STAT 153 Final Project

Wako Morimoto
wakomorimoto@
SID: 3032639716

Max Zhang
max.zhang@
SID: 3034535635

Jiyeon Seo
ppori98@
SID: 3036015267

Yunjae Cho
jyj960814@
SID: 3033617604

Raymond Wang
raymondwang@
SID: 3034017055

Spring 2022

Abstract

In this project, we analyzed air pollution data from Data.World that includes information on five main pollutants in the U.S. After the close examination of the data set and the model, we have decided to make a prediction of O_3 (Ground-level Ozone) concentration using a Sinusoid + SARIMA$(2,0,0) \times (1,0,1)_{12}$ model. This analysis and prediction are crucial in deepening the understanding of physical and environmental health in cities across the country. We hope that this research would contribute to further analysis in air pollution that plays an integral part in climate change.

1 Introduction

Phoenix is known to be one of the cities with the worst level of air pollution. The various pollutants in Phoenix are associated with negative health consequences, specifically lung and heart health, for residents. In this project, our group analyzes and predicts one of the pollutant indexes in the city to capture the pattern. We hope that this will achieve two goals. First, we want to have a better understanding of the causes of pollution, especially any causes related to events or times of the year. Second, we want to identify any patterns that can help minimize future pollution.

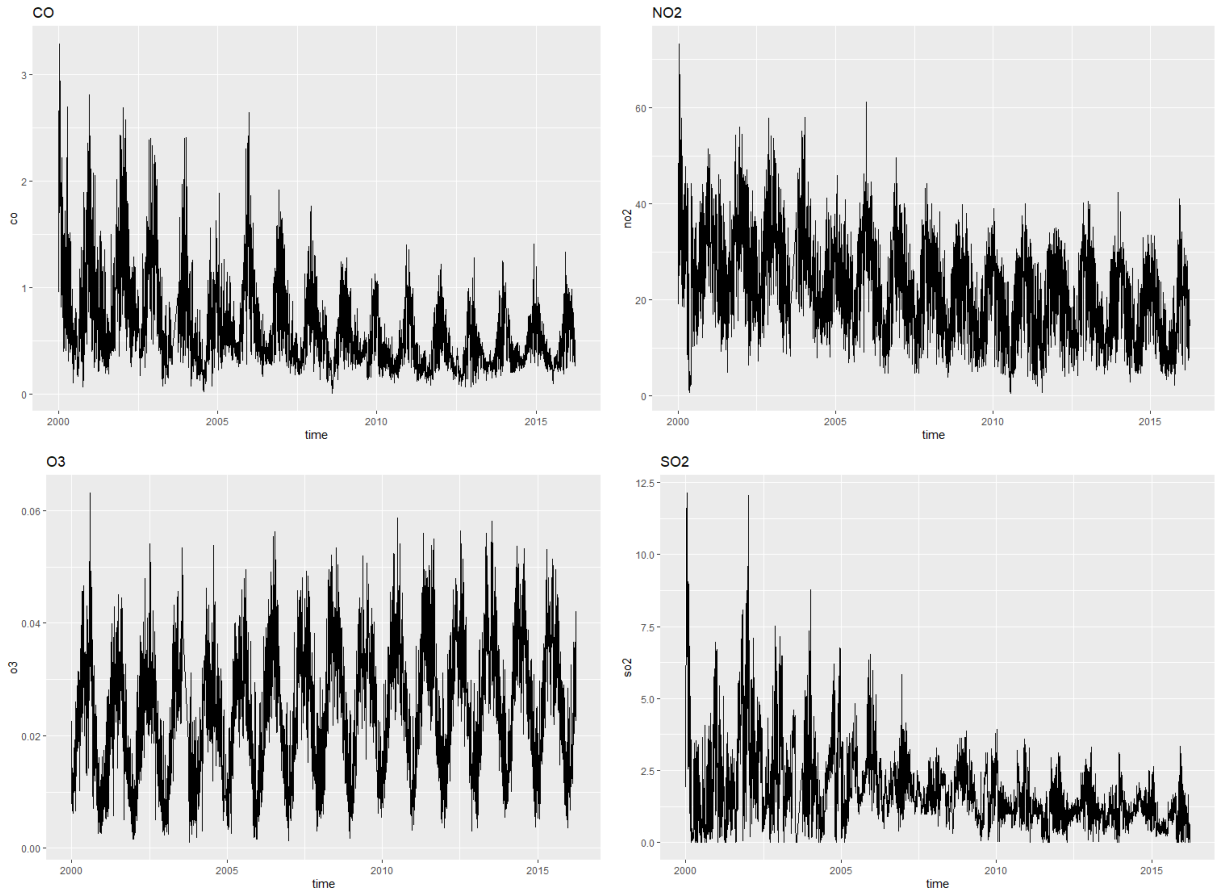
We acquired pollutant data from Data.World (https://data.world/data-society/us-air-pollution-data). This dataset consists of seven geolocation columns (State Code, State, County Code, County, Site Num, City, Address), one date column (Date Local), and five specific columns each for the four pollutants (NO_2, O_3, SO_2, and O_3) observed. For example, for O_3:

- O_3 Units: The units measured for O_3.
- O_3 Mean: The mean of concentration of O_3 within a given day.
- O_3 AQI: The calculated air quality index of O_3 within a given day
- O_3 1st Max Value: The maximum value index of O_3 within a given day
- O_3 1st Max House: The house where the O_3 1st Max Value is recorded

The granularity of our data is essentially one observation per day by location. For our project, we focus only on data points recorded in Phoenix, AZ. Observations range from January 2000 to December 2016.

