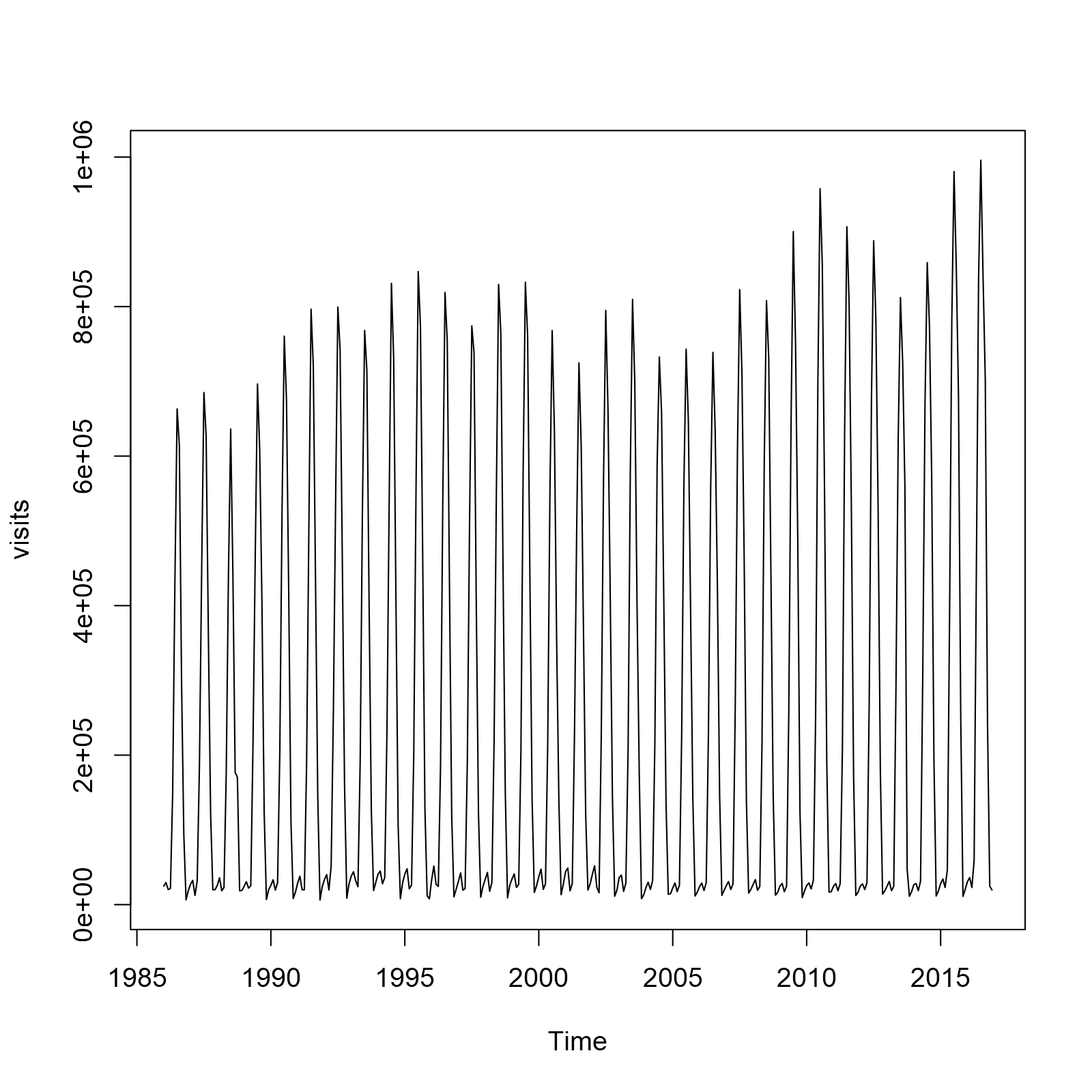
Time Series Analysis Project: Yellowstone Visitor Trends

