**Mauna Loa CO2 Prediction Report**

Alec Wright, Maharshi Pandya, Dev Pandya, Kevin Bowerman

Abstract: Our task was to use 120 months of accurate co2 data from the Mauna Loa observatory to predict the co2 levels for November 2019 as accurately as possible to two decimal places. Given this task we tried to understand our dataset through scientific literature as to precisely be able to remove trend and seasonality from the data and work with the best residuals we could. The seasonality was a strong component of the data as co2 levels are heavily dependent on the seasons but the trend was a bit harder to estimate. After multiple models and learning from existing literature we determined that an autofitted ARIMA(2,1,5) best modeled our data and we obtained a predicted November co2 level and a 95% and 99% confidence interval around our prediction.

Prior to attempting to understand the trend and seasonality of the Mauna Loa time series data, we first had to understand the scientific reasoning behind the data’s behavior. Regarding trend, prior research since the 1970s has resolved that the underlying trend of the CO2 data reflects an accelerating growth in the presence of atmospheric carbon dioxide. As a result of anthropogenic causes, average monthly atmospheric carbon dioxide readings have not only increased but have done so at an increasing rate. This suggests that the time series data bears a trend that is possibly quadratic, and not linear. With this understanding of the science, we were able to take steps that chose models or evaluated models based on their effectiveness to address a potentially non-linear trend.

