FINAL PROJECT

Jin Kweon and Clover Jiyoon Jeong

17/04/2018

Goal

First goal is to find the right model for our dataset, and pitch that Hawaii is under threat by huge amounts of Carbon Dioxide.

Abstract & Motivation & Introduction

- Explain the motivation for studying the particular dataset of interest. What do we care about the dataset?

Hawaii is a beautiful place/island (one of the places we would really like/wish to visit). And, we hope to find out the time when people are mostly travel to Hawaii, so we could plan out our trip more efficiently. However, we found that carbon dioxide emission rates in Hawaii has been significantly increasing, and we thought that it would be great if we could back up that argument with time series technique/skills and knoweldge to tell people how serious the problems are.

Outline

Here is our rough draft for our project:

Intro

1. Make descriptions of the data: what it is and where it came from
2. Describe what questions we are addressing
3. Perform EDA

Analysis (Body)

4. Plot data

- Is there a trend and seasonality?
- Are there any cyclical patterns?
- Are there any outliers? (tso and tsoutliers)
- Chasing stationarity? (detrend - regression, spline, smoothing/filtering, differencing, boxcox, adf.test -> pvalue is less than 0.05 implies stationarity)

5. Scatterplot matrix
6. Split into train and test sets
7. plot ACF, PACF, EACF And, check sarima$ttable p-value to see how significant they are. (significant meaning 0.05, then its closer to be the right one to use)
8. Build and estimate parameters with Yule-Walker (sarima)
9. Model diagnostics (residual test, qqnorm, qqline, shapiro.test, turning.point.test,)
10. Compare models with AIC, AICC, BIC, etc
11. Forecasts (sarima.for and get MSE)
12. Spectral Analysis if seasonal trend is not obvious and to get frequency

- Identify key frequencies and cycles (analyze the seasonal behaviors)
- Regression terms cosine and sines
- Spectral densities and Periodogram (mvspec, periodogram, spec.pgram - with and without log, spec.ar, arma.spec)
- For periodogram, try different window sizes, kernels, somoothing parameters, tapers, etc and get the best looking ones
- Smoothings
- Confidence interval of spectral density

- Check peaks are significant

Data

- Why is the dataset interesting? The purpose of the analysis?

As you might already find out, pollusion is getting worse and worse. As we are really interested in environmental science, we hope to pitch out that one of the most popular visiting places, Hawaii is under threat by seious pollusion. After doing analysis, we hope to contribute for saving the Earth.

- Elaborate on the background of the problem/dataset. Where does the dataset come from? How are the data collected?

First, the data sets is collected through January 1959 to December 1990 by monthly. The dataset is collected by Rachel Passmore (Royal Society Teacher Fellow, Department of Statistics, University of Auckland), and they collected the data by conducting tests of CO2 (with the unit of parts per million - ppm).

Later, we might need to divide into training and testing sets (leave 12 data out) for forecasting. We decided to leave 12 data out, because a year has 12 months, and it is more reasonable for us to forecast the entire year.

Preprocessing and Exploratory Data Analysis

This EDA was performed before any analysis, to learn about the data. Our team spent a lot of time on this part.

1. Find missing value/NA

The data had no missing values. We used the function called "anyNA" to see whether the is any NA/missing values.

2. Units (check whether it is equally spaced), observation period/duration

The data is equally spaced and the data is collected monthly, from January 1959 till December 1990.

3. Handling outliers/weird data

For these kind of continuous variables, we did not know whether the so called "outlying datas" were bad (wrongly inputed) or not. So, our logical thoughts were really important, and we relied on it, here. We looked at boxplots and density plots to help us make decisions whether we needed to remove outlying observations or not.

First of all, we wanted to say that we printed out the lists of outlying datas we got from usual mathematical way (we usually say outliers were the observatoins above or below Q1/Q3 $\pm$ IQR * 1.5); however again, since we did not know the distribution assumption, we were going to approach this problem with our logics.

By plotting our time series plot, boxplots, and density plots, we decided that there is no outlying observation. (no additive outlier or level shift)

Also, just to be extra safe, we used "tso" and "tsoutliers" functions, and we could conclude that there is no outlying data. (did similarly as chapter 11.1 - 11.2 from Cryer and Chan did)

5. Do str and summary to find out data structures and fix if necessary

When we read the structure and summary, nothing looked abnormal, so we did not take action (or modifying data in other way). The minimum of difference is 313.2, median is 330.6, mean is 332.2, sample standard deviaion is 11.76, and maxis 356.9.