2 Exploratory Data Analysis

The raw daily time series for all four pollutants is shown below:



Since 2003 and 2016 do not contain all 12 months worth of data, we have decided to truncate the time frame of interest to 2004-2015 to maintain uniformity.

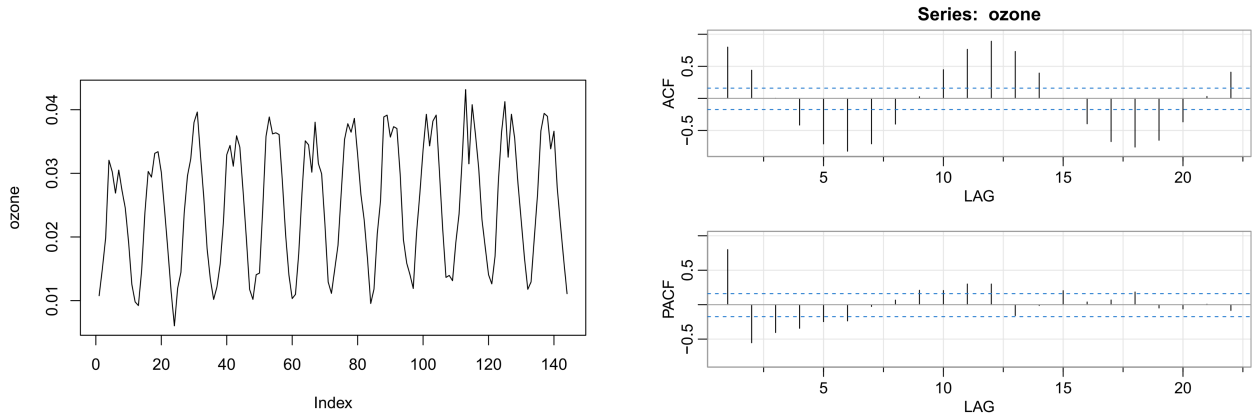
All four pollutants seem to have strong seasonality, though only CO and SO_2 have visible heteroscedasticity. On the other hand, NO_2 and O_3 seem to have slight trend throughout time while the rest of the two may also show trend. Among the four pollutants, Nitrogen Dioxide (NO_2) is considered to be the most harmful one as it is directly associated with lung and heart disease[1]. Out of the four, O_3 visibly has the strongest seasonality and is also a great indicator of severity of pollution since NO_2 causes the formation of O_3. For these reasons, we have decided to focus on modeling levels of O_3.

We chose to aggregate observations at the monthly level, since we are more interested in longer-term trends than attempting to predict O_3 levels on a day-to-day basis. Since September and August are missing from 2003, we decided to look only at the series from January 2004 to December 2015, resulting in 144 observations. It is important to note that the O_3 is measured in parts per million.

3 Models Considered

Our monthly aggregated data for ozone levels shows seasonality, which is also evident in the ACF and PACF plots.

[1] "5 Dangerous Pollutants You're Breathing in Every Day." UNEP, https://www.unep.org/news-and-stories/story/5-dangerous-pollutants-youre-breathing-every-day.

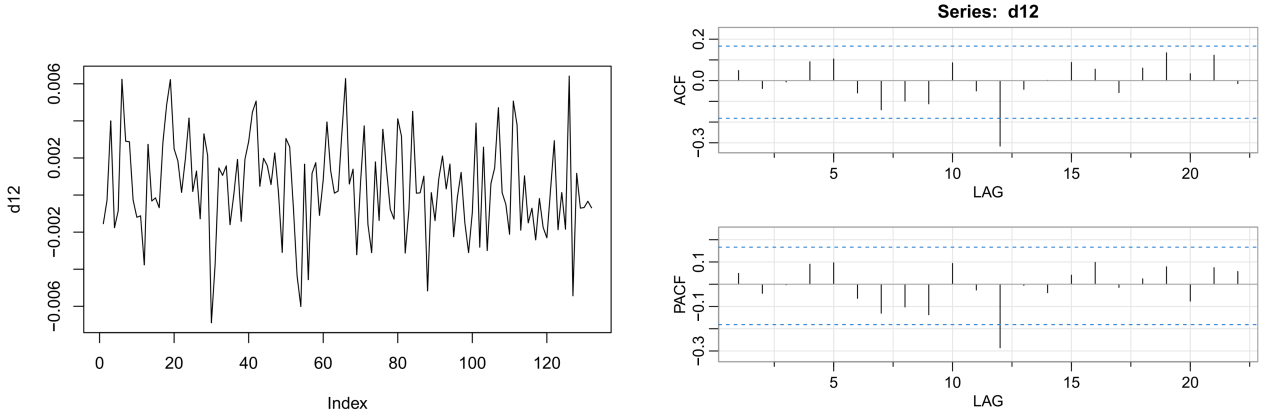


We can observe seasonality, which is also evident in the ACF and PACF plots.

