**Carbon Dioxide Monthly Prediction for April, 2018**

Mackenzie Bogiages, Jackson McKenzie,

Peter Oliveira, Courtney Robinson, Marc Rovner

Stor 556

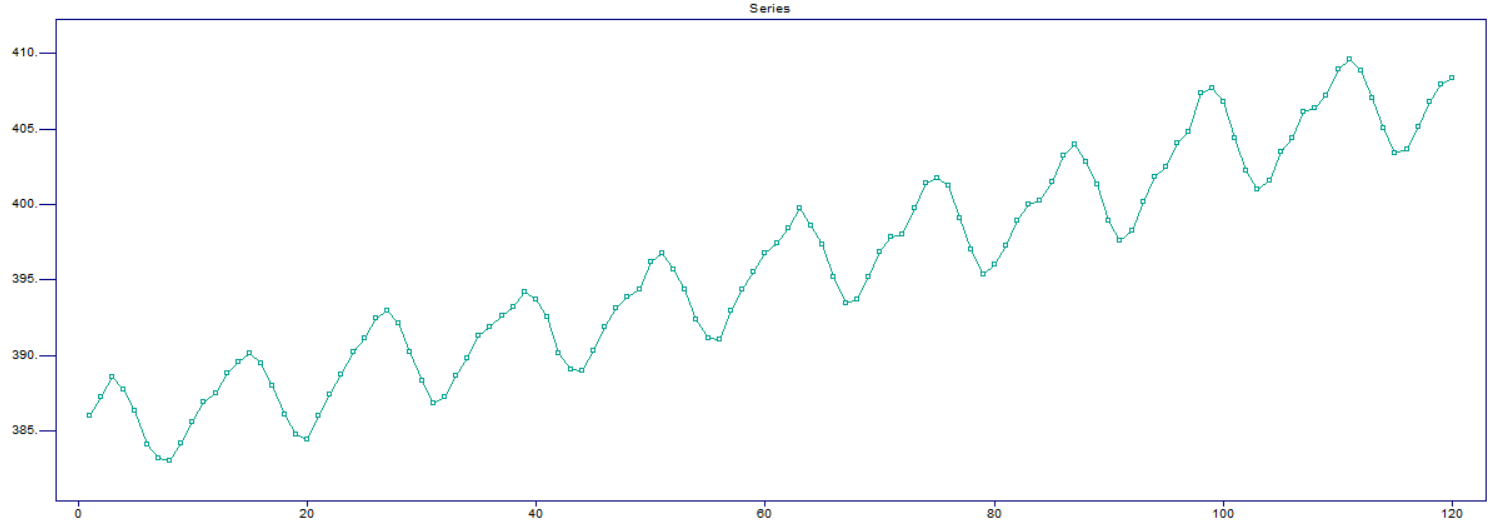
Group Project

Background:

Carbon dioxide levels have been fluctuating for centuries, but the human activities of recent decades have caused an increase in CO2 ppm levels beyond the “safe” carbon dioxide concentration level of 350 ppm (Bassett). In addition to manmade causes of CO2 increases, such as burning of fossil fuels, natural events like el Niño cause increases in CO2 rates by producing droughts, which lead to stressed vegetation and reduced photosynthesis, so less CO2 is absorbed. Additionally, many plants are killed off during severe droughts, which also contributes to higher CO2 levels (Wolter). Many methods have been used to measure CO2 levels; however, the most well renowned is the Keeling Curve, a graph that plots the CO2 change in concentration and, thus far, shows an increase since the 1950s (Monroe). In this paper, we will create a model to predict CO2 levels for April, 2018 using least-squares regression, the ARMA process, and tests of randomness.

Choosing Number of Data Points:

To begin, our group, the Random Radical Residuals, constructed our prediction model with different group members using either ITSM and R software. This was due to technology limitation. Despite the different tools used, the methodology was the same. ITSM limits the number of datapoints to 250 and because of that restriction, we decided to use the resultant R model for the final model, but have included graphs from the subset of data used in ITSM. Our prediction model is based on 720 data points or 60 years worth of data. For unrecorded or incomplete data points, we used interpolated data as put forth by the National Oceanic & Atmospheric Administration. The original graph of the last 10 years of data looks as follows:



Quadratic and Harmonic Regression Trend Functions:

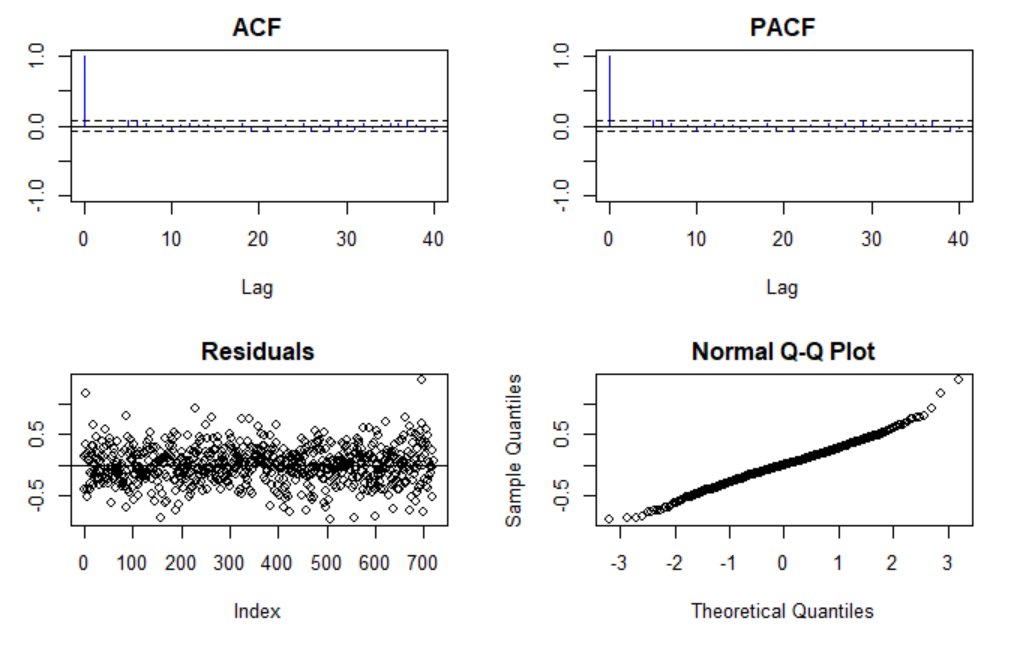
Plotting CO2 emission levels over time create a visual indicating undeniable trend and seasonal components. In order to achieve reliable estimation, we account for this pattern through quadratic and harmonic regression components. Initially, before considering the seasonal pattern evident in the graph, we determined that CO2 emission levels are generally increasing over time in a quadratic manner. Therefore, we fit a polynomial regression of second order to the data. This on its own produces an estimation line that increases following the general trend of the model.

CO2 emission levels are heavily dependent on seasonally varying factors, such as the weather, which follow an annual periodic pattern peaking around May and hitting troughs around September. We represent this seasonal effect with the following simple model , with the assumption we have fully addressed the trend component through the previously established quadratic fit. In order to remove this seasonal component from our observation we simply used differencing to subtract a seasonal component of period 12 from our model. After addressing both seasonality and trend we obtain a fit that closely resembles the pattern in our data, indicating that the series appears to be the result of the sum of a quadratic trend and a period 12 seasonal component.

Fitting ARMA Model to data:

After removing both the trend and seasonality components of the data, it is appropriate to fit an ARMA model to the residuals, since the residuals are a weakly stationary time series. More specifically, the Autoregressive-Moving Average model is generated based on past observations and past variation. Using R’s autofit function, we arrive at a best ARMA model of ARMA(6,8). Specifically, we arrive at the following model:

The white noise variance for this model is 0.08607737 and the graph of ACF, PACF, post-regression residuals, and Q-Q Plot look as follows:



Randomness:

After fitting an ARMA(6, 8) model to the stationary residuals from the least-squares regression, the new residuals pass all tests of randomness. All the graphs above indicate the regression has IID noise. Starting with the ACF and PACF tests, only one bar on each graph passes the dotted lines, which signify a 95% confidence level. Since we would expect 720\*0.05 =36 or fewer bars to cross the significance level if we assume the null hypothesis (that the data is random). Both the residual plot and Q-Q plot suggest a normal distribution, which contributes to our assumption that the data is random. Thus all of our initial tests fail to reject the null hypothesis.

