Time Series Prediction of the Monthly Average CO_2Level from MAUNA LOA OBSERVATORY in November, 2019

Stor 556 - Fall 2019

By Group NaOH
Yiran Zhang, Yuxiao Yao, Ziyi Liu
Hanzhao Yang, Yanchu Zhou

1. Introduction

The gradually rising concentration of CO_2 in the atmosphere and the consequential climate change have been two of the most difficult problems that the world is facing today. The Mauna Loa island was originally chosen as a monitoring site because of its location and elevation. It is located far away from any continent as the nearest city is 40 miles away, and is 13,679 feet above sea level. The Mauna Loa Observatory is known for its continuous monitoring and collecting data of the emission level of carbon dioxide since 1958. The rapid rise of atmospheric carbon dioxide continues in 2019, and the average for May peaked at 414.7 parts per million at NOAA's Mauna Loa Atmospheric Baseline

The objective of our project is to develop a model that would predict the monthly average of CO_2 level from Mauna Loa Observatory in November, 2019.

2. Data

The data we used to develop our model is published by the National Oceanic and Atmospheric Administration Earth System and Research Laboratory Global Monitoring Division. The data contains the monthly emissions data from April, 1958 to October, 2019. We found that there are a few missing values in the data so we decided to use the interpolated data provided. Below is the time series plot for raw monthly CO_2 emission data from 1958-2019. (Figure 1)

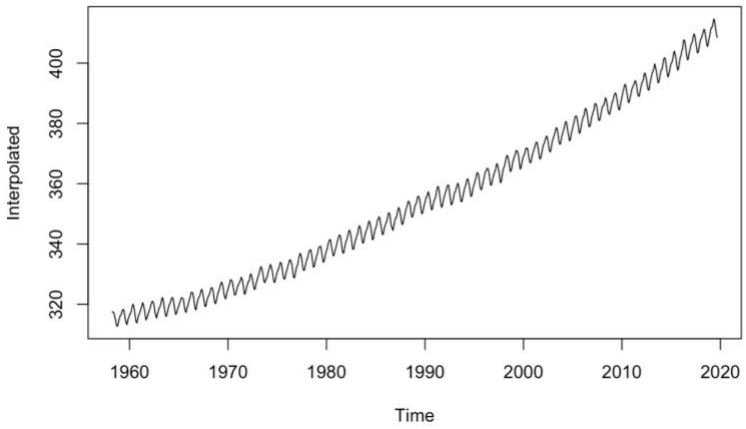

Figure 1: Time Series Plot on CO2 Emission 1958.4-2019.10

3. Features and preprocessing

As we have learned in class, in a classical decomposition model of a time series,

$$X_t = m_t + s + Y_t$$

$\{m_t\}$ stands for trend, $\{s_t)$ stands for seasonality, and $\{Y_t\}$ stands for residuals. Figure 2 shows the estimated trend, seasonal, and residuals of the monthly carbon dioxide concentration on Mauna Loa.