Michael Dunn, Jacob Klein, and Evan Waldmann

Introduction

Yellowstone National Park is the world’s first national park that was established in 1872 by president Ulysses Grant. Spread over Wyoming, Montana, and Idaho, 96% of Yellowstone is located in Wyoming. The rest of the area of Yellowstone is split between 3% in Montana and 1% in Idaho. With a 45-mile-wide, activate super volcano, Caldera, and a plethora of geysers, it is a popular tourist attraction for people of all ages. Past events in its history include, the U.S army takeover of management for the park from 1886 until 1918, when the management of the park was transferred to National Park Service. In 1915, the U.S army allowed private automobiles the ability to enter the park. During 1929-1932, President Hoover expanded the park multiple times, and in 1948, Yellowstone hits a huge milestone they receive one million visitors. Most importantly, in 1988, an uncontrollable fire hit Yellowstone by a lightning strike and that fire destroyed 1.2 million acres of Yellowstone, which as we will discuss later created a notable decrease in park attendance. Now, most people visit Yellowstone National Park for its plethora of wildlife, and it hosts an impressive average of 3.8 million people per year. With the intent of forecasting the number of visitors the park attracts each year, we set out on this project to analyze and fit a model to this time series of the number of visits to the park each month.

General Trends

The first step the team took was to look at the overall trend in the data. Looking at the plot of the data, it is trivial to see the seasonal cycle with a stationary trend over the years. There looked to be a possible slightly positive linear slope to the data, so we explored taking a logarithmic transformation, which is more easily seen in time series decomposition. Through the decomposition, we divided the data into categories displaying the trend, seasonality, and random error in graphs.