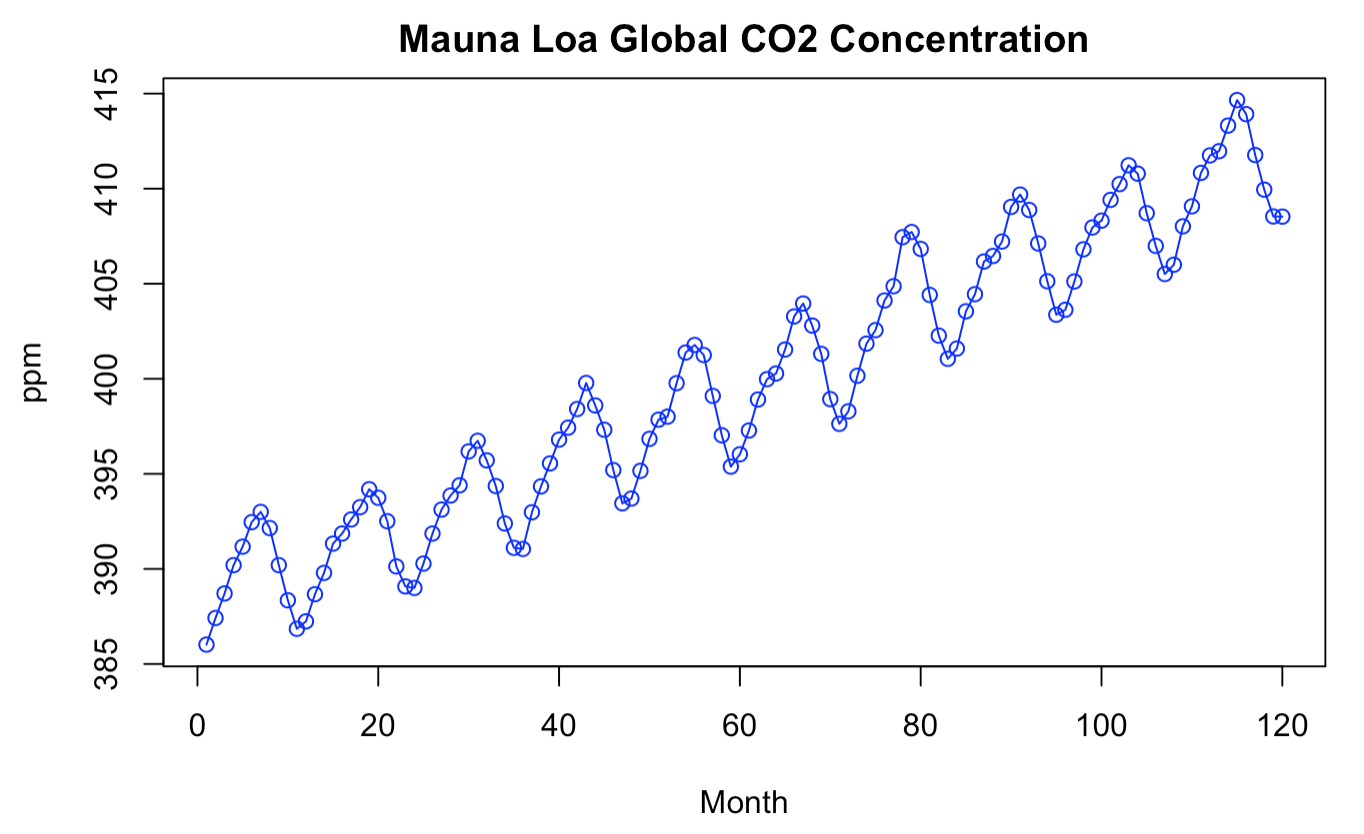


Figure 1: Average Monthly CO2 Concentration, starting from October 2009

The Mauna Loa data also reflects a notable seasonal component. It appears that atmospheric carbon dioxide readings increase from September/October and peaks in May, after which average readings descend back to the valleys seen in the following September and October. Scientifically, this seasonality can be attributed to carbon dioxide absorption by plant life, particularly in the northern hemisphere. Beginning in May, as global temperatures reach annual highs, plants begin to consume more carbon dioxide and sequester the gas in the earth as a resource for growth. This growth, which takes place in late Spring and through the Summer, contributes to a decrease in atmospheric carbon dioxide. On the other hand, as Fall and Winter approach, plants begin to die off and release that stored carbon dioxide back into the atmosphere, contributing to the seasonal rise in atmospheric CO2. This natural process closely follows a cycle over a calendar year, and as a result, we can approximate this seasonality with a period of 12 data points. By understanding the scientific reasoning behind the seasonal variation, we can better create a model that approximates and eliminates seasonality based on the 12-term period seen.

After thinking about the best way to approach the data we decided to choose the classical method with a season of 12 and a trend of 1 (linear) to model the data and then after interpreting the results try to find good ways to tweak the model, using this as a building block. After getting residuals from eliminating trend and season we saw some concerning signs with the residuals. Firstly, The ACF shows slow exponential decay making us conclude that there was still potentially some trend and the cyclicality made us think there could also be some seasonality as well within the residuals showing non-stationarity. This can also be observed in the Residuals vs Time plot in which after t = 60 there appears to be some quadratic piece we originally missed with the linear fit. This was also corroborated by two tests of randomness being significant the Li-Jung Box and McLeod-Li displaying that these classically estimated residuals were not ideal(see next page).

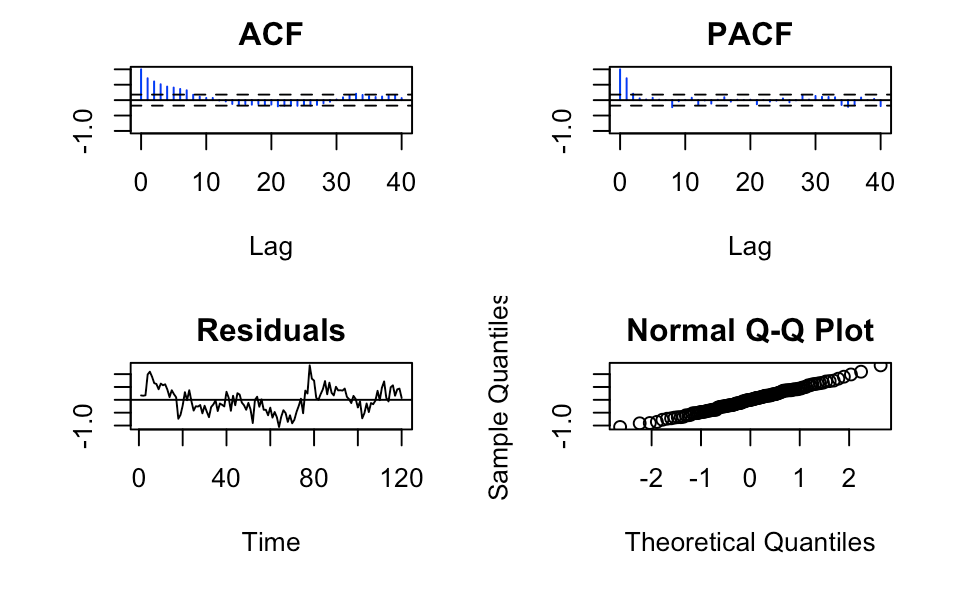


Figure 2: Test of residuals from classical estimation method season 12 trend 1

Due to these residuals not being stationary, the autofit function would not comply, showing us that an ARMA model with the classical estimation method was not correct in dealing with this method.

