Analysis of San Francisco Mean Sea Level

Suoyi Yang

3/17/2020

1 Introduction

The rise in global sea level is currently a pressing environmental issue. It could have potentially devastating effects on coastal plant life as well as increase the intensity of floodings, storms, and damages to coastal areas [3]. The four main causes of sea level rise are ice melt, thermal expansion, land sinkage, and slowing Gulf Stream, most of which are also associated with global warming [5].

Rises in local sea levels are also of particular concern to the state of California. California has more coastal residents than any other state in the United States with over 25 million people living near the sea. In addition to the $100 billion worth of property along with the California coast, natural coastal ecosystems are also at risk. Coastal communities are already experiencing major flood events, and around two-thirds of the beaches in Southern California are projected to disappear. As a result, the state is planning to allot over $6 billion to find solutions to the challenges posed by sea level rise [3].

Similar to global sea level rise, the two main causes of the rise in sea level of California are likely to be ice melt and thermal expansion. Ice melt simply refers to the melting of glacial ice from land into the sea. Thermal expansion is the warming of the ocean, which causes water to expand as it warms, resulting in sea level rises. On a more local level, sea level rises and flooding in California can also be impacted by El Niño weather events. In addition to the increased rainfall which causes flooding, the El Niño also causes warmer temperatures in the Pacific Ocean, resulting in thermal expansion of the ocean onto the West Coast of the United States [5].

A new study from the U.S. Geological Survey predicted that of the projected $150 billion worth of property and about 600,000 coastal residents that could be flooded by the end of the century, about two-thirds of those at risk are in and around San Francisco [1]. As a result, this paper will specifically focus on looking at mean sea level data of San Francisco over a period of several decades.

2 Dataset

I obtained my dataset from the National Oceanic and Atmospheric Administration (NOAA) Tides and Currents database [4]. The data consisted of several different monthly measurements of sea levels (all of which measured in units of feet) in San Francisco from the period of January of 1950 to December of 2016. These monthly measurements include mean high water, mean sea level, mean tide level, and several more. In my analysis, I am only interested in looking at the mean sea level (MSL) data over this period of time. I've also converted the MSL data values from feet to centimeters in my analysis as the metric system is much more standard in science. There were only two missing data points for MSL in my dataset, which were August 2012 and September 2012.

I decided to impute these missing data points by setting their value to the average MSL of July 2012 and October 2012. I believe these two data points accounts for such a small percentage of my total data that my method of imputing these missing values will not significantly alter the results of my analysis.

I set aside two years of data (which are the 24 months of 2015 and 2016) for testing the prediction values that would later be generated from my fitted model. The remaining 780 months were used for training.

3 Analysis of Data

A plot of the MSL of San Francisco from January of 1950 to December of 2014 is shown in Figure 1. Just by looking at the plot, there seems to be an upward trend in the data. To confirm that there is indeed a statistically significant upward trend in the data, we can try to run a linear regression on the data. Figure 2 shows the MSL data being regressed on time, with the red line indicating the regression line. The results of the linear regression are shown in Table 1 (with coefficient estimates, the standard error in parenthesis, and an indication of the significance of p-value indicated by the number of asterisks while the actual p-values are not displayed). The results indicate that both the intercept and slope of the regression are significant, indicating there is a statistically significant upward trend in the data.

From some research, it was found that the highest tides in California occur during winter. This is likely due to the frequent winter storms that are experienced in California that push more water to the coast, increasing sea level [5]. This is combined with the effects of El Niño which heats up the ocean and causes thermal expansion and rise in sea level. So from this information, there is likely to be a seasonality component to the data. Looking at the data from Figure 1 as well as a decomposition of the data shown in Figure 3, there does seem to be evidence of seasonality. We can also take a look at the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) of the original data shown in Figure 4. In these graphs, each lag is a multiple of 1/12 (a lag of 1.0 is one year/12 months). We can see that the ACF plot indicates a very slow decay. However, the cyclical patterns in the plot add to our belief that there is monthly seasonality in the data.

4 Model Fitting

4.1 Fitting an ARIMA Model

After identifying that seasonality and trends exist in the data, it is important to try and remove these from the data in order to fit an ARIMA model. I detrended the data by taking the residual of the linear regression. Then I attempted to de-season the data by using first-difference. The data after taking these steps are shown in Figure 5.

A sample ACF and PACF plot of the detrended and de-seasoned data is shown in Figure 6. The sample ACF and PACF plots both tails off relatively quickly, but spikes in the graphs do seem to indicate that there could be some seasonal cycles left in the data despite the attempt to remove monthly cycles. The peaks that occur around lag 1 and 2 indicate that there could potentially be some annual cycles left in the data. After looking at the ACF and PACF, I tried to fit several seasonal ARIMA models to the data. I attempted to find the best fit model by looking for the lowest AIC. In the end, the ARIMA model with the lowest AIC was the seasonal ARIMA$(1,0,1) \times (2,0,1)_{12}$.

Seasonal ARIMA$(1,0,1) \times (2,0,1)_{12}$ has an AIC of 4565.88 with $\sigma^2 = 20.76$. The coefficients of this model are shown in Table 2. Both the seasonal and non-seasonal MA coefficients are negative. This could suggest that the MSL of the following month is negatively correlated to peaks in MSL in the previous months. A large MSL peak one month could result in smaller MSL in the following month. Meanwhile, both the seasonal and non-seasonal AR coefficients are positive, which could suggest that the MSL for the following month is positively correlated to the current MSL.

