MAT375: Final Project

Lincoln Stewart

May 25, 2025

Contents

List of Figures

List of Tables

1.0 Introduction

Description of data

The data set that I am using is the Mauna Loa co2 record. It contains monthly measurements of the atmospheric carbon dioxide concentration. This time series has been recorded over decades and is one of the most important long term records of co2 level in Earth's atmosphere. It was first established by Dr. Charles David Keeling in 1958 at the Mauna Loa Observatory in Hawaii. The motivation was to monitor long term changes in the atmosphere. It is all a part of climate change research to help scientists understand human contribution to climate change. The Mauna Loa site was chosen because of its remote location and high altitude. Its low local pollution makes it an ideal place for measuring background atmospheric co2 level. The data set was obtained from the US National Oceanic and Atmosphere administration and the Scripps institution of Oceanography.

Scientific questions

1. Long term trends in co2 concentrations. Based off this data interesting things to look at will be things like how has atmospheric co2 changed over time? Is there a clear upward trend in co2 levels ? Can we model the trend in the co2 ?

2. Seasonality in co2. Does the co2 fluctuate by season, and if so what are the ranges of said fluctuations? Do all seasonal patterns compare to historic patterns. Does de-trending affect the seasonality of the data.

2.0 Exploratory data analysis

```r
library(astsa)
library(forecast)
library(tseries)
library(tidyr)
library(knitr)
```

Time Series and description

```r
print({
plot(cardox,
     ylab = "Co2 level (ppm)",
     xlab = "Time",
     xaxt = "n",
     xlim = c(1955, 2030),
     ylim = c(300,430),
     yaxt = "n"
     )
title("Mauna Loa Co2 Record")
axis(1, at = seq(1955, 2030, by = 5))
axis(2, at = seq(300,430, by = 10), las = 1)
time_vals <- time(cardox)
legend("topleft", legend = c("CO2 levels"),
```

```r
        col = c("black"), lty = c(1, 2), lwd = c(1, 2))
})
```

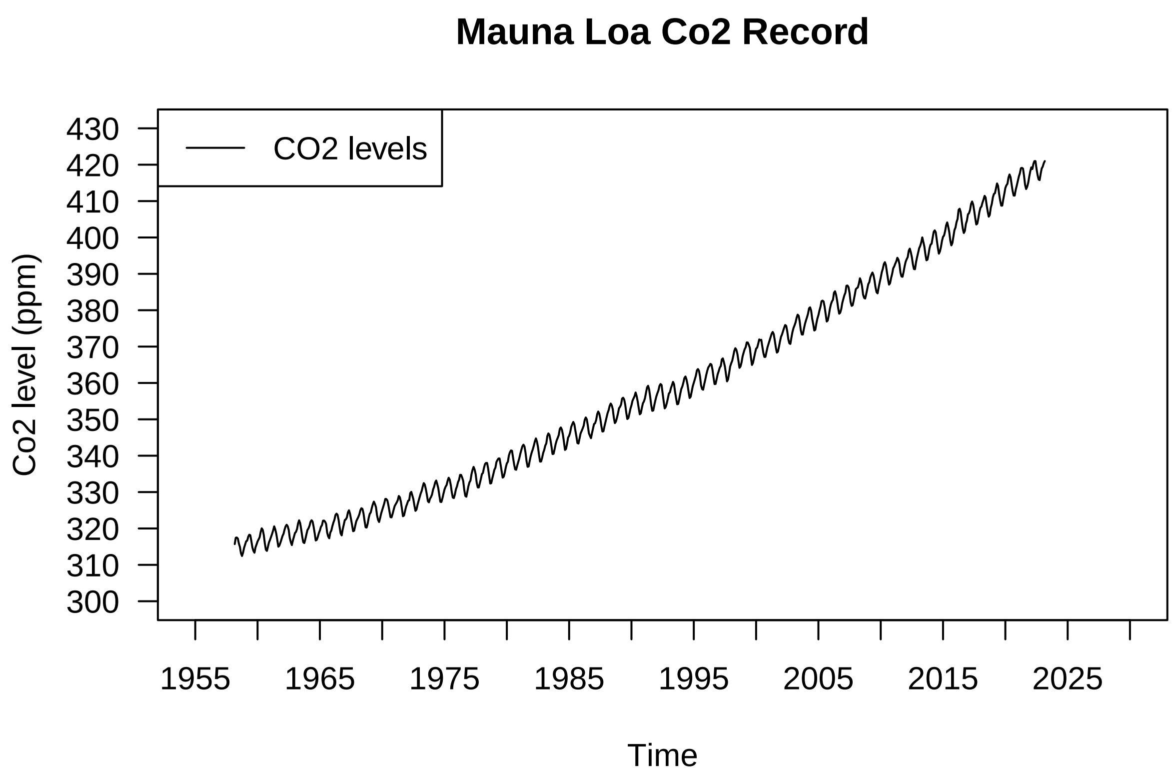


Figure 1: Mauna Loa atmospheric Co2 concentrations recorded monthly from 1958 to 2023. The time series shows a clear upward quadratic trend and strong seasonal oscillations.

```
## $rect
## $rect$w
## [1] 22.80371
##
## $rect$h
## [1] 21.11278
##
## $rect$left
## [1] 1952
##
## $rect$top
## [1] 435.2
```

```
##
##
## $text
## $text$x
## [1] 1961.24
##
## $text$y
## [1] 424.6436
```

Here we can see the time series plotted from the years that it has been recorded. It has a clear quadratic trend upwards, and very regular variability throughout. Some slight changes can be seen that deviate from the trend, like around the time of 1990, but it mostly follows it.