Methodology & Data Analysis

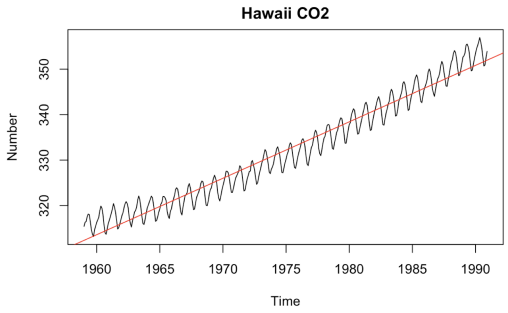

Figure 1: Original plot for Hawaii CO2

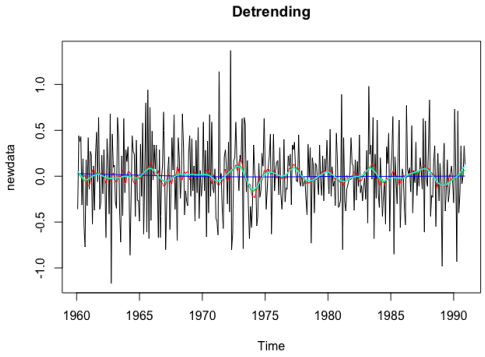

Figure 2: More detrendings added

We plotted our original data first, **(Fig 1)** and in general, and we could easily find out that there is a very consistent and obvious seasonal pattern with polynomial trend. (Our data does not seem follow a linear model, but more of a polynomial model as the line does not pass through the middle of the data well)

1. Check seasonality or trend

When we investigate the data, there is a obvious pattern each year that CO2 is increasing at the beginnin of the year upto May, decreasing from May to September/October, and later increasing again at end of the year. **(Fig 2)** So, we end up saying that there will be an overall increasing trend and seasonal effect. Also, it became more obvious when we performed Dickey-Fuller Test, as the p-value of the test was 0.3964.

To remove the seasonal trend, we take seasonal difference of differencing lag of 12, and then, it looked pretty stationary (and Dickey-Fuller test also back this up); however, we found that as we differenced more, we could make the model more stationary. So, we ran the for-loop for differencing lag of 1 to 11, and found that when diffencing an additional 1st difference operation makes the model the most stationary. (but, remember that since the data is trasnsformed, some of the years were cut-off/removed)

We also tried to use the function "decompose" (both additive and multiplicative models) and "stl" to see whether we could determine the trend using a moving average and lowess. Here, since the seaosonal variation is relatively constant, it is better for us to use additive model. From these two, we could easily found the trend is increasing.

Last but not least, we tried to plot a scatterplot matrix to back-up what we did above. As it can be see, every sample autocorrelation is really high and we found really strong positive linear relationships at every lag. **(Fig 3)** And, after we transformed the data (chased the staionary), we could finally conclude that our transformation was the right decision, as we have seen less autocorrelation and lowess fit looks much less linear. **(Fig 4)**

2. Check stationarity (if not, determine a transformation that makes the series look stationary)

After removing seasonality and trend, it looks staiontary. (Doing extra detrending techniques such as kernel smoothing, lowess, and smoothing splines can even remove white noises, which is not really good for ARIMA model, since ARIMA model works on stochastic datasets. So, we are going to just continue ARIMA analaysis after differencing only) **(Fig 2)**

3. ARIMA

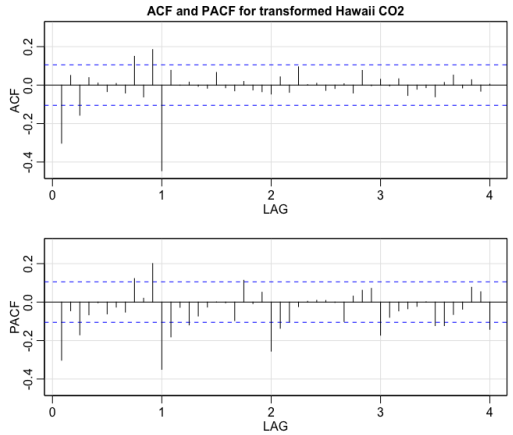

Figure 3: ACF and PACF for transformed Hawaii CO2

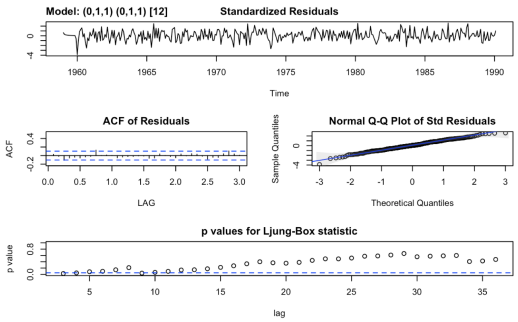

Figure 4: Third model

Before we actually built the model, we decided to split the data into two sets. We would leave 12 data out. (from original dataset, not from the dataset differencing applied since they removed some data due to differencing) We do not want to want to use entired data set to build a model, since we do not want to re-use them when we forecase, as it would be biased and might cause over-fitting.

