Introduction

Mauna Loa is one of five volcanoes that form the island of Hawaii. It is located 40 miles from the nearest city of Hilo and extends 13, 679 feet above sea level (West Hawaii Today, 2015). Mauna Loa is home to the world's oldest continuous air monitoring station, known as Mauna Loa Observatory (Butler, 2015). This observatory is specifically known around the world for monitoring the level of the most important greenhouse gas, carbon dioxide (CO_2), as well as acting as the global benchmark site for the increase of this global-warming gas.

The monitoring of CO_2 levels began in 1958 through work done by Charles David Keeling at Mauna Loa Observatory (Monroe, 2015). Once other researchers learned of Keeling's work, they began to measure CO2 levels at other sites around the world. Comparing various measurements around the world led to the realization that seasonal cycles of CO_2 at different latitudes follow predictable patterns. It was found that there is an increasing, upward trend in CO_2 levels at all measurement sites, supporting the concept of global warming (Monroe, 2015). The seasonal cycles of CO_2 at different latitudes can be attributed primarily to photosynthesis in plants. As plants start the photosynthesis process in the spring and summer, they absorb CO2 from the atmosphere thus resulting in an annual decrease in CO_2 levels beginning in May (Monroe, 2015). Once summer has ended, plants reduce the rate of photosynthesis in order to save energy and the seasonal cycle begins again.

The goal of our study is to develop a statistical model that best predicts the CO_2 level in Mauna Loa using historical monthly average data from the observatory. The CO_2 data that we used to develop our model included the recorded monthly average levels from October 2000 to February 2015. We did not include monthly levels from January through September 2000 because there was an aberration in the data set in the beginning of 2000, which is shown in Figure 1.1. This abnormality may have been caused by a volcanic eruption, massive hurricane, or some other isolated environmental occurrence.

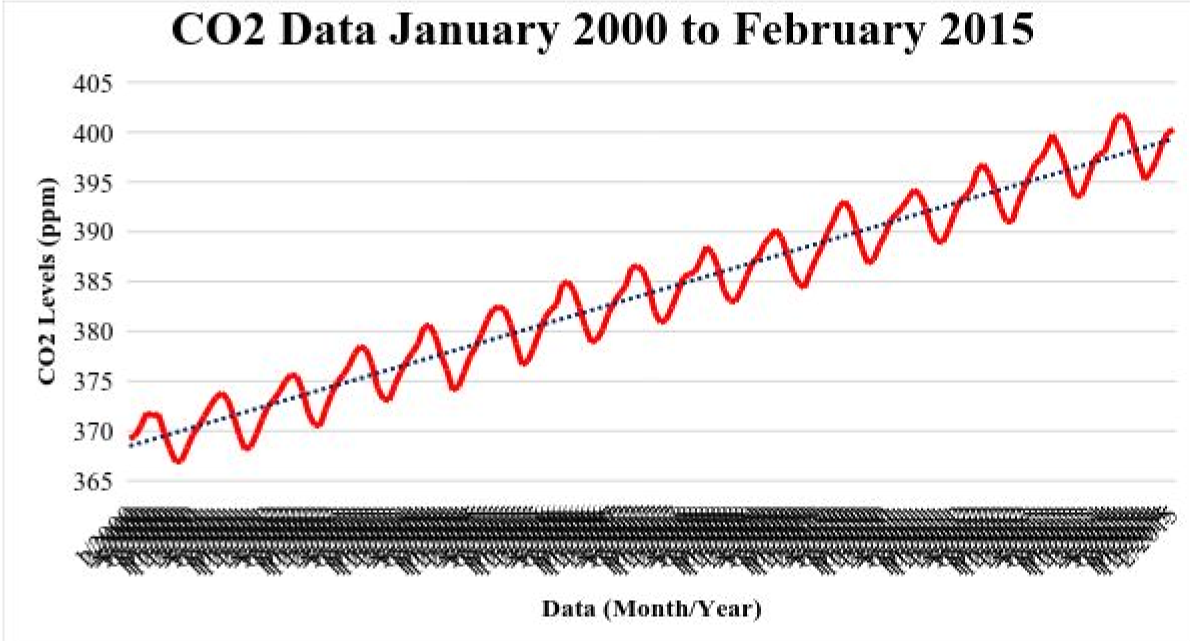

Figure 1.1 CO_2 Data from January 2000 to February 2015

Additionally, the data before 2000 does not fit a linear trend. A graphical representation of the historical CO_2 levels dating back to March 1958 is shown in Figure 1.2. This non-linear trend could be due to less precise measuring capabilities in the past or some other unknown factors. Given the shift from a non-linear to a more linear trend within the data set, we felt that including the data prior to 2000 would be potentially harmful and extraneous at best. Furthermore our sample size from October 2000 onward is sizable, so we wanted to avoid over-fitting our model by not including too many data points.

