ANALYSIS OF CO_2 CONCENTRATION TIME-SERIES DATA COLLECTED AT MAUNA LOA, OBSERVATORY, HAWAII AND RESULTS GENERATED BY MODEL FITTING

Prepared by Liang Chen

April 11, 2021

Abstract

Climate change has been a global issue in recent years, one of these phenomena is global warming. As one the main greenhouse gases, Carbon Dioxide draws more attention and taking action to reduce its atmospheric concentration back to normal level becomes even more urgent. In this case, building a reliable model to predict CO_2 concentration would be considered as a useful tool to help environmental scientists investigate and find solutions to global warming. In this report, we took advantage of CO_2 concentration data collected in Hawaii to build a seasonal ARIMA model. Except for conducting significance tests on the fitted model, we used this model to predicted future values as well. The good performance of the fitted model tells us that: there is an upward trending of atmospheric CO_2 concentration, which might be caused by industrialization. Besides, seasonal variation of CO_2 concentration indicates huge impacts of human activities on the nature. This implies that action has to be taken to reduce the upward trending of CO_2 concentration to save creatures at the edge of extinction. It also proved that regulation on human activities will effectively solve this issue. Besides, the model fitted is rough and there are many other factors

might be associated with CO_2 concentration that should be taken into consideration to generate more reliable models.

Introduction

It is known that excessive Carbon Dioxide in the atmosphere is tightly associated with global warming, which poses threats to all living creatures on the Earth. It would be helpful if we could build a relatively accurate CO_2 concentration model and use it to predict climate changes and prevent occurrence of extreme weather. To build the model, we downloaded the data set of CO_2 concentration collected roughly every week at Mauna Loa, Observatory, Hawaii from the official website of Scripps institution of Oceanography. This data was collected from March 1961 to the January 2021. Considering rapid modernization in recent decades, we take only recent five years' observations with sample size equals to 240 into consideration to fit a more accurate model.

Statistical Methods

Firstly, we start by generating a weekly time series plot of CO_2 concentration. From Figure 1, we can tell that there exists a significant upward trend with slightly increasing variance in both the five-decades' observations and recent 5 years' data.

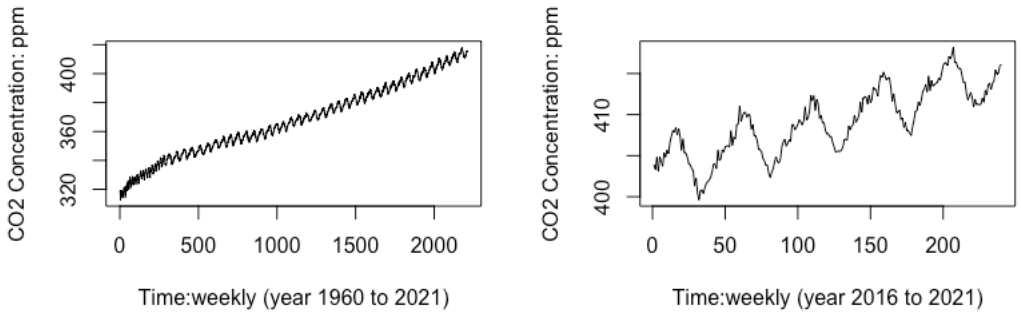

Figure 1. *weekly time series plot of CO2 concentration at Mauna Loa, Observatory, Hawaii*

Therefore, we will do transformation on recent 5 year's CO_2 concentration x_t, which is a large enough data set to investigate in. By differencing the logged data, we remove the trend upward trend and get a time series $\nabla \log x_t$. Besides, it is clear that there exists persistence in seasons as well, and a fifty-two-order difference on the differenced log data is necessary as well. It can be told from Figure 2 that, after these transformations, the time series data $\nabla_{52}\nabla \log x_t$ looks stationary with constant mean and variance around zero.

