Forecasting Carbon Dioxide Levels in Mauna Loa, Hawaii: A Study of the Moving Averages Smoothing Method

Abstract

In a NOAA dataset from Mauna Loa, Hawaii, average carbon dioxide levels have been steadily increasing for the past few decades. Additionally, the monthly levels vary seasonally, with relatively higher levels in the winter and relatively lower levels in the summer, due to photosynthetic plants. Since this dataset has a secular and seasonal component, it would serve as a good example to perform a time series model with the moving averages smoothing method. To analyze its effectiveness, we compared its results to the results of a linear regression model with seasonal variation and a time series model with first order autoregressive errors. We compared all the models' forecasted results to the actual measured result of October 2018. While the third model gave the most accurate result, it does not include a seasonal component, for which the moving averages model would be the most useful.

Research Questions
How useful is the Moving Averages Method in forecasting future results? How does its utility compare to the linear regression model with seasonal variation and the time series model with a first order autoregressive error component?

Introduction
The purpose of this study is to forecast monthly carbon dioxide levels in Mauna Loa, Hawaii. While the Earth's carbon dioxide is largely stored in the Earth's oceans, it is a trace gas in our atmosphere that varies seasonally due to photosynthetic plants. In the spring and summer, the plants are active and performing photosynthesis, meaning that less carbon dioxide is in the air. In the fall and winter, the plants die off or are dormant, meaning that there is less carbon dioxide in the air. However, carbon dioxide is also a greenhouse gas, and scientists have concluded that anthropogenic emissions such as the burning of fossil fuels and agriculture have been steadily increasing the amount of atmospheric carbon dioxide, and therefore contributing to global warming. It is important to track the growth of the carbon dioxide percentage in the atmosphere and predict future amounts to be able to try to reduce our carbon emissions and keep our planet healthy.

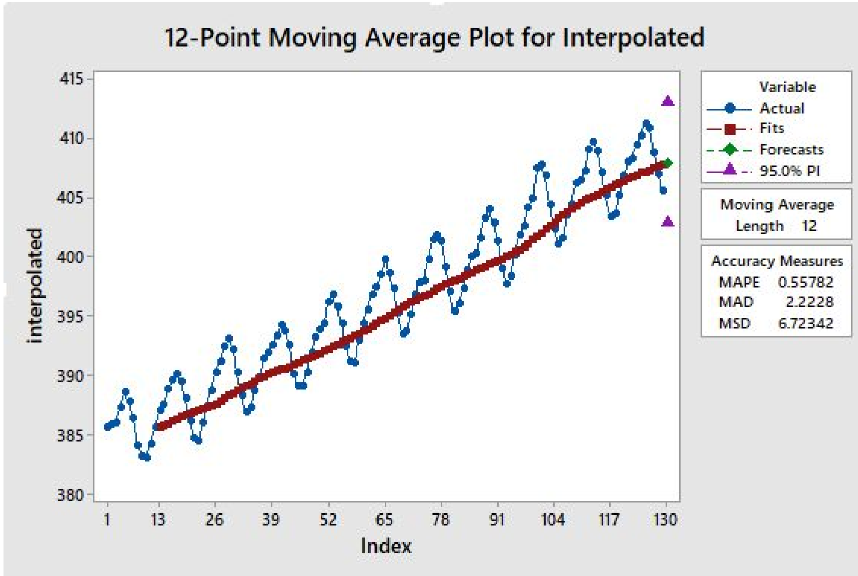
Our dataset is from the National Oceanic and Atmospheric Administration's Earth System Research Laboratory Global Monitoring Division (NOAA ESRL Global Monitoring Division), which records the monthly mean carbon dioxide level as a mole fraction from March 1958 to October 2018 (Tans and Keeling). Since monthly carbon dioxide level is recorded over time, one approach to modeling the carbon dioxide level is to use a time series model. In this study, our team will identify and model the data with time series methods, analyze the long-term trend, and utilize the moving average method to "smooth out" these seasonal fluctuations and forecast future carbon dioxide levels.

Data Summary
We generated scatter plots with MINITAB to explore average carbon dioxide trends across recent years. The graph shows a clear long-term increasing trend of carbon dioxide levels during the ten year period of interest **(Figure 1)**. From the visual, we observed that average carbon dioxide levels rise earlier in the year, decrease around mid-year, and then increase into the next calendar year. This trend continues across the years in this subsample and the complete dataset and is consistent with our expected pattern of carbon dioxide emission level **(Figure 2)**.