Detrending

```
dtn_diff <- diff(cardox)
```

```
##  [1] 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020 2025
## [16] 2030
```

Based off the plot of the detrended time series, no transformation was deemed necessary based off the fact that the variance appears constant after detredning.

Stationary check

$H_0 = $ The detrended series has a unit root, making it non stationary $H_a =$ The detrended series is stationary

```
adf.test(cardox, alternative = "stationary")
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  cardox
## Dickey-Fuller = -0.63799, Lag order = 9, p-value = 0.9755
## alternative hypothesis: stationary
```

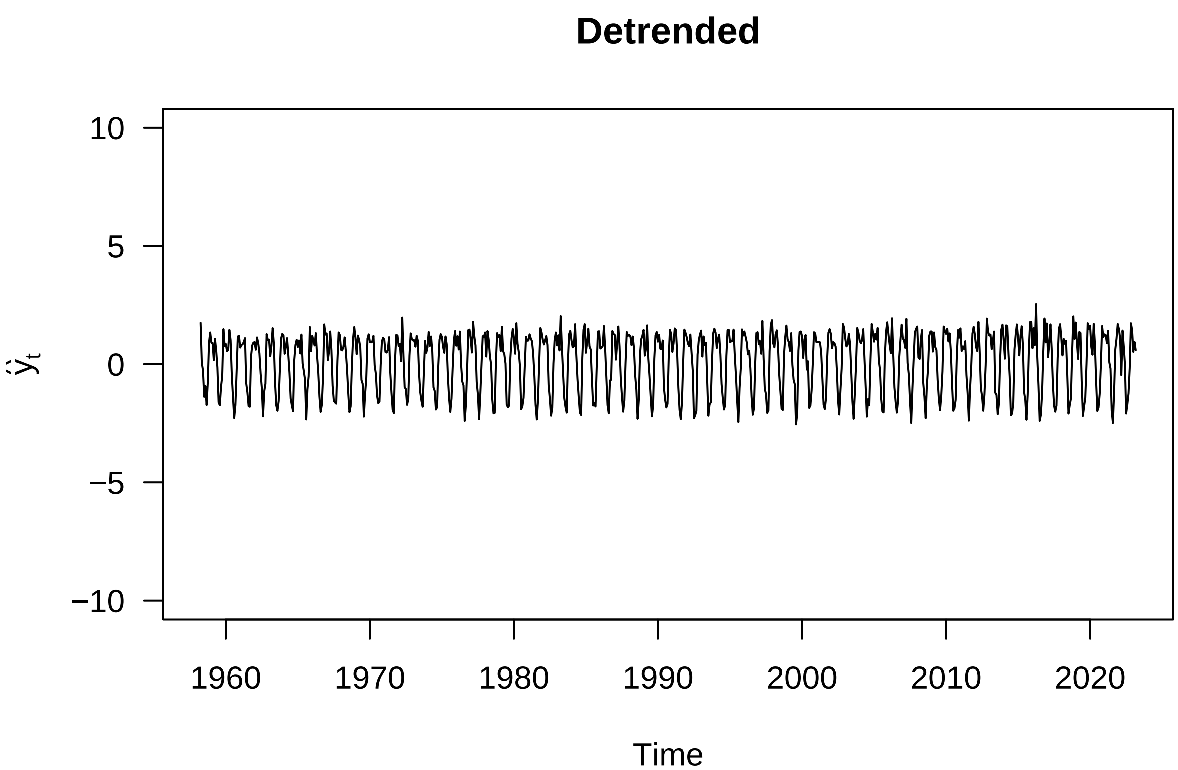


Figure 2: Detrended Co2 time series after applying first-order differencing. Variance appears stable, and seasonality remains visible.

Based off the ADF test, we fail to reject H_0. Therefore we will try differencing once.

$H_0 = $ The detrended and once differenced series has a unit root, making it non stationary $H_a =$ The detrended and once differenced series is stationary

```
adf.test(dtn_diff, alternative = "stationary")
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  dtn_diff
## Dickey-Fuller = -32.424, Lag order = 9, p-value = 0.01
## alternative hypothesis: stationary
```

Because the p values is much lower this time, .01, we can now reject H_0.