4.2 ARIMA Model Diagnostics

Now that we have managed to fit an ARIMA model to our detrended and de-seasoned data, we want to look at the diagnostics of the model using the parameters selected in the previous subsection to see if the model is a good fit for the data. The plots and results of the diagnostics are shown in Figure 7.

Looking at the top plot showing the standardized residuals of the model, we can see that they appear to be more or less white noise and randomly scattered around zero with mostly constant variance. The ACF plot of residuals is also consistently below the threshold for significance. In the QQ-plot, there does seem to be some deviations from normality around the larger theoretical quantiles. However, overall the QQ-plot looks to be more or less reasonably normal. Finally, the Ljung-Box plot indicates that most of the lags are above the significance threshold.

All this information from the ARIMA model residual diagnostics seems to indicate that the model is valid and is a fairly good fit for our dataset.

4.3 Forecasting

Now that we have our model and decided that it is a pretty decent fit, we can try to forecast the MSL of San Francisco from January 2015 to December 2016. The results of the forecasted data are shown in the graph in Figure 8.

The red line in the graph indicates the forecasted data, while the blue line indicates the actual data that I initially set aside as testing data. The region with the lighter shade of grey indicates the 80% confidence interval, while the region with the darker shade of grey is the 95% confidence interval. While the predicted and actual data points for the next 24 month are not exactly the same, they do seems to match fairly closely. Much of the actual values fell within the 95% confidence interval created by the forecasting. This suggests that the model that was fit might be a useful tool to forecast the monthly MSL of San Francisco, at least a year or two out.

5 Spectral Analysis

Now that we have found a good ARIMA model for the data, we can now take a look at spectral analysis of the detrended data with monthly seasonality removed.

A raw periodogram of the data is shown in Figure 9. We can see that there are several peaks in the graph, which can make identifying major/significant peaks difficult visually. It is found that there are strong peaks at frequencies of around 0.34375 and 0.59375. These correspond to about a 3-year cycle and 1.7-year cycle respectively. However, since there are so many peaks in the data, it is difficult to both determine the significance of the peaks and determine the predominant period in the data.

In order to try and remedy this issue, I smoothed the periodogram by fitting an AR model to the data. This is done by using an autoregressive spectral estimator using the minimum AIC model selection method. The AR model that been to best fit the data is an AR(26) model, the spectral density of which is shown in Figure 10. Indeed, we see there are a lot fewer peaks and the approximate frequency of the largest peak is easily identifiable even visually. The largest peak in the graph is at frequency 0.312, corresponding to a 3.19-year cycle. This predominant period seems to be pretty close to the results found in the raw periodogram.

The data indicating a predominant cycle of around 3.2 years could potentially coincide with El Niño cycles that we previously discussed as they occur on average every two to seven years.

6 Conclusion

We saw that the seasonal $\text{ARIMA}(1,0,1) \times (2,0,1)_{12}$ model that was fit for our data and did a decent job of predicting the MSL in San Francisco two years ahead. From a spectral analysis of the data, it was found that the predominant cycle of the detrended data with monthly seasonality removed seemed to be around 3.2 years.

While we did find a seemingly good model, there are still some ways to potentially improve the model even more. A possible solution could be to look at other potential covariates and datasets. Like previously mentioned, we could potentially find and use data on El Niño as there could potentially be some correlation between the frequency of El Niño storms and some yearly cycles found in the spectral analysis. We could also look at winter storm data to potentially improve the predictive abilities of the model since some monthly seasonalities in mean sea level of San Francisco could be attributed to winter storm patterns in California.

References

[1] Barnard, P.L., Erikson, L.H., Foxgrover, A.C. et al. Dynamic flood modeling essential to assess the coastal impacts of climate change. Sci Rep 9, 4309 (2019). https://doi.org/10.1038/s41598-019-40742-z

[2] Heberger, Matthew & Cooley, H. & Herrera, P. & Gleick, Peter & Moore, E.. (2009). The Impacts of Sea Level Rise on the California Coast.

[3] National Geographic Society. (2019, March 27). Sea Level Rise. Retrieved from https://www.nationalgeographic.org/encyclopedia/sea-level-rise/

[4] National Oceanic and Atmospheric Administration. (n.d.). Water Levels - NOAA Tides & Currents. Retrieved from https://tidesandcurrents.noaa.gov/waterlevels.html?id=9414290&units=standard&bdate=19500101&edate=20161231&timezone=GMT&datum=MSL&interval=m&action=data

[5] SeaLevelRise.org. (n.d.). California's Sea Level Has Risen Over 6" Since 1950. Retrieved from https://sealevelrise.org/states/california/