Our model is ARIMA(p, 1, q) $\times$ $(P, 1, Q)_{12}$, as we made seasonal and 1^{st} differencings before. After, we identified the model by using ACF, PACF, and EACF built from the training set. And, we came up with a few options: **(Fig 3)**

1) Seasonal Component: At the season, it seems like the ACF is cutting off a lag 1s (s = 12), while PACF is tailing off at lags 1s, 2s, 3s, ... From there, we could come up with a SMA(1) in the season where s = 12.

2) Non-Seasonal Component: For the nonseasonals, we could inspect ACF and PACF at the lower lags. Both of them are tailing off, and we could first try an p = q = 2. And, since they are not really obvious, we would rather perform model diagnostics and use AIC, BIC, AICc to find the best model fit. (which seems pretty reasonable as this is what it seems like many people and the textbook used)

For paramter estimtions, model significance, model comparions, and model diagnostics will be computed using "sarima" function. We have compared 16 different models (please refer to the codes we have made), and came up with three good models that can fit well:

Model 1: sarima(train, 2, 1, 1, 0, 1, 1, 12)

Model 2: sarima(train, 1, 1, 1, 0, 1, 1, 12)

Model 3: sarima(train, 0, 1, 1, 0, 1, 1, 12) **(Fig 4)**

There is (are) one or two outliers in standardized residual plots, and residuals do not have trend. All of ACF of residuals are within the dotted line for every lag. (check ACF individually) Normality assumption for standardized residuals seem reasonable. (just a bit off at the tails) Ljung-box statistic shows the model is adequeate as there is no correlation between residuals (check ACF groupwise) So, we could say the residuals seems like they are white noise.

After model comparisons, we came up with three good models: but everyone has its own strength - for example, one has the lowest AIC, AICC, BIC but have higher p-value for model coefficients (variable is not significant), the other one has the lowest p-value (variable is the most significant) but have higher AIC, AICC, BIC and ljung-box looks little bit worse.

So, we decided to perform forecasting with these three models, and would go for the one that has the lowest MSE. (choose the best model in terms of forecasting)

4. Forecast

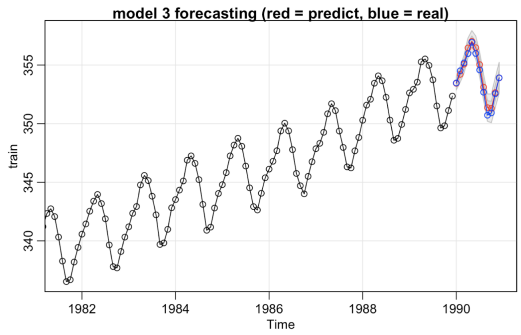

Figure 5: Third model forecasting

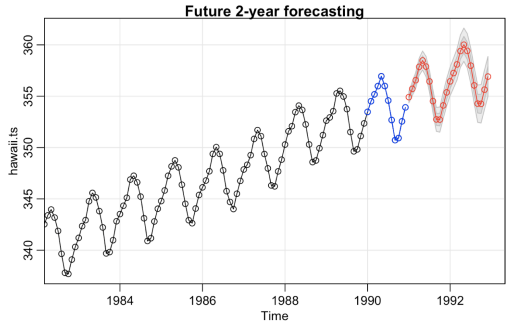

Figure 6: Forecast 2-year with prediction interval

Now, it is time for us to talk about forecasting. As we have mentioned earlier, we leave out the last 12 observations, and built models from there. By looking at the three models' forecastings, they all look **reasonable** as the actual 12 values are within confidence interval. (grey area) However, we would compare MSE for these three models, to decide which model would be the best. (it seems pretty reasonable although this will not guarantee the chosen model would have the lowest MSE again) We made two tables to compare actual and predicted values: one is for summary of predictions and standard errors for the three models, and the other one is for summary of MSE and mean MSE for the three models. **(Fig 7)** And, we found that the model 3, ARIMA(0, 1, 1) $\times$ $(0, 1, 1)_{12}$ **(Fig 5)** has the lowest mean MSE. Last but not least, we made next 2-year forecasts with model 3, and as we already expected, CO2 level would keep go up. **(Fig 6)**

5. Spectral Analysis

Detrending

We will use the dataset without the last 12 observation in order to select the best model for spectral analysis as we did in ARMA modeling. However, unlike we used differencing to detrend the data in ARMA analysis, we used STL decomposition (Seasonal Decomposition of Time Series by Loess) in spectral analysis. We found that when we used the detrended data by differencing to pick the best frequencies and use them to generate features in cosine and sine terms, the generated data(red dotted line) did not grasp clear periodic trends in the original data. It rather looked like a linear line than periodic graph. **(Fig 7)** Englarged image is in **appendix (Fig)**

