Climate Change Contributor: Predicting Motor Vehicle Emissions with Time Series Techniques

Odell, Christopher [1]

February 22, 2020

Introduction

CO_2 emissions are known to be a key contributor to global climate change. A significant source of CO_2 emissions in the United States is from gas powered motor vehicles. To better understand emissions trajectory, I conducted a time series analysis of monthly motor vehicle CO_2 emissions in metric megatonnes from gasoline (excluding ethanol) in the United States for the time period between January 1973 and June 2013. I used four different time series techniques to gain a deeper understanding of the data and to set myself up to predict the next 24 months of emission levels. The techniques that I used were ARMA, (S)ARIMA, Holt-Winters Exponential Smoothing, and Spectral Analysis. Each method has a different approach to examining the time series, and my aim was to understand how they can be used to help with prediction analysis.

Data Description

The data for this analysis comes from the United States Energy Information Administration. The focus is on carbon dioxide emissions (CO_2) from energy consumption by source with an emphasis on motor gasoline, excluding ethanol. The CO_2 emissions are reported in million megatonnes (Mt). The data was a monthly measurement over time from January 1973 to June 2013. An initial plot (Figure 1) of the data appeared to show a change in mean for the CO_2 over time as well as seasonality. Another detail that I noticed while examining the series was that there appeared to be a polynomial pattern over time. The data has 486 observations which meets the length requirement for time series analysis.

---

[1] The initial analysis of this data was originally performed for a group project in ST_566, at Oregon State University in Winter 2019. The analysis was changed and updated to reflect statistical accuracy and new objectives in Winter 2020.

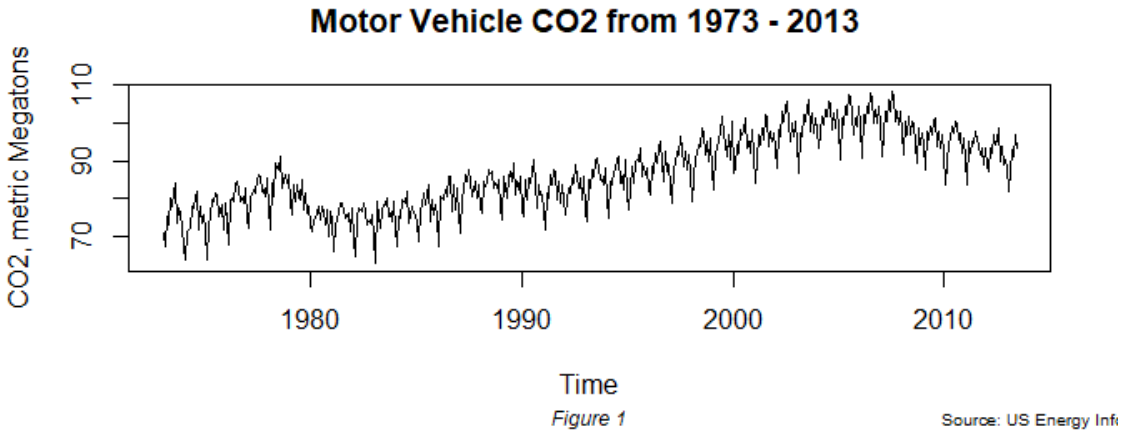

Figure 1 Source: US Energy Inf

Model Description:

Since the goal of the analysis was to predict 24 months of future carbon dioxide emission levels using different time series analysis tools, I needed to evaluate each of these tools separately. The different tools that were used are: ARMA, (S)ARIMA, Holt-Winters Exponential Smoothing. The ARMA model was chosen since it is one of the simpler time series models and would be a useful comparison versus more complex models. The (S)ARIMA model was chosen because of its capability to handle seasonality which was observed in the CO_2 data. Holt-Winters Exponential Smoothing is one of the few exponential smoothing models that can handle both change in mean and seasonality.

I wanted to look to see how well the (S)ARIMA model fit in the frequency domain and therefore also used Spectral Analysis. I used the Spectral Analysis to evaluate whether the AR portion of the (S)ARIMA model was an appropriate fit in the frequency domain.

ARMA

The ARMA model, which is comprised of the Autoregressive model (AR) and the Moving-average model (MA) was the first model I used. The model equations are shown below.

$$X_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \, \epsilon_{t-i}$$

This can be broken down into two parts:

AR model:

$$\phi(B) = 1 - \alpha_1 B - \alpha_2 B^2 - \ldots - \alpha_p B^p$$

MA model:

$$\theta(B) = 1 - \beta_1 B - \beta_2 B^2 - \ldots - \beta_q B^q$$

Where the p, q are the terms of the ARMA model, the ϕ is the Autoregressive operator, and θ is the Moving-average operator. Additionally, α is the coefficients for the Autoregression portion of the model and β is the coefficients for the Moving-average portion of the model.

(S)ARIMA

The $(S)ARIMA(p,d,q) \times (P,D,Q)_s$ model was the second model evaluated in the analysis. The AR and MA components were the same as in the ARMA model. The additional component S was added to handle the seasonality portion of the data. The (d, D) are differencing parameters that can help with trend (d) and seasonality (D). The (p,d,q) represented the short term correlations (non-seasonality) and the (P,D,Q) was the correlations over multiple seasons (seasonality). The subscript s was used to control the type of seasonality, which I recognized as yearly.

The model is given by:

$$\Phi(B^2)\phi(B) \bigtriangledown_s^D \bigtriangledown^d X_t = \Theta(B^s)\theta(B)W_t$$

This is made up of the non-seasonal operators: $\phi(B)$ and $\theta(B)$ with the seasonal operators $\Phi(B^s)$ and $\Theta(B^s)$, the portion of the model that is $\bigtriangledown_s^D \bigtriangledown^d X_t$ was the seasonal differencing $ARMA(p,q) \times (P,Q)_s$

Holt-Winters Exponential Smoothing

The third method mentioned is the Holt-Winters - Exponential smoothing. The model is broken down into three parts. Each model part is a smoothing method, one for the level ℓ_t, one for the trend b_t , and one for the seasonal component s_t , with corresponding smoothing parameters α, β^* and γ. I use m to denote the frequency of the seasonality, i.e., the number of seasons in a year.

$$\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + hb_t + s_{t+h-m(k+1)} \\
\ell_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \\
s_t &= \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}
\end{aligned}$$

where k is the integer part of $(h-1)/m$, this ensures that the estimates of the seasonal indices used for forecasting come from the final year of the sample. The level equation shows a weighted average between the seasonally adjusted observation $(y_t - s_{t-m})$ and the non-seasonal forecast $(\ell_{t-1} + b_{t-1})$ for time t. The beta equation $\beta^*(\ell_t - \ell_{t-1})$ shows an adjustment for the previous observation with the $(1-\beta^*)b_{t-1}$ being the adjustment for moving average trend. The seasonal equation shows a weighted average between the season, $(y_t - \ell_{t-1} - b_{t-1})$, and the seasonal before it last year.

Spectral Analysis

The fourth portion of the analysis I looked at the data on the frequency domain instead of the time domain through Spectral Analysis. The equation for spectral analysis can be found below:

$$\gamma(k) = \int_{\pi}^{0} cos(wk)dF(w)$$
$$f(w) = \frac{dF(w)}{dw}$$

$\gamma(k)$ is the auto-covariance function, $F(w)$ is the spectral distribution function and is the contribution to the variance from the frequency of 0 to w. And $f(w)dw$ is approximately the contribution on the variance with frequencies in the range $(w, w + dw)$.

