SROR556 Group Project: Prophet

Zining Han, Meiyi Long, Yimeng Hua, Fengyi Chen, and Shi-Ting Lu

Introduction

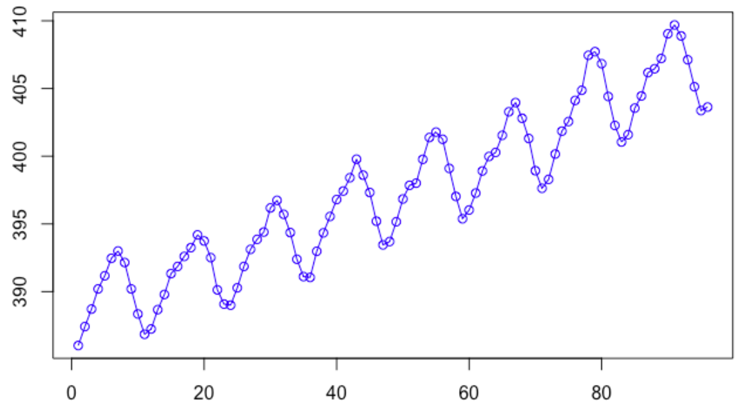
The American scientist Charles D. Keeling monitored carbon dioxide at Mauna Loa Observatory to first alert the world to the possibility of anthropogenic contribution to the "greenhouse effect" and global warming. “Greenhouse effect” has become a hot topic. Our group tried to predict the monthly average of CO_2 levels from Mauna Loa Observatory in Nov. 2019 based on historical data. We used the dataset of monthly mean carbon dioxide from Nov. 2009 to Oct. 2019 measured at Mauna Loa Observatory, Hawaii. The carbon dioxide data measured as the mole fraction in dry air with unit parts per million.

Heuristic of Approach and Assumptions of Model

There are conspicuous seasonality and trend observed in the plot of data. Therefore, the method of classical decomposition is considered as a good choice. The model for the time series is assumed as $X_t = m_t + s_t + Y_t$. Where $\{m_t\}$ is the trend, $\{s_t\}$ is seasonal component and with known period d (i.e., $s_{t+d} = s_t$). Then, estimate the seasonal pattern and then form deseasonalized data: $d_t=X_t-s_t$. Next, use deseasonalized data to get an estimate of trend $\{m_t\}$. Then, remove trend and seasonality of the X_t based on estimation and form R_t. Next, analyze the residuals of: $R_t = X_t - m_t - s_t$. If the residuals are close to be iid, accept the estimated trend and seasonal pattern. Next, fit the X_t with models and do residuals analysis. The models with the iid residuals and lowest AICC standard will be selected for prediction. To check the accuracy of our data, we decided to use the method of cross validation, dividing the available data into 80% train set and

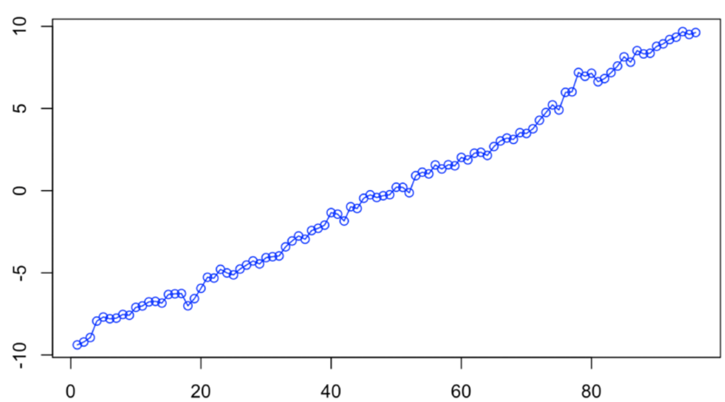
20% test set, using the model obtained from training set to predict testing set and get the residuals to test our model's accuracy.