Methodology
For the focus of this report, we will explore the moving averages smoothing method. The main assumption of forecasting with a moving averages model is that the model is constant (Jensen). This implies that the parameters don't change as time changes, and the mean response value shows a linear trend during the period of interest. Also, it is important that the time variable under consideration is consistent, meaning that if the monthly data is of interest each observation used for the moving average should be the monthly data. For our analysis, monthly index values were calculated to estimate the effect of month on average carbon dioxide levels. These monthly index values can be utilized to forecast future average carbon dioxide levels across months with consideration of the long-term trend and the seasonal variation trend.

Analysis

Model 1: 12-point Moving Averages



The forecasted value for average carbon dioxide level in October 2018 was of 407.8940 with a 95% prediction interval of 402.8120 to 412.9780. As we have observed a seasonal trend, the final forecasted value is calculated as 407.894 multiplied weighted with the seasonal index for October. Our Model 1 forecast with the seasonal index is 405.43.

Model 2: Linear Regression Model with Seasonal Variation

Model 2 was developed as a linear time series model with monthly dummy variables (x_1, ..., x_{11}) to account for seasonal variation we observed during exploratory data analysis. From the Model 2 regression results, the observed p-value on the global F-test below the 0.05 threshold led us to reject the null hypothesis which suggests that at least one of the included variables is statistically significant for predicting monthly carbon dioxide level **(Table 1)**. The October 2018 prediction from Model 2 yielded a forecast of 405.6228 and a 95% prediction interval of 404.4995 to 406.7462. Model 2 relies on the assumption of independent errors, so to check for residual correlation, we conducted a Durbin-Watson test and found the p-value for negative autocorrelation is 1.000 which suggests there is autocorrelation in the error term **(Table 2)**.

Model 3: Time Series Model with First Order Autoregressive Errors

Model 3 assumes that the average carbon dioxide level in month t will be highly correlated with the average carbon dioxide level in the previous month, month t-1. However, different than the other two models, Model 3 doesn't take seasonal variation into account, but accounts for the autocorrelation in the error term. For model assumptions, we checked the lack of fit and constant variance assumptions by visual analysis of the residual plot **(Figure 3)**. From the plots, we determined that the residuals don't exhibit a distinctive pattern or trend, and appear to be evenly distributed around zero. As well, the lack of a "fanning" in or out pattern and the spread of residuals around the mean is consistent across time, so the assumption of homoscedasticity is also satisfied. We then affirmed the normality assumption by examining the residual histogram, which showed that the distribution is pretty symmetric **(Figure 4)**. Therefore, all assumptions for Model 3 are satisfied. From a SAS regression, the forecasted value of average carbon dioxide level in October 2018 based on Model 3 is 406.0895 with a 95% confidence interval of 403.5892 to 408.5898 **(Table 3)**.

Conclusion

Best Model and Interpretation

Since the distinct modeling methods make it difficult to compare model utility with a statistical test, our principal decision for goodness of fit is based on comparing each model's forecast for October 2018. The actual measured carbon dioxide level for October 2018 is 406.00, and the closest model was Model 3. Model 3 assumes that the average carbon dioxide level in any given month will be closely related with the average carbon dioxide level in the previous month. Unlike Model 1 and Model 2, Model 3 measures the increase and decrease in carbon dioxide levels and does not provide explanation for seasonal variation across months. As well, Model 3 saw a smaller 95% prediction interval when compared to Model 1. Although Model 2 had an even smaller range of possible values than the third, it violated the regression error assumption independence, and thus we opted not to proceed with the model. From our exploration, we believe that considering the seasonal trend is very important to predicting the mean monthly carbon dioxide levels. However, Model 3 only serves as a good predictor of the next month's values given current data. From a forecasting perspective, a model's predictions will become less accurate the further the projection, and this is an important consideration when assessing our models.

Recommendations for Future Research

This study was successful in assessing the positives and negatives of a variety of forecasting methods. Future researchers should be aware of the problem of extrapolation, as any model that makes predictions about the future based on current data will run into the issue of extrapolation. In working with the models assessed in this study, it is important that the models are updated with current data to note any change in the observed trends. There is also the issue of negative autocorrelation in the data, as the measured carbon dioxide levels are not independent. This must be taken into consideration when building a model to predict carbon dioxide levels, as shown from our research, as the carbon dioxide levels of one month are closely related to the previous month. For our study, although we used first-order autoregressive errors to account for the dependence, there might be a better autoregressive model than the one we used.

Our study focused on measuring average carbon dioxide levels across months and observing the clear seasonal and long-term trends. Thus, our model can be utilized to note any changes in the trend which would suggest a detrimental impact on the environment. For future studies to diagnose the question of the relationship of climate change and carbon dioxide levels, it is crucial to dig deeper into the causes of changing carbon dioxide levels and explore our question from various perspectives. For the sake of our ecosystem and society, it is critical to analyze carbon dioxide levels, identify contributors to change in the carbon dioxide levels, and implement societal changes to keep our planet healthy.