With the classical method falling short of allowing us to run ARMA model due to the AR potion being non-stationary, we decided to refer to the class examples and use the differencing method with a trend of 1 and season of 12 which does not need stationary residuals to work, and thus implement an ARMA model. Post-differencing method residuals were tested for randomness giving us these results(see next page).

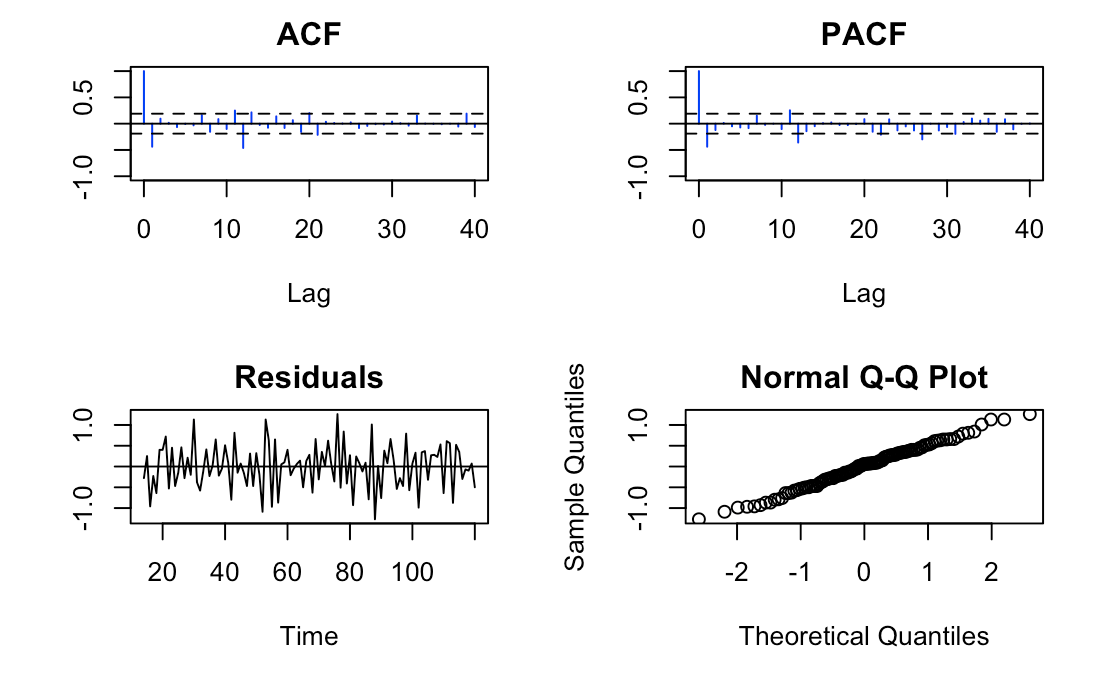


Figure 3: Test of residuals from differencing method season 12 trend 1

The residuals look much better than the classical estimation in their appearance in the time graph(Figure 3 bottom left) but show signs of negative autocorrelation. The ACF and PACF also improved and fall more in line with what we would expect from IID noise residuals but by eye test appear to almost be negatively autocorrelated a bit as well. However, with this model is that the turning point and Ljung-box tests both showed significance which caused us to worry if the residuals are truly IID noise, even though the non-stationarity issue was fixed. Regardless, we decided to autofit the model to an ARMA and got values of p = 3 and q = 4, then running an ARMA(3,4) we forecasted the November co2 level at 410.75 with a 95% prediction interval of [409.92, 411.58] ppm. While this result looked promising, we decided to be thorough and try to run a model that would account for the seemingly quadratic piece that we observed in the classical estimation.