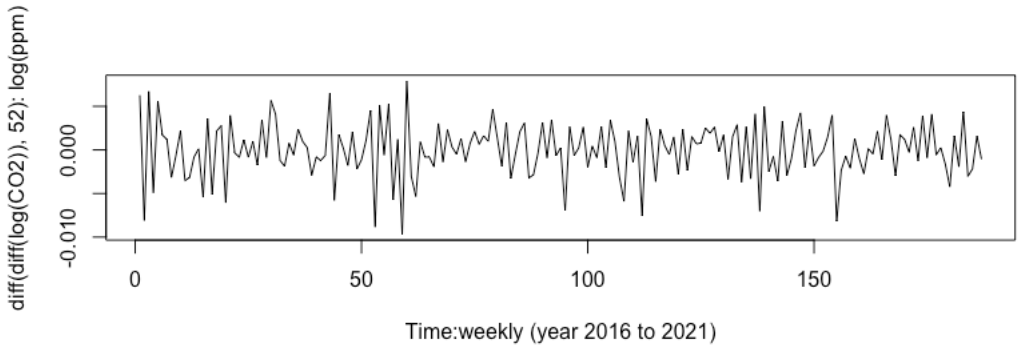

Figure 2. *weekly time series plot of* $\nabla_{52}\nabla \log x_t$ *at Mauna Loa, Observatory, Hawaii (2016-2021)*

Now, we can use the preprocessed data to fit models. The sample ACF and PACF of $\nabla_{52}\nabla \log x_t$ are shown in Figure 3.

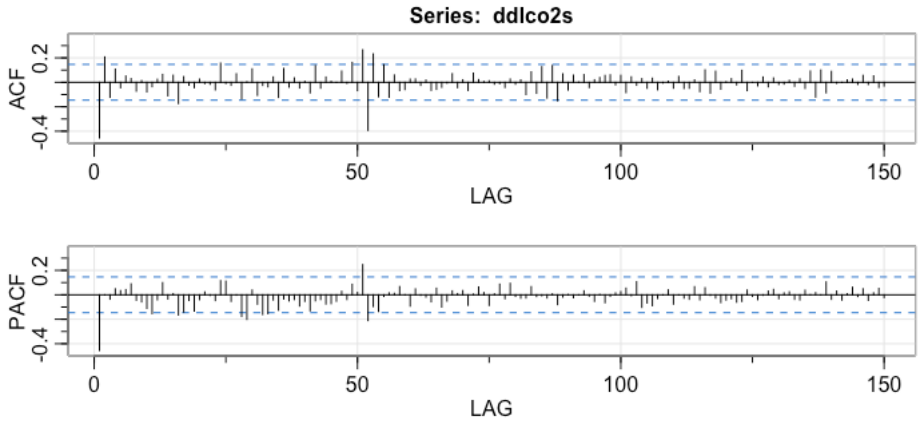

Figure 3. *Sample ACF and PACF of ddlco2 (*$\nabla_{52}\nabla \log x_t$)

Since we consider fit a SARIMA model for logged data $\log x_t$, we have to identify both seasonal and non-seasonal components of this model. It seems that at seasons, ACF cut off at $lag = 1s$ where $s = 52$. PACF also appears to cut off at $lag = 1s$. Therefore, we may consider try $P = 1$, $Q = 1$, $s = 52$, and $D = 1$. As for the non-seasoning components, we can tell that ACF cuts off at $lag = 1$ or 2. PACF cuts off at $lag = 1$ though the AR parameter is insignificant. Thus, we consider $p = 1$, $q = 1$ or 0, and $d = 1$. Then, we try to fit two models: $ARIMA(1,1,1) \times (1,1,1)_{52}$ and $ARIMA(1,1,0) \times (1,1,1)_{52}$. Figures 4 and 5 show the residual analysis for these two possible models.