Figures

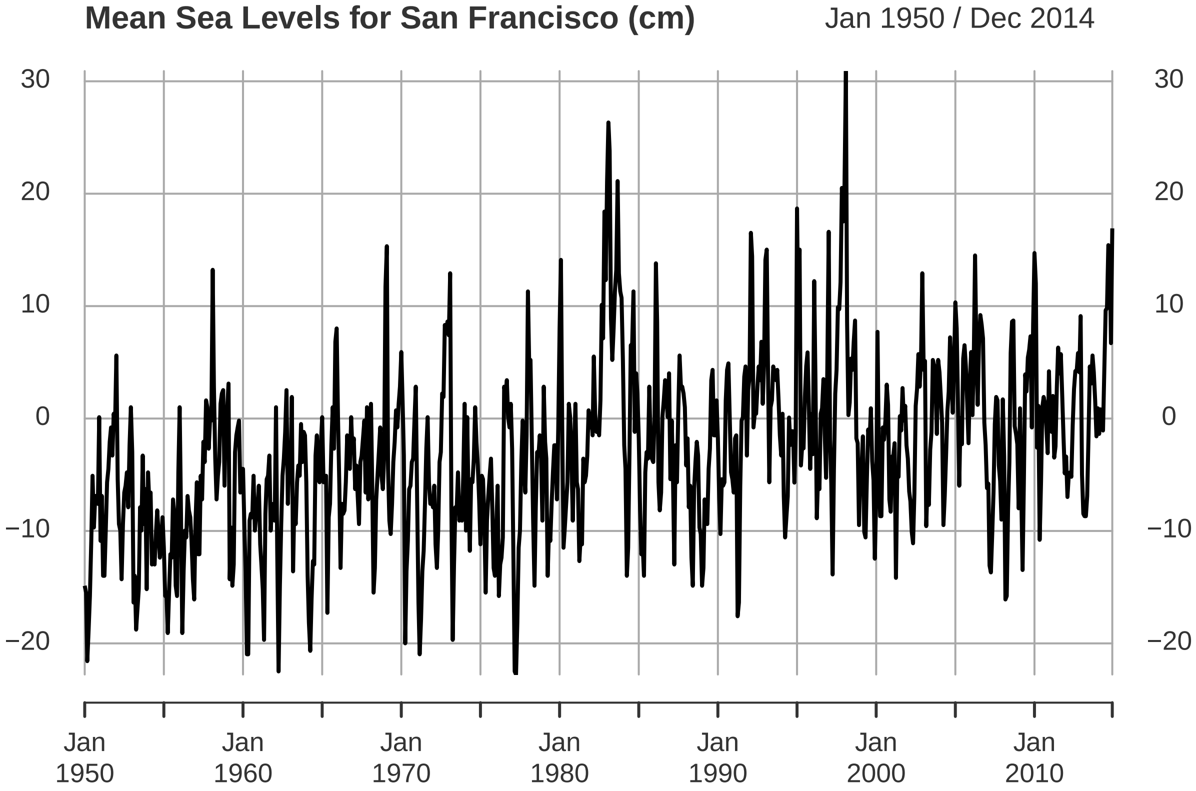

Figure 1: Plot of the MSL of San Francisco

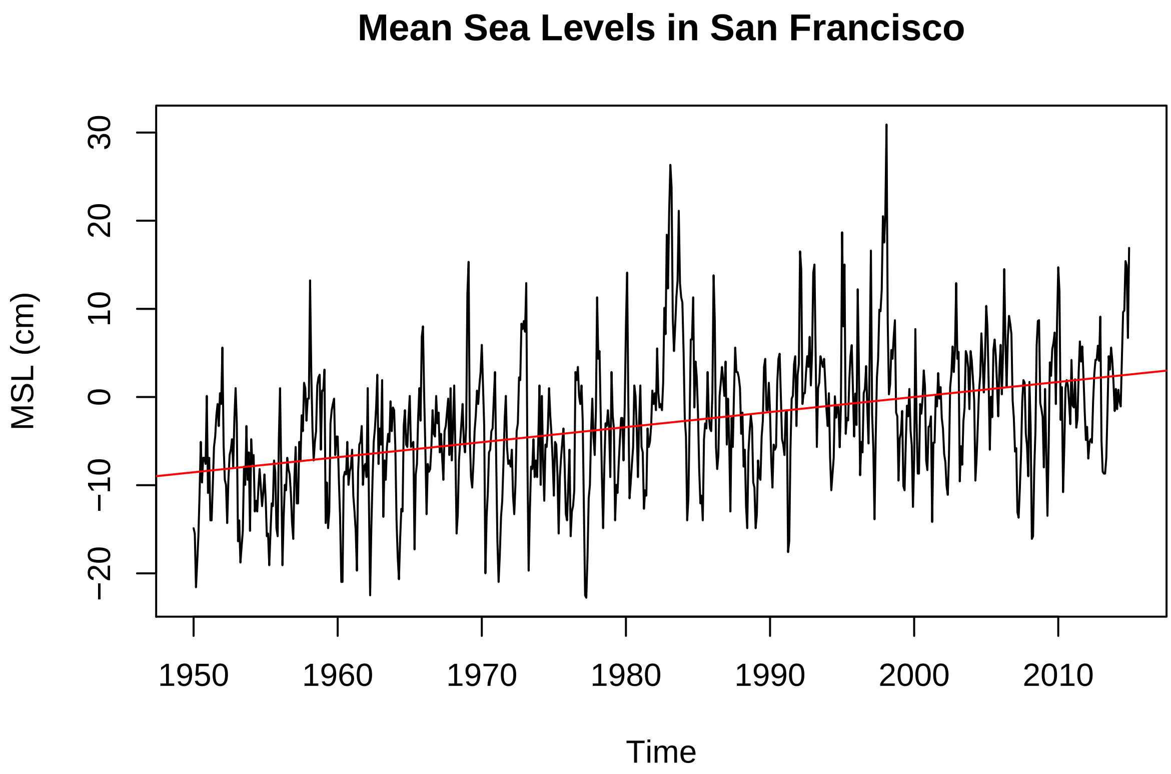

Figure 2: MSL data Regressed on Time

Table 1: Original Data with Regression Line

	Dependent variable:
	trainTs
time(trainTs)	0.171^{***}
	(0.013)
Constant	-341.431^{***}
	(26.757)
Observations	780
R^2	0.171
Adjusted R^2	0.170
Residual Std. Error	7.073 (df = 778)
F Statistic	159.998^{***} (df = 1; 778)
Note:	*p<0.1; **p<0.05; ***p<0.01