The next model that we attempted was differencing with a seasonality of 12 and a trend component of 2. The trend of 2 means that this model will attempt to remove a quadratic trend from the data, rather than a linear trend. The reason why we decided to test this model is due to the scientific research we had done on the problem before, which suggested a quadratic trend may exist in the data. From running this model, we got the following plots of the residuals.

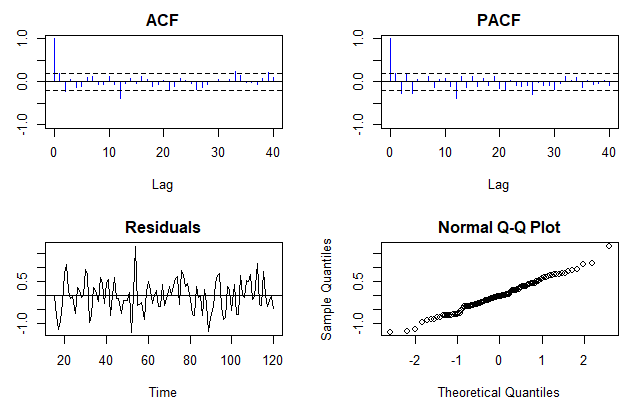


Figure 4: Results of running autofit on the differencing (12,2) model

As shown above, the ACF and PACF plots both show little autocorrelation(possibly some negative autocorrelation) which is in line with the previous model but still show slight signs of a leftover trend. The Residual Plot and the Q-Q Plot both do not have any clear improvement over the previous model, but still seem to raise some issues, due to a potential seasonal trend in the residuals. This model was implemented similarly to the previous differencing model, but simply changing the trend value to 2 rather than 1, and computing new values for both p and q based on a quadratic trend fit instead of linear. This new model gave us a prediction of 410.84 with a 95% prediction interval of [409.9822, 411.7044]. The size of the bounds is much larger than that of the previous model, which means that this model has a larger MSPE and thus margin of error. Although this model does appear to be fairly appropriate, there are still some clear issues, so we looked to find a new candidate model.

We did some research, and looked at documentation online about how others solved this problem, to see if there was another model that may be better. From the literature search, we decided to implement an ARIMA model over the Mauna Loa data and see if there are significant changes compared to the previous two approaches. ARIMA process is a generalization of ARMA process as it incorporates a wide range of nonstationary series.

Using the auto.arima() function of the ITSMR package, we were able to figure out the hyperparameters p,d, and q on which our univariate time-series can be fitted. Normally, the standard pipeline for implementing ARIMA would involve making series stationary and determining ARIMA parameters from ACF/PACF plots. However, the auto.arima() function uses AIC and BIC values over a grid of specified intervals to predict the best combination of model parameters. The auto.arima() will choose from optimal p, d, q parameters from 6 total combinations whereas autofit() function will choose p, q only from 2 combinations giving us a larger range in which to choose optimal parameters. In our case, the optimal parameters were ARIMA(2,1,5).