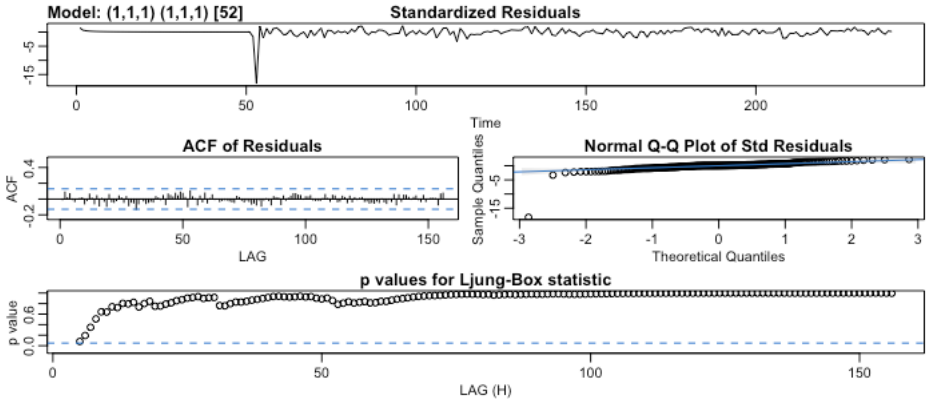

Figure 4. *Residual analysis for* $ARIMA(1,1,1) \times (1,1,1)_{52}$ *fit for logged CO2 concentration (*$\log x_t$*)*

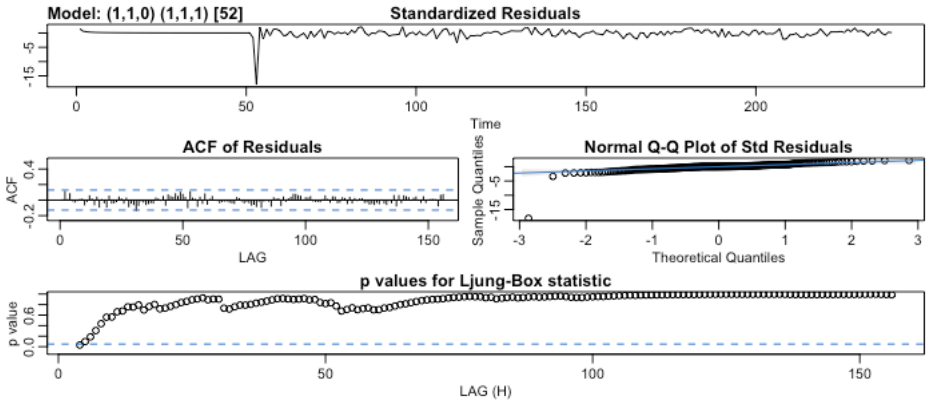

Figure 5. *Residual analysis for* $ARIMA(1,1,0) \times (1,1,1)_{52}$ *fit for logged CO2 concentration (*$\log x_t$*)*

It can be seen that both models fit well except for a few outliers: both standard residuals look like white noise except for the fifty-third or fifty-fourth observation. The ACF of both series show no departure from the normal assumption. Both normal Q-Q plots apparent to be straight lines except for tolerable outliers, which meets the normality assumption as well. The outlier at very small quantile in the Q-Q plot meets the outlier shown in the standard residual plot. The p-values for Ljung-Box statistics are above the significance level for most lags except for a few points close to the dashed line, which means that we do not reject the null hypothesis that residuals are independent. To pick a better one among the two models, we calculated AIC, BIC and AICc and results are shown below in Table 1.

	AIC	BIC	$AICc$
$ARIMA(1,1,1) \times (1,1,1)_{52}$	-7.128896	-7.061015	-7.128174
$ARIMA(1,1,0) \times (1,1,1)_{52}$	-7.131548	-7.077244	-7.131117

Table 1. *AIC, BIC and AICc for two fitted models*

From the table, we can tell that $ARIMA(1,1,0) \times (1,1,1)_{52}$ has smaller values of AIC, BIC and AICc. Combined with what we found earlier that these two models have similar good performance in residual analysis, we choose to fit an $ARIMA(1,1,0) \times (1,1,1)_{52}$ with smaller AIC, BIC and AICc values as the following format:

$$(1 - \Phi_1 B^{52})(1 - \phi_1 B)\,\nabla_{52}\nabla \log x_t = (1 + \Theta_1 B^{52}) w_t$$

Result

By fitting the model selected in previous section, we got parameter estimations and corresponding p-values as listed in Table 2:

	point estimation	$p-value$
ϕ_1	-0.3540	0.0000
Φ_1	-0.0531	0.8268
Θ_1	-0.5401	0.0363

Table 2. *Parameter estimation and p-values of fitted* $ARIMA(1,1,0) \times (1,1,1)_{52}$ *model*