Description and Estimation for the model



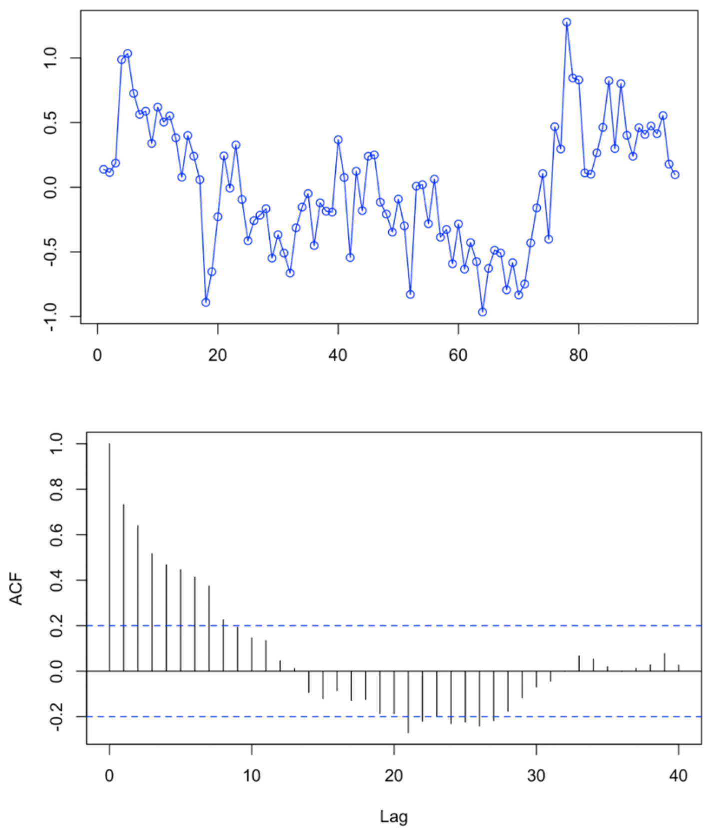
(Figure 1)

At first, we plotted the training set of the original time series (Figure 1), we can see clearly that there exist a trend component and a seasonal component, so our goal is to use classical elimination to remove the trend and seasonality component, and then fit residuals into a model to ideally obtain an iid noise. First, we counted there are roughly twelve data entries in each period, so we believe the period is twelve months (one year). Then, we obtain the deseasonalized data which is the original series with the estimated seasonal component removed.



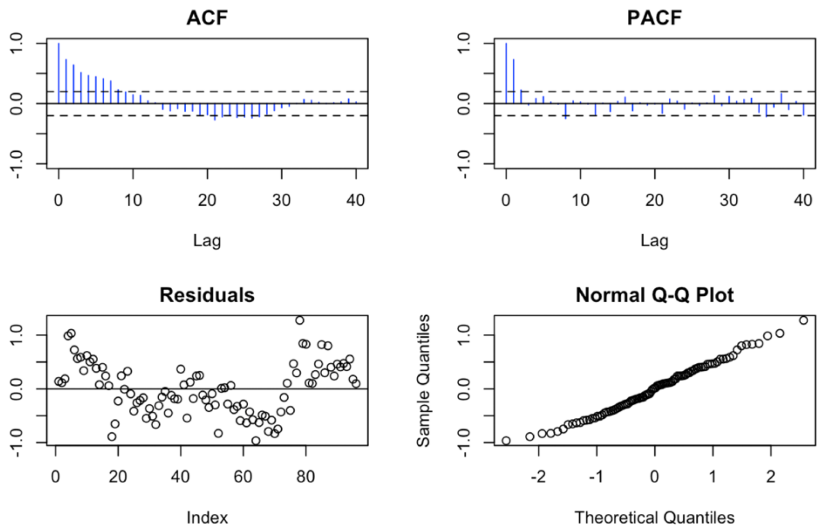
(Figure 2)

The graph of the deseasonalized data (Figure 2) suggests the presence of an additional linear trend function. In order to fit such a trend to the deseasonalized data, we removed the linear trend and obtained residuals called R_t.