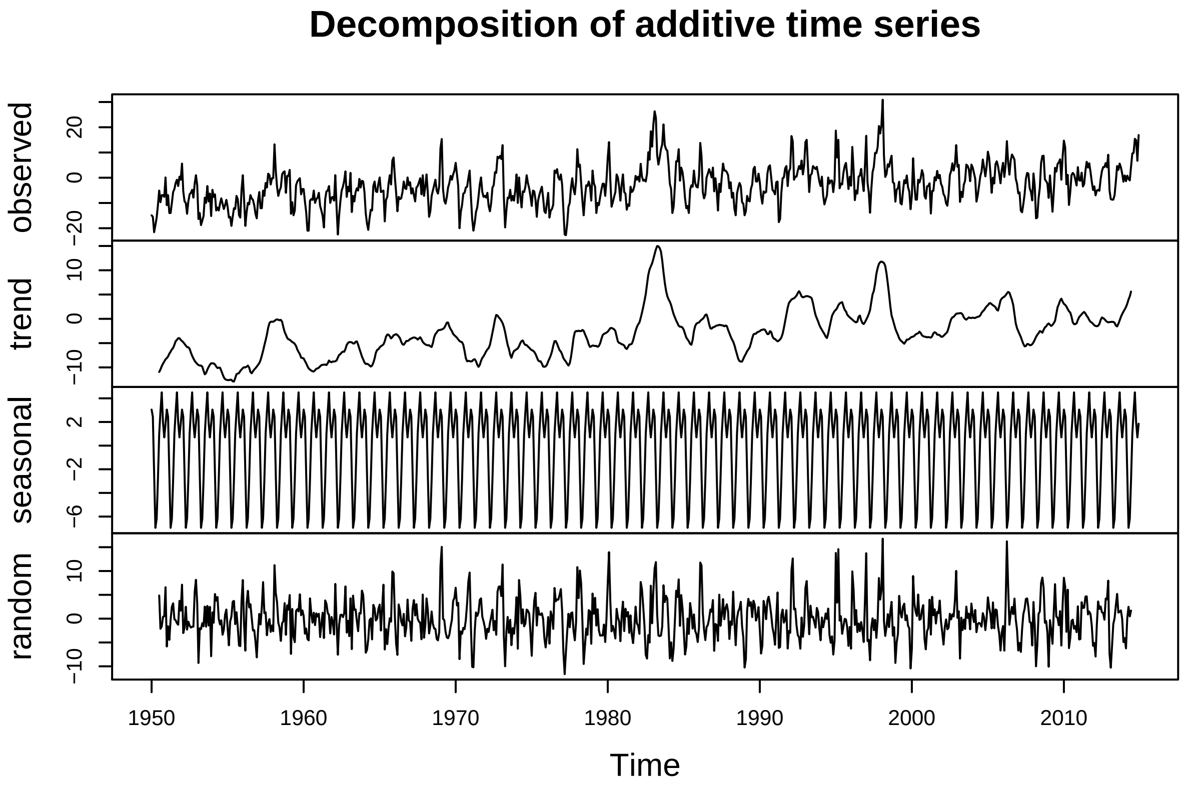

Figure 3: Decomposition of Dataset

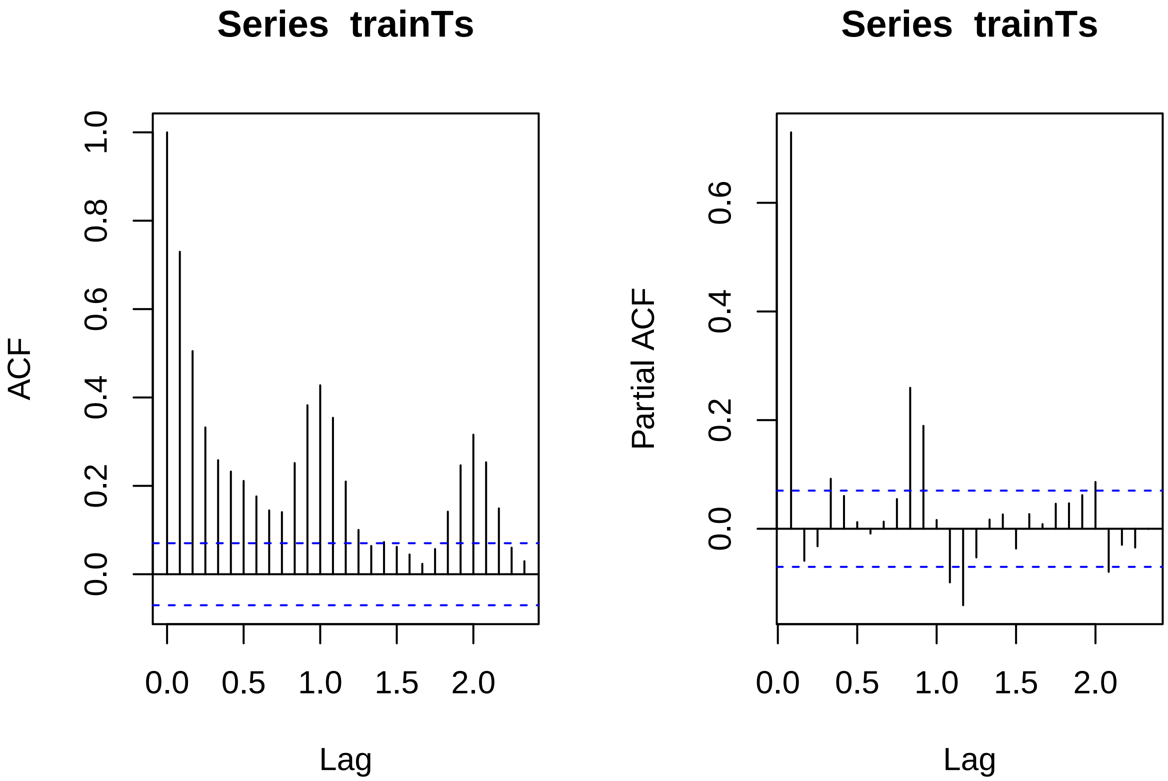

Figure 4: ACF (left) and PACF (right) of Dataset

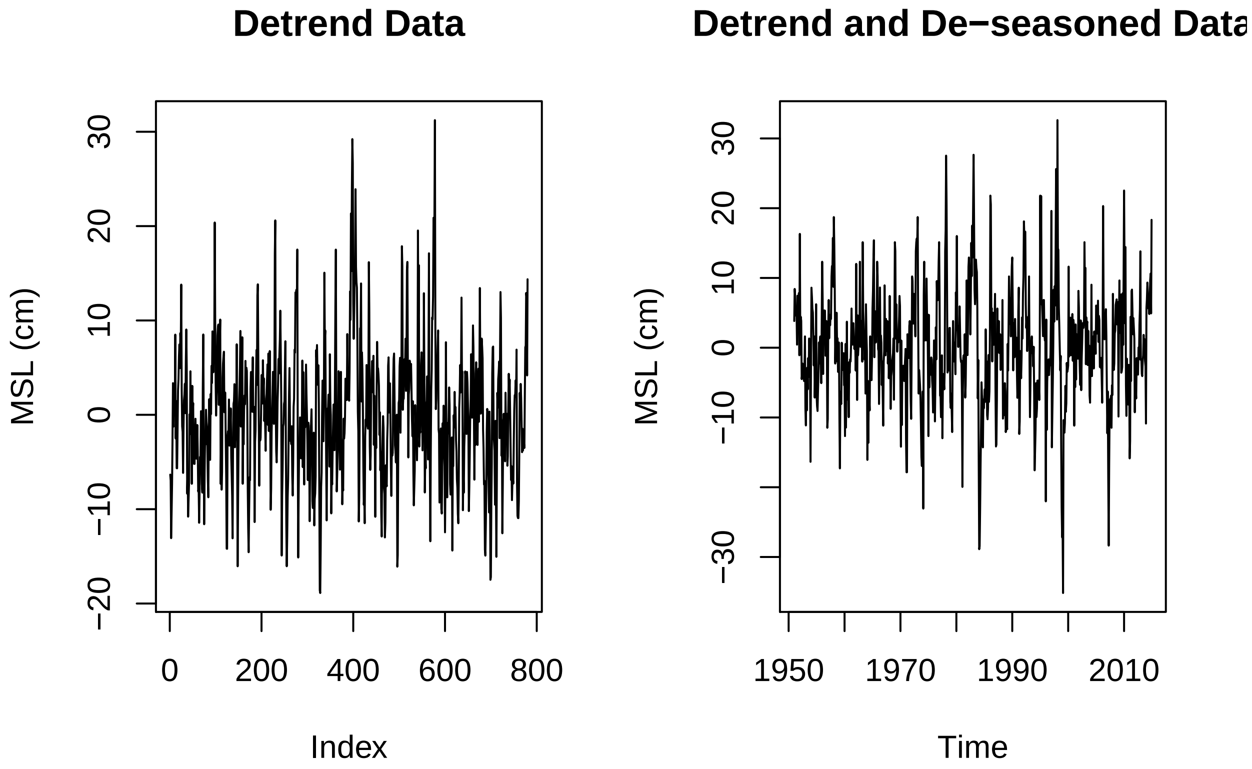