These four different model methods were used to gather a deeper understanding of the data, and to find a model for prediction.

ARMA Modelling

Decomposing Trend

I observed both trend and seasonality components in the emissions data. To fit an ARMA model, I needed to first estimate the trend using a locally weighted regression (loess) on a monthly frequency and then subtract it from the observations. I selected the loess, a type of smoothing method, after also considering a linear model due to two advantages: first, the non-parametric features of the loess smooth are appropriate for both normal and non-normally distributed data, and second, a linear model would not be appropriate for this dataset due to large fluctuations in the long term mean change in the emissions. The controlling mechanism in the loess is the span which gave me the ability to control a level of fit to remove the trend from the data.

I needed to address the challenge of finding the appropriate span to sufficiently estimate the trend without overfitting the data. I addressed this issue by comparing eight different model fits. I started with the span of 0.30 and worked down in increments of 0.025. I found that 0.30 was not sufficient to remove the trend, as it left the auto-correlation function (ACF) plot with a pattern that still showed a trend. When I got as low as 0.15 in the span, I noticed the overfitting of the data with the pattern in the ACF increasing correlation versus lag 1. The most appropriate span I found for the loess model was 0.20 which gave me a model fit that seemed to control the results of the ACF and partial autocorrelation function (PACF).

The plot series below (Figure 2) shows a smooth trend, with the blue line showing the estimated trend and the black points and line showing the emissions time series. The loess on a monthly frequency was a close fit to the actual observations. Therefore, I removed the estimated trend from the observed emissions data and proceeded with the analysis. The second plot (Figure 3) below shows the time series excluding the estimated trend.

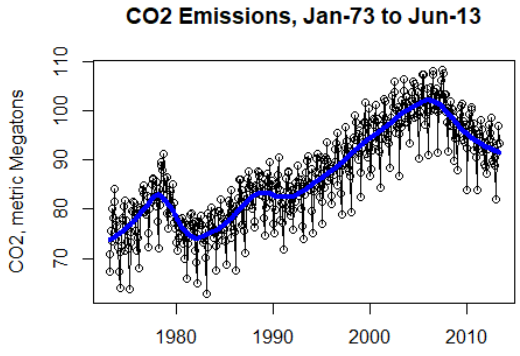

Figure 2

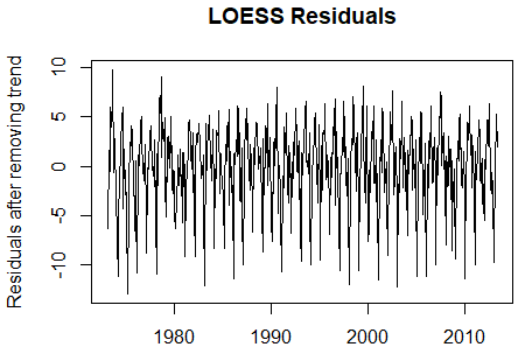

Figure 3

Decomposing Seasonality

As seen in the residuals plot (Figure 3) of the data after removing the trend, there is still strong annual seasonality. This is not surprising, as driving habits typically fluctuate with the seasons. Therefore, to understand the seasonality of the CO_2 emissions, I calculated the 12-month moving average of the emissions of the time series without the trend. Five years of the data are plotted below (Figure 4) to demonstrate the shape of the seasonality. The resulting fit shows the decrease in driving during the winter months which appears to have a greater impact on the emissions than the increase in driving during the peak season. The impact is shown by how far the top peak is from zero and how far the bottom peak is from zero.

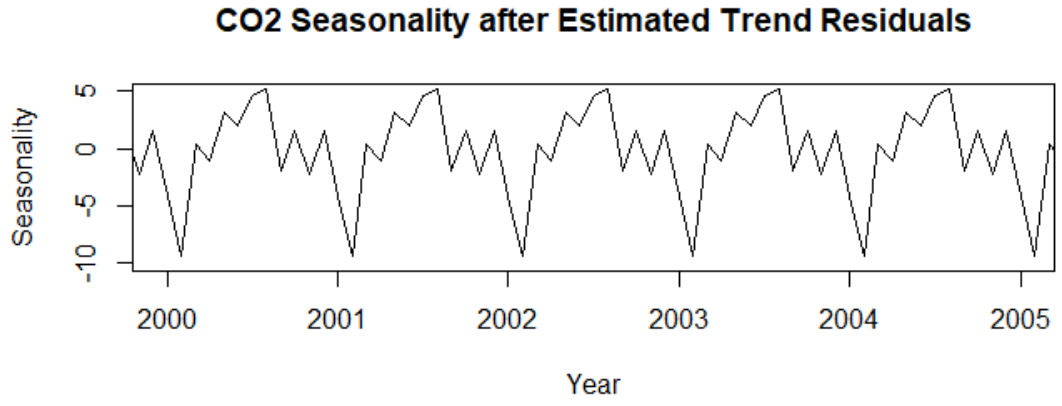

Figure 4

Residual Analysis (Stationary Series)

The final step in fitting the ARMA model was to derive the stationary series (residual noise) by subtracting the seasonal fit from the trend-adjusted emissions data. To confirm the effectiveness of the ARMA model, I examined the mean, ACF, and PACF.

Looking at the mean, it appears that the ARMA model effectively removed non-stationarity, because the mean was practically zero at 1.480402e-17. The stationary series has a low point in 2009 which could be an outlier, with a couple of high points early on. However,

looking at the ACF (Figure 6), where a unit lag is equal to one month, there is a strong correlation with previous months, indicating that the result after fitting the ARMA model is not stationary. The ACF is outside of the standard deviation after the initial decrease to zero, which occurs at a lag of 4 months. The ACF for the previous three months is the strongest, with a correlation value above 0.25. Further, there is a negative correlation beyond a lag of 7. The PACF (Figure 7) also shows a strong regular correlation with values above zero at regular intervals. This also indicates a non-stationary series, due to the fact that there is still regular correlation after removing the trend and seasonality. Therefore, I found the ARMA model to not be a sufficient fit to this time series and proceeded forward to examine the fit of a Seasonal ARIMA model without conducting prediction.