Figure 1 Plot of Yellowstone National Park visits from 1986 to 2017

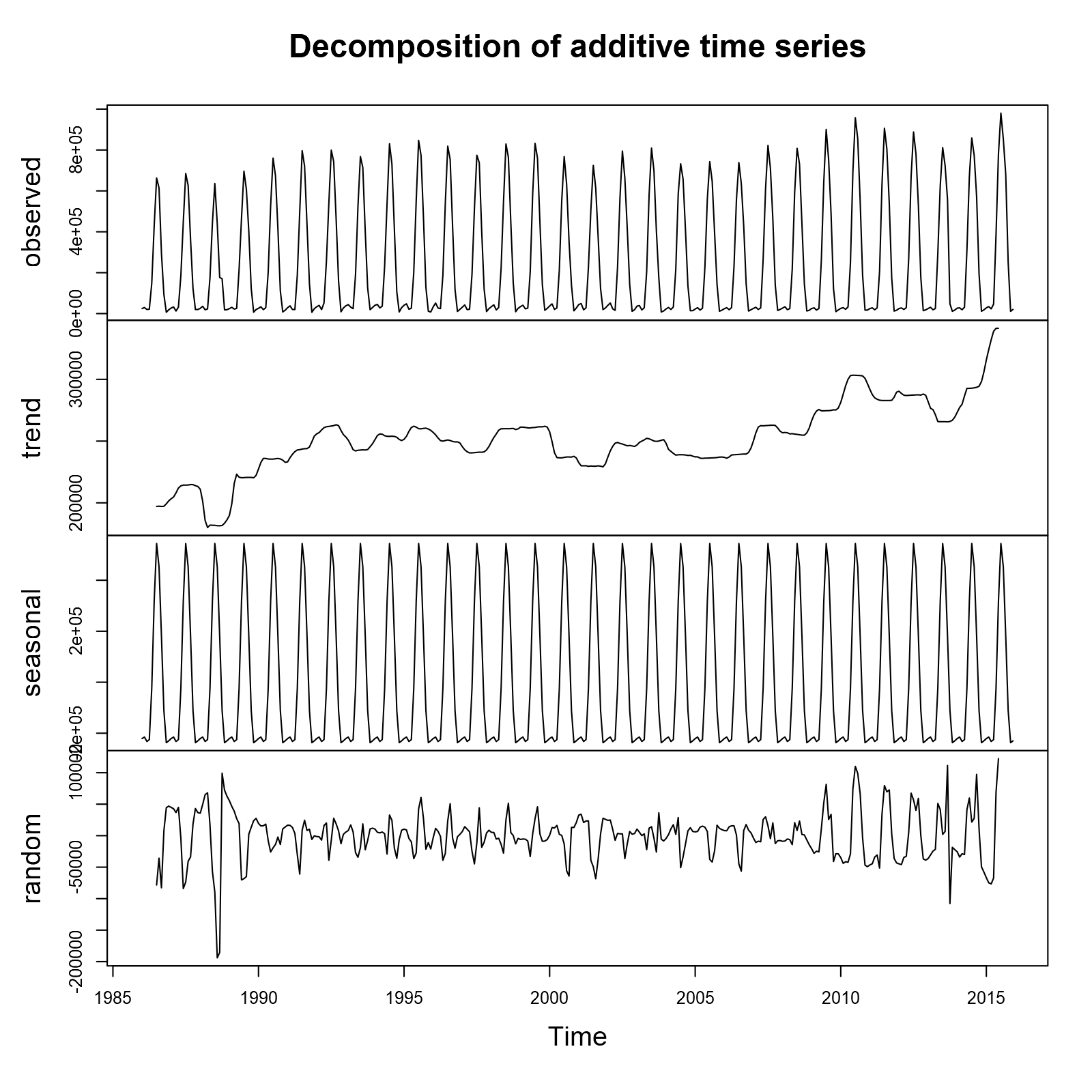


Figure 2 Decomposition of Time series prior to log transformation

From the decomposition, it appears as if the time series will be a good subject and display a stationary seasonal process once the linear trend is removed. Another thing to note here is that the plot of the random error seems to have no explicit pattern except for a larger drop around 1988, caused by the rampant wildfires in Yellowstone that year causing a major drop in visits at the park. Additionally, towards the end of the series, we observe that the random error seems to grow while still being center around zero; this led us to believe that a log transformation might help make these errors seem more random. We will explore this situation when we both take the transformation of the data as well as explore the processes of deleting the additive outlier by replacing the 1988 data with its respective five-year average.

Data Transformation and Seasonality

After analyzing the general trends in the data, we decided to see how a log transformation would affect our series, hoping that it would help our random error appear slightly more random. Looking at the decomposition for the transformed series, we can still see our slight linear trend, although it is weakened. In all, we now see that our random error plot appears a bit more like random noise. This leads us to believe that the log transforms added to the stationarity of the series and should be used going forward.