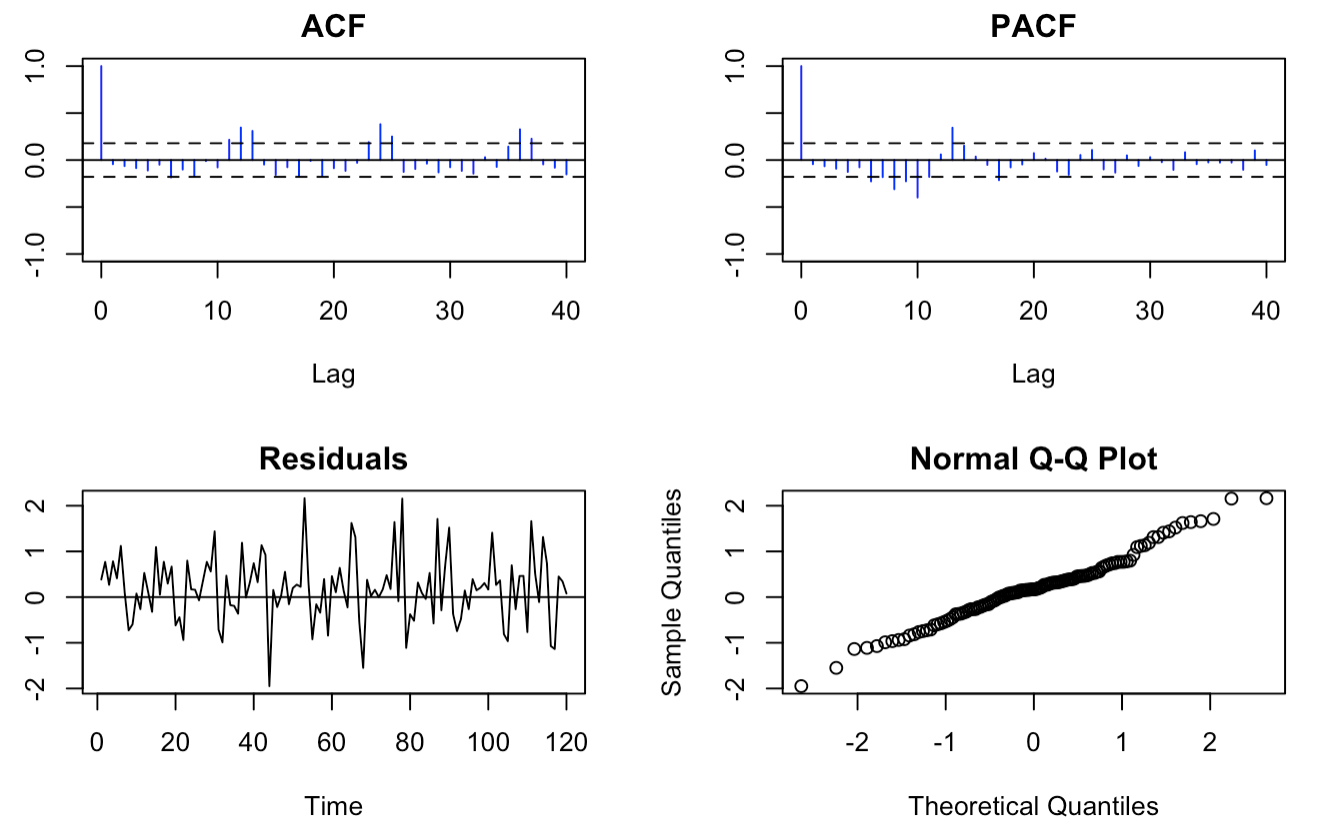


Figure 5: Test of Residuals from ARIMA(2,1,5) model.

Test on the ARIMA(2,1,5) model resulted in plots shown in Figure 5. While the ACF and PACF plots in Figure 5 seem to be not completely free of any patterns, since auto.arima() takes in account all combinations of p,d,q over the univariate time series, we expect the prediction from ARIMA to be better than the prediction from differencing model with a seasonality of 12 and a trend component of 2. Also, model with higher ordered differencing (e.g. 2) tend to assume a local time-varying trend, the predictions tend to be more variable. Therefore, automatically predicted ARIMA(2,1,5) model with only one nonseasonal difference seems to be the most practical.

.