Figure 5: Detrended and De-seasoned Data

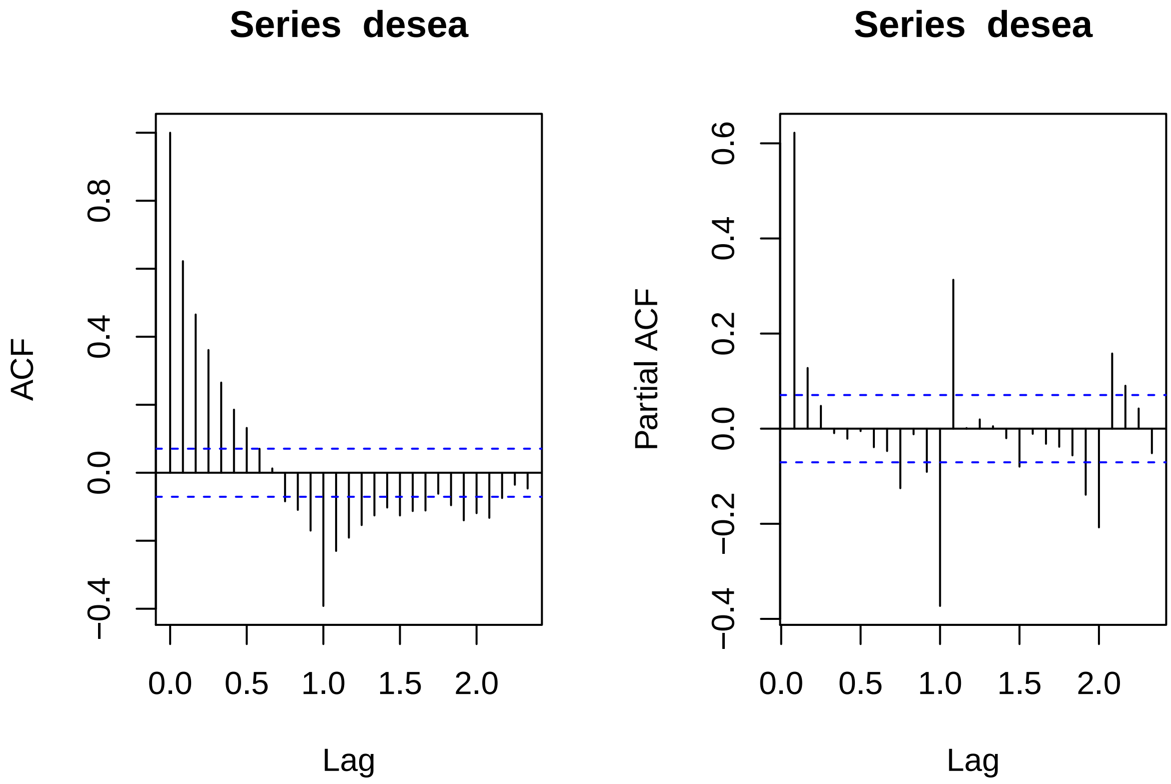

Figure 6: ACF (left) and PACF (right) of Detrended and De-seasoned Data

Table 2: ARIMA Coefficients

	Estimate	SE	tValue	pvalue
ar1	0.7666	0.0360	21.3193	0.0000
ma1	-0.1812	0.0566	-3.2025	0.0014
sar1	0.0345	0.0390	0.8855	0.3762
sar2	0.0012	0.0391	0.0317	0.9747
sma1	-0.9904	0.0620	-15.9727	0.0000
xmean	0.0070	0.0317	0.2206	0.8254