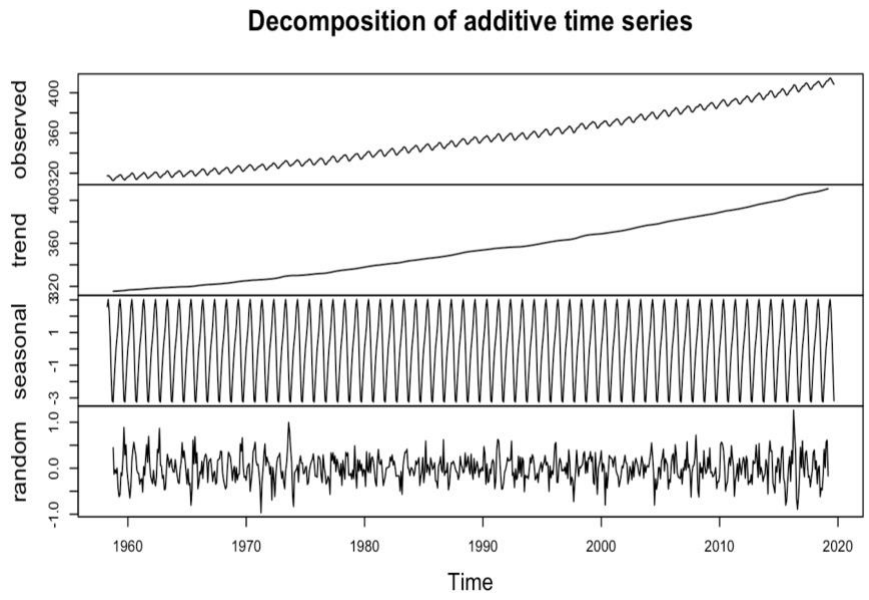

Figure2: Decomposition of additive time series of CO2

From figure 2, we can clearly detect a trend and a seasonal pattern in our CO_2 data. This provides us with incentives to choose different models. To make the prediction of CO_2 emission, we tried four different models, which are **ARIMA**, **Holt's Exponential Smoothing Model**, **LOESS Smoothing Model**, a**nd TBATS Model**. From the statistic test of residuals and the predicted data for recent months, we analyzed the differences among these models and finally chose the **ARIMA model** to prediction the CO_2 emission in November, 2019.

Before we experimented with the four different models, we used differencing to remove the trend and seasonality in the data. We used the dif() function and did a first-differencing to remove the linear trend. Then, we used lag=12 to remove the seasonality since we are dealing with a 12 months period. We ended up having a time series that appears to be random errors without any obvious trend or seasonal components. But it turned out that the four models we chose to experiment with have their own algorithms that will do the differencing job on its own, so this step is not used later in our models.

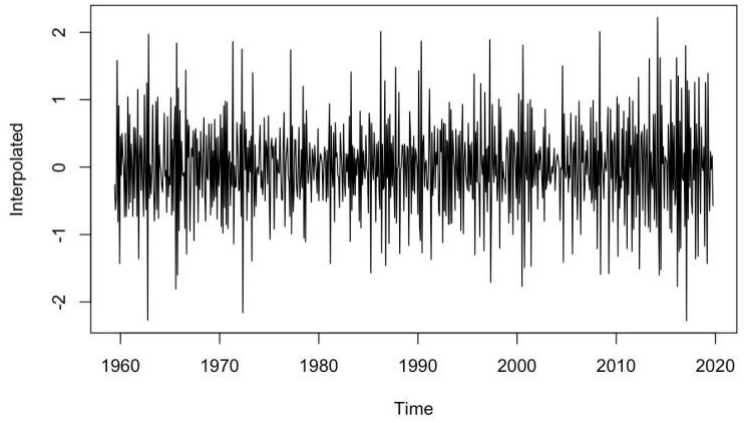

Figure 3: Time series of the lag-12 difference of the differenced atmospheric CO2

4. Prediction Models

ARIMA Model

ARIMA represents AutoRegressive Integrated MovingAverage model, with notation ARIMA(p,d,q)×(P,D,Q). The I in between AR and MA refers to a "integrated" I series. ARIMA is a stochastic time series model that we can use to forecast future time points. ARIMA models rely on past values to do regression, and it can obtain complex relationships since it takes error terms and observations of lagged terms. The AR property of ARIMA is referred to as P, degree of differencing is referred to as D, and MA is referred to as Q. ARIMA models have degrees of differencing that can eliminate seasonality. We used Akaike's information criterion, AIC, to select the appropriate model. Our criteria is to select the model of the (p,q,P,Q) that produces the smallest AIC. Below is the equation for ARIMA models.