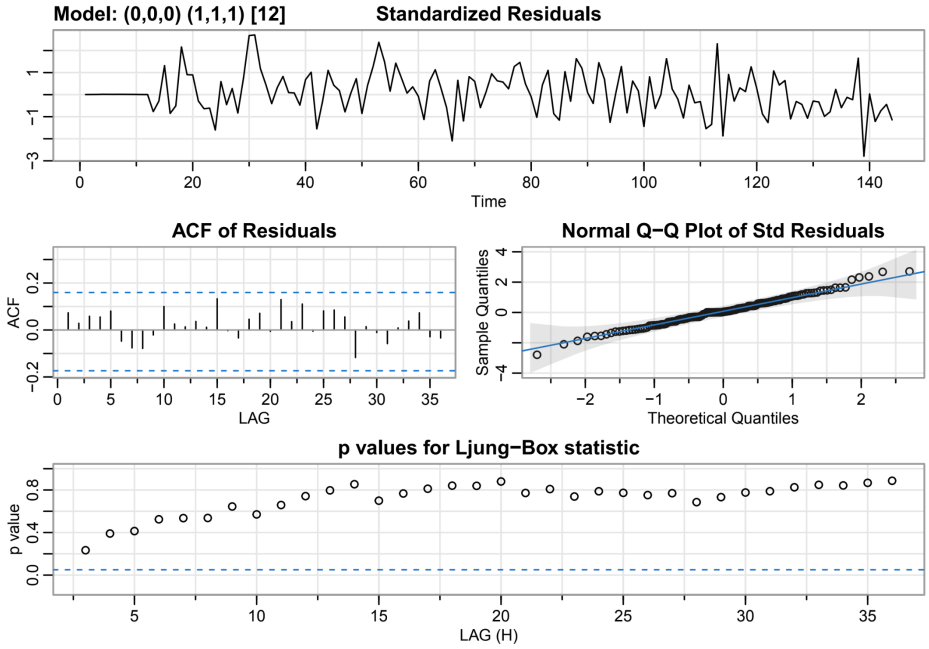
We break down our model exploration into two sections, Non-parametric and Parametric in the following 2 sections.

3.1 Non-parametric Model

To achieve stationarity, we apply a lag 12 difference to account for yearly seasonality. The ACF and PACF plots are shown below. The only significant autocorrelation/partial autocorrelation is at lag 12, which suggests trying a SARIMA model with a seasonal MA and/or AR component.

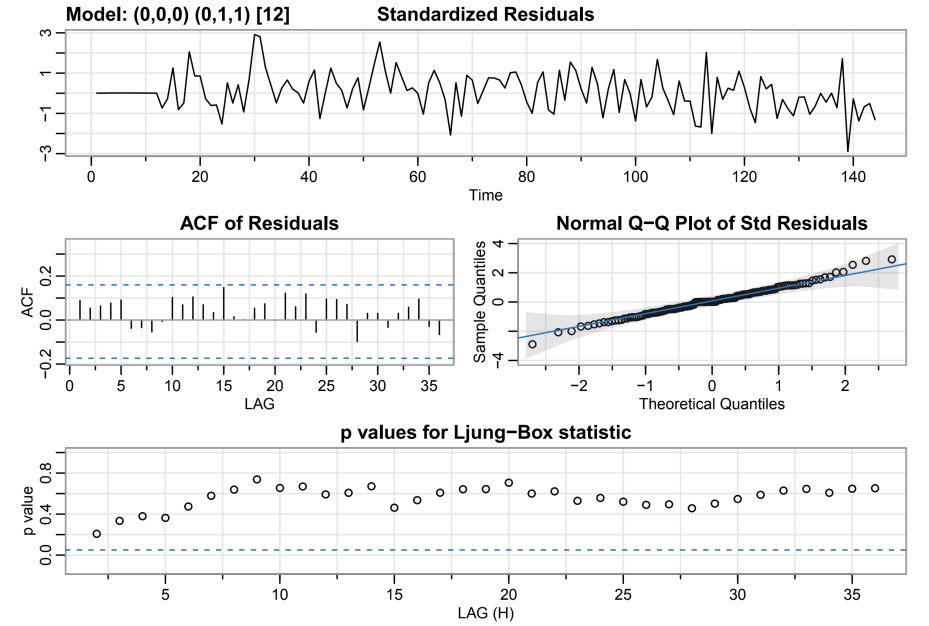


3.1.1 Model 1: SARIMA$(0,0,0) \times (1,1,1)_{12}$



The standardized residuals appear to exhibit stationary behavior over time. We can also be confident in assuming stationarity after looking at the ACF plot of the residuals, as the sample autocorrelations for every displayed lag are within the 95% confidence bands. The Q-Q plot also indicates that the normality assumption of the residuals is satisfied. Finally, we can see that none of the Ljung-Box statistic p-values are significant at the 5% level, meaning that we fail to reject the null hypothesis that the residuals are white noise generated from the proposed ARMA model. All of these diagnostic plots suggest that this SARIMA model is a good fit to the data.