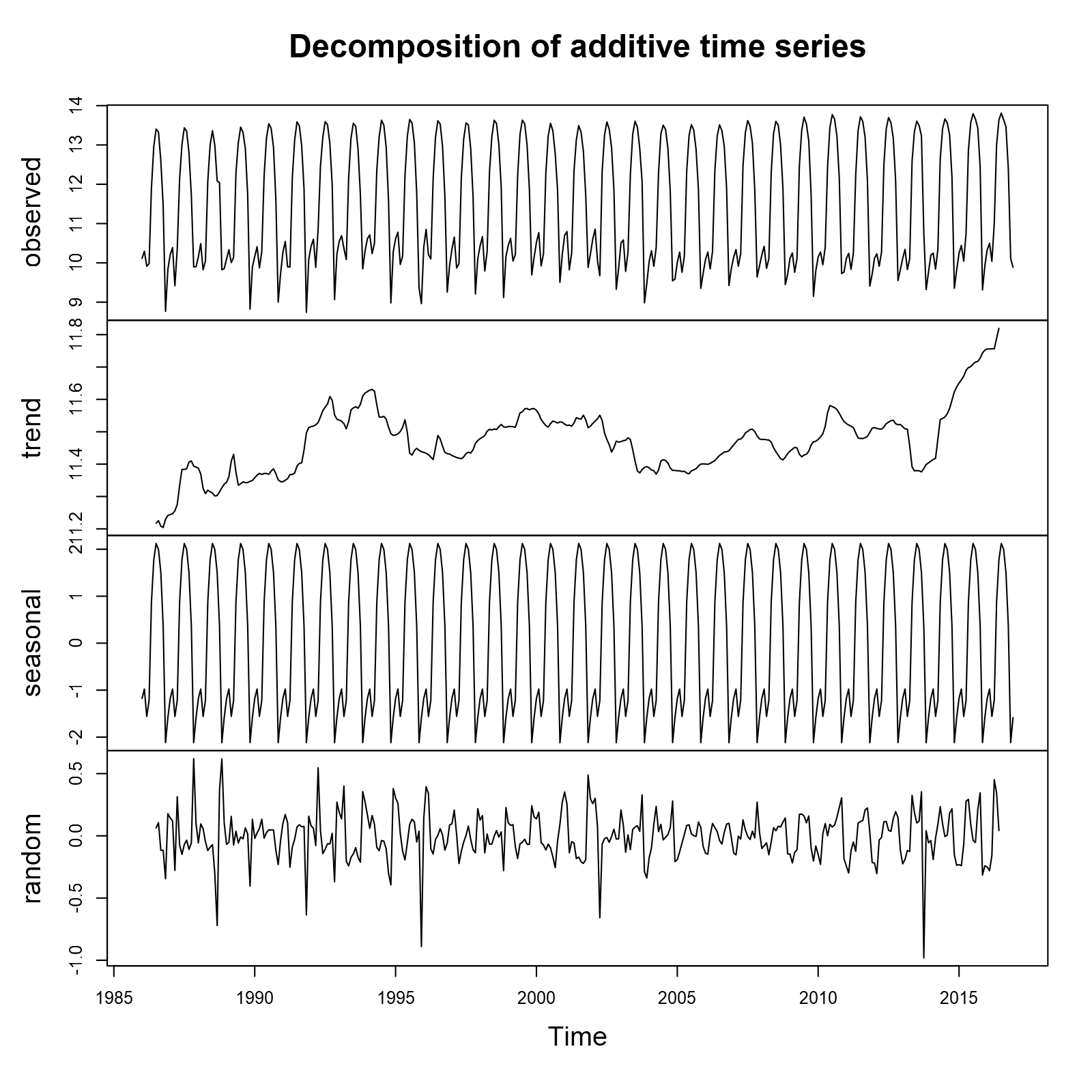
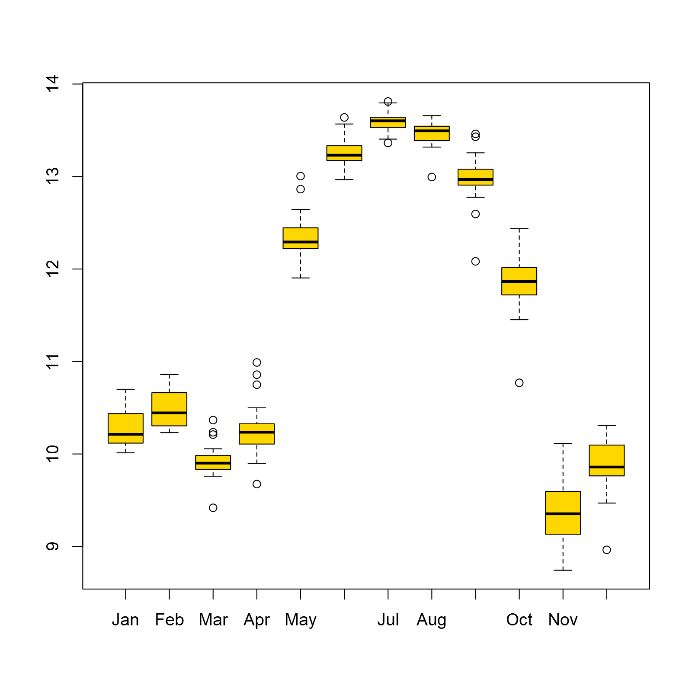
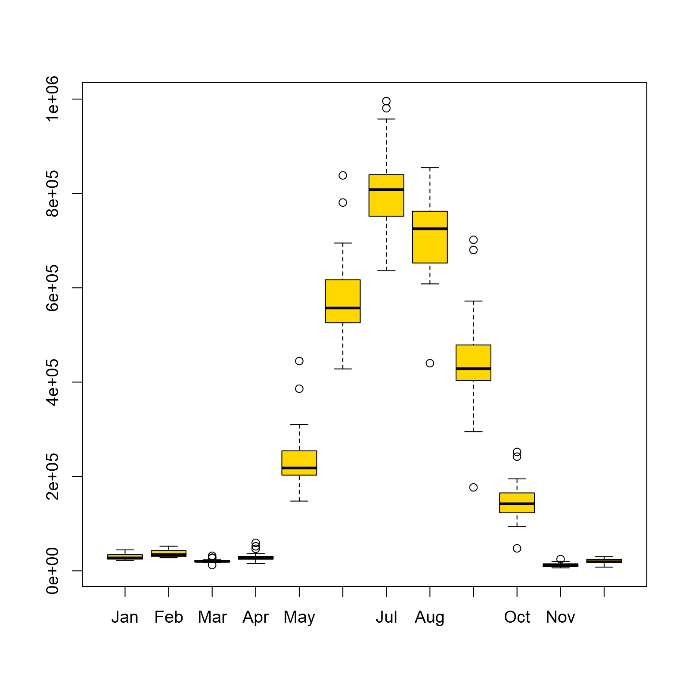