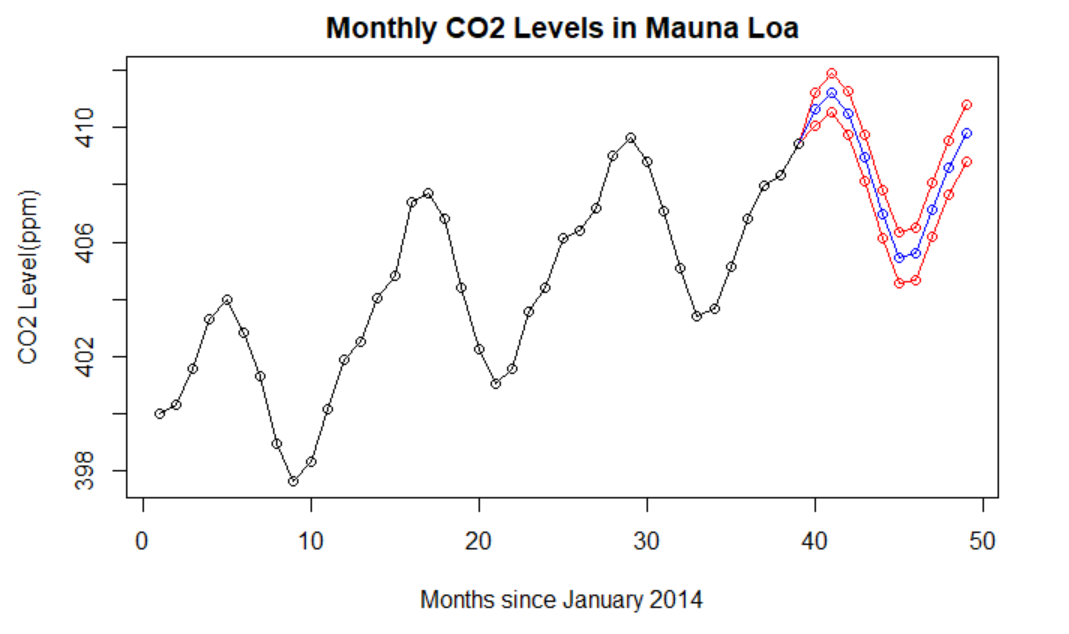
Now turning to the other tests of randomness, here is a table with each randomness test and its corresponding P-value:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Randomness Test: | Ljung-Box | McLeod-Li | Turning Point | Difference Sign | Rank | Jarque-Bera |
| P-Value: | .696 | .0463 | .3018 | .9486 | .5413 | .24279 |

All of these tests, excluding the McLeod-Li test, result in a P-value above .05 which means that they are all non-significant at the 95% confidence level and therefore we cannot reject the null hypothesis that our residuals are random. From these tests, we know that our ARMA model is effective.

Final Prediction:

Using the regression along with the ARMA(6,8) model we can predict the CO2 emission level for April, 2018 using R’s forecast function. These result in a final prediction of 410.67 with a 95% confidence interval of [410.10, 411.25]. Additionally, the 99% confidence interval is [409.92, 411.43]



For the above graph, the blue line represents the predicted value for the next 10 months and the two red lines represent the 95% confidence interval bounds.

Conclusion:

In summary, our group was able to predict the CO2 recorded emissions level at the NOAA Mauna Loa Observatory for April 2018 by using least square regression and an ARMA process. When initially viewing the data, it's clear that there are both trend and seasonality components in the data and in conjunction with our methodology we were able to eliminate these components and create a stationary time series plot. We then concluded that our residuals were random at the 5% significance level by using different tests of randomness. In conclusion our final prediction for April, 2018 is: 410.67.

Notes:

(Following analysis with ITSM)

* The data obviously appears to have both trend and seasonality components. Removing the seasonality component, the data appears to have a quadratic trend.
* Initially running the analysis with 250 data points (20.8 years) the results had a larger confidence interval as opposed to running the analysis with 120 data points (10 years). For that reason, we ran the analysis with 120 data points, which seems like ample enough data.
* Removing both the seasonality and trend with least squares regression results in a stationary time series with the following equation: Text

  Description automatically generated
* Using the informal sample ACF test for randomness, the residuals don’t appear random as more than 5% (2) bars are outside of the lines: Chart

  Description automatically generated
* Fitting an ARMA process to the data seems appropriate. Outsourcing the work to the machines, we arrive at an ARMA(4,2) process:Text

  Description automatically generated
* The residuals + the ARMA model passes the ACF testChart

  Description automatically generated with medium confidence:
* Additionally the model passes all other tests of randomness: Text, letter

  Description automatically generated
* Using the model results in a final predicted value for March of: 409.8 with a 95% confidence interval of [409.23, 410.37].

R Model

* Remove seasonality and quadratic trend
* Use autofit to find the best ARMA model for the data.
* ARMA(6,8) gives coefficients

phi: 0.91873036 0.36387597 0.01328009 -0.37803691 -0.91085103 0.95876030

theta: -0.2614188 -0.3790018 -0.3103944 0.1390596 1.0004936 -0.3015216 -0.0494876 -0.0770085

* ACF, PACF, and Residual Plots for the ARMA(6,8) model

Diagram

Description automatically generated

* Test of Randomness

A picture containing text, receipt

Description automatically generated

* 10 Step Prediction

A picture containing text, receipt

Description automatically generated

* Graphing the Results

Chart, line chart

Description automatically generated

* Final Prediction for April 2018, 410.67 with a 95% confidence interval (410.10,411.25)

Citations

1. Bassett, Carol Ann. *Galapagos at the Crossroads: Pirates, Biologists, Tourists, and*

*Creationists Battle for Darwin's Cradle of Evolution*. National Geographic, 2009.

1. Monroe, Rob. “The Keeling Curve.” *The Keeling Curve*, 28 Mar. 2018.
2. Wolter, Klaus, ESRL Web. “Physical Sciences Division.” *ESRL News*.