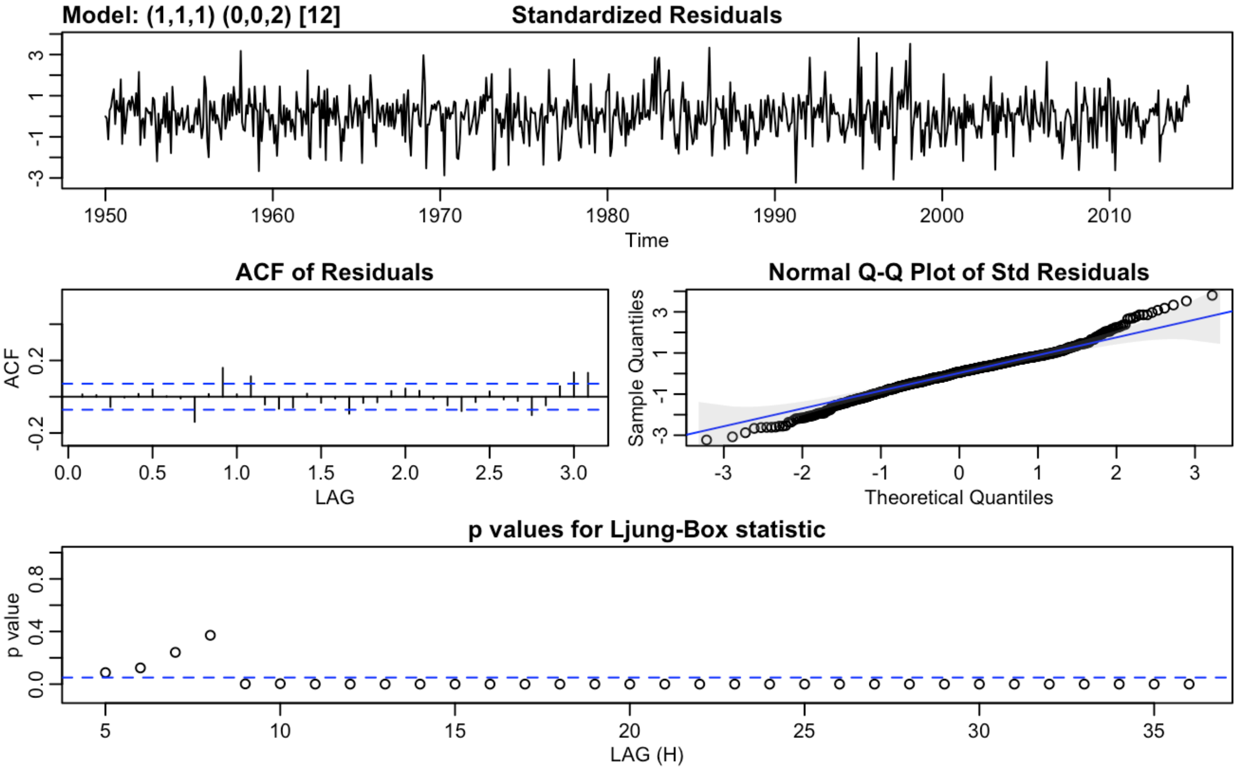

Figure 7: ARIMA Diagnostics

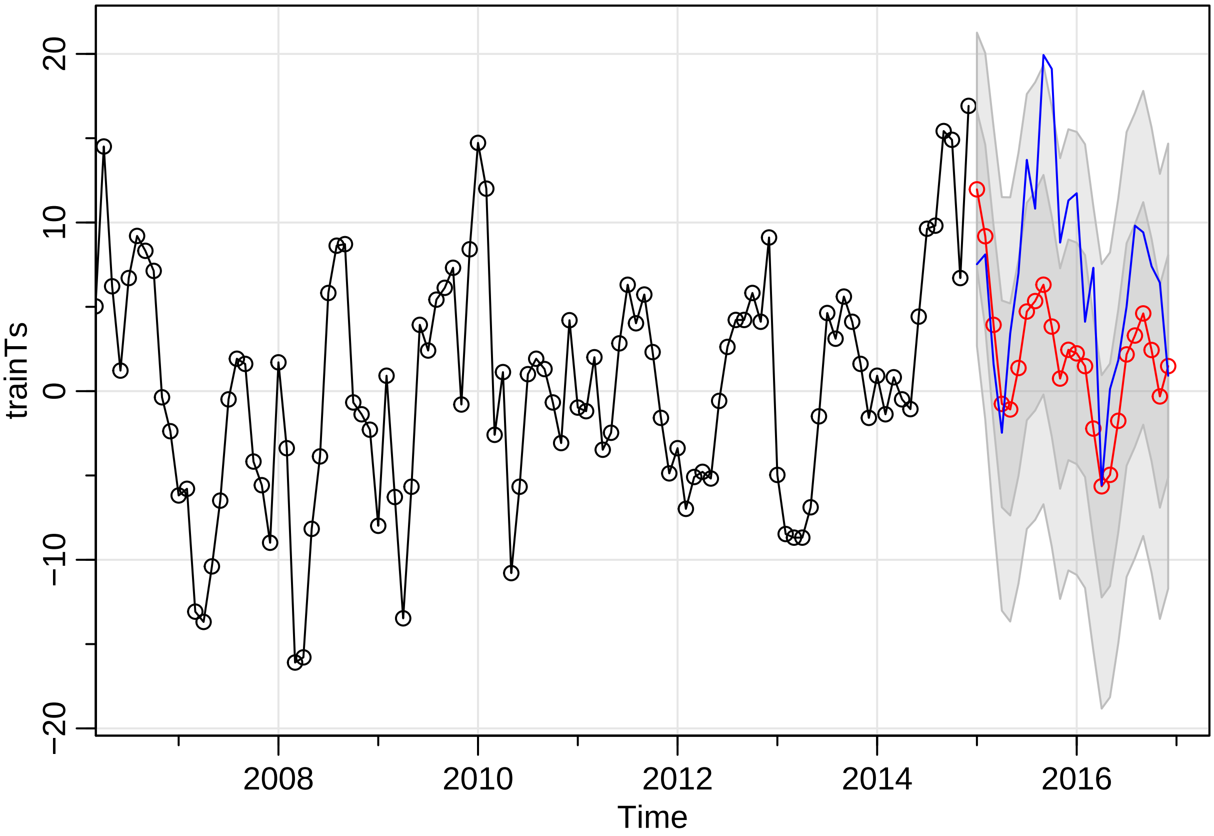

Figure 8: Forecasting 2 Years Ahead