Figure 3 Decomposition of Time series after log transformation

We can also observe the distribution of our series over its months. The decomposition has helped to reveal the heavy seasonality in our data, most likely related to heavy travel and vacations during the summer months, but now we can observe the distribution plots below. Below on the left we have our original series, and on the right, we have our transformed series. Through this we confirm that Yellowstone has had a drastic increase in visitors beginning in May and continuing on through October, which corresponds to school vacation schedules, wildlife appearances such as moose and bears in the park, and more comfortable temperatures that attract campers. These plots also exhibit the squishing effect of our log transform, which also us to see that summer months are quite severely separated from the winter months, or off season.



*Figure 4 Distribution plots of Yellowstone visits by month for both transformed data (right) and raw data (left)*

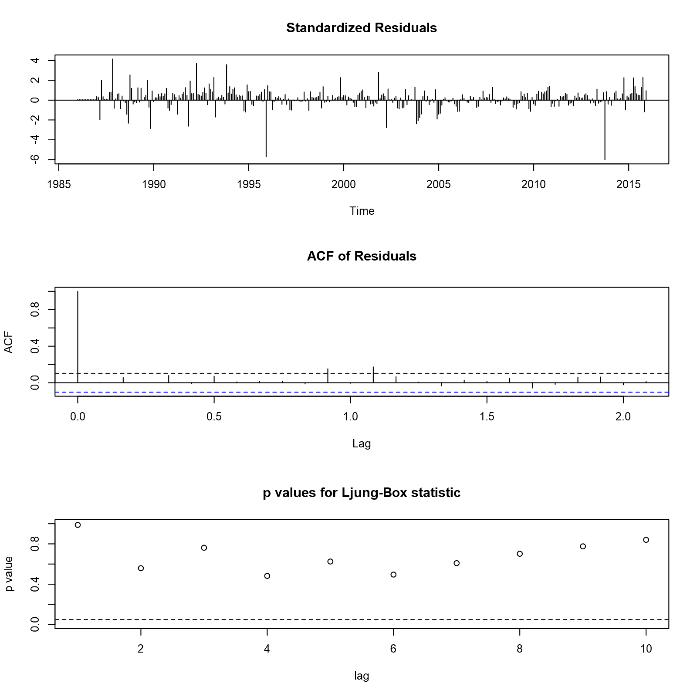
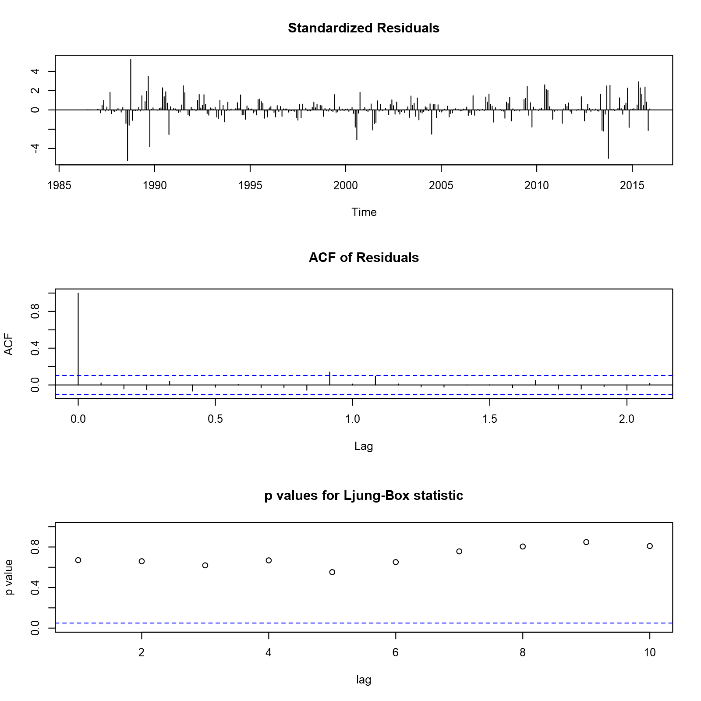
Model Choice and Fitting