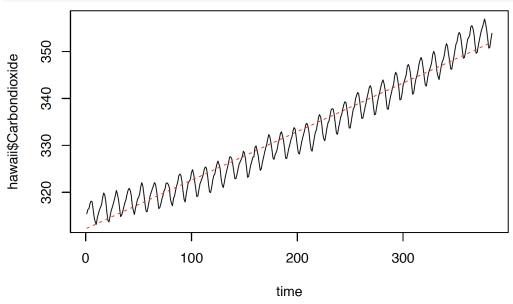

Figure 7: Generated features by using detrended data by differencing

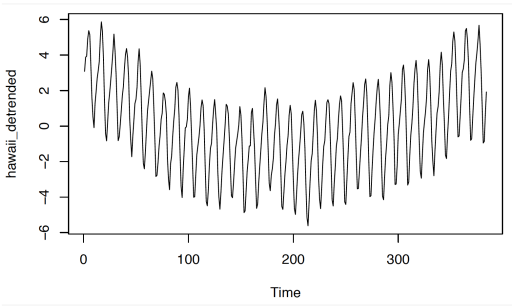

Figure 8: Detrended data by lm function - first order

The distance between each peak is not fairly a constant **(Fig 2)** and we can conclude that differencing adds more noises to the original dataset and gave us inefficient detrended data which is not adequate for spectral analysis. Therefore, we tried to detrend it using lm function to remove the linear trend in the dataset because it looked like it had a linearly increasing trend when we glanced our eyes over it.

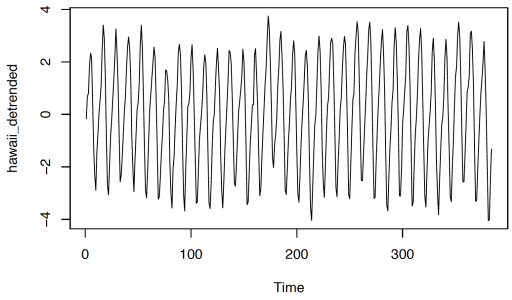

Figure 9: Detrended data by lm function - second order

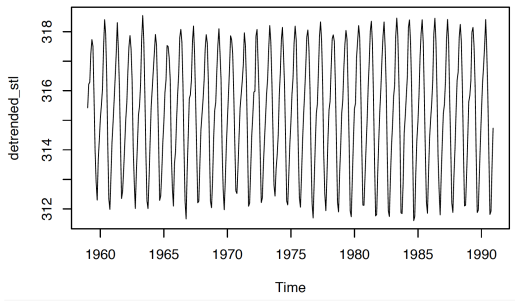

Figure 10: Detrended data by STL decomposition

However, surprisingly, the detrended data from lm looked like a parabola which means it is obviously not a stationary series.**(Fig 8)** Therefore, we tried lm function with the second order polynomial, and the resulted periodogram looked like a stationary time series.**(Fig 9)** However, we want to explore more method for detrending and it led us to use stl function in R, and it worked successfully on our dataset. Basically we used STL decomposition to estimate the trend, then subtracted it from the original dataset. **(Fig 10)**

The fundamental problem of the periodogram is that the periodogram doesn't get more reliable as we collect more data unlike the other estimates such as mean or a coefficient of regressions. We are just adding more noise to periodogram as we collect more data. Therefore, we assume that there exists some underlying curve of spectral values which is called spectral density function or power spectrum. In order to smooth the estimates of spectral density function, we calculated moving averages of the periodogram by giving high weight to close frequencies and low weight to distant frequencies. The shape of the weights can be triangle, rectangle, or bell shape and it determines the type of the kernel functions. The smoothing method of spectral density function is a nonparametric method because it doesn't assume any parametric model for the underlying time series process.