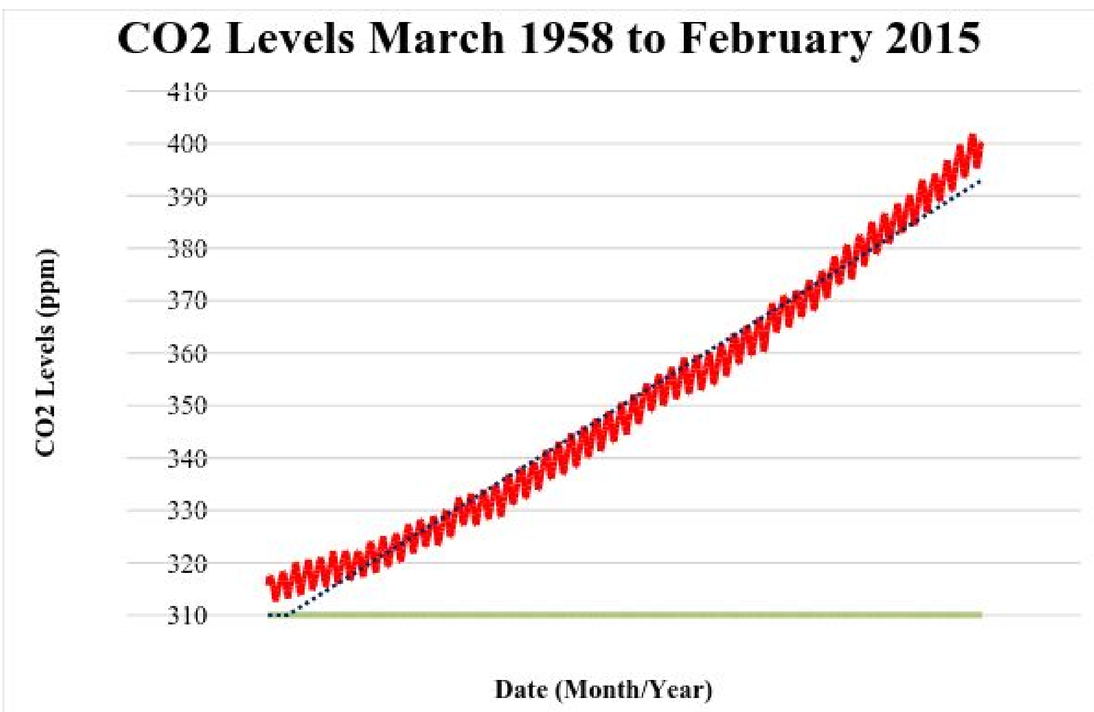

Figure 1.2. CO_2 Levels March 1958 to February 2015

Theory and Assumptions

The two main characteristics of the CO_2 levels are that the data has an increasing mean and that it is seasonal. The behavior of this data agrees with our current knowledge of CO_2 levels. CO_2 levels are the on rise globally, and this is no different in Mauna Loa (in fact most of the data for the global trend is measured at Mauna Loa). The largest portion of this increase comes from the rise of industry and automobile usage. The seasonality of the data can be explained by the seasons themselves. In spring and summer plant life, and thus photosynthesis, is at it's peak. This means that there is an entire eco system absorbing CO_2 from the atmosphere. In the fall and winter months as as plants die there is no such absorption mechanic and we see levels rise.