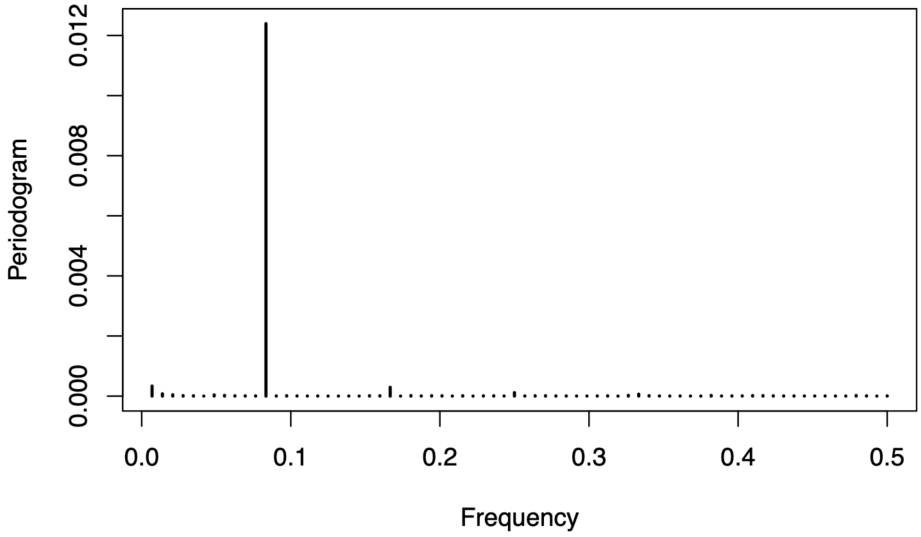
3.1.2 Model 2: SARIMA$(0,0,0) \times (0,1,1)_{12}$



The diagnostics for our second non-parametric SARIMA model are very similar to our first model. The standardized residuals appear to exhibit stationary behavior over time. We can also be confident in assuming stationarity after looking at the ACF plot of the residuals, as the sample autocorrelations for every displayed lag are within the 95% confidence bands. The Q-Q plot also indicates that the normality assumption of the residuals is satisfied. Finally, we can see that none of the Ljung-Box statistic p-values are significant at the 5% level, meaning that we fail to reject the null hypothesis that the residuals are white noise generated from the proposed ARMA model. All of these diagnostic plots suggest that this SARIMA model is a good fit to the data.

3.2 Parametric Model

In our search for a parametric model, we observed a lot of seasonality in the data, which suggested that a sinusoidal model could be a good candidate. We display the periodogram below:



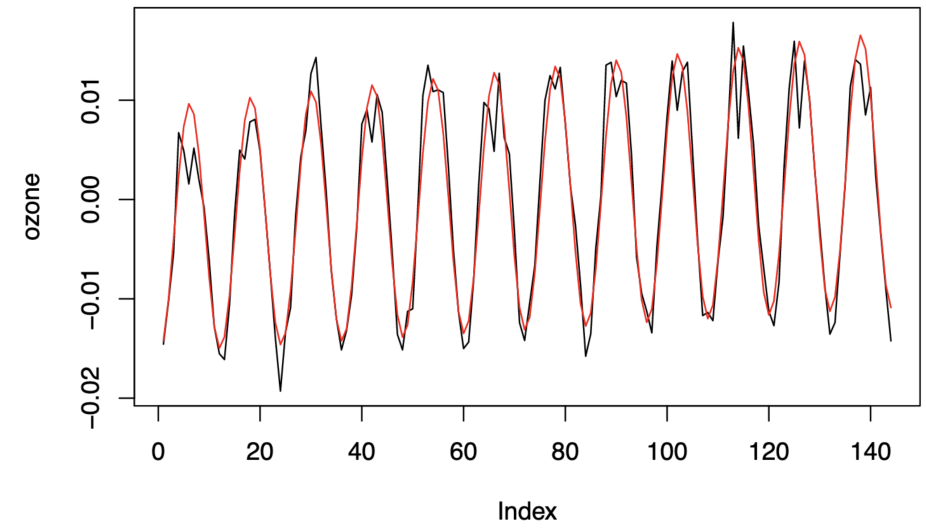
We clearly observe that there is one significant frequency, around 0.0833, which corresponds to a seasonal component of 1/12, or a yearly season since the data is monthly.

Accordingly, we define our parametric model as

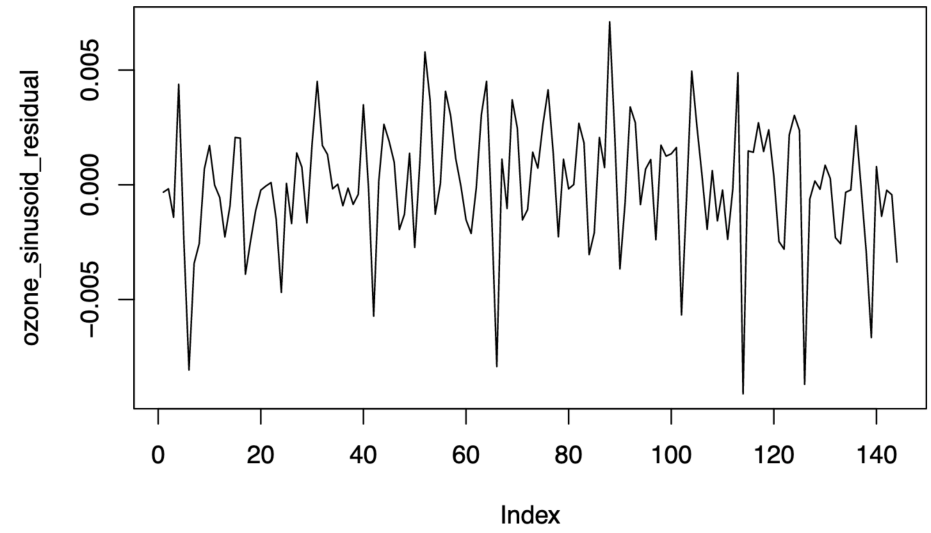
$$f(t) = \beta_0 + \beta_1 t + \beta_2 \sin(t) + \beta_3 \cos(t) + \beta_4 t \sin(t) + \beta_5 t \cos(t)$$