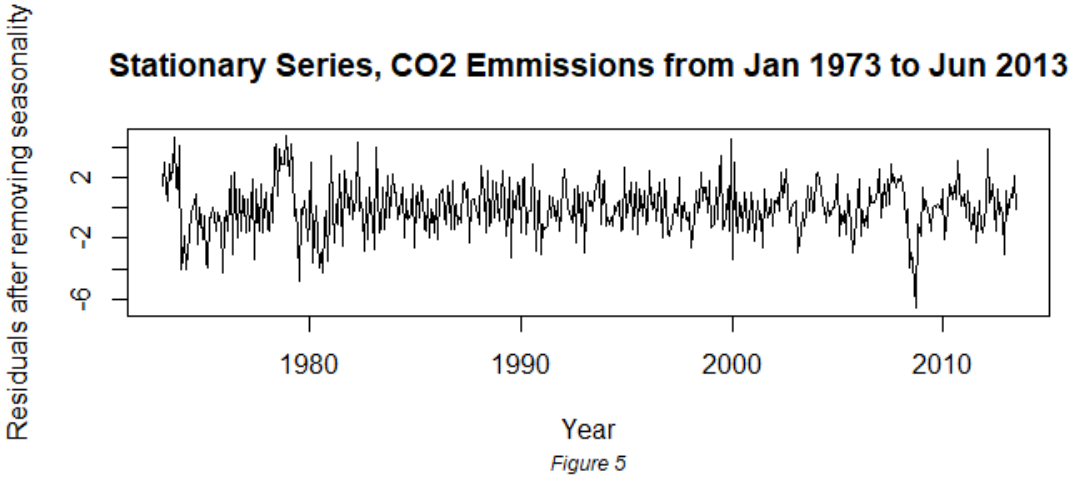

Figure 5

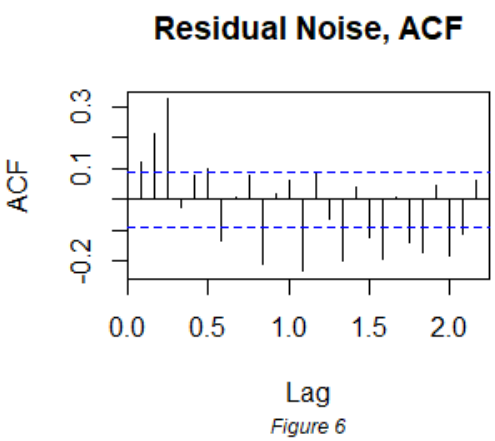

Figure 6

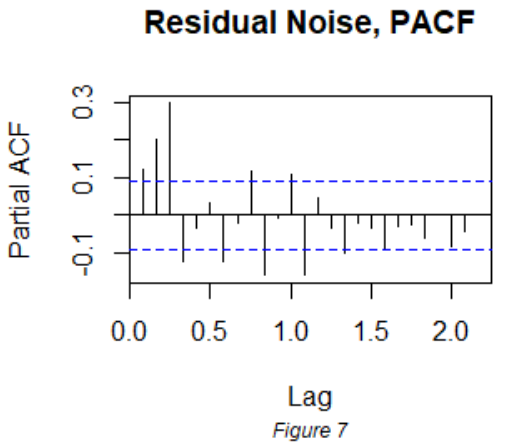

Figure 7

Seasonal ARIMA Modeling

To fit the (S)ARIMA model, I first performed differencing (d, D) to produce a stationary series that removes trend and seasonality. I used differencing parameters of lag 1 to remove the trend and lag 12 to remove the monthly seasonality. The results of differencing are shown below in Figures 8 and 9. The plot of the 1st order difference (Figure 8) is centered around zero and there is no evidence to suggest that the trend was not removed. The same can be said for the 1st and 12th order differencing plot (Figure 9), although I was

still suspicious about the area around 2009 where the variance decreases. In general, the differencing suggested an appropriate fit for the time series.

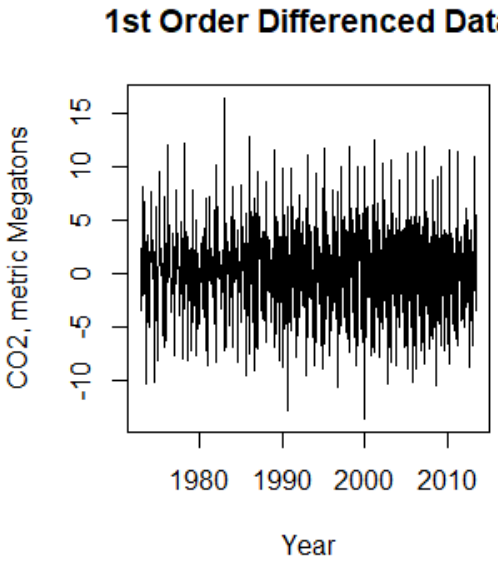

Figure 8

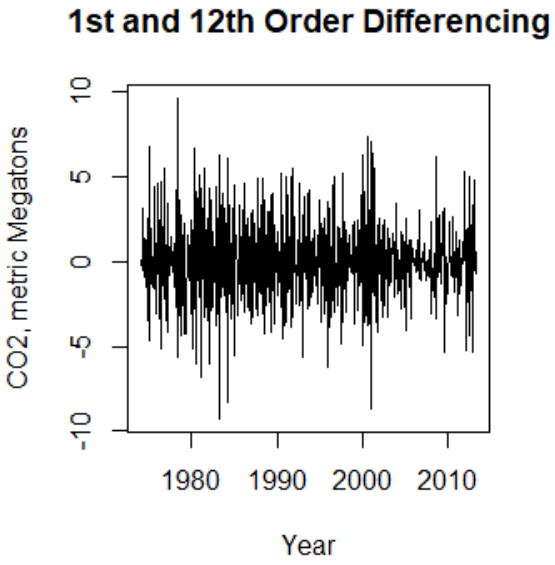

Figure 9

Selecting a Time Series Model

To ensure the appropriateness of the model, I started by examining the ACF and PACF of the differenced data. These plots are harder to interpret than the ARMA ACF and PACF plots and have a little more subjectivity in the interpretation. I started by looking at the non-seasonal ARIMA model portion. The ACF (Figure 10) shows a cutoff after lag 1, which is indicative of a MA(1) model. The PACF (Figure 11) was a bit more complicated and I saw either a cutoff after lag 2 (suggesting AR(2)) or a decay that is indicative of an ARMA model. Looking at the seasonal portion of the model (Figure 10), the ACF seems to show a gradual decay which is indicative of an ARMA model. The PACF (Figure 11) shows what might be an AR(2) behavior or perhaps nothing at all.

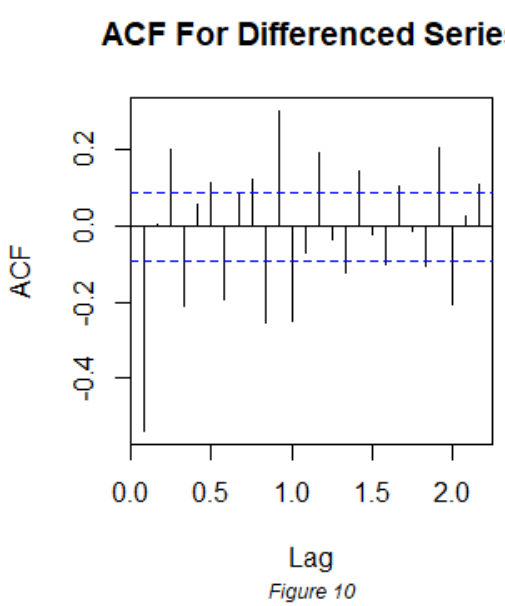

Figure 10

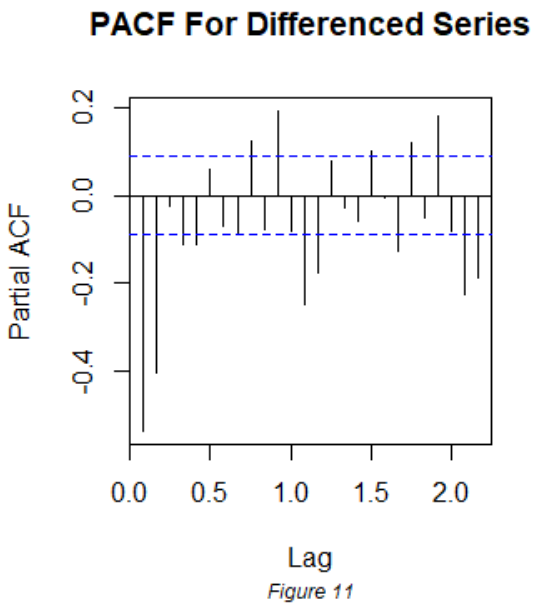

Figure 11

Given the results of the plots above, I examined the suitability of the following models:

- Model 1: $ARIMA(1,1,1)x(1,1,1)_{12}$
- Model 2: $ARIMA(1,1,2)x(1,1,1)_{12}$
- Model 3: $ARIMA(0,1,1)x(1,1,1)_{12}$
- Model 4: $ARIMA(2,1,0)x(1,1,1)_{12}$
- Model 5: $ARIMA(1,1,1)x(2,1,1)_{12}$
- Model 6: $ARIMA(2,1,2)x(2,1,1)_{12}$
- Model 7: $ARIMA(0,1,1)x(2,1,1)_{12}$
- Model 8: $ARIMA(2,1,0)x(2,1,1)_{12}$

The models as mentioned earlier have the components of $(p,d,q) \times (P,D,Q)$ with the (p,P) as the autoregressive portion, (d,D) as the differencing of trend and seasonality respectively, and (q,Q) as the moving-average portion. The subscript 12 which was indicated earlier as m is the seasonal component that I had choosen and in this case would was one year. I compared the models using the least squares, with a rank based on their AIC value. The model with the lowest AIC was the model that I would proceed forward with as the best fit.

```
##     Model1   Model2  Model3   Model4   Model5   Model6   Model7   Model8
## 1 1890.221 1889.134 1901.09 1868.352 1880.397 1865.273 1888.871 1862.578
```

Model 8, $ARIMA(2,1,0)x(2,1,1)_{12}$ was the best fit, with an AIC score of 1862.578.

SARIMA Model Diagnostics

Next, I explored some diagnostics to assess if the model fit represented the data in a meaningful way. I started with the basic model diagnostic plots (Figure 12). The residuals did not appear to have any trend and the variance looked consistent throughout. The ACF

showed no residual trend to the data and was sufficiently zero after one lag. Looking into the p-values for the Ljung-Box statistic the only area of concern might be some of the p-values that hovered near 0.05. Since there were some p-values around the 0.05 area, I wanted to ensure that I had selected the best model. I ran some diagnostics on several of the other models and found their diagnostics to be worse for various reasons.

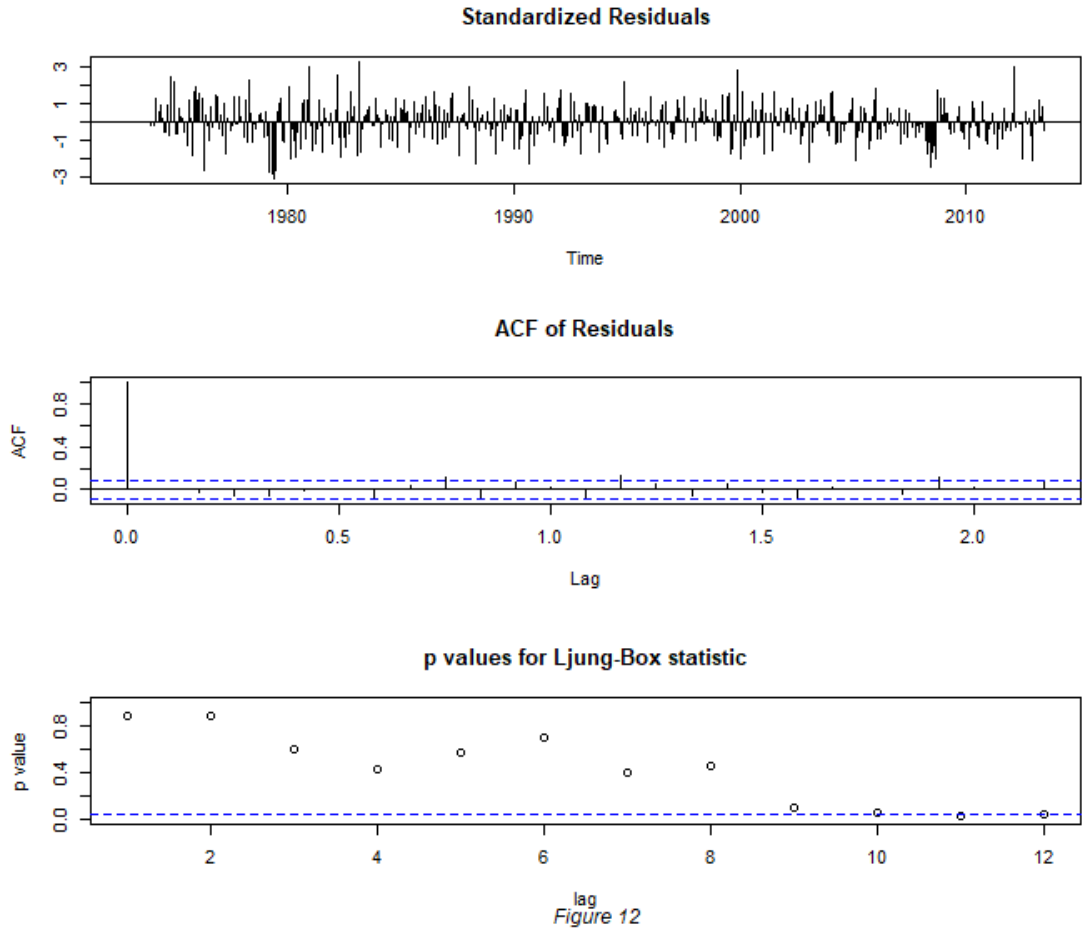

Figure 12

Next, I looked at the distribution of residuals; if the model was properly fitted, I knew that I should find that the residuals were relatively normal. I found the residuals to look normally distributed as shown in the QQ plot (Figure 13) below.

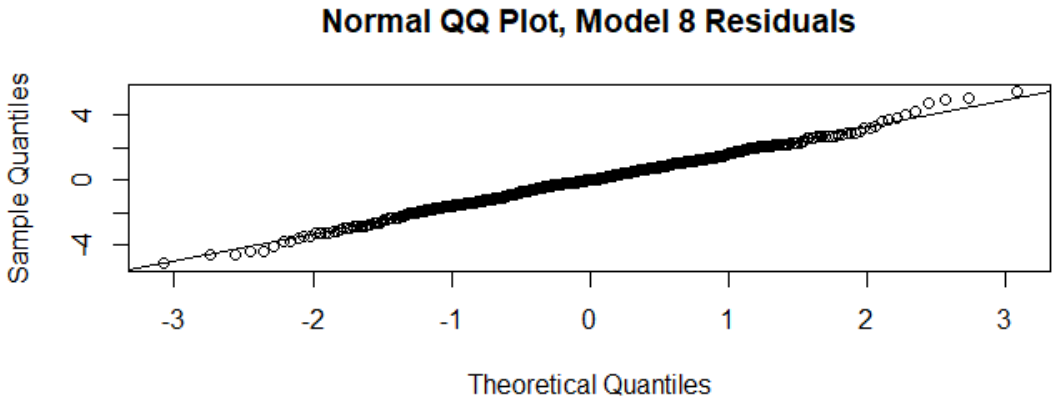

Figure 13

Next, I checked model 8 versus that of a model generated using the auto.arima() function in R. This function dynamically fitted a model with numerous possibilities. I used the criteria of AIC, to stay consistent with how my model was selected by hand. I also wanted to limit the complexity of the model to ensure that the computation time was reasonable, and I did this by limiting the variables of (p, P, q, Q) to 4.

```
## Series: emission
## ARIMA(2,1,2)(4,1,2)[12]
##
## Coefficients:
##           ar1      ar2     ma1      ma2    sar1     sar2     sar3     sar4
##       -0.8481  -0.4004  0.0985  -0.1011  -0.165  -0.1366  -0.0683  -0.0307
## s.e.   0.1165   0.0829  0.1187   0.1046   0.616   0.1129   0.1429   0.0843
##          sma1     sma2
##       -0.5907  -0.2507
## s.e.   0.6170   0.5322
##
## sigma^2 estimated as 2.869:  log likelihood=-925.55
## AIC=1873.1   AICc=1873.68   BIC=1918.85
```

As shown above, R selected a model fit of $ARIMA(2,1,2)(4,1,2)_{12}$ with the AIC comparable to model 8. The diagnostic plots (not shown) are similar. Because the model selected by R dynamically is more complex, and computationally longer, and the results are visually the same I proceeded forward my selected model (model 8).