ACF & PACF

```
acf(dtn_diff, lag.max = 48, main="ACF of Co2", las = 1)
```

```
pacf(dtn_diff, lag.max = 48, main="PACF of Co2", las = 1)
```

Based off the ACF there are strong spikes at 1.0, 2.0, 3.0, indicating that there are seasonal spikes with period 12. Based off the PACF there is a large spike at lag 1, and then cuts off. There are spikes around month 12, which indicates seasonal AR terms. We will examine multiple arima models and also explore some other types that are often used for this type of data.

3.0 Statistical data analysis

Fitting suggested Models

lets look at an ARIMA model using the parameters based off the ACF and PACF. That being (1,1,0)(1,1,0) with the seasonality being [12]

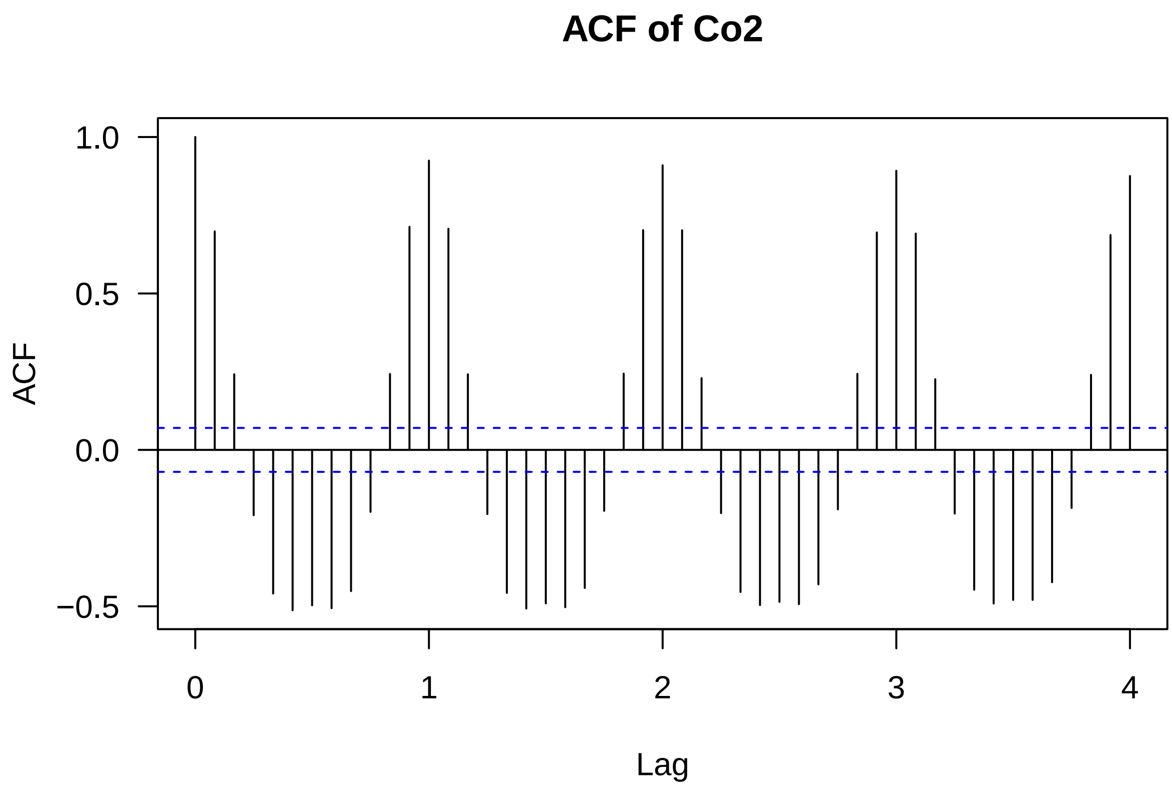


Figure 3: Autocorrelation Function (ACF) of the differenced Co2 series. Strong seasonal autocorrelation is visible at lags 12, 24, 36, suggesting annual seasonality.