Based on the significance test result, we can tell that p-values of parameters estimations of ϕ_1 and Θ_1 are smaller than significance level $\alpha = 0.05$. However, p-value of parameter estimation of Φ_1 is much greater than 0.05. It shows that the fitted $ARIMA(1,1,0) \times (1,1,1)_{52}$ model parameters are statistically significant except for Φ_1, and the final fitted model is as following:

$$(1 + 0.3540B)\,\nabla_{52}\nabla \log x_t = (1 - 0.5401B^{52})w_t$$

Except for point estimation of parameters ϕ_1, Θ_1, Φ_1 and their corresponding p values, we can also tell that the difference in log (CO_2 concentration) is almost entirely due to what the level of log (CO_2 concentration) was one year ago with minor influence from log (CO_2 concentration) of previous weeks. In detail, $p = 1$ and seasonal $P = 1$ refers to the number of autoregressive term is 1; $d = 1$ and seasonal $D = 1$ means that one differencing must be done to stationarize the series; $q = 0$ and seasonal $Q = 1$ refers to the number of moving average terms separately; and $s = 52$ indicates seasonal length of 52 weeks in the data.

Taking advantage of the fitted model, we are able to predict future values of CO_2 concentration. Figure 6 shows the next ten weeks' log (CO_2 concentration) forecasting in red dots and corresponding prediction intervals in gray areas.

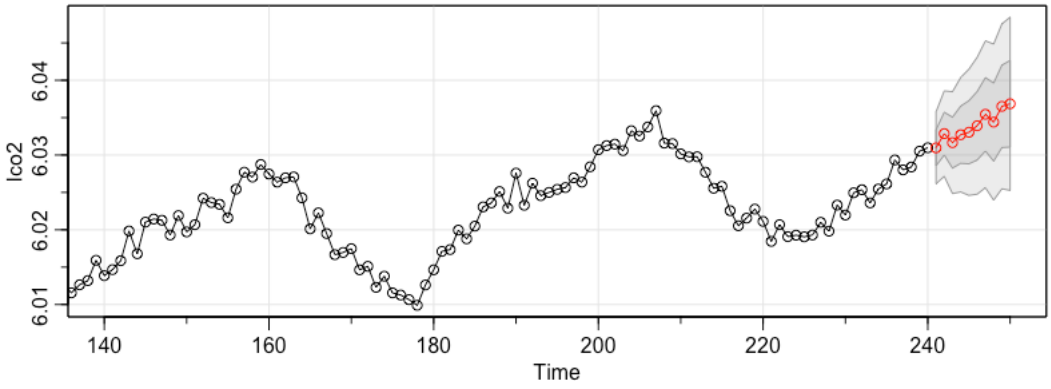

Figure 6. *Ten-week forecast using* $ARIMA(1,1,0) \times (1,1,1)_{52}$ *model on logged CO2 concentration (*$\log x_t$*)*

The detailed prediction values of log (CO_2 concentration) and the 95% prediction intervals are listed in Table 3 as well.

	Prediction	Lower bound of 95% PI	Upper bound of 95% PI
Week 1	6.030957	6.026250	6.035665
Week 2	6.032866	6.027262	6.038470
Week 3	6.031655	6.024977	6.038333
Week 4	6.032715	6.025212	6.040218
Week 5	6.033045	6.024768	6.041323
Week 6	6.033918	6.024943	6.042894
Week 7	6.035450	6.025824	6.045076
Week 8	6.034403	6.024168	6.044637
Week 9	6.036512	6.025703	6.047321
Week 10	6.036856	6.025501	6.048210

Table 3. *Ten-week forecast with 95% prediction interval using* $ARIMA(1,1,0) \times (1,1,1)_{52}$ *model on logged CO2 concentration (*$\log x_t$*)*

From the forecasting results, we can tell a few things. In the next 10 weeks, there exists an upward trending of point estimation, which means that the log (CO_2 concentration) will become higher in the next 10 weeks. If we make prediction on a larger time period, such as 52 weeks, we would be able to see a full cyclical seasonal trend of log (CO_2 concentration) including a peak and a valley with a general upward trend, which means that for the same week, the log (CO_2 concentration) in this year would be higher than that in last year.

As for the range of 95% prediction interval, we can tell that as time goes, the range of prediction interval becomes wider. However, compared with the value of point estimation, the range is still narrow, which means that we can make relatively precise prediction.