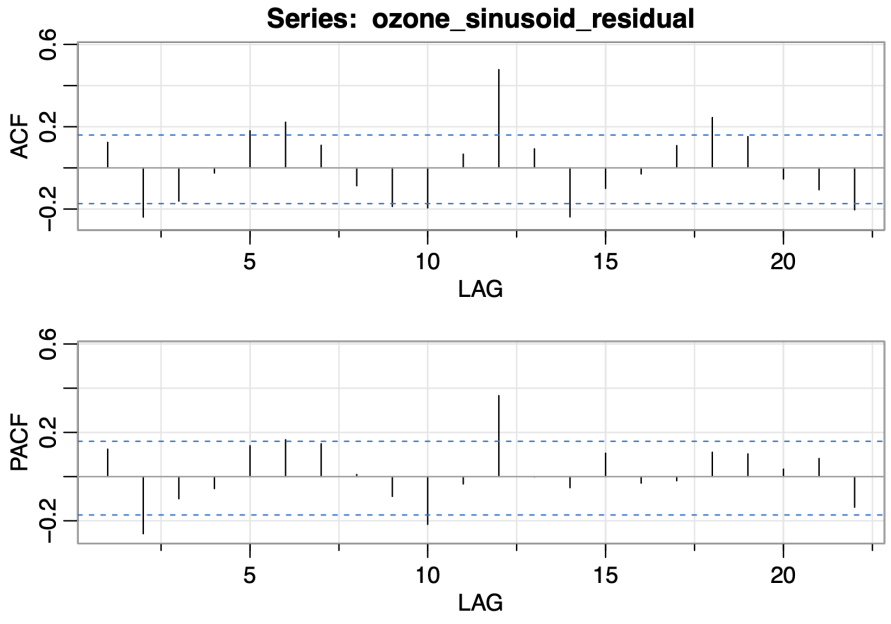
Since we want to achieve stationary, first let V_t denote the ozone level at time t. Then $X_t = V_t - f(t)$ will be stationary.

Our sinusoidal model captures the seasonality quite well:



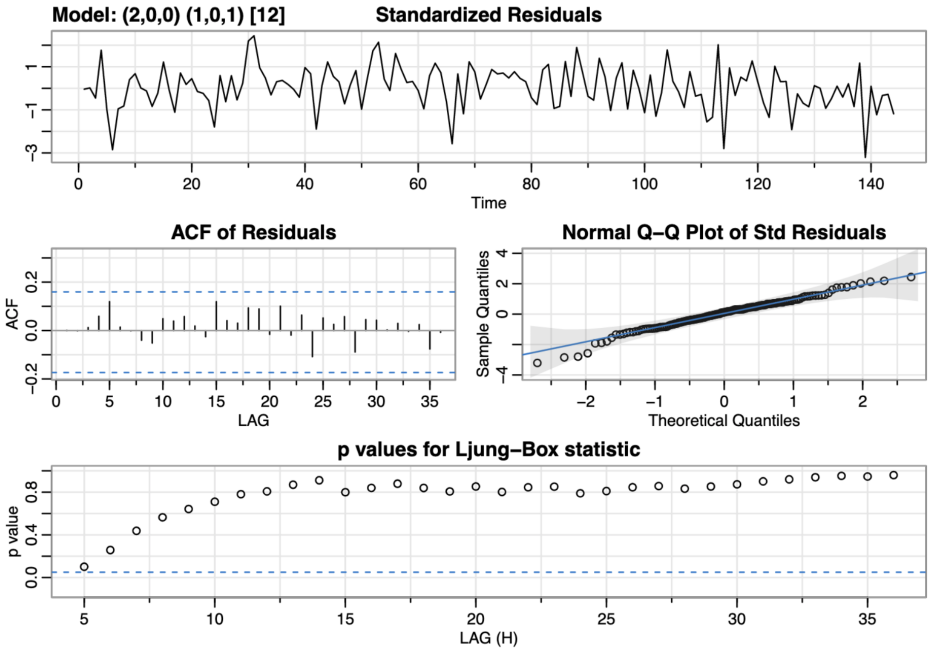
The residuals, along with their ACF and PACF, are shown below:





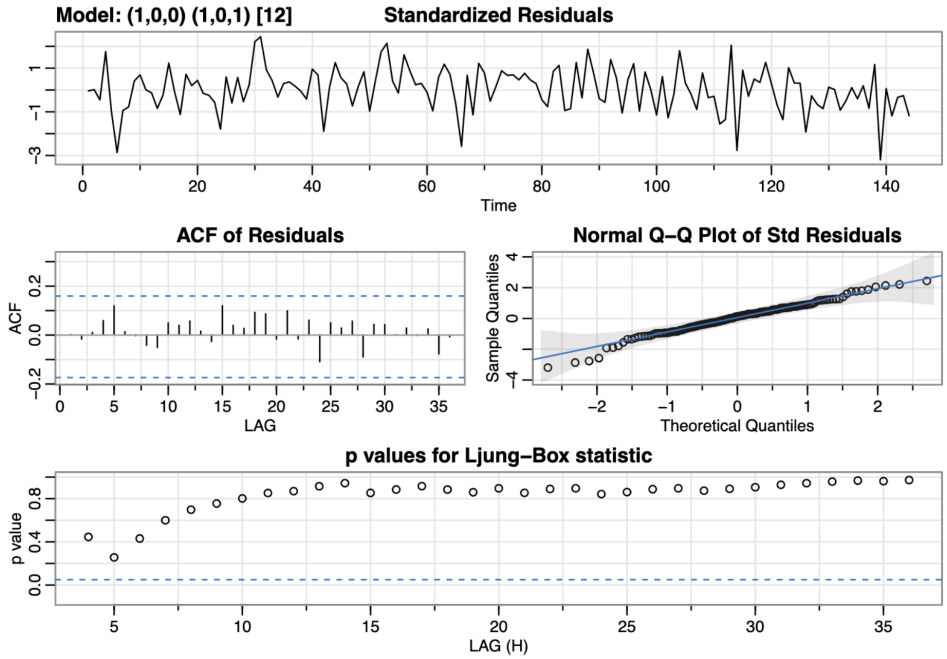
The residuals appear to be stationary, so we now attempt to fit a SARIMA model on the residuals. We used the ACF/PACF plots of the residuals, along with experimentation with different ARMA and seasonal ARMA parameters, to develop the following two models.

3.2.1 Model 3: Sinusoid + SARIMA$(2,0,0) \times (1,0,1)_{12}$



The plot of the standardized residuals shows that they exhibit stationary behavior over time. We are confident in assuming stationarity after looking at the ACF plot of the residuals as well, as the sample autocorrelations for all the lags displayed are within the 95% confidence bands. The Q-Q plot also indicates that the normality assumption of the residuals is satisfied. Finally, we can see that the Ljung-Box statistic p-values are all not significant at the 5% level, meaning we fail to reject the null hypothesis that the residuals are white noise generated from the proposed ARMA model, suggesting a good fit to the data.

3.2.2 Model 4: Sinusoid + SARIMA$(1,0,0) \times (1,0,1)_{12}$



Our second parametric model has diagnostic plots similar to our first parametric model. The plot of the standardized residuals shows that they exhibit stationary behavior over time. We are confident in assuming stationarity after looking at the ACF plot of the residuals as well, as the sample autocorrelations for all the lags displayed are within the 95% confidence bands. The Q-Q plot also indicates that the normality assumption of the residuals is satisfied. Finally, we can see that the Ljung-Box statistic p-values are all not significant at the 5% level, meaning we fail to reject the null hypothesis that the residuals are white noise generated from the proposed ARMA model, suggesting a good fit to the data.

4 Model Comparison and Selection

To choose our final model, we performed cross-validation and also compared AIC, AICc, and BIC to select our model. In our cross-validation, we predict mean monthly ozone in parts per million for all 12 months of the year for the years 2010-2015. In predicting monthly ozone for a given year, we use all years previous all the way back to 2004. After predicting, we calculate the SSE for each year, and for each model we use the sum of the six SSE values as our final cross-validation score. CV scores and criteria values are reported below:

Model Name	**CV Score**	**AIC**	**AICc**	**BIC**
SARIMA$(0,0,0) \times (1,1,1)_{12}$	0.000385	-9.26	-9.26	-9.17
SARIMA$(0,0,0) \times (0,1,1)_{12}$	0.000391	-9.25	-9.25	-9.19
Sinusoid + SARIMA$(2,0,0) \times (1,0,1)_{12}$	0.000347	-9.28	-9.28	-9.16
Sinusoid + SARIMA$(1,0,0) \times (1,0,1)_{12}$	0.000350	-9.29	-9.29	-9.19

Overall, non-parametric models generally have higher CV scores than parametric models. Looking at the rest of the three evaluations, there are only slightly difference among the four models. With these taken into consideration, Sinusoid + SARIMA$(2,0,0) \times (1,0,1)_{12}$ model seems to be the best one due to its lowest CV score.

5 Results

5.1 Estimation of model parameters

Our final model is $X_t = V_t + f(t)$, where

$$f(t) = \beta_0 + \beta_1 t + \beta_2 \sin(t) + \beta_3 \cos(t) + \beta_4 t \sin(t) + \beta_5 t \cos(t)$$

with parameters

$$f(t) = -0.003 + 0.000042 * t - 0.0012 * \sin(t) - 0.0123 * \cos(t) + 0.000002 * t \sin(t) - 0.000011 * t \cos(t)$$

and

$$V_t = \phi_1 V_{t-1} + \phi_2 V_{t-2} + \Phi V_{t-12} + \Theta W_{t-12}$$

We choose this model as we found the cross validation scores were the lowest along with good fits for the residuals of the sinusoidal model. This is an interesting phenomenon as we found that the Sinusoid + SARIMA$(1,0,0) \times (1,0,1)_{12}$ was underfitting the data slightly as it had more errors in the predictions.