References

Mendenhall, William, Terry Sincich, and William Mendenhall. *A Second Course in Statistics: Regression Analysis*. Upper Saddle River, N.J: Prentice Hall, 1996. Print.

Jensen, Paul A. "Forecasting Theory - Moving Average." *Department of Mechanical Engineering*, University of Texas at Austin

Tans, Pieter and Ralph Keeling. "Trends in Atmospheric Carbon Dioxide." NOAA/ESRL, Scripps Institution of Oceanography. 9 December, 2018.

Appendix

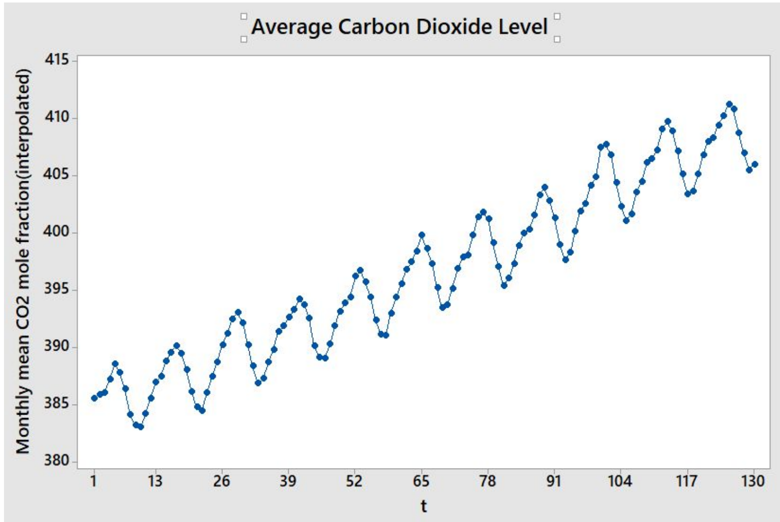


Figure 1: MINITAB scatter plot of monthly average carbon dioxide level from January 2008 to October 2018. The time variable is indicated as t. A long-term trend is apparent across the most recent 10 years, as well as seasonal variation.

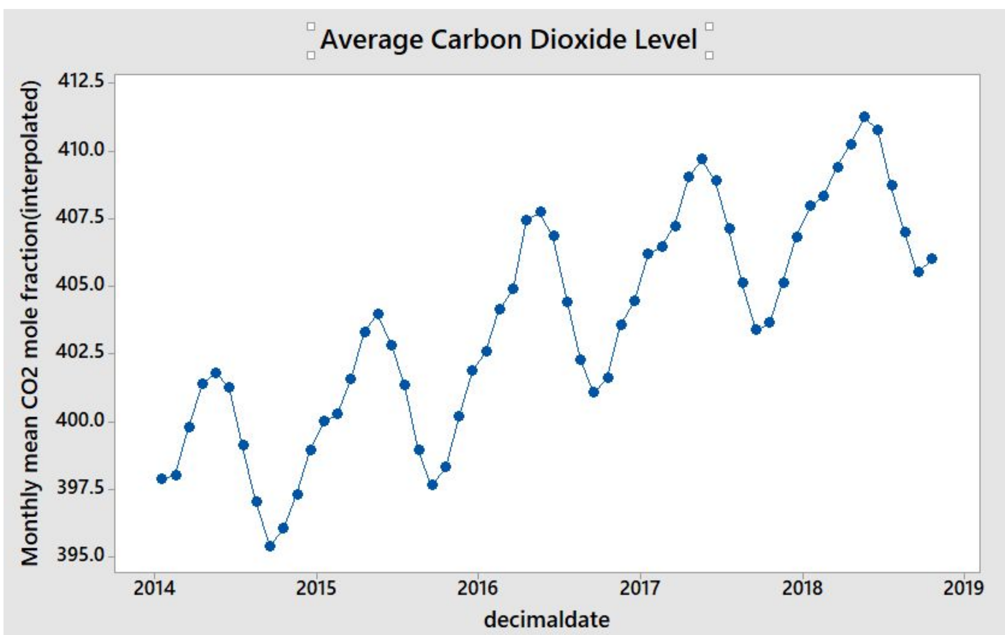


Figure 2: MINITAB scatterplot of average carbon dioxide level from January 2014 to October 2018, focusing on a smaller portion of the data. Time is indicated in years. The long-term and seasonal variation trends continue. This plot gives us a clear idea of carbon dioxide fluctuation across months with the sample pattern across years.