$$X_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Auto.arima function in the forecast package helps us to identify the best fit ARIMA model. It uses a variation of the Hyndman-Khandakar algorithm, which combines unit root tests, minimization of the AICc and MLE to obtain an ARIMA model. This automatic function gave us the model ARIMA(0,1,2)(2,1,2)[12] (shown on the left below). We also manually tried more than ten sets of (p,q,P,Q) and chose ARIMA(1,1,1)(0,1,1)[12] with the smallest AIC as our final model(shown on the right below).

```
> d.arima <- auto.arima(ts) #auto ARIMA model (0,1,2)(2,1,2)
> d.arima
Series: ts
ARIMA(0,1,2)(2,1,2)[12]

Coefficients:
          ma1      ma2     sar1     sar2     sma1     sma2
      -0.3751  -0.0679  -0.3579  -0.0338  -0.5028  -0.3027
s.e.   0.0373   0.0364   0.6401   0.0440   0.6396   0.5532

sigma^2 estimated as 0.0991:  log likelihood=-188.89
AIC=391.77   AICc=391.93   BIC=423.86
```

```
> model0<-Arima(ts,order=c(1,1,1),seasonal=list(order=c(0,1,1),period=12))
> model0
Series: ts
ARIMA(1,1,1)(0,1,1)[12]

Coefficients:
         ar1      ma1     sma1
      0.2023  -0.5678  -0.8629
s.e.  0.0907   0.0763   0.0189

sigma^2 estimated as 0.09865:  log likelihood=-188.75
AIC=385.51   AICc=385.56   BIC=403.84
```

Figure 4: AIC results from auto.arima model and arima modal

Then we tested residuals of the model and below are the results:

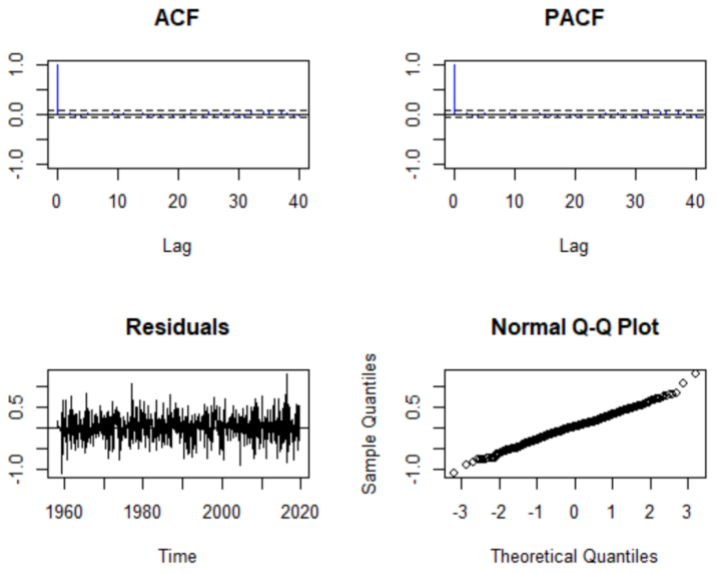

Figure 5: Analysis of Residuals using ACF, PACF, and Normal QQ Plot

We used Ljung-Box test, Turning Point test, and Rank test to test the null hypothesis that there is independence in the time series. The p-values for the three tests are all greater than 0.05, so we conclude that IID hypothesis is reasonable. Finally, we predicted the CO_2 amount in August, September, and October in 2019 using the best fit ARIMA model and we got 409.8063, 408.3671 and 408.6742.

Holt-Winters Exponential Smoothing Model

The model of time series is used to make predictions and to forecast future observations. When a time series has both seasonal and trend components, we can use Holt-Winters exponential smoothing to make short-term forecasts. This model is built to adjust for seasonality in the data and it uses additive seasonal adjustments.

The additive seasonal model is shown here:

$$y_{y+1} = L_t + T_t + S_{t+1-m} + e_{t+1}$$
$$F_{t+k|t} = L_t + kT_t + S_{t+k-m}$$
$$L_t = L_{t-1} + T_{t-1} + \alpha(y_t - L_{t-1} - T_{t-1} - S_{t-M})$$
$$L_t = L_{t-1} + T_{t-1} + \alpha(e_t)$$
$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$
$$S_t = \gamma(y_t - L_t) + (1-\gamma)S_{t-M}$$

First, we estimated the additive trend and seasonality and it turned out that the estimated values of α, β, and γ for the CO_2 data are 0.5512, 0.0087, and 0.1022. The α here means that the estimate of CO2 concentrations is based on recent observations and also some observations from the past, β means that the estimate of the slope of the trend stays almost the same throughout the process of time series, and the γ here means that the estimates of the seasonality component are based off observations from the past instead of the more recent observations. Then, we used the estimated data to fit the original time series data. The graph shows below.