We could also identify the first three predominant periods by performing a periodogram analysis. Figure 7 shows the periodogram of CO_2 concentration.

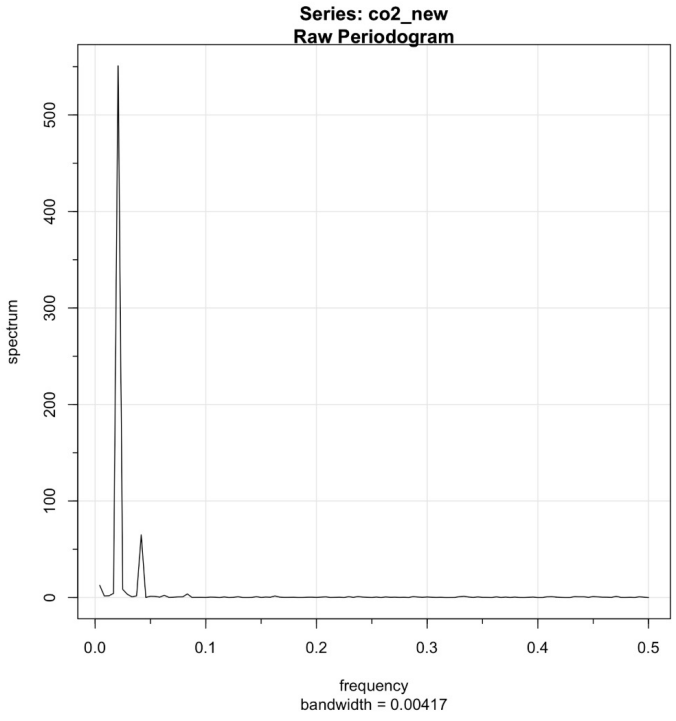

Figure 7. *Periodogram of CO2 concentration*

We can tell from Figure 7 that there is a peak at around 0.02, closing to the value of 1/52. This consistent with previous seasonal analysis with $s = 52$. The first three predominant spectrums with the 95% confidence intervals are shown in Table 4 below.

	spectrum	frequency	Period	Lower bound	Upper bound
period 1	550.8557	0.0208	48.0000	149.32873	21757.6380
period 2	64.9138	0.0417	24.0000	17.59716	2563.9581
period 3	12.5779	0.0042	240.0000	3.40968	496.8005

Table 4. *First three predominant periods of CO2 concentration data set*

From the table above, we can tell that the 95% confidence intervals for all three periods are large, which implies that it is helpless and cannot tell much information about significance. In detail, we get the following information: we cannot establish significance of the first and second peaks, but we are able to establish significance for the third peak. For the first peak, its periodogram ordinate is 550, lying in the 95% confidence interval of the second peak; for the second peak, its periodogram ordinate is 65, lying in the 95% confidence interval of the third peak. However, for the third peak, its periodogram ordinate is 12.6, which does not lie in the 95% confidence intervals of any other two peaks. Thus, we can only establish the significance of the third peak among all the three peaks. Also, we can tell that the 95% confidence intervals for all three periods are large, which implies that it is helpless and cannot tell much information about significance.

Discussion

Based on the combination of prediction of the fitted model and the original data, we can tell that there exists an upward trending of atmospheric CO_2 concentration in recent decades. This can be explained by excess emission of greenhouse gases into the air caused by industrialization and globalization. Seasonal trend of CO_2 concentration is also significant, and peaks usually appear in fall and winter and drop back to bottoms in summer. That might be explained by living habits of human beings that people have to heat houses by burning fossil fuels in winter times. Although the model provides reasonable prediction, there still exists some limitations. One thing is that we did not deal with outliers before fitting the model, and trimmed data sets might help improve the performance of fitted model. Besides, we may consider mixture of multiple periods in building the model. This is reasonable assumption since there exist other periods such as economic inflation and recession cycles that also impact CO_2 concentration. Meanwhile, when we generate the model, we didn't take the special situation of COVID-19 into consideration. The lockdown worldwide caused by this pandemic dramatically shocked global economy and CO_2 concentration dropped significantly as well. Part of the data used in model-fitting are collected during the COVID-19, and thus it may impact the reliability of fitted model.