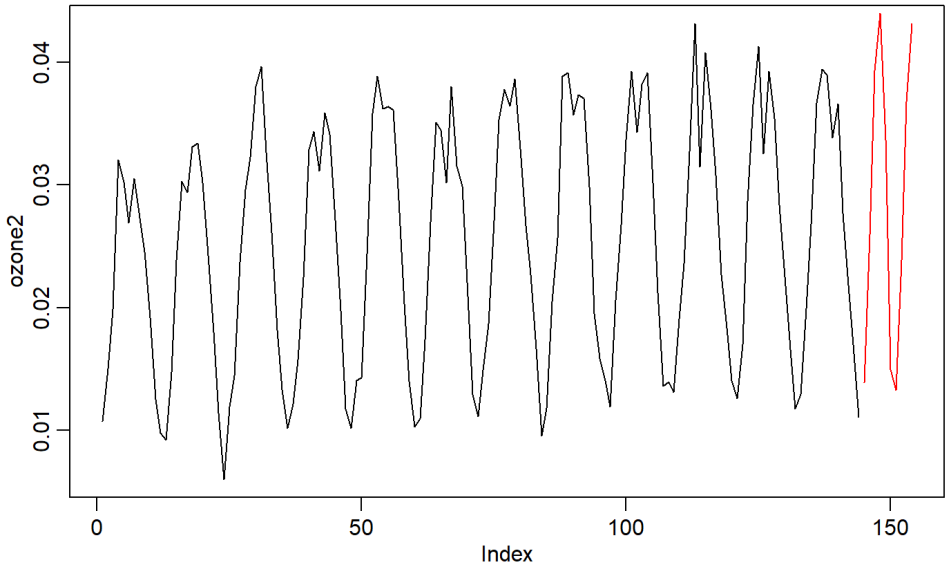
Here are the parameters for the ARIMA model along with their standard errors.

Parameter	Estimate	SE
ϕ_1	0.0903	0.0837
ϕ_2	-0.0160	0.0881
Φ	0.9462	0.0545
Θ	-0.7176	0.1439

5.2 Prediction

Using our best model, we predict the ozone levels of the next 10 months.

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
0.0139	0.0260	0.0393	0.0440	0.0337	0.0150	0.0133	0.0239	0.0368	0.0432



6 Conclusion

After comparing the evaluation scores, we have concluded that parametric models fit better than non-parametric models for this specific data-set. Two of the Sinusoid + SARIMA models have lower CV Scores and other evaluations (AIC, AICc, and BIC) than non-parametric SARIMA models. Comparing the two parametric models, the one with phi value of 2 is better than the other because of its lower CV Score. In addition, the model with phi value of 1 may be susceptible to underfitting, leaving us with the conclusion that our model of Sinusoid + SARIMA$(2,0,0) \times (1,0,1)_{12}$ is the best among the four models tested.

Going back to our scientific problem, we have learned that model creation on time series is exceedingly helpful in capturing data's pattern and predicting future values. These predictions give us insights into the future of our environment, especially greenhouse gases. It is not an ominous future if we take action now. After all, these are predictions based on past behavior, if we change our behavior now, we can reverse the effects of climate change and protect the world we live in.

7 Code Appendix

7.1 Data Cleaning

```
library(tidyverse)
library(lubridate)
library(forecast)
library(astsa)

pollution <- read_csv(file = 'pollution_us_2000_2016.csv')

phoenix <- pollution %>%
  filter(State == 'Arizona', City == 'Phoenix') %>%
  select(`Date Local`, `NO2 Mean`, `O3 Mean`, `SO2 Mean`, `CO Mean`) %>%
  mutate(year = year(`Date Local`),
         month = month(`Date Local`),
         day = day(`Date Local`)) %>%
  filter(year %in% 2004:2015) %>%
  group_by(year, month) %>%
  summarize(no2 = mean(`NO2 Mean`),
            o3 = mean(`O3 Mean`),
            so2 = mean(`SO2 Mean`),
            co = mean(`CO Mean`)) %>%
  ungroup()
```

7.2 Analysis

```
# raw time series
ozone <- phoenix %>%
    pull(o3)
plot(ozone, type = 'l')

# yearly difference
d12 <- diff(ozone, lag = 12)
plot(d12, type = 'l')
acf2(d12)
```

7.2.1 Non-parametric Models

```
# model 1: SARIMA(0,0,0)x(1,1,1)_12
model1 <- sarima(ozone, p = 0, d = 0, q = 0, P = 1, D = 1, Q = 1, S = 12)
model1$AIC
model1$AICc
model1$BIC

# model 2: SARIMA(0,0,0)x(0,1,1)_12
model2 <- sarima(ozone, p = 0, d = 0, q = 0, P = 0, D = 1, Q = 1, S = 12)
model2$AIC
model2$AICc
```

```
model2$BIC

# cross-validation
sse <- matrix(NA, nrow = 2, ncol = 6)  # 2 models, test on 2010-2015

for (i in 1:6) {
  train <- ozone[1:(12 * (i + 5))]
  test <- ozone[(12 * (i + 5) + 1):(12 * (i + 6))]

  m1 <- sarima.for(train, n.ahead = 12,  # 30
                   p = 0, d = 0, q = 0, P = 1, D = 1, Q = 1, S = 12)
  m2 <- sarima.for(train, n.ahead = length(test),
                   p = 0, d = 0, q = 0, P = 0, D = 1, Q = 1, S = 12)

  sse[1,i] <- sum((test - m1$pred)^2)
  sse[2,i] <- sum((test - m2$pred)^2)
}

# CV score: sum of SSEs
rowSums(sse, na.rm = TRUE)
```