For our model we took those factors into account and used a seasonal autoregressive integrated moving average, also known as an SARIMA model. The SARIMA model or Seasonal ARIMA model is a generalization of the ARIMA or AutoRegressive moving average model which are fitted to time series models to get a better understanding of the data or in our case, predicting future values. The autoregressive part indicates that the regression is based on values from itself and therefore predicted values are formed from past values and/or past predicted values. The moving average part uses past forecasting errors and white noise as a means for generating predicted values rather than using past values. This integration of the moving average allows for the estimation to become more accurate as more predictions are made. In our case, given the length of time over which we have CO2 data, this works in our advantage. The integrated part takes into account differencing, or specifically takes into account the trending aspect of the our data. This stabilizes the mean to allow for more accurate future predictions. The integrated aspect of ARIMA can also account for seasonality, but considering the overwhelming influence that seasonality has on the data, it was decided to make the ARIMA model specifically seasonal. The seasonality aspect is clear from the provided data but also from a scientific aspect. Due to seasonal blooming of plants, it makes sense that CO_2 levels would have a trend that follows the cycle of plant life/growth. This is the basic rationale for the choice of SARIMA model to fit and predict the data. A SARIMA model is defined by parameters p, d, q and P, D, Q displayed as ARIMA(p,d,q)x(P,D,Q). For this model p is the order given to the autoregressive process, d is the order given to the regular differencing, and q is the order given to moving average process. The capital letters, P, D, and Q are simply the same definitions, but given to the seasonal processes. The last parameter that requires definition will be the seasonal subindex s. Given that CO_2 fluctuates based on yearly cycles we determined that the optimal value for s would be 12.

To generate our chosen model we first set our differencing factor to a positive integer, given the fact that our data is not stationary. From there we set the range of our p, q, P, and Q to 0-2. Given that range of possible models we tested each one and calculated the AIC of each model and used those values to rank the fit and predictive power of them.

The final model that was chosen to predict the CO_2 levels from Mauna Loa was a weighted system using the SARIMA as the predicting model. After testing out multiple statistical models we concluded that SARIMA model provided the most accurate prediction for the month of April given testing of February and March. It was decided to use a weighted average for a couple of reasons. After testing multiple p, d, q values for SARIMA, the resulting models seemed to provide conflicting information based on standard metrics. The p,d,q model that most accurately predicted the month of February had one of the worst BIC and AIC values from all the models tested. While the models with the highest BIC and AIC provided accurate estimations but not nearly as accurate as the previously stated model. Seeing as the BIC and AIC are criterion measures used to select appropriate models, it was decided that these models could not be ignored. As such, a weighted model was used to incorporate the models with the three highest BIC/AIC criterion with the most accurate predicting model. The most accurate model was given the highest weight seeing as it has the highest probability of providing the true value for April. The remaining three models were split evenly among the remaining weight.

Description of Model

The statistical model that we found that best predicts the monthly average CO_2 levels in Mauna Loa is a weighted-average of four seasonal ARIMA models.(). To create our final model, we had to accurately select a p, d, q for the SARIMA model to fit the CO2 values from Mauna Loa. Initially a by-hand trial and error method to test p, d, q values was used with expected values believed to provide to best forecasting. After a few models, an automated method was found for testing large ranges of p, d, q values using SAS Jump to test ranges of p, d, q from 0 to 3. This provided a large range of models to compare BIC, AIC and most importantly, the predicted CO2 value for April. Looking at the second set of P, D, Q values for the SARIMA, it should be noted that this part of the model across all 4 models that were selected are very similar. This is the seasonal aspect of the SARIMA model and as such leaves little room for interpretation. The rationale for choosing the AR, differencing and MA values for the seasonal aspect can be found above in the description of the model. The rationale for the ARIMA will translate to the seasonal component. The variation in the selected models can be seen in ARIMA part of our models. As stated in the description, BIC and AIC had a large influence on the which model was chosen. The top three BIC/AIC values also provided a good range of interpretations for the data. The final model selected entirely removes the ARIMA aspect and focuses solely on the seasonal aspect of this data. This model provides the most accurate prediction for February and March. As such this model provides the best possibility for accurate predictions of April based on past predictions. We felt that all four of these models gave a wide interpretation for the AR, Differencing, and MA values and should all be included as predictors for April. Therefore we decided to use a weighted average of all these models. The breakdown is as follows: Model 4, the most accurate would count for 40% of the predicted value of April, the remaining three would be each count 20%. We feel the most accurate model past predicted values should weighted more due to its past accuracy. The remaining models, given their almost equally high likelihood values from BIC and AIC should count equally into the model.