Spectral Analysis

In contrast to the above methods that focused on modelling in the time domain, spectral analysis focuses on the frequency domain. I performed a brief exploratory spectral analysis to assess the appropriateness of the ARIMA portion of the SARIMA model above (model 8). In Figure 14, I plotted a periodogram to confirm that the seasonal component was annual, using the first order difference series from the SARIMA model above. This periodogram of the first-order differenced series shows peaks at frequencies that are multiples of 1/12,

which confirmed the presence of an annual trend. These peaks are rather narrow, suggesting the seasonality is quite regular in this time series.

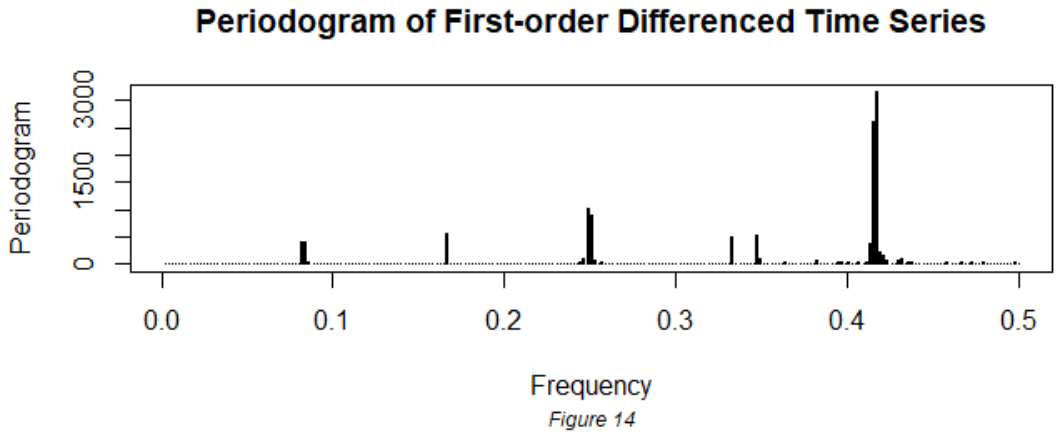

Figure 14

Next, I used the combination of first and twelfth order differencing to plot the periodogram (Figure 15) of the stabilized residuals. This periodogram shows an increasing spectral density up to a frequency of approximately 0.36, after which the density drops off while remaining greater than at the lower frequency values. Smoothing of the periodogram was not used so that I did not introduce any more complexity or bias to the fit.

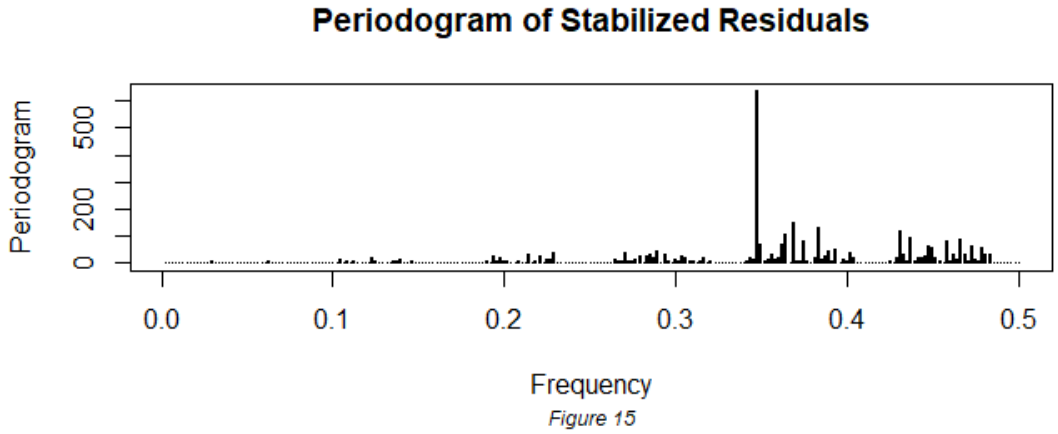

Figure 15

Now that I had an idea of what the spectrum for the residuals of the time series looked like, I compared the residual spectrum to the true spectrum from the non-seasonal part of the SARIMA model. Recall that model 8 took the form $SARIMA(2,1,0) \times (2,1,1)_{12}$. Since I had removed both trend and seasonality from this time series to generate a stable spectrum, I expected the AR(2) part of this model to reflect a similar spectral density as is shown in the periodogram of the residuals above. By calculating the spectrum of the AR(2) part of the model using the coefficients from model 8, I got the following spectral density (Figure 16):

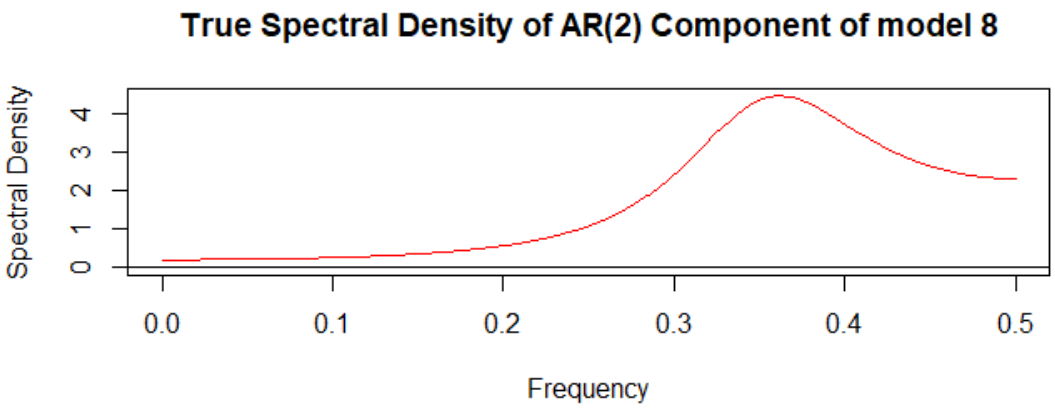

Figure 16

This appeared to match what the periodogram in Figure 15 showed, but to be sure I chose to simulate the AR(2) component 1000 times and compare that average spectral density from the simulated periodograms to the true spectrum of the AR(2) model. This is shown below in Figure 17.