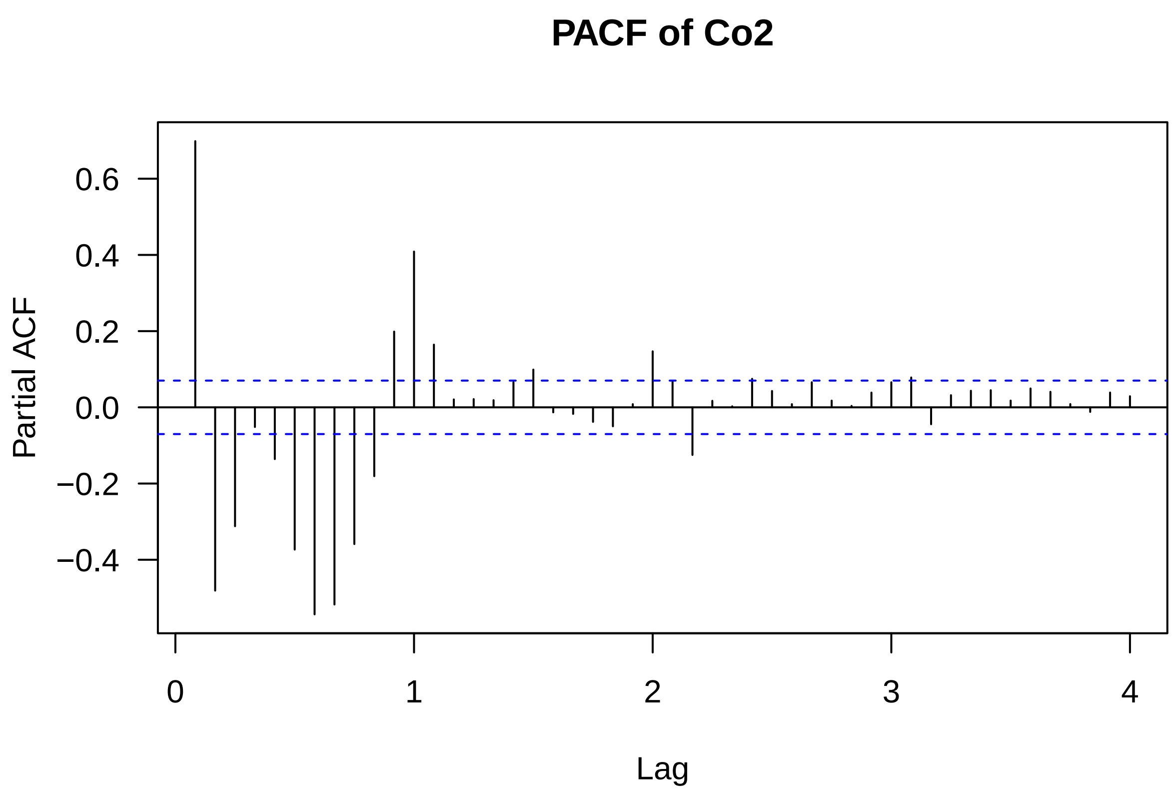
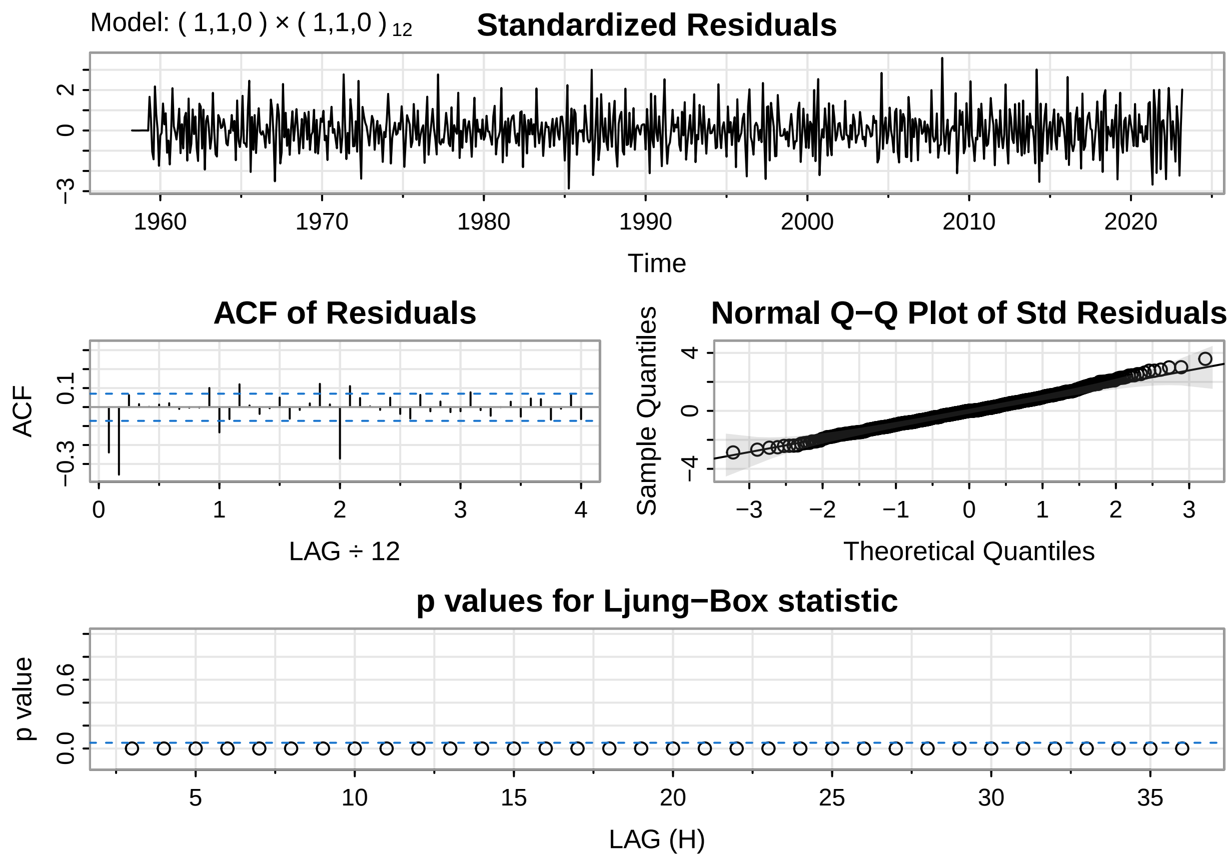