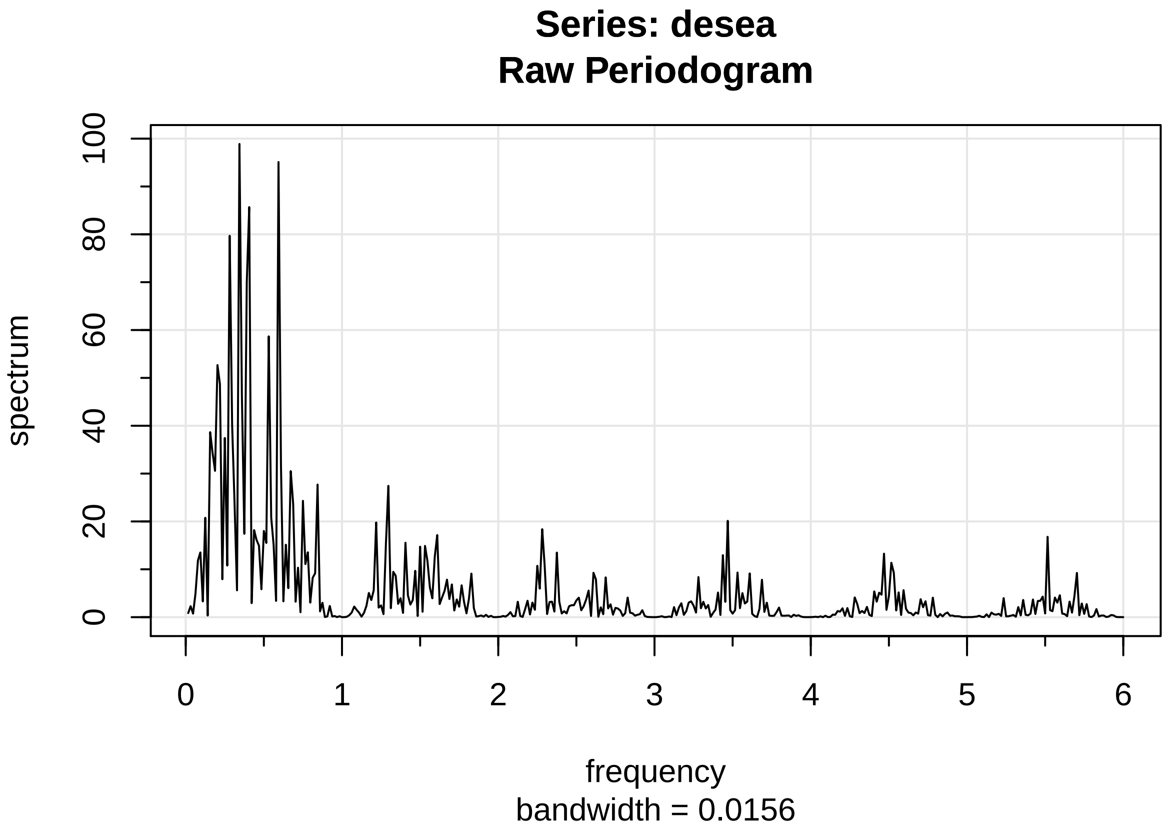

Figure 9: Raw Periodogram

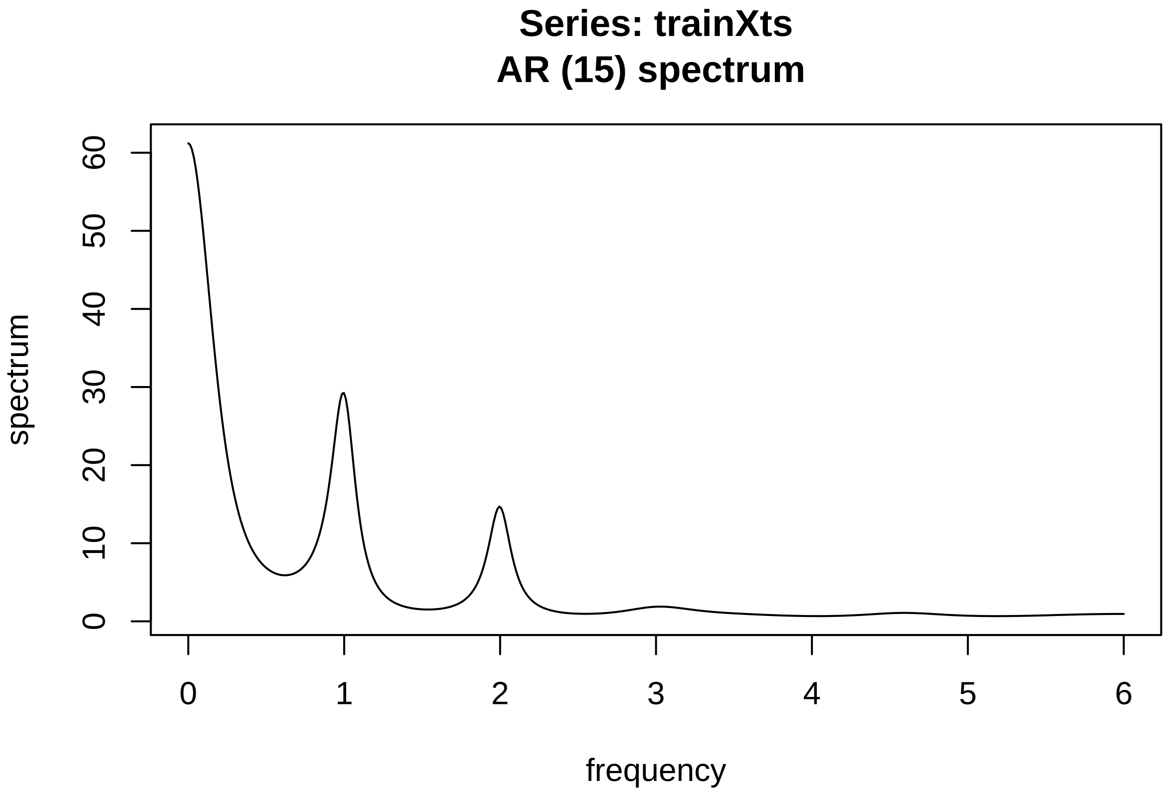

Figure 10: Smoothed Periodogram