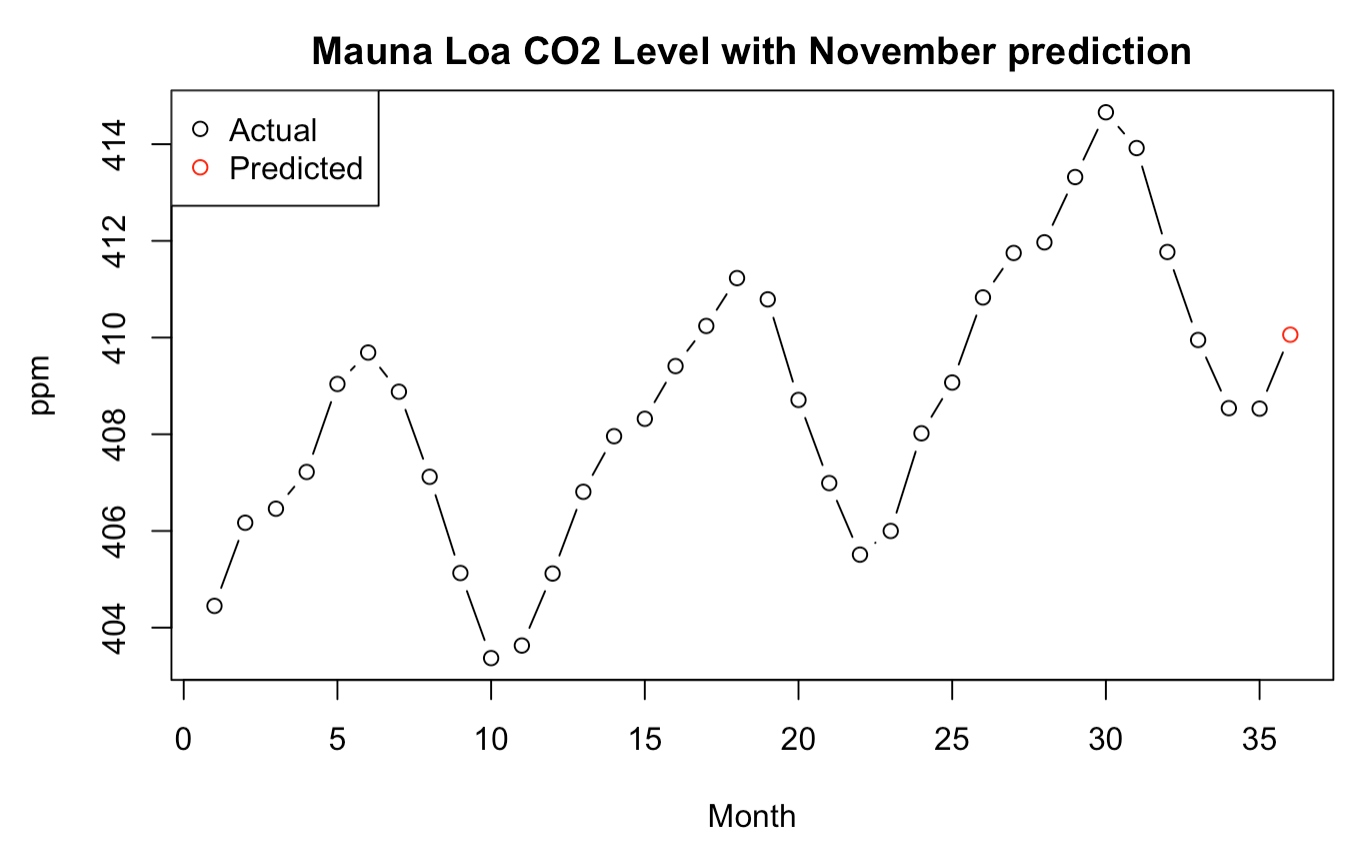


Figure 6: CO2 levels for the last 3 years and a prediction for November 2019, indicated with a red dot.

Our implementation of auto.arima() generated following parameters: ARIMA(2,1,5). Using forecast.Arima(), we were able to make CO2 level predictions for the month of November. Our forecast for November 2019 is 410.06 ppm with 95% confidence interval of (408.54, 411.58) and 99% confidence interval of (408.06, 412.06). We settled on this model vs the previous two differencing models as we felt that the autofitted differencing values chosen by auto.arima() would prove to be more accurate chosen by software than by humans who are prone to error.

The prediction matches scientific expectations, as November marks the end of warmer months and thus an expected return of carbon dioxide to the atmosphere as plant life in the Northern Hemisphere begins to die as the season changes. The predicted interval also falls in the range of what we would expect given historical seasonality, and the greater carbon dioxide level relative to last year’s November average backs the theory that a trend of rising average carbon dioxide levels exist.

*Cited Literature and Sources*

*STOR 556 Class Lecture Slides*

Nau, Robert. *The Mathematical Structure of ARIMA Models*. Fuqua School of Business, Duke University,http://people.duke.edu/~rnau/Mathematical\_structure\_of\_ARIMA\_models--Robert\_Nau.pdf.

*https://en.wikipedia.org/wiki/Autoregressive\_integrated\_moving\_average* - more ARIMA literature

Monroe, Rob. *Why are Seasonal CO2 Fluctuations Strongest at Northern Latitudes?.* Scripps Institution of Oceanography, University of California at San Diego, https://scripps.ucsd.edu/programs/keelingcurve/2013/05/07/why-are-seasonal-co2-fluctuations-strongest-in-northern-latitudes/