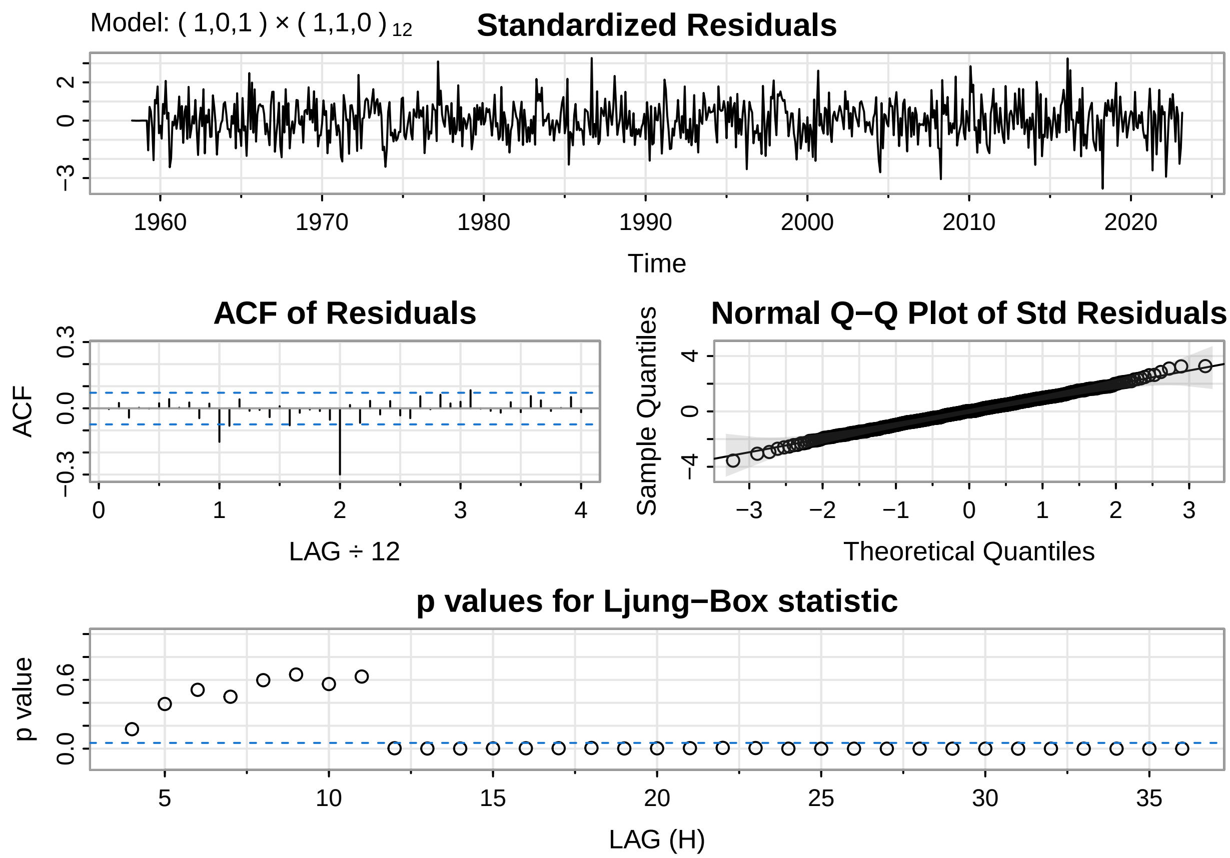
Figure 4: Partial Autocorrelation Function (PACF) of the differenced Co2 series. A sharp spike at lag 1 and decay afterward suggests an AR(1) component.

```
fit1 <- sarima(dtn_diff, 1, 1, 0,  1, 1, 0, 12)
```