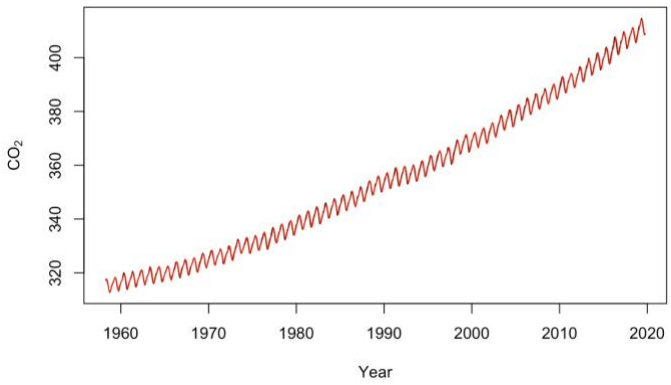

Figure 6: Fitted graph of the estimated time series (red) and original time series(black) of CO_2

To prove the feasibility of our HoltWinters model, we ran several tests. The null hypothesis is that there is an independent relationship between the residuals and they are identically distributed in the time series. The p-values of McLeod-Li Q, Diff signs S, Ljung-Box Q and Rank P tests were all greater than 0.05, so they failed to reject the null hypothesis at level alpha = 0.05 and we concluded that the residuals were independent to each other. Moreover, the Holt-Winters Exponential Smoothing model ACF (blue) is 1 at lag=0 and 0 for lag> 1. The sample ACF is 1 at lag=0 and insignificant for lag> 1. IID noise should be a reasonable assumption here.

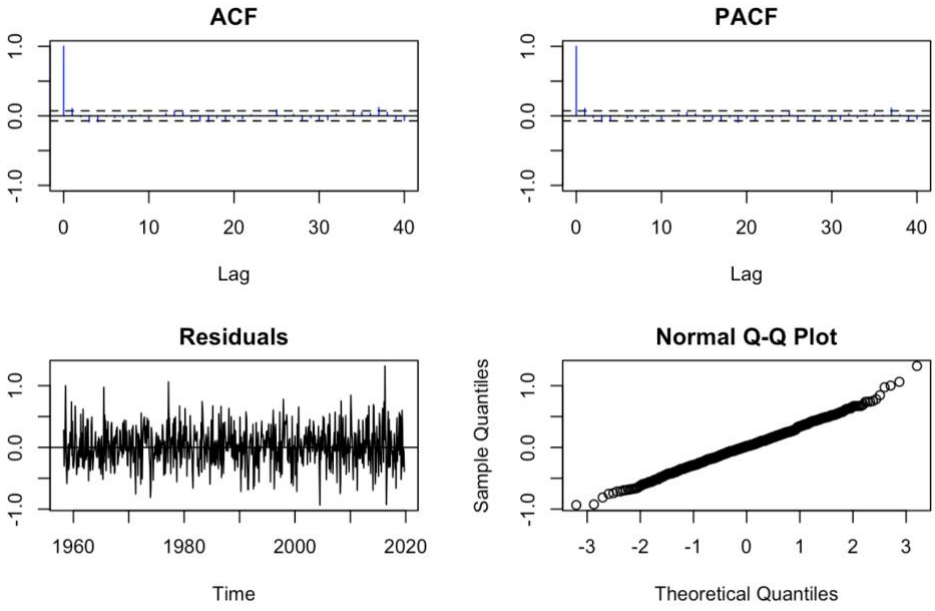

Figure7: Analysis of Residuals using ACF, PACF, and Normal QQ Plot

In order to test the validity of this model, we used the data from April, 1958 to July, 2019 to make predictions for the CO_2 data in August, September, and October 2019. The forecast results are 409.6060, 408.1678, 408.4154, which are all close to the real CO_2data.