Model	DF	Variance	AIC	SBC	AIC Rank out of 4096	SBC Rank out of 4096	RSquare	-2LogLH	Weights	MAPE	MAE
Seasonal ARIMA(2, 0, 1)(0, 1, 1)12	156	0.079619513	86.44450145	101.8515233	1	4	0.998637807	76.44450145	0.06363553	0.061902	0.238854383
Seasonal ARIMA(1, 0, 2)(0, 1, 1)12	156	0.079594074	86.51621503	101.9232369	2	5	0.998638139	76.51621503	0.121378448	0.061947	0.239051197
Seasonal ARIMA(1, 0, 1)(0, 1, 1)12	157	0.079968012	86.62522734	98.9508448	3	2	0.998622577	78.62522734	0.114939647	0.062364	0.240696595
Seasonal ARIMA(0, 0, 0)(1, 1, 1)12	158	0.150957654	183.1694135	192.4136266	1475	1212	0.99742546	177.1694135	2.25046E-21	0.089552	0.345210119

Seasonal ARIMA(2, 0, 1)(0, 1, 1)12	Date	Prediction	Standard Error	Residual	Upper .99 CI	Lower .99 CI	Weighted Prediction .2	Weighted Upper CI .2	Weighted Lower .2
20% weight	Feb-15	400.639633	0.28215878	-0.379633004	401.3664259	399.9128402	80.1279266	80.27328517	79.98256803
	Mar-15	401.285125	0.28215878		402.0119179	400.5583322	80.25702501	80.40238358	80.11166644
	Apr-15	402.6338711	0.321784929		403.4627341	401.805008	80.52677421	80.69254682	80.3610016

Seasonal ARIMA(1, 0, 2)(0, 1, 1)12	Date	Prediction	Standard Error	Residual	Upper .99 CI	Lower .99 CI	Weighted Prediction .2	Weighted Upper CI .2	Weighted Lower .2
20% weight	Feb-15	400.6284936	0.282118812	-0.368493547	401.3551835	399.9018036	80.12569871	80.27103669	79.98036073
	Mar-15	401.2843602	0.282118812		402.0110501	400.5576703	80.25687205	80.40221003	80.11153407
	Apr-15	402.6435653	0.321612618		403.4719845	401.8151461	80.52871306	80.6943969	80.36302922

Seasonal ARIMA(1, 0, 1)(0, 1, 1)12	Date	Prediction	Standard Error	Residual	Upper .99 CI	Lower .99 CI	Weighted Prediction .2	Weighted Upper CI .2	Weighted Lower .2
20% weight	Feb-15	400.6297892	0.282782837	-0.369789212	401.3581895	399.9013889	80.12595784	80.27163791	79.98027778
	Mar-15	401.2808432	0.282782837		402.0092435	400.5524429	80.25616864	80.40184871	80.11048858
	Apr-15	402.5844367	0.319847956		403.4083105	401.760563	80.51688735	80.6816621	80.3521126

Seasonal ARIMA(0, 0, 0)(1, 1, 1)12	Date	Prediction	Standard Error	Residual	Upper .99 CI	Lower .99 CI	Weighted Prediction .4	Weighted Upper CI .4	Weighted Lower .4
40% weight	Feb-15	400.2559807	0.388528965	0.004019333	401.256765	399.2551964	160.1023923	160.502706	159.7020785
	Mar-15	401.3178573	0.388528965		402.3186416	400.317073	160.5271429	160.9274566	160.1268292
	Apr-15	402.709708	0.388528965		403.7104923	401.7089237	161.0838832	161.4841969	160.6835695

Weighted Prediction	Date	Prediction	Upper .99 CI	Lower .99 CI
	Feb-15	400.4819754	401.3186658	399.6452851
	Mar-15	401.2972086	402.133899	400.4605183
	Apr-15	402.6562578	403.5528027	401.7597129

Figure 3.1 Weighted-Average Final Seasonal ARIMA Model

Forecast of Monthly Average of CO2

Based on our weighted-average SARIMA model, our final forecast of the monthly average of CO_2 in April is 402.6563 ppm. A graph showing the forecast for April can be seen in Figure 4.1. The 99 percent prediction bound for the April forecast is [401.7597, 403.5528].