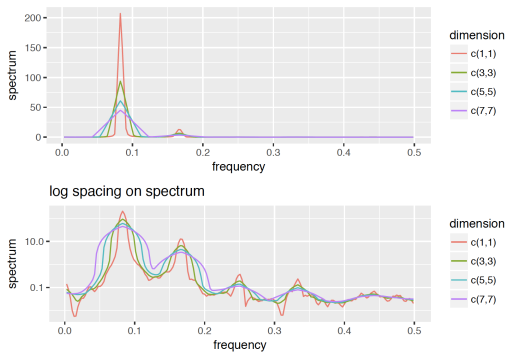

Figure 11: Periodograms by different span of the kernel

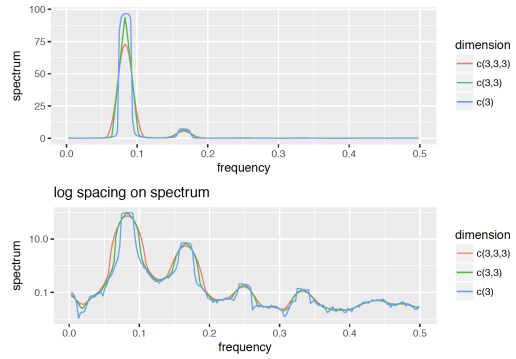

Figure 12: Periodograms by different span(dimension) of the kernel

As we expected from the periodic trends in the raw data, the raw periodogram shows clear peak frequencies. **appendix Fig** However, we still want to see other peak frequencies which might help give us more information on the periodic trends so we implied log spacing on spectrum and examined it. By looking at the raw periodogram and periodograms by different span of the kernel **Fig** , we could see that spans = c(3, 3) gives the best estimates since it did not do over smoothing nor missed peaks. We could also examine that smoothing with larger span was inaccurate since it combined the small peaks into one big flat peak even though the actual spectrum had multiple peaks. Also, oversmoothing makes the difference between peaks smaller as we can see in the periodogram of span c(5, 5) and span c(7, 7) above.

Also, we set up the different dimension of the span in kernel and checked that the smoothed periodograms are *not* fairly sensitive to dimension of the span in kernel because there was not much difference between different smoothed periodograms. **Fig 23**. The general shape of the periodogram did not change significantly, and we concluded that more than two dimension of span is unnecessary.

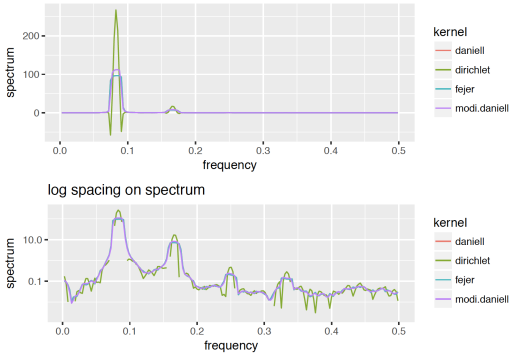

Figure 13: Periodograms by different type of the kernel

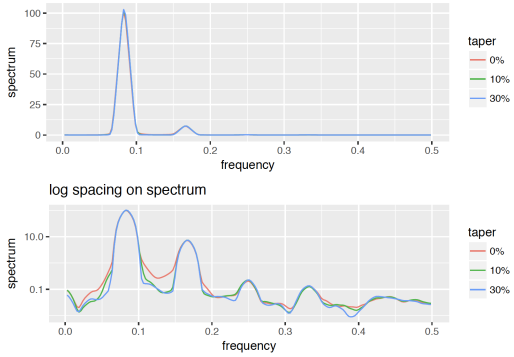

Figure 14: Periodograms by different tapering

Next, we were curious about the effect of different type of kernel in smoothing so we set up all parameters the same but only changed the type of the kernel and plotted each smoothed periodograms. As we could see in **Figure 24** above, there was not much difference between daniell and modified daniell kernels, and dirichlet and fejer kernels did not smoothe periodogram very well because the former one produced unnecessary peaks and the latter flatten original peaks excessively. Therefore, we chose modified daniell kernel as our final type of the kernel for smoothing.

Lastly, we checked the effect of different tapering. When we estimate a periodogram, we implicitly assume that our time series is circular. However, there will be a jump where the end meets the start again if we wrap the time series around. Resulted jump is unreasonable but will effect itself through all the frequencies and contaminate them. Therefore, one solution can be downweighting the beginning and end of the data and it will give more weight to the middle, and less weight to the ends when we calculate the periodogram. Even though there is still the jump at the end, it will not have great influence on periodogram due to the very little weight. This downweighting method is called tapering.

We set up all parameters the same but only changed the portion of the tapering and plotted each smoothed periodograms like we did in the previous graph. As we could see in **Figure 25**, we can barely detect the difference between three different tapered smoothed periodograms. Therefore, we concluded that our smoothed periodogram is not sensitive to different tapering setup. We found that 10% of tapering gave us enough smoothing since it eliminated insignificant small bumps in the log spaced periodogram. By the result of previous observations, we decided our final periodogram smoothing model as modified daniell with span c(3,3) and 10% tapering.

