## Coefficients:
##           ma1     sma1
##       -0.3661  -0.8736
## s.e.   0.0419   0.0204
##
## sigma^2 estimated as 0.0919:  log likelihood = -146.66,  aic = 299.32
##
## $RMSE
## [1] 2.956572

# Best model chosen: ARIMA(p=1,d=1,q=1,P=0,D=1,Q=1)
# 1st and last are comparable, but if we go by AIC, we should choose 1st one.
```