Taking the data without applying a transformation, we find that the stationary assumption is satisfied with the Augmented Dickey-Fuller Test. The test statistic found is -20.33 with a p-value less than 0.01. We first attempt to fit a time series model to the data before applying a transformation and find that a seasonal AR(1) model fits best, as it is the model with the lowest AIC value. We fit the model to the data taking out the last year in order to test our model forecast against the actual data. When looking at the plot of residuals of this model, we find a randomness in the residuals with no general pattern and determined that a seasonal model would fit the data best due to the weather aspect of planning a trip to Yellowstone. From plot in the previous sections, it is apparent that the top of the seasonal trend tends to occur around July and August which makes sense to us due to the temperature being ideal for outdoors activities. Similarly, the bottom of the seasonal trend appears to be in the deep winter months of December and January which also makes sense due to cold snowy weather as well as road closures.

We will also look at a model fit of the transformed data as it will reduce the impact of the linear trend in the data. When we take the log of the data the outlier in 1988 as well as the general linear trend will be diminished. Looking at the decomposition of the plots after the transformation there is a diminished linear trend but there is still a strong seasonal trend. We also found that the randomness showed much less influential points and appears to be more random around 0 leading us to believe that the log transformation of the data will fit better. We will fit the model to the transformed data and determine the fit of that model as well as the complexity whether it is worth the transformation or not. After taking the log the Augmented Dickey-Fuller Test for stationary the test statistic is even better at -23.802 with a p-value of less than 0.01 meaning that the stationary assumption is satisfied. We find that a seasonal MA(1) model with the seasonal component being of the order (2,1,1), by choosing the model with the lowest AIC value. This is an interesting change since before we took the transformation, we found an autoregressive model while after we see a moving average process. Next, we will run model diagnostics in order to figure out whether taking the transformed data will be helpful or if taking the non-transformed data will produce a better fitting model.