LOESS Smoothing Model

LOESS (locally weighted smoothing) is a tool that creates a smooth line through a timeslot to help see the relationship between variables and foresee trends. It can be applied using the loess() on a numerical vector to smoothen it and to predict the Y locally. The weight function is shown below:

$$w(x) = \begin{cases} (1 - |x|^3)^3 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \geq 1 \end{cases}$$

The most important step will be to choose a smoothing parameter, s, between 0 and 1, which represents the proportion of observations to use for local regression. And we have chosen to apply 10%, 25%, and 50% smoothing span to the model. And it turns out that at 42% smoothing span, our prediction is closest to the actual number. The predicted November value is 412.1585. As we did backtesting for October, September, and October, we got 411.9437, 411.7291, and 411.5149. The predicted value is way off from the actual one. Besides, even though Loess has many advantages, such as ease of use and flexibility, it did not use cross-validation to select a span and sometimes requires some guesswork. Thus, we decided not to proceed with Loess Smoothing.

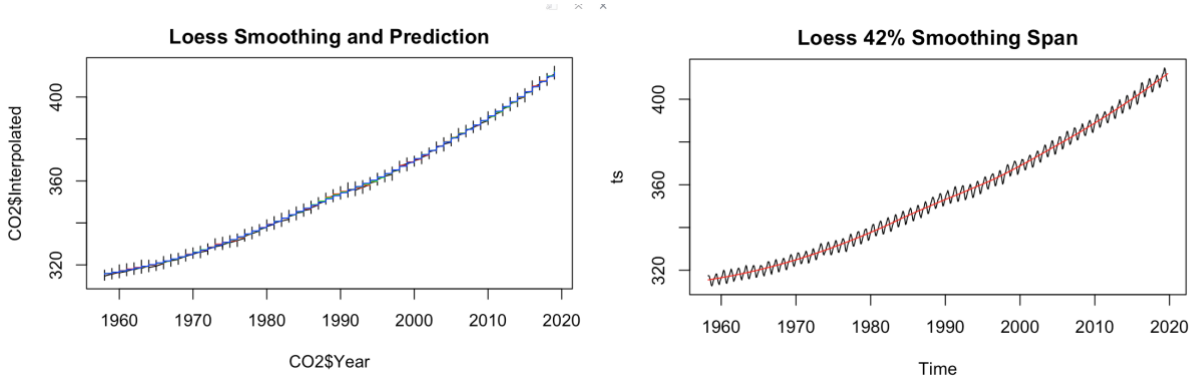

Figure 8: CO_2 level with different smoothing span in the given time span

TBATS Model

The last model we used was a TBATS model. This model is designed to deal with multiple cyclic patterns (e.g. daily, weekly and yearly patterns) in a single time series. It should be able to detect complex patterns in our CO_2time series.

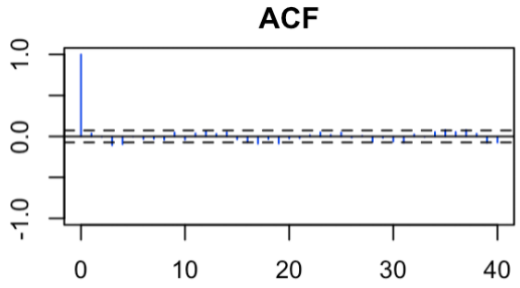

Figure 9: Analysis of Residuals using ACF

The TBATS model ACF (blue) is 1 at lag=0 and 0 for lag> 1. The sample ACF is 1 at lag=0 and insignificant for lag> 1. So IID noise should be a reasonable assumption here.

In the hypothesis test (test function in R), our null hypothesis is that residuals are iid noise. Since the p-value of McLeod-Li test, 0.6934, and the p-value of the rank test, 0.8967, were both larger than the 5% significance level, we concluded that there is enough evidence to say that the forecast error terms are independent.

The backtesting data for August, September, and October are 406.61, 404.80, and 408.53 respectively. The recently released data for October, 2019 is 408.53 while our model predicts 405.67. Due to the difference between the predicted and actual data, we chose to move forward and use it with other models for better predictions.

5. Results & Conclusion

After analyzing the time series data with 4 models, we found that each model passed the significance tests and seemed to be accurate. Considering the fact that we could not compare AICs of different models, we decided to compare four models' prediction values of the past three months with the actual value to verify which model did a better job in terms of predicting the CO_2 level.