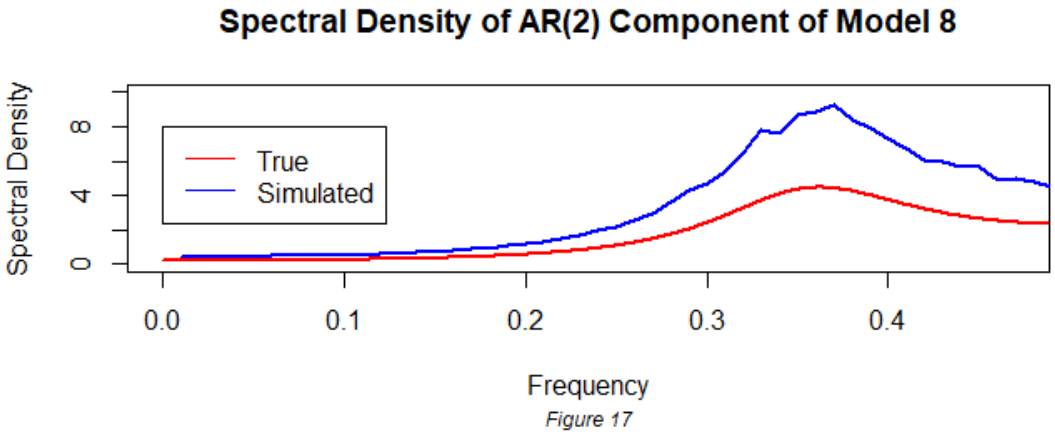

Figure 17

The simulated spectral density of the AR(2) component takes a similar shape as the true spectral density, albeit somewhat overestimated. This similarity in the distribution helped me to support the notion that model 8 was a good fit for this time series.

Forecasting

As mentioned earlier, that the purpose of this analysis was to create a prediction of the next 12 months for the CO_2 emissions on gasoline powered motor vehicles. I generated forecasts for the next 24 months of the CO_2 emissions with two methods: prediction from the SARIMA model 8 and Exponential Smoothing via the Holt-Winters approach.

SARIMA Forecast

I began with the forecast of CO_2 emissions using SARIMA model 8, along with confidence bands representing two standard errors. This prediction is shown below in Figure 18. After

examining the visual output, I did not see anything that appeared to be concerning about the model fit. The prediction seemed to follow a reasonable output with the confidence intervals expanding as time goes on which was to be expected.

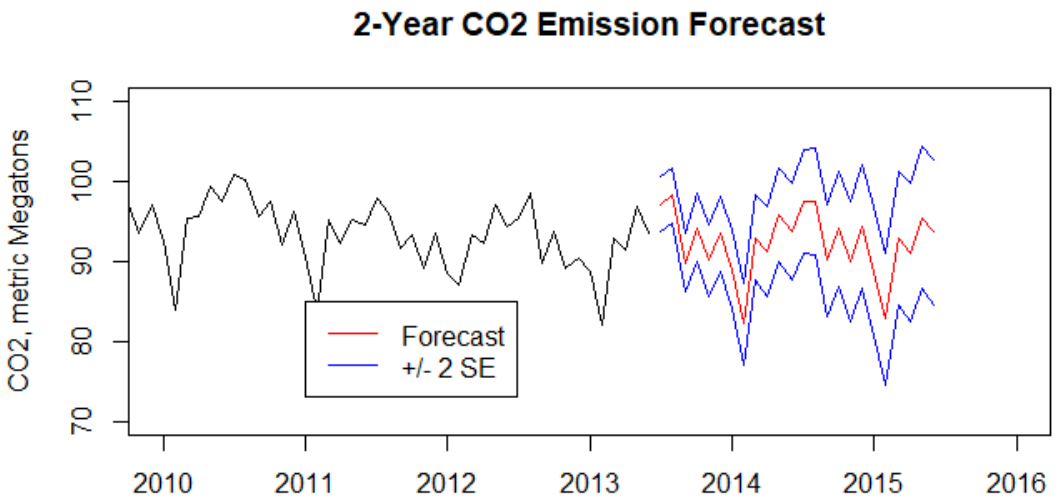

Figure 18

Exponential Smoothing: Holt-Winters

In contrast to the SARIMA method above, I performed long-term prediction analysis with exponential smoothing. I selected the Holt-Winters seasonal smoothing because it incorporates aspects to handle both trend and seasonality as seen in the time series data. Holt-Winters has the ability to optimize by two updating equations and can be fit with either a multiplicative or additive model. Since the seasonal components of the time series showed stable variance over time, I decided to use the additive model.

In R, I used the `forecast` library to calculate the smoothing estimates for the β (trend) and γ (seasonal) components of the Holt-Winters model. I specified a prediction level of 0.95 and a prediction period of 24 months. As shown in the plot below (Figure 19), the predicted values follow the most recent trend and seasonality, with a greater increase in upper and lower bounds over time. This predictive analysis looked almost identical visually to that produced by the SARIMA fit of model 8.

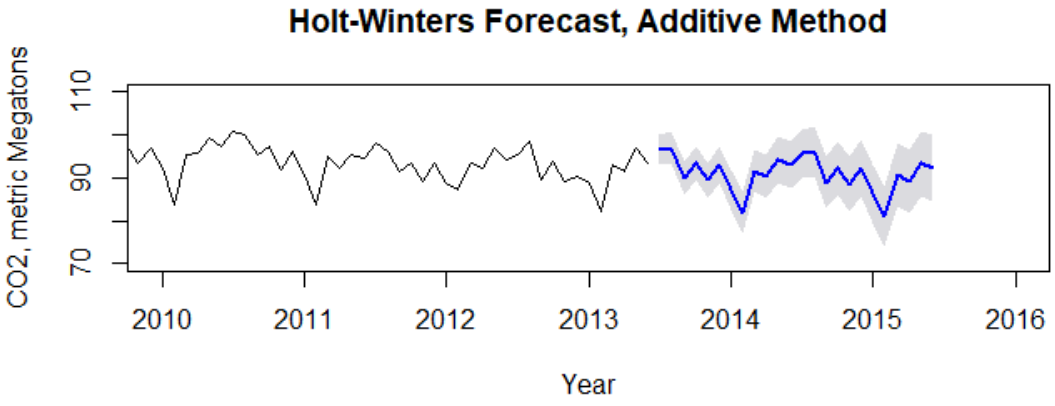

Figure 19

Forecast Comparison

Visually, both forecasts seemed to emulate the time series rather well. The Holt-Winters method projected the localized decreasing trend, which can be seen when looking at the earlier years of prediction. As time went on it seemed to level off with no trend. To see if there was a difference between the two predictions, I plotted the two predictions together as shown below in Figure 20.

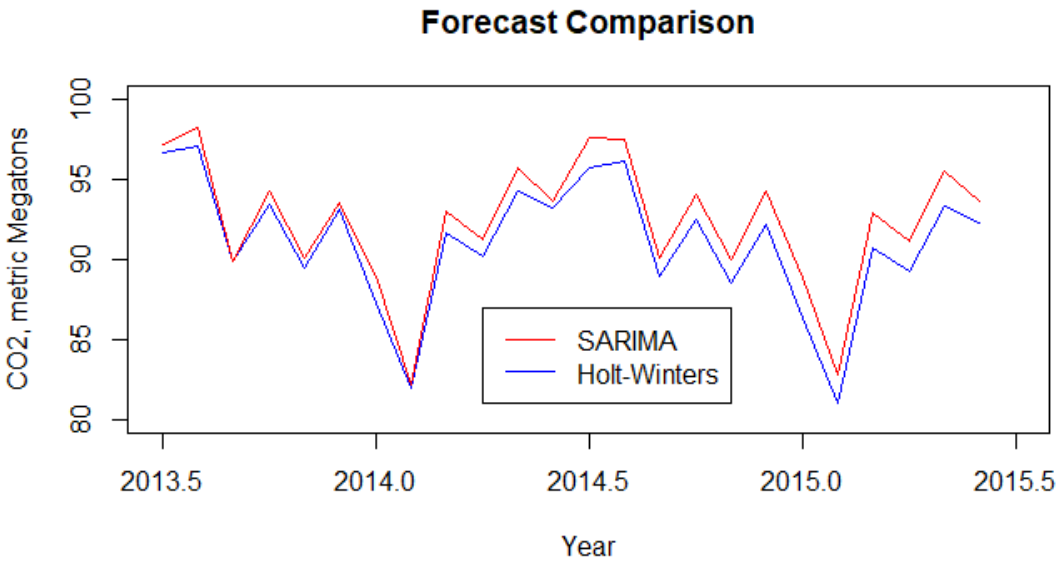

Figure 20

After I had looked at Figure 20, it does show that there is a slight difference between the two model predictions. When looking closely I saw that the predictions actually separated more as time increased. This was not unsurprising, given that the further a prediction is from the most recent observation, the greater the uncertainty will be. Next, I looked at the actual predictions to see if there is a numerical difference between these two predictive models. I created a dataframe that is shown below for all 24 predictive points. Column 1 is the prediction from the Holt-Winters model, column 2 is the prediction from the SARIMA