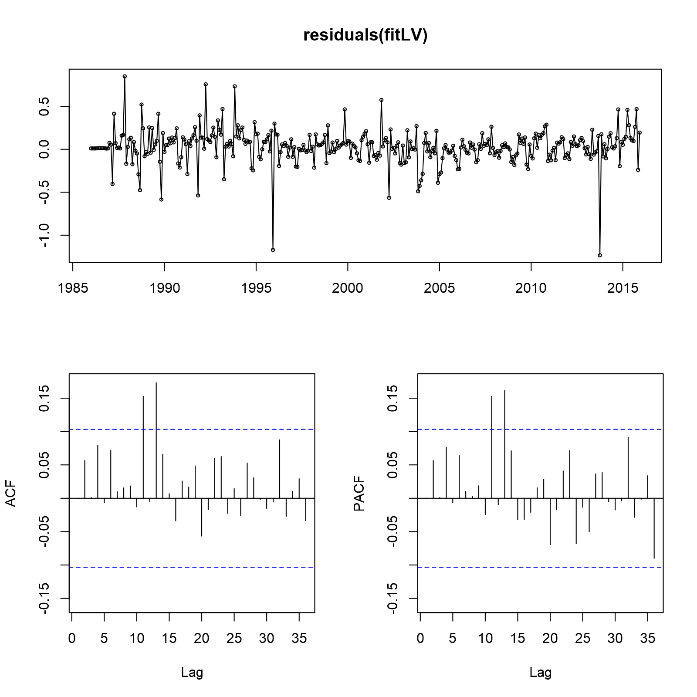
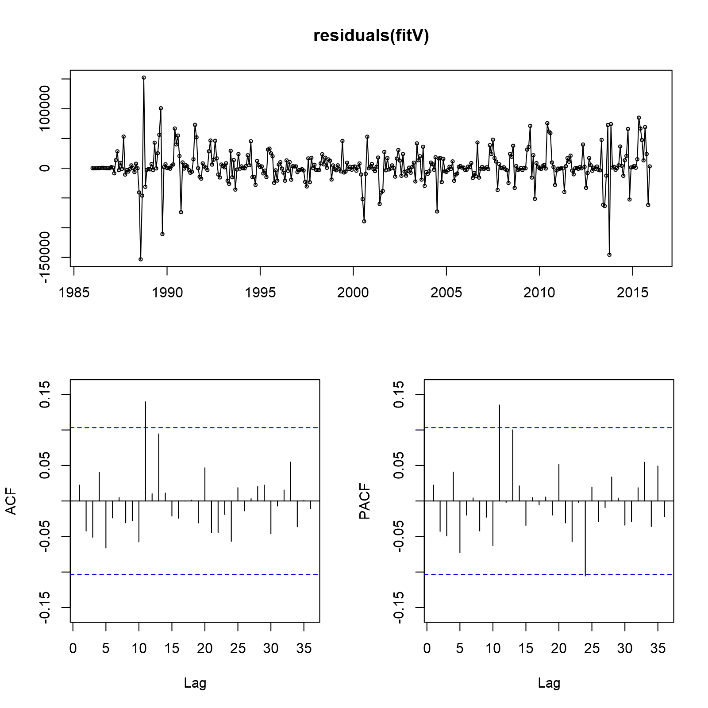
Model diagnostics

Now that we have two models, one without a log transformation and one with a transformation applied, we found using the augmented Dickey-Fuller test and found that both were statistically significant enough to conclude stationarity. It is worth noting that after taking the log transformation the stationarity was stronger. The next thing that we looked at was the residuals given from taking our model. In the model applied to the raw data, a seasonal AR(1) model, we found that our model tended to over predict at first, corrected in the middle, and under predict towards the final years. This suggests a possible positive linear trend in our data, so we analyzed the log transformation of the data as well. In the model for the transformed data we found that the seasonal MA(1) model still over predicted in the early years but by much less after the first two observations, was right on in the middle years, and still under predicted in the later years. That being the case we concluded that, in terms of residuals, the log transformed data provided a model that more accurately predicted for most of the years than the model from the raw data.



*Figure 5 Time series diagnostics for models prior to transformation (left) and after log transformation (right)*