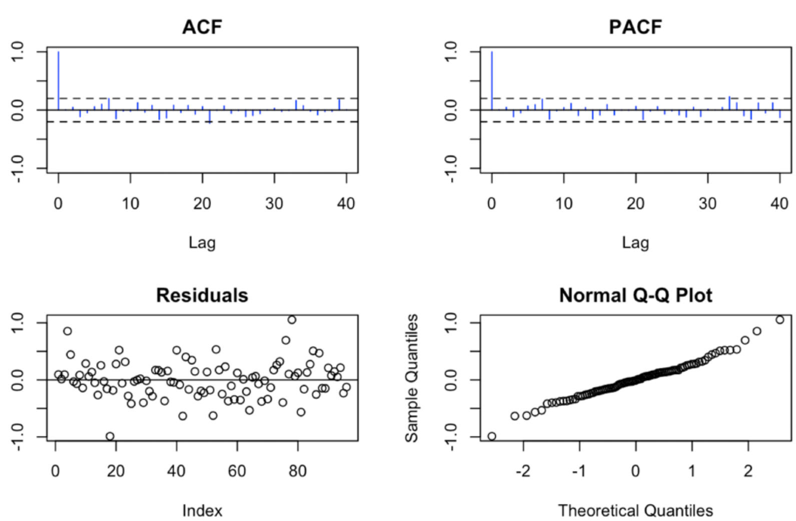
(Figure 3)

Yet based on the ACF plot (Figure 3), the residuals are not iid so we decided to further fit R_t into a model to finally obtain the expected iid noise. Having tried the ARMA(0,0) model, we can clearly see an AR(1) signature in PACF plot (Figure 4), so we decided to try fitting R_t into AR(1) and ARMA(1,1) model.



(Figure 4)

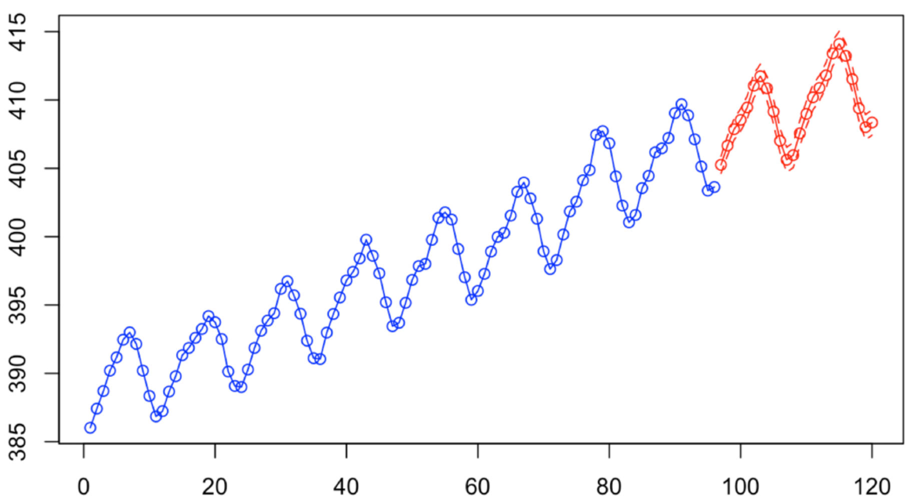
Fitting both AR(1) and ARMA(1,1) generated good ACF and PACF plot suggesting iid noise, they also pass all the tests of randomness. However, ARMA(1,1) has an AICC (59.88862) that's smaller than AR(1)'s AICC (62.42509) which indicates that ARMA(1,1) is a better model for R_t. For the final residuals obtained by subtracting the predicted value based on ARMA(1,1) model from R_t, seen from the output (Figure 5), the null hypothesis which suggests that our final residuals are iid noise is not rejected by all tests since the p-values are all greater than 0.05, including Ljung-Box test, McLeod-Li test, Turning points test, Different sign test, and Rank test.



```
Null hypothesis: Residuals are iid noise.
Test                        Distribution Statistic   p-value
Ljung-Box Q                Q ~ chisq(20)     20.21    0.4448
McLeod-Li Q                Q ~ chisq(20)     13.32    0.8631
Turning points T   (T-62.7)/4.1 ~ N(0,1)        65    0.5685
Diff signs S       (S-47.5)/2.8 ~ N(0,1)        45    0.3792
Rank P            (P-2280)/158 ~ N(0,1)      2266    0.9294
```

(Figure 5)

We intended to use cross validation for the model we selected to predict the latest 20% of the data (i.e., data from November 2017 to October 2019). The reason why we included cross validation is that we want to test our model with the existing dataset to make sure that the predicted values are feasible. Using the forecast function in R with the first 80% of the data (i.e., data from November 2009 to October 2017) to predict the next 24 months average of CO_2 (Figure 6).