	August	September	October	3-MonthTotal Difference
Arima Model	409.8063	408.3671	408.6742	-0.16
Holt-Winters Exponential Smoothing	409.61	408.17	408.42	-0.82
LOESS Smoothing Model	411.94	411.73	411.51	8.16
TBATS	406.61	404.80	405.66	-9.95
Real	409.95	408.54	408.53	-

In the table above, we noticed that ARIMA Model predicted more accurately among all four models. Therefore, we decided to use the value from Arima Model to predict the Monthly Average CO_2 Level in November, 2019.

ARIMA Model Prediction:

	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
Nov 2019	410.2287	409.8268	410.6307	409.6140	410.8435
Dec 2019	411.6557	411.1797	412.1317	410.9278	412.3836
Jan 2020	412.9342	412.4076	413.4608	412.1288	413.7396

	Point Forecast <dbl>	Lo 99 <dbl>	Hi 99 <dbl>
Nov 2019	410.2287	409.4208	411.0366
Dec 2019	411.6557	410.6990	412.6123
Jan 2020	412.9342	411.8758	413.9927

Figure 10: Forecasting results by ARIMA Model

Therefore, our predicted Monthly Average CO_2 Level in November, 2019 is **410.23.**

At 95% prediction bound: (**409.61, 410.84**)

At 99% prediction bound: (**409.42, 411.04**)

7. References

"Forecasting Time Series With R." Dataiku, https://www.dataiku.com/learn/guide/code/r/time_series.html

"Forecasting: Principles and Practice." 8.7 ARIMA Modelling in R, https://otexts.com/fpp2/arima-r.html.

Holmes, E. E., et al. "Applied Time Series Analysis for Fisheries and Environmental Sciences." 4.3 Differencing to Remove a Trend or Seasonal Effects, 28 Mar. 2019, https://nwfsc-timeseries.github.io/atsa-labs/sec-tslab-differencing-to-remove-a-trend-or-seasonal-effects.html?nsukey=8Fk/VZh9AWqUowQPiVWGZxIx1ic0/iKS3TfXtl0C2ur8HumfaZRixUh+prctATZz944En3Z6oQEMDFio9a3Xmnd5oXAsH8lXSpnN7ds/XMdMzNLJEOpdYYZ422po/Pgbd42tjKXKyBPt93IoIGEjT6BSktPR5cvJKG1qQdivf49Cj/Y/yYvJKHofGw6kYbLDJPoFTvvZfhCUc+VI0aQrZw==

Krzaczek , Lauren A, and Philip A Yates. "A Statistical Analysis of Atmospheric CO2 Levels at Mauna Loa." Ball State Undergraduate Mathematics Exchange, 2017, pp. 14–27.

Malik, Farhad. "Understanding Auto Regressive Moving Average Model - ARIMA." Medium, FinTechExplained, 2 Oct. 2018, https://medium.com/fintechexplained/understanding-auto-regressive-model-arima-4bd463b7a1bb.

Shih, Shou Hsing, and Chris P Tsokos. "Prediction Models for Carbon Dioxide Emissions and the Atmosphere ." The International Journal Neutral, vol. 16, 2008.

Stein, Theo. "Carbon Dioxide Levels Hit Record Peak in May." Welcome to NOAA Research, Welcome to NOAA Research, 3 June 2019,

https://research.noaa.gov/article/ArtMID/587/ArticleID/2461/Carbon-dioxide-levels-hit-record-peak-in-May

"Trends and Seasonalities¶." Trends and Seasonalities - Econ/Fin250a: Forecasting In Finance and Economics, http://people.brandeis.edu/~blebaron/classes/fin250a/filtering/trendseason.html.

Vink, Ritchie. "Ritchie Vink." Algorithm Breakdown: AR, MA and ARIMA Models, https://www.ritchievink.com/blog/2018/09/26/algorithm-breakdown-ar-ma-and-arima-models/.

Malik, Farhad. "Understanding Auto Regressive Moving Average Model - ARIMA." *Medium*, FinTechExplained, 2 Oct. 2018, https://medium.com/fintechexplained/understanding-auto-regressive-model-arima-4bd463b7a1bb.