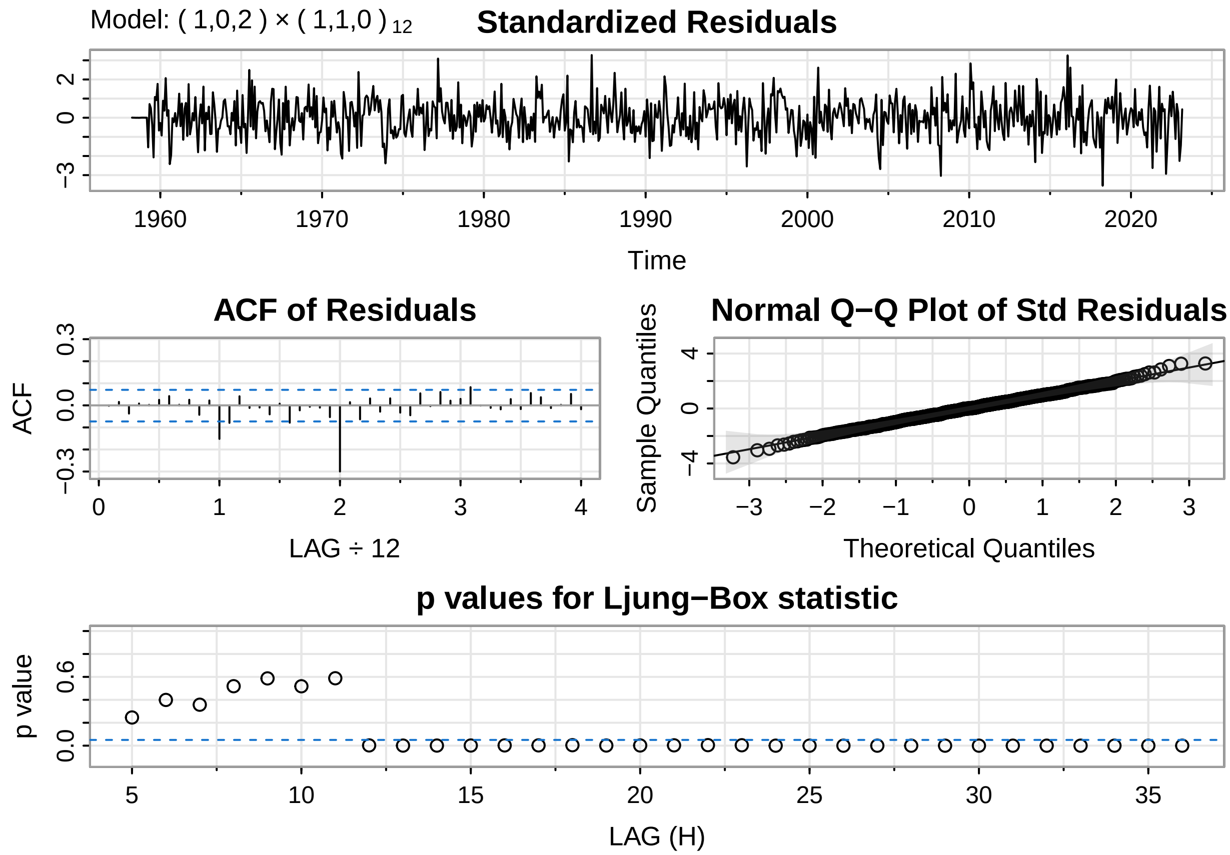
Most of the plots look good except for the p values, they are all 0. The seasonality seems to be a good fit, so we'' keep that, lets try a different one.

```
fit2 <- sarima(dtn_diff, 1,0,1, 1,1,0, 12)
```



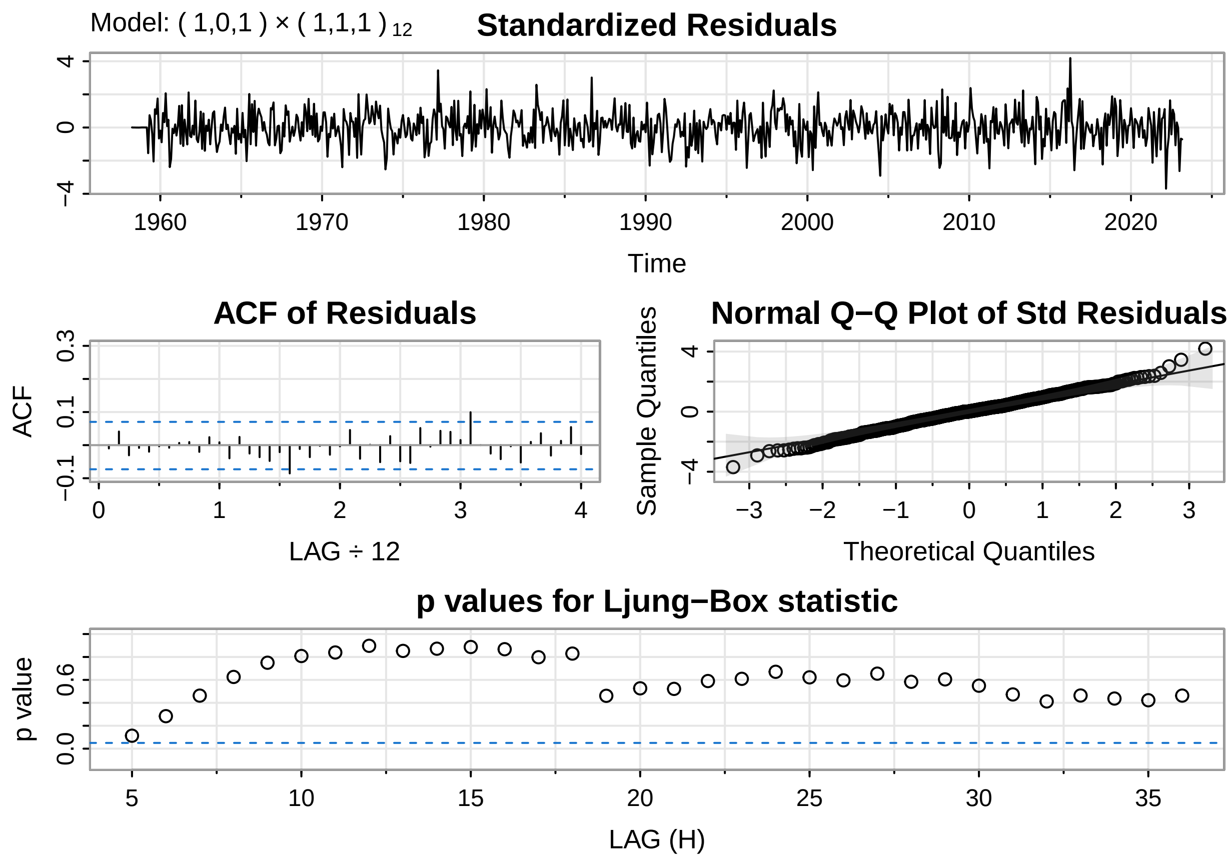
This makes the p values for the first 12 lags somewhat better, but still not the best. The other plots still look good, so lets try and change the q value for a better fit.

```
fit3 <- sarima(dtn_diff, 1,0,2, 1,1,0, 12)
```