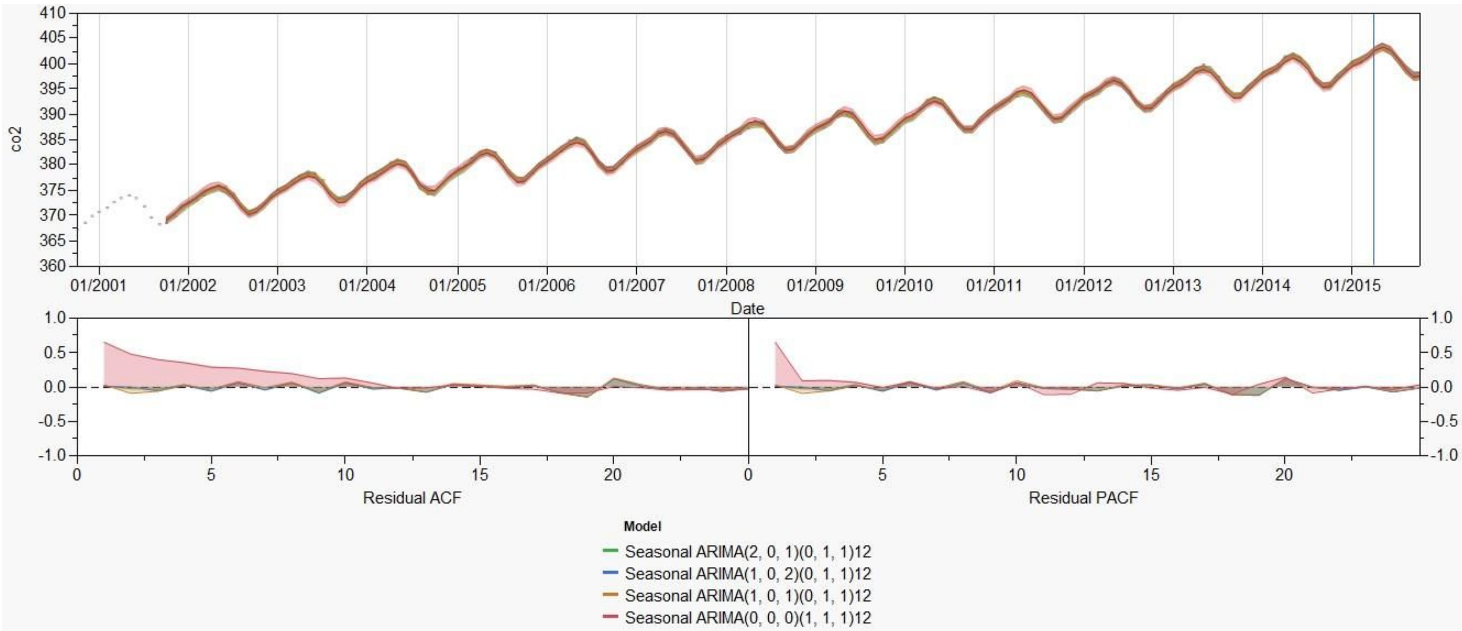

Figure 4.1 CO_2 Levels October 2000 to April 2015

Conclusion

We created a weighted-average seasonal ARIMA model that we believe best characterizes the underlying trend behind the monthly average CO_2 levels. We evaluated different orders of SARIMA models and several combinations of weighted averages in order to arrive at this model. We used data from February 2015 and March 2015 to analyze the accuracy of our model and concluded that the (p,d,q)x(P,D,Q) values and weighted average that we ultimately chose led to the model with the best fit.

References

Butler, James. "CO2 at NOAA's Mauna Loa Observatory Reaches New Milestone: Tops 400 Ppm." *NOAA Research*. N.p., 10 May 2013. Web. 28 Mar 2015. <http://www.esrl.noaa.gov/gmd/news/7074.html>.

"Facts about Mauna Loa." *West Hawaii Today*. N.p., 4 Feb 2015. Web. 28 Mar 2015. <http://westhawaiitoday.com/news/volcano-update/facts-about-mauna-loa>.

Monroe, Rob. "Why are seasonal CO_2 fluctuations strongest at northern latitudes." *Scripps Institute Of Oceanography*. N.p., 7 May 2013. Web. 28 Mar 2015. <https://scripps.ucsd.edu/programs/keelingcurve/2013/05/07/why-are-seasonal-CO2-fluctuations-strongest-in-northern-latitudes/>.

"8.4 Moving Average Models." *OTexts*. N.p., n.d. Web. 28 Mar 2015. <https://www.otexts.org/fpp/8/4>.

"8.1 Stationarity and Differencing." *OTexts*. N.p., n.d. Web. 28 Mar 2015. <https://www.otexts.org/fpp/8/4>.