model (model 8). Columns 3 and 4 are the upper and lower bands (one standard deviation error) of the SARIMA (model 8) respectively. Year of the prediction is column 5. Column 6 is an indicator of whether the Holt-Winters prediction lands within the confidence bands of the SARIMA model. All values equated were true, indicating that these two models are very similar in their prediction on this time series. From this analysis, these methods both seem to perform equally well in forecasting the 24-month period from July 2013 to June 2015 and there is no evidence to support that they are statistically different.

Prediction Comparison Table

Holt-Winters	Model 8 Prediction	Model 8 Lower	Model 8 Upper	Years	Holt-Winters in Range
96.61654	97.15448	95.47712	98.83185	2013	TRUE
97.03865	98.21069	96.48465	99.93673	2013	TRUE
89.86840	89.86084	88.00804	91.71364	2013	TRUE
93.37895	94.24528	92.14056	96.34999	2013	TRUE
89.42961	90.10781	87.90740	92.30823	2013	TRUE
93.15426	93.49117	91.15753	95.82480	2013	TRUE
87.24195	88.99124	86.51687	91.46561	2014	TRUE
82.00308	82.18336	79.60296	84.76376	2014	TRUE
91.69922	92.99267	90.29644	95.68890	2014	TRUE
90.19189	91.19090	88.38364	93.99815	2014	TRUE
94.23171	95.74318	92.83457	98.65180	2014	TRUE
93.17187	93.65067	90.64012	96.66121	2014	TRUE
95.68417	97.54803	94.31709	100.77898	2014	TRUE
96.10628	97.49416	94.14907	100.83925	2014	TRUE
88.93603	90.13486	86.66176	93.60796	2014	TRUE
92.44658	94.06072	90.44375	97.67768	2014	TRUE
88.49724	90.01625	86.28301	93.74949	2014	TRUE
92.22189	94.25510	90.40095	98.10925	2014	TRUE
86.30958	88.80726	84.83341	92.78112	2015	TRUE
81.07071	82.82316	78.73815	86.90817	2015	TRUE
90.76685	92.87863	88.68266	97.07460	2015	TRUE
89.25953	91.12141	86.81740	95.42542	2015	TRUE
93.29934	95.46224	91.05388	99.87060	2015	TRUE
92.23950	93.65574	89.14465	98.16683	2015	TRUE

Conclusion.

In this analysis, I explored multiple methods for modeling the CO_2 emissions from motor vehicles between January 1976 and June 2013. I found that the decomposition of the ARMA approach did not yield a stationary series and instead deferred to a SARIMA model that appeared to better fit of the data. The spectrum analysis of the AR(2) component of this SARIMA model was then shown to agree with the spectrum of the CO_2 data within a reasonable margin, serving as a validation that the SARIMA model was appropriate for the modelling of this time series.

The SARIMA model was further tested by forecasting 24 months into the future and comparing the results to predictions from the Holt-Winters exponential smoothing method. These methods agreed in their forecast of the future with high precision, both in terms of trend and seasonality. I therefore conclude that the monthly motor vehicle CO_2 emissions time series can successfully be modelled with a $SARIMA(2,1,0) \times (2,1,1)_{12}$ and that of the Holt-Winters. This predictive analysis could lead to further examinations of other models, more complex models or longer predictions. The next step would be to assess the actual predictions and see if the actual CO_2 emissions from July 2013 to June of 2015 actually landed in the ranges provided in the model.