No change in the p values, lets try changing the Q value then.

```
fit4 <- sarima(dtn_diff, 1,0,1, 1,1,1, 12)
```



Changing the Q values helped, so then having the model as (1,0,1) (1,1,1) (12) is one our of best looking models, the p values, QQ, and ACF looks great. Lets then look at our AIC and BIC and compare the different models we fitted.

```
ics1 <- fit1$ICs
ics2 <- fit2$ICs
ics3 <- fit3$ICs
ics4 <- fit4$ICs

tab_ICS <- data.frame(
  Model = c("Fit 1", "Fit 2", "Fit 3", "Fit 4"),
  AIC = c(ics1["AIC"], ics2["AIC"], ics3["AIC"], ics4["AIC"]),
  AICc = c(ics1["AICc"], ics2["AICc"], ics3["AICc"], ics4["AICc"]),
  BIC = c(ics1["BIC"], ics2["BIC"], ics3["BIC"], ics4["BIC"])
)
kable(tab_ICS, caption = "Table 1: AIC, AICc, and BIC values for four SARIMA models fittec
```

Table 1: Table 1: AIC, AICc, and BIC values for four SARIMA models fitted to the differenced Co2 time series. Lower values indicate better model fit, with Model 4 showing the best overall performance.

Model	AIC	AICc	BIC
Fit 1	1.4269073	1.4269278	1.4450657
Fit 2	0.8040683	0.8041365	0.8343013
Fit 3	0.8062866	0.8063891	0.8425662
Fit 4	0.5384247	0.5385272	0.5747043

Based off the AIC and BIC it seems that the fourth model was definitely the best. That it expected as as we progressed in building our model to be the best that it could be. Lets look at it forecasted ahead 3 years.

```
sarima.for(cardox, 36, 1,0,1, 1,1,1, 12)
```

Estimated models

Based off our best model, this is our written out model.

$$x_t = 0.0002 + 0.2632\, x_{t-1} - 0.6237\, w_{t-1} + 0.0096\, x_{t-12} - 0.8818\, w_{t-12} + w_t$$

as well as its coefficients

```
coef_table <- data.frame(
  Coefficient = c("AR(1)", "MA(1)", "SAR(1)", "SMA(1)", "Constant"),
  Estimate = c(0.2632, -0.6237, 0.0096, -0.8818, 0.0002),
  StdError = c(0.0902, 0.0740, 0.0409, 0.0194, 0.0001),
  pValue = c(0.0036, 0.0000, 0.8145, 0.0000, 0.0161)
)

kable(coef_table, caption = "Table 2: Estimated Coefficients for SARIMA(1,0,1)(1,1,1)[12]
```