7.2.2 Parametric Models

```
t = 1:length(ozone)
# Check the periodogram: one significant peak
periodo = periodogram(ozone, plot=TRUE,
                      ylab = "Periodogram", xlab = "Frequency")

# get the high magnitudes in descending order
order_spec = sort(periodo$spec, decreasing = TRUE)

# get the frequency that gives max magnitude
first_max = order_spec[1]
first_maximizing_freq = periodo$freq[periodo$spec == first_max]
first_sin_max = sin(2 * pi * first_maximizing_freq * t)
first_cos_max = cos(2 * pi * first_maximizing_freq * t)

# max sinusoidal fitting
ozone_sinusoid_model = lm(ozone ~ first_sin_max*(1 + t) +
                            first_cos_max*(1 + t))
print(ozone_sinusoid_model$coefficients)

# overlay the sinusoidal fitting over the original plot
plot(ozone,type = "l")
lines(t, ozone_sinusoid_model$fitted.values, col = "red")

# get the residual, hoping to remove seasonality
ozone_sinusoid_residual = ozone_sinusoid_model$residuals
plot(ozone_sinusoid_residual, type = "l")  # seems to be stationary
acf2(ozone_sinusoid_residual)
```

```
# model 3 : SARIMA(2,0,0)x(1,0,1)_12
model3 <- sarima(ozone_sinusoid_residual,
                 p=2, d=0, q=0, P=1, D=0, Q=1, S=12)
coeff_table <- as.data.frame(model1$ttable)
coeff_table <-coeff_table %>% mutate(ci_lower = Estimate-1.96*SE,
                                     ci_upper =  Estimate+1.96*SE)
coeff_table  # show estimated coefficient and its ci

# model 4: SARIMA(1,0,0)x(1,0,1)_12
model2 <- sarima(ozone_sinusoid_residual,
                 p=1, d=0, q=0, P=1, D=0, Q=1, S=12)
coeff_table <- as.data.frame(model2$ttable)
coeff_table <-coeff_table %>% mutate(ci_lower = Estimate-1.96*SE,
                                     ci_upper =  Estimate+1.96*SE)
coeff_table  # show estimated coefficient and its ci

# AIC, AICc, BIC
eval<- function(model){
  return (c(model$AIC, model$AICc,model$BIC))
}

m1_evaludation = eval(model1)
m2_evaludation = eval(model2)
m3_evaludation = eval(model3)
m4_evaludation = eval(model4)

eval_matrix = rbind(m1_evaludation,
                    m2_evaludation,
                    m3_evaludation,
                    m4_evaludation)
rownames(eval_matrix) = c("SARIMA(2,0,0)(1,0,1)12",
                          "SARIMA(1,0,0)(1,0,1)12",
                          "SARIMA(2,1,1)(1,0,1)12",
                          "SARIMA(2,0,1)(1,0,1)12")
colnames(eval_matrix) = c("AIC","AICc","BIC")
eval_matrix

# cross-validation
sse1 = c()
sse2 = c()
sse3 = c()
sse4 = c()
test_years = seq(10,15,1)
for (year in test_years) {

train_index = 1:(12*(year-4))
test_index = (12*(year-4)+1):(12*(year-4+1))

train <- ozone_sinusoid_residual[train_index]
test <- ozone_sinusoid_residual[test_index]
```

```
m1_forecast <- sarima.for(train, n.ahead=12,
                          p=2, d=0, q=0, P=1, D=0, Q=1, S=12)$pred
m2_forecast <- sarima.for(train, n.ahead=12,
                          p=1, d=0, q=0, P=1, D=0, Q=1, S=12)$pred
m3_forecast <- sarima.for(train, n.ahead=12,
                          p=2, d=1, q=1, P=1, D=0, Q=1, S=12)$pred
m4_forecast <- sarima.for(train, n.ahead=12,
                          p=2, d=0, q=1, P=1, D=0, Q=1, S=12)$pred

sse1 = c(sse1,sum((m1_forecast - test)^2))
sse2 = c(sse2,sum((m2_forecast - test)^2))
sse3 = c(sse3,sum((m3_forecast - test)^2))
sse4 = c(sse4,sum((m4_forecast - test)^2))
}

sse = rbind(sum(sse1),sum(sse2),sum(sse3),sum(sse4))
rownames(sse) = c("SARIMA(2,0,0)(1,0,1)12",
                  "SARIMA(1,0,0)(1,0,1)12",
                  "SARIMA(2,1,1)(1,0,1)12",
                  "SARIMA(2,0,1)(1,0,1)12")
colnames(sse) = c("SSE")
print(sse)
```

7.3 Prediction

```
# predict the next 10 points
m1_forecast <- sarima.for(ozone_sinusoid_residual, n.ahead=10,
                          p=2, d=0, q=0, P=1, D=0, Q=1, S=12)$pred

ts = as.data.frame(145:154)
predict_t = 145:154
prediction = c()
coef = ozone_sinusoid_model$coefficients

for (t in predict_t) {
  pred = coef[1] + coef[2]*sin(t) + coef[3]*t +
         coef[4]*cos(t) + coef[5]*sin(t)*t + coef[6]*cos(t)*t
  prediction = c(prediction, pred)
}
prediction

m1_forecast = as.vector(m1_forecast)
prediction = as.vector(prediction)
predictions = m1_forecast+prediction

ozone2 = c(ozone,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA)

ts2 = 1:154
preds = rep(NA, 144)
preds = c(preds,predictions)
```

```
plot(ozone2, type ="l")
lines(ts2, preds, col = "red")
```