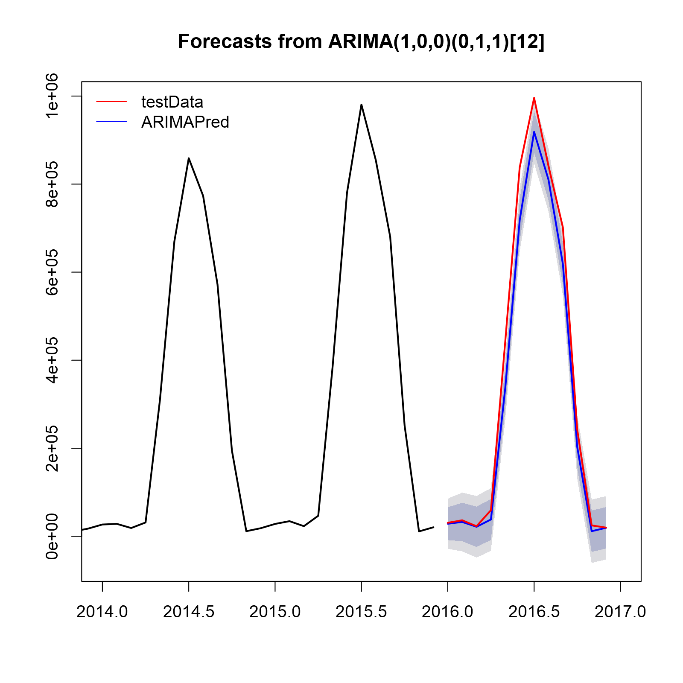
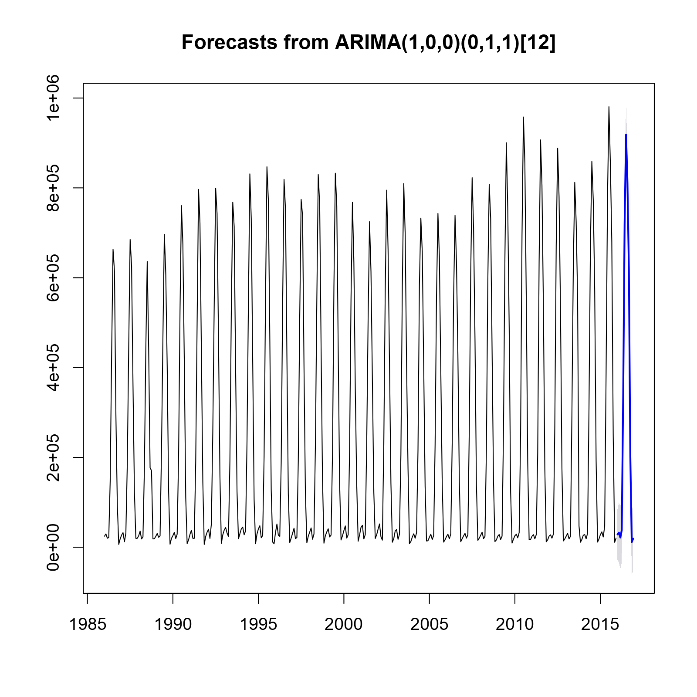
Next, we analyzed the autocorrelation function graphs for both. For the raw data there is a significant observation at lag 11 which makes sense for a seasonal AR(1) model. However, at lag 23 and 24 we do not find statistically significant observations for residuals. The autocorrelation for the seasonal MA(1) model we find significant residuals at lag 11 and 13. We looked at the assumption of independence of error terms. In order for it to be met there needs to be a lack of statistically significant evidence to reject a null hypothesis of independence of error terms. We used the Ljung-Box test and compared it to a chi-squared distribution with K-p-q degrees of freedom. For the raw data the Ljung-Box statistic is significant up to a lag of 15 while the Ljung-Box statistic for the transformed data is only significant up to a lag of 12 after which the assumption of independent error terms is rejected. The seasonal AR(1) model holds the assumption of independent error terms to a higher degree and with higher degrees of freedom.



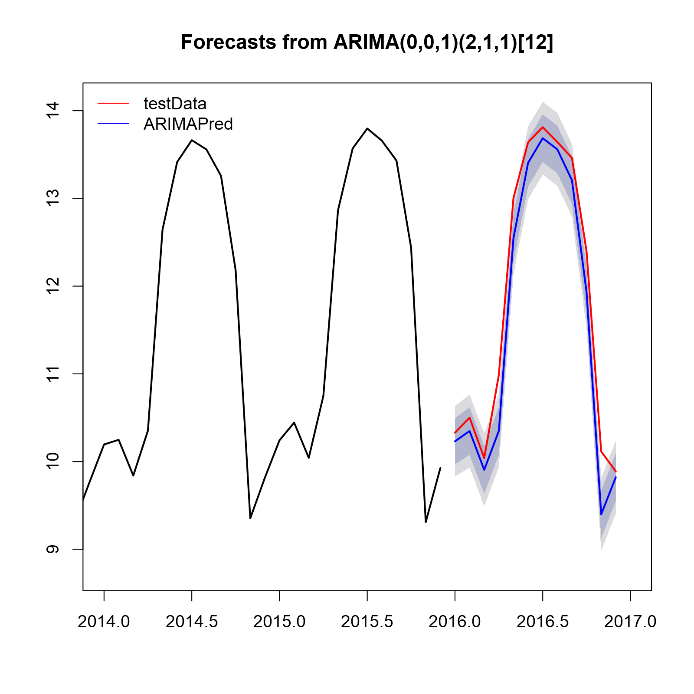