Appendix A - Source Code

```
knitr::opts_chunk$set(echo=FALSE, message = FALSE,
warning = FALSE,  fig.height = 4, fig.width = 7)

library(TSA)
library(forecast)
library(tidyr)
library(dplyr)
library(knitr)

#Read in Data
df <- read.csv("environment-carbon-dioxide-emiss.csv", head = T,
stringsAsFactors = FALSE)

# Rename a column in R
colnames(df)[colnames(df)=="Environment..carbon.dioxide.emissions.from.energy
.consumption.by.source"] <- "CO2"

df_ts <-  ts(df$CO2[-487], start = c(1973, 1), frequency = 12) # remove NA at
observation 487
#
# INTRODUCTION
#
#Plot the time series
plot(df_ts, ylab = "CO2, metric Megatons", main= "Motor Vehicle CO2 from 1973
- 2013 " )
log_df_ts <- log(df_ts)
```

```r
title(sub = "Figure 1",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")
mtext("Source: US Energy Info. Admin.", side=1, adj=1.21, line=4, cex=0.65,
font=1);

#
# ARMA
#

#Estimate trend using a LOESS fit
df_ts.time <- time(df_ts) # Get the time span for the loess regression
df_ts.loess <- loess(df_ts ~ df_ts.time, span = 0.20)
df_ts.loess1 <- loess(df_ts ~ df_ts.time, span = 0.30)
df_ts.loess2 <- loess(df_ts ~ df_ts.time, span = 0.20)
df_ts.loess3 <- loess(df_ts ~ df_ts.time, span = 0.25)
df_ts.loess.pred <- predict(df_ts.loess,newdata=df_ts.time) # Predict the
trend
df_ts.loess.trend <- ts(df_ts.loess.pred, start = c(1973, 1),  frequency =
12)# Monthly Freq

par(mfrow = c(1, 2))

# overlay the trend on the time plot
plot(df_ts, xlab = "Year", ylab = "CO2, metric Megatons",
     main = "CO2 Emissions, Jan-73 to Jun-13",type='o') # Simple Time-Series
lines(df_ts.loess.trend, col = "blue", lty = 1, lwd = 4)
title(sub = "Figure 2",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")

trend.res <- df_ts.loess$residuals
# change it into time series
trend.res <- ts(trend.res, start = c(1973, 1), deltat = 1/12)
# plot it
plot(trend.res, xlab = "Year", ylab = "Residuals after removing trend",
main="LOESS Residuals")
title(sub = "Figure 3",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")

#anova(df_ts.loess,df_ts.loess1,df_ts.loess2, df_ts.loess3)



mon <- cycle(df_ts)
# calculate the means at each month
res.lm <- lm(trend.res ~ factor(mon))
# deduct each point by the corresponding month mean
co2.season <- ts(res.lm$fitted.values, start = c(1973, 1) , deltat = 1/12)
# plot the seasonality, 5 years to get a feel for the shape
plot(co2.season, xlab = "Year", ylab = "Seasonality", main = "CO2 Seasonality
```

```
after Estimated Trend Residuals", xlim = c(2000, 2005), ylim = c(-10, 5))
title(sub = "Figure 4",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")


# remove the seasonality
co2.random <- ts(residuals(res.lm), start = c(1973, 1), deltat = 1/12)
#mean(co2.random)
# plot the final series
plot(co2.random, xlab = "Year", ylab = "Residuals after removing
seasonality", main = "Stationary Series, CO2 Emmissions from Jan 1973 to Jun
2013")
title(sub = "Figure 5",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")

# correlation of residuals
par(mfrow = c(1, 2))
co2_acf <- acf(co2.random, na.action = na.pass, main = "Residual Noise, ACF")
title(sub = "Figure 6",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")
co2_pacf <- pacf(co2.random, na.action = na.pass, main = "Residual Noise,
PACF")
title(sub = "Figure 7",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")

#
#Seasonal ARIMA
#

#Reload data for SARIMA section of paper to avoid polluting other sections by
re-using parameter names.
emission <-ts(df_ts, start=c(1973,1), frequency=12)

par(mfrow = c(1, 2))

#First order differencing
diff1 <- diff( emission, lag=1, differences=1)
plot(diff1,
     ylab="CO2, metric Megatons",
     xlab="Year",
     main="1st Order Differenced Data",
     type='l')
title(sub = "Figure 8",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")

#Difference with lag 12 to address seasonality
diff12 <- diff(diff1, lag=12)
plot(diff12,
     xlab="Year",
```

```
     ylab="CO2, metric Megatons",
     main="1st and 12th Order Differencing",
     type='l')
title(sub = "Figure 9",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")

#ACF and PACF of differenced data

par(mfrow = c(1, 2))

acf(diff12,
    main="ACF For Differenced Series")
title(sub = "Figure 10",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")
pacf(diff12,
     main="PACF For Differenced Series")
title(sub = "Figure 11",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")

#Fit candidate models
mod1 <- arima( emission,
               order=c(1,1,1),
               seasonal=list(order=c(1,1,1),period=12))

mod2 <- arima( emission,
               order=c(1,1,2),
               seasonal=list(order=c(1,1,1),period=12))

mod3 <- arima( emission,
               order=c(0,1,1),
               seasonal=list(order=c(1,1,1),period=12))

mod4 <- arima( emission,
               order=c(2,1,0),
               seasonal=list(order=c(1,1,1),period=12))

mod5 <- arima( emission,
               order=c(1,1,1),
               seasonal=list(order=c(2,1,1),period=12))

mod6 <- arima( emission,
               order=c(2,1,2),
               seasonal=list(order=c(2,1,1),period=12))

mod7 <- arima( emission,
               order=c(0,1,1),
               seasonal=list(order=c(2,1,1),period=12))

mod8 <- arima( emission,
```

```
               order=c(2,1,0),
               seasonal=list(order=c(2,1,1),period=12))


#Produce the results
data.frame( Model1 = mod1$aic,
            Model2 = mod2$aic,
            Model3 = mod3$aic,
            Model4 = mod4$aic,
            Model5 = mod5$aic,
            Model6 = mod6$aic,
            Model7 = mod7$aic,
            Model8 = mod8$aic  )




#Model 8 diagnostic plots
tsdiag(mod8, gof.lag=12)
title(sub = "Figure 12",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")


#Plot the QQ plot
qqnorm(mod8$residuals, main = "Normal QQ Plot, Model 8 Residuals")
qqline(mod8$residuals)
title(sub = "Figure 13",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")
#Let R choose a model for us
mod.auto <- auto.arima(emission,
           d=1, D=1,  max.p = 4,
           max.q = 4, max.P = 4,
           max.Q = 4, ic="aic")

mod.auto
per.diff1 <- periodogram(diff1, main = "Periodogram of First-order
Differenced Time Series")
title(sub = "Figure 14",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")
per.diff12 <- periodogram(diff12, main = "Periodogram of Stabilized
Residuals")
title(sub = "Figure 15",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")
spec.AR2 <- ARMAspec(model = list(ar = c(mod8$coef[1], mod8$coef[2])), plot =
T, col = "red", main = "True Spectral Density of AR(2) Component of model 8")
title(sub = "Figure 16",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")
```

```r
n <- 100
y <- arima.sim(n = n, model = list(ar = c(mod8$coef[1], mod8$coef[2])))
per.y <- periodogram(y, log = 'no', main = "AR(2), n = 500", plot = FALSE)

nsim <- 1000
spec.sim <- matrix(NA, nsim, length(per.y$freq))

for(isim in 1:nsim){
  y <- arima.sim(n = n, model = list(ar = c(mod8$coef[1], mod8$coef[2])))
  spec.sim[isim, ] <- periodogram(y, plot = F)$spec
}

spec.avg <- apply(spec.sim, 2, mean)
spec.sd <- apply(spec.sim, 2, sd)

plot(spec.AR2$freq, spec.AR2$spec, main = "Spectral Density of AR(2)
Component of Model 8", xlab = "Frequency", ylab = "Spectral Density", type =
"l", col = "red", lwd = 2, ylim = c(0, 10), xlim = c(0, 0.47))
lines(per.y$freq, spec.avg, lwd = 2, col = "blue")
legend(0,8, lty=c(1,1), col=c('red','blue'), legend=c("True","Simulated"))
title(sub = "Figure 17",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")


#Forecast 24 months
pred <- predict( mod8, n.ahead=24 )

plot(emission,
     xlim=c(2010,2016),
     ylim=c(70,110),
     xlab="Year",
     ylab="CO2, metric Megatons",
     main="2-Year CO2 Emission Forecast")
lines(pred$pred,
      col='red')
lines(pred$pred + 2 * pred$se,
      col='blue',
      lty=1)
lines(pred$pred - 2 * pred$se,
      col='blue',
      lty=1)
legend(2011,85,
       lty=c(1,1),
       col=c('red','blue'),
       legend=c("Forecast","+/- 2 SE"))
title(sub = "Figure 18",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")
fore_df_ts <- hw(df_ts, seasonal = "additive", h = 24, level = .95)
plot(fore_df_ts, xlim=c(2010,2016), ylim=c(70,110), main = "Holt-Winters
```

```
Forecast, Additive Method", ylab = "CO2, metric Megatons", xlab = "Year")
title(sub = "Figure 19",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")

#fore_df_ts$upper
#fore_df_ts$lower
#fore_df_ts$mean


plot(fore_df_ts$mean, col = "blue", ylim = c(80, 100), xlim = c(2013.5,
2015.5), main = "Forecast Comparison", ylab = "CO2, metric Megatons", xlab =
"Year")
lines(pred$pred, col = "red")
legend(2014.25,87, lty=c(1,1), col=c('red','blue'), legend=c("SARIMA","Holt-
Winters"))
title(sub = "Figure 20",
      cex.sub = 0.75, font.sub = 3, col.sub = "black")


first_6 = cbind(replicate(6,"2013"))
second_12 =  cbind(replicate(12,"2014"))
last_6 = cbind(replicate(6,"2015"))


years = rbind(first_6, second_12, last_6)

data = as.data.frame(fore_df_ts$mean) %>% mutate(prediction = pred$pred,
lower = pred$pred - pred$se, upper = pred$pred + pred$se , year = years)

names(data) = c("Holt-Winters", "Model 8 Prediction", "Model 8 Lower", "Model
8 Upper", "Years")

kable(data %>% mutate("Holt-Winters in Range" = data$`Model 8 Lower`<
data$`Holt-Winters` & data$`Holt-Winters` < data$`Model 8 Upper`), caption =
"Prediction Comparison Table")

##
```