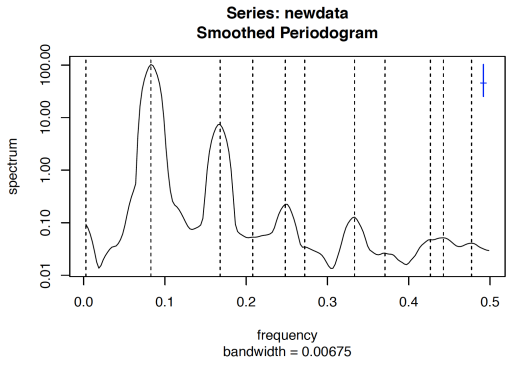

Figure 15: (log) Periodograms with peaks of dotted lines

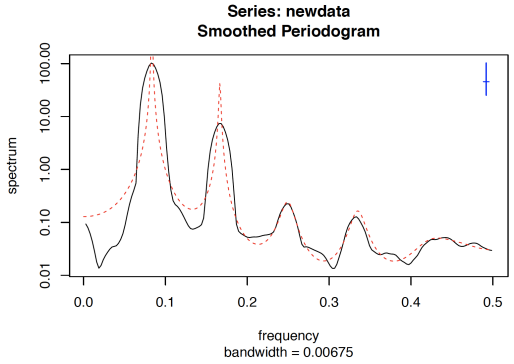

Figure 16: Smoothed Periodograms VS parametric spectral estimator

The dotted lines in Figure 26 above indicates the peak frequencies of periodograms except the first one. We picked two significant frequencies because they were the most obvious ones in the raw periodogram and had the great difference between the other peaks. The rounded best two frequencies from our data were 0.083 and 0.168 which are equal to approximately 12 months($1/0.083 = 12.04819$) and 6 months($1/0.168 = 5.952381$).

Figure 27 is comparing smoothed periodogram with the parametric spectral estimator which estimates spectral density of a time series from AR fit which is the red dotted line. We could easily see that the AR estimation fits fairly well with the logged raw periodogram.

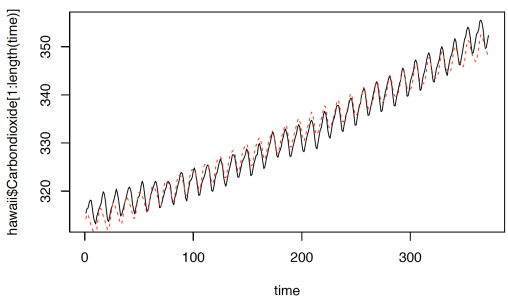

Figure 17: Data generated by peak frequencies

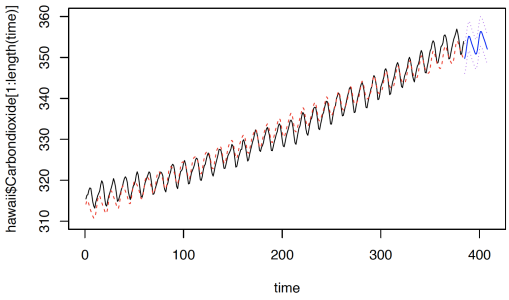

Figure 18: Forecast 2-year with prediction interval

As we previously mentioned, we used the two frequencies to generate features in terms of sine and cosine functions. The red dotted line is Figure 28 generated by sine and cosine terms, and it approximates the general periodic trend in raw data well enough. Therefore, we used this model to forecast 12 observations and calculated MSE by comparing them with the original data points which we have left as the test dataset in the beginning. The MSE was 11.25757, and we predicted 24 future observations and prediction interval for each month in Figure 29.

Result/Key Findings

ARIMA: ARIMA(0, 1, 1) $\times$ $(0, 1, 1)_{12}$ **(Fig 5)** is the best model (lowest mean MSE with the leave 12 out forecasting) we could come up with.

Spectral Analysis:

Conclusion

- What are the key takeway messages from the project?

As we noticed when we were detrending the raw data, the data looked like it had a linearly increasing trend but it turned out to have a quadratic increasing trend which is quite unexpected. It reminded us that we should not merely assume that the data itself is a linear because it seems following a linear trend pretty well.

In spectral analysis, the significant general periodic trends are 1 year and 6 months which indicates that the data is likely to be related to the actual seasons. We can see that the local maximums(peaks) are generally in May and June and the local minimums are in October and November, so it corresponds to the frequencies that we had found. Therefore, we can conclude that the CO_2 level gets higher when the temperature gets higher and vice versa.

We might need different types of techniques for ARIMA and spectral analysis to make them stationary.

- Highlight the most interesting findings from your data analysis.

In spectral analysis, we were quite surprised that the actual best frequencies are almost exactly 6 months and a year because it means that the CO_2 level fluctuates according to the exact time schedule. It also means that there exists a annual circular trends of CO_2 which is not related to the industrial by-product of mankind.