(Figure 6)

By subtracting the original values of the test dataset from our predicted values for test dataset, we got the residuals. As shown in the following output, the one sample t-test has a p-value of 0.1652, which is greater than 0.05. Therefore, we do not reject the null hypothesis that the true mean is equal to 0. Consequently, the cross-validation step affirms that our model selection is a good fit.

```
        One Sample t-test

data:  resid_final
t = -1.4335, df = 23, p-value = 0.1652
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.28909571  0.05243445
sample estimates:
 mean of x
-0.1183306
```

Forecast

After selecting our model for the data, we used the forecast function in R with the whole dataset to predict the monthly average of CO_2 in November. The predicted number we have is 410.23 (more precisely, 410.2343) parts per million. The 95% confidence interval is from 409.61 to 410.8587 while the 99% confidence interval is from 409.4138 to 411.0549. These numbers seem quite reasonable given the newly released daily CO_2 in the first few days of November.

Step	Prediction	sqrt(MSE)	Lower Bound	Upper Bound
1	410.2343	0.3185491	409.4138	411.0549
Step	Prediction	sqrt(MSE)	Lower Bound	Upper Bound
1	410.2343	0.3185491	409.61	410.8587

Scientific Justification

The data shows an obvious trend pattern for the CO_2 level at Mauna Loa. The average annual CO_2 level continues to increase in the past few decades. This pattern is consistent with the pattern of the average global CO_2 level. According to the Intergovernmental Panel on Climate Change (IPCC), studies have shown that "Anthropogenic greenhouse gas emissions have increased since the pre-industrial era, driven largely by economic and population growth", which leads to atmospheric concentrations of carbon dioxide and the effects are "extremely likely to have been the dominant cause of the observed warming since the mid-20th century". Many researchers also support the theory that human actions like industrial production are the major factor for the increase of CO_2 level (Zhang and Mi. 2016, Baum et al. 2012, etc.).

There is an obvious seasonality pattern for the CO_2 level at Mauna Loa. The level is usually at its maximum in May and at its minimum around September or October. Researchers have tried to link the pattern with plant growth (Cleveland et al. 1983). One possible explanation is that in spring, the maxima were linked to an increase in rainfall during spring, contributing to vegetation growth at the site, thus increasing respiration rates at night-time. When respiration is larger than photosynthetic, plants consume oxygen and eliminate carbon dioxide. When summer arrives, the rapid growth of plants somewhat causes more intense photosynthetic activity and make the CO_2 level reaches its minimum around September (Fernández-Duque et al, 2019). Furthermore, a study based on the Atlantic Ocean points out that the influence of fossil fuel consumption is greater during the cold months, which may lead to the increase of CO_2 level during winter and cold spring. By contrast, clean air masses approaching the monitoring station are more frequent in summer, which may lead to a decrease of CO_2 level from May to September (García et al. 2016).

Reference

Baum, Seth D., Jacob D. Haqq-Misra, and Chris Karmosky. "Climate change: Evidence of human causes and arguments for emissions reduction." Science and engineering ethics 18.2 (2012): 393-410.

Cleveland, William S., Anne E. Freeny, and T. E. Graedel. "The seasonal component of atmospheric CO2: Information from new approaches to the decomposition of seasonal time series." Journal of Geophysical Research: Oceans 88.C15 (1983): 10934-10946.

Fernández-Duque, Beatriz, et al. "Annual and seasonal cycles of CO 2 and CH 4 in a Mediterranean Spanish environment using different kernel functions." Stochastic Environmental Research and Risk Assessment 33.3 (2019): 915-930.

García, M. Ángeles, et al. "Influence of atmospheric stability and transport on CH4 concentrations in northern Spain." Science of the Total Environment 550 (2016): 157-166.

Pachauri, Rajendra K., et al. Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change. Ipcc, 2014

Zhang, Tiantian, and Hong Mi. "Factors influencing CO2 emissions: a framework of two-level LMDI decomposition method." Revista Ibérica de Sistemas e Tecnologias de Informação E10 (2016): 261.