Variable	Estimate	SE	p-value
Time	0.1944	0.0013	<.0001
January	1.1606	0.2234	<.0002
February	1.6499	0.2234	<.0003
March	2.3692	0.2234	<.0004
April	3.5557	0.2234	<.0005
May	4.1186	0.2234	<.0006
June	3.1252	0.2234	<.0007
July	1.2436	0.2234	<.0008
August	-1.0245	0.2234	<.0009
September	-2.607	0.2234	<.0010
October	-2.5993	0.2234	<.0011
November	-1.2476	0.2392	<.0012
Intercept	382.9552	0.1887	<.0013

Table 1: The results for Model 2 are summarized above. The global F-test (p<.0001) led us to reject the null hypothesis that all parameter estimates were equal to zero. From the output, parameters for Model 2 are significant at alpha = .05 level.

The REG Procedure
Model: MODEL1
Dependent Variable: interpolated

Durbin-Watson D	0.484
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	129
1st Order Autocorrelation	0.736

Table 2: A Durbin-Watson test was conducted to test for correlation of the error term in Model 2. From the results, the p-value for negative autocorrelation is 1.000, suggesting there is correlation in the error term and an autocorrelation term should be added to the model.

The ARIMA Procedure

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	384.76524	1.00026	384.67	<.0001	0	interpolated	0
AR1,1	0.84263	0.04935	17.07	<.0001	1	interpolated	0
NUM1	0.18047	0.01442	12.52	<.0001	0	time	0

Forecasts for variable interpolated				
Obs	Forecast	Std Error	95% Confidence Limits	
130	406.0895	1.2757	403.5892	408.5898

Table 3: Included above is the ARIMA output of Model 3, our time series model with a first order autocorrelation term. The ARIMA procedure includes the estimated parameters for the prediction equation and the forecasted carbon dioxide level in October, 2018: 4.0895.

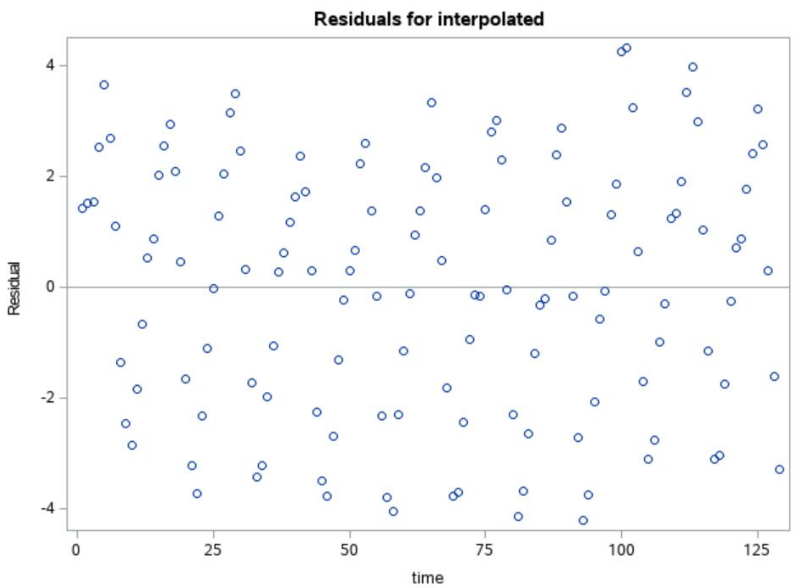


Figure 3: The SAS residual plot for carbon dioxide level against time from Model 3. From visual analysis, there is no apparent pattern in the residuals.

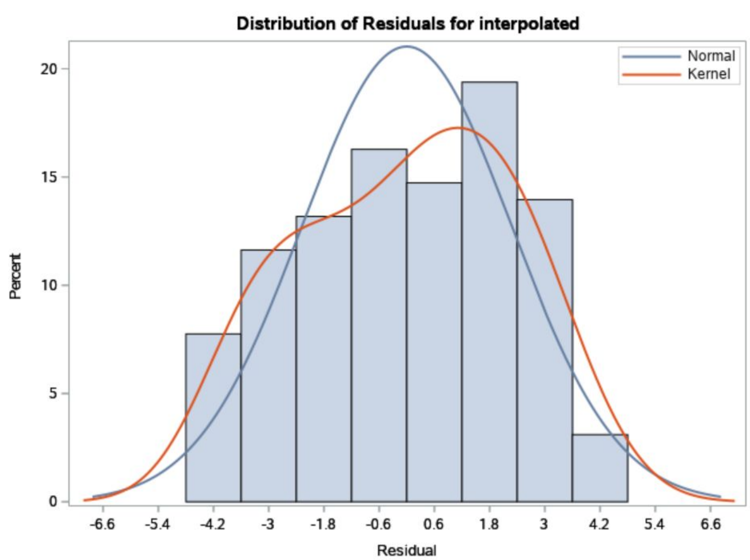


Figure 4: Above is the Model 3 residual histogram. The distribution of residuals shows a symmetric bell shape which is very close to normal distribution. We conclude that the residuals sufficiently follow a normal distribution and our Model 3 can be assumed normal.