Moreover, despite the fact from the artical in NASA and other reports that the recent relentless rise in CO_2 shows a remarkably constant relationship with fossil-fuel burning, we found that the CO_2 level is not increasing in a linear trend but in a quadratic trend. It warns us that the CO_2 level was increasing slower in back in 1960's but more rapidly increasing in 1990's and possibly nowadays from our previous modeling results. These findings alert us the reason why technology companies need to develop eco-friendly products and natural sustainable energy in order to survive next century.

ARIMA does a better job on forecasting than spectral analysis, as the ARIMA MSE is much lower.

- What are some possible future analysis of the dataset?

Based on our models, we could expect to have way more CO_2 recently with the higher increase rate; however, people get acknowledged of the issue, and many government puts more regulations on CO_2 emmision rate. So, we expect to see some level-shifts (outlier) in recent year.

How it can be further developed

It would be much better if we could attain more recent data, then we might be able to better persuade how serious the pollutions are. Because we know that we might not be able to make good long-term forecastings with ARIMA models and more data lower the variance, so we hope to collect more data.

Appendix

EDA and ARIMA

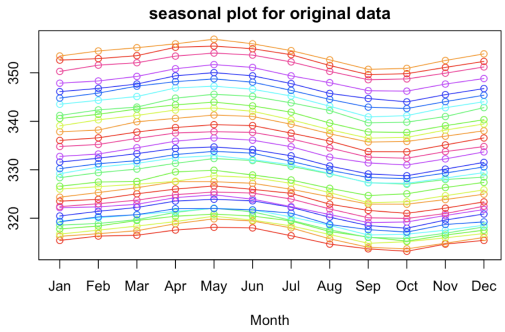

Figure 19: Seasonal plot

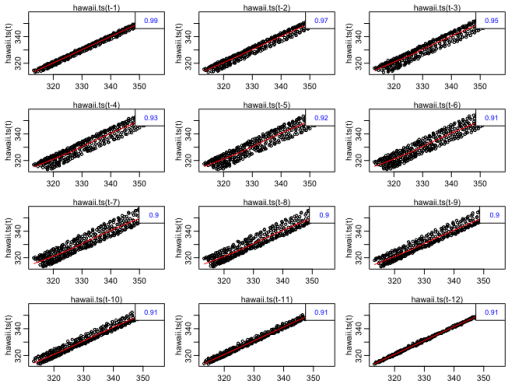

Figure 20: Scatterplot matrix before transforming

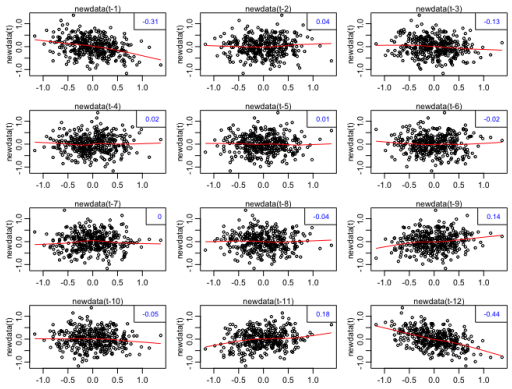

Figure 21: Scatterplot matrix after transforming

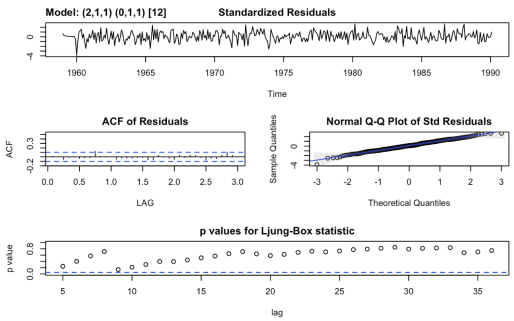

Figure 22: First model

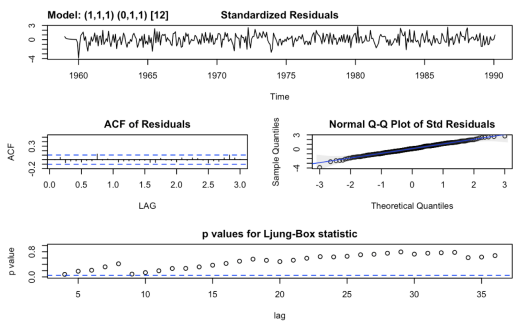

Figure 23: Second model

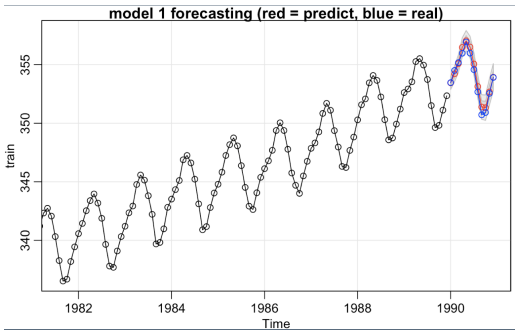

Figure 24: First model forecasting

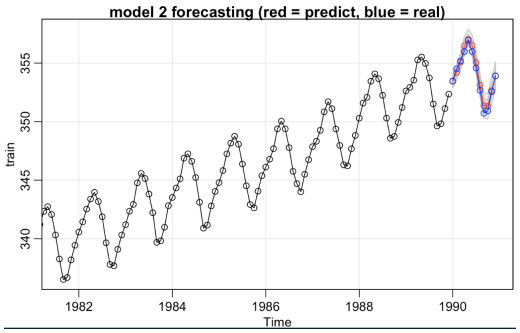

Figure 25: Second model forecasting

	model1_prediction	model1_sd	model2_prediction	model2_sd	model3_prediction	model3_sd
1	353.451	0.282	353.441	0.283	353.439	0.282