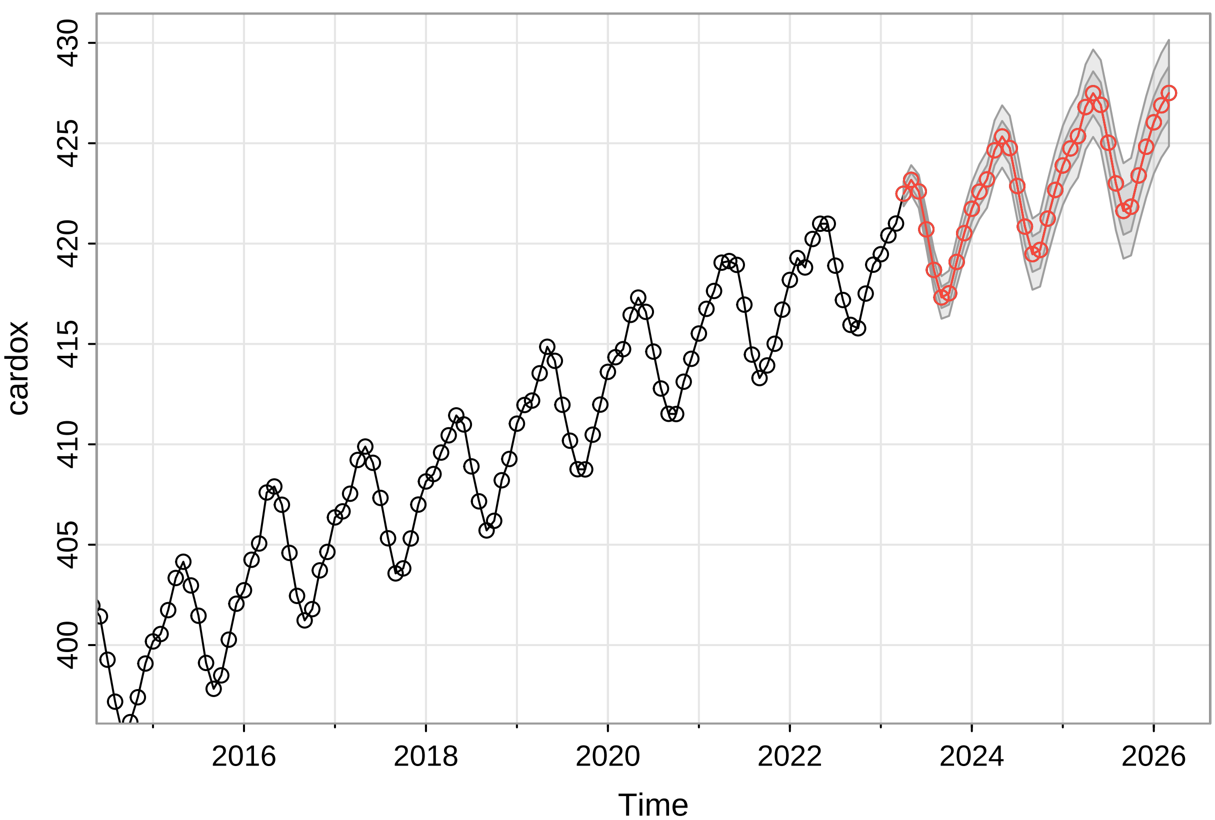


Figure 5: Forecast of atmospheric Co2 for 36 months using the fitted SARIMA(1,0,1)(1,1,1)[12] model. The forecast continues the trend and seasonality observed in the historical data.

Table 2: Table 2: Estimated Coefficients for SARIMA(1,0,1)(1,1,1)[12] Model

Coefficient	Estimate	StdError	pValue
AR(1)	0.2632	0.0902	0.0036
MA(1)	-0.6237	0.0740	0.0000
SAR(1)	0.0096	0.0409	0.8145
SMA(1)	-0.8818	0.0194	0.0000
Constant	0.0002	0.0001	0.0161

4.0 Conclusions

Results of analysis

Based off our analysis of the multiple models, the SARIMA (1,0,1)(1,1,1)[12]) was chosen because of multiple reasons. The first being the residual analysis and that it had the lowest AIC and BIC values out of all the others. The Ljung-Box p values showed that the residuals were similar to white noise, and the QQ plot looked normal with no strange tailing off at the fringes. The model was able to capture the short and long term changes as well as the seasonal dynamic effectively, which was the 12 month seasonal structure. The model when forecasted against the original time series continues the upward trend of the atmospheric CO2 level. The oscillation of the levels seasonally are expected to continue, with a rise in the spring months and the tapering off during the late summer months.

Answers to scientific questions

1. In regards to if there is a upward trend in the co2 levels in the atmosphere, the answer would be yes. Based off the analysis of the time series and the forecasting of said time series, there is a clear upward trend in the back round levels of it. Using forecasting we can see that that trend continues to climb in both level and trend line.

2. There are clear indicators of seasonality in our analysis. As far as looking into the extent of that seasonality, it looks to be a constant rate, one that has been the same since the beginning of the recording. The seasonality is a 12 month cycle that is reflected every year.

Limitatios of analysis

While the model did a good job and was a good fit, there are several limitations to the analysis. One of them being that the model assumes that the seasonality is constant. While in reality the magnitude of it could change over time due to factors like human activity or ecological change. An additional one is that the data is from a single location, the Mauna Loa observatory. Because of this, it may not reflect the global Co2 variability. Another problem could be that human and global interventions (volcanic eruption, COVID, the gas crisis, global recession) are not explicitly represented. From the modeling that was done it could have just smoothed over those things.

Potential future work

If I were to further this research there are a couple things that would be incorporated in the further analysis of this topic. I would try and use outside variables to better understand the factors that drive the levels of Co2. It would be things like GDP, oil price, global energy consumption, and temperature anomalies. I would also strive to use more complex models that better capture the other complexities of the model. I would also look at global policies and laws that were concerned with climate change and see if there any changes in the level in the following years.