2	354.222	0.335	354.210	0.339	354.206	0.335
3	355.094	0.378	355.078	0.375	355.075	0.378
4	356.484	0.409	356.468	0.406	356.462	0.409
5	357.048	0.437	357.030	0.434	357.024	0.437
6	356.509	0.461	356.491	0.461	356.485	0.461
7	355.061	0.483	355.042	0.486	355.037	0.483
8	353.136	0.504	353.117	0.509	353.112	0.504
9	351.386	0.524	351.367	0.532	351.362	0.524
10	351.331	0.543	351.311	0.554	351.306	0.543
11	352.671	0.561	352.652	0.575	352.646	0.561
12	353.931	0.579	353.911	0.595	353.905	0.579

Figure 26: Prediction and SE for three models

	Model1	Model2	Model3
MSE: 1990, 1	0.0003610	0.0008410	0.0009610
MSE: 1990, 2	0.0829440	0.0900000	0.0924160
MSE: 1990, 3	0.0073960	0.0104040	0.0110250
MSE: 1990, 4	0.2540160	0.2381440	0.2323240
MSE: 1990, 5	0.0116640	0.0081000	0.0070560
MSE: 1990, 6	0.2693610	0.2510010	0.2450250
MSE: 1990, 7	0.2313610	0.2134440	0.2088490
MSE: 1990, 8	0.2079360	0.1909690	0.1866240
MSE: 1990, 9	0.4435560	0.4186090	0.4121640
MSE: 1990, 10	0.1689210	0.1528810	0.1489960
MSE: 1990, 11	0.0146410	0.0104040	0.0092160
MSE: 1990, 12	0.0004410	0.0000010	0.0000250
Mean MSE	0.1410498	0.1320665	0.1295567

Figure 27: MSE and Mean MSE for three models

Spectral Analysis

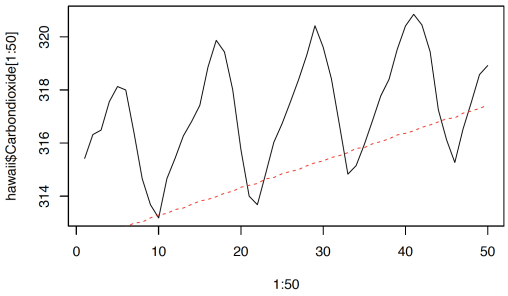

Figure 28: enlarged Figure 7 for better visualization

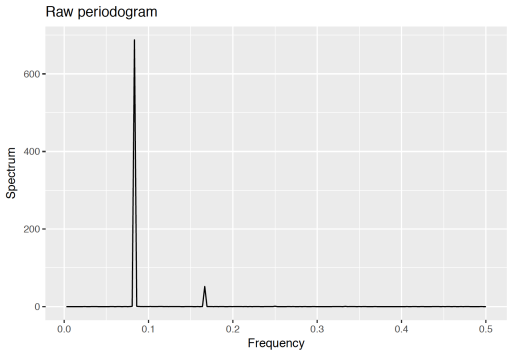

Figure 29: Raw periodogram

Reference/Citations

Shumway, Robert H., and David S. Stoffer. Time Series Analysis and Its Applications: with R Examples. Springer, 2017.

Chan, Jonathan D. CryerKung-Sik. "Time Series Analysis." With Applications in R | Jonathan D. Cryer | Springer, Springer-Verlag New York, www.springer.com/us/book/9780387759586.

Passmore, Rachel. "Data Sets." Time Series, timeseries.weebly.com/data-sets.html.

Passmore, Rachel. "15 Time Series Datasets (2012)." CensusAtSchool New Zealand, 14 July 2017, new.censusatschool.org.nz/resource/time-series-data-sets-2012/.

Garziano, Giorgio. "Outliers Detection and Intervention Analysis." DataScience+, 4 Dec. 2017, datascienceplus.com/outliers-detection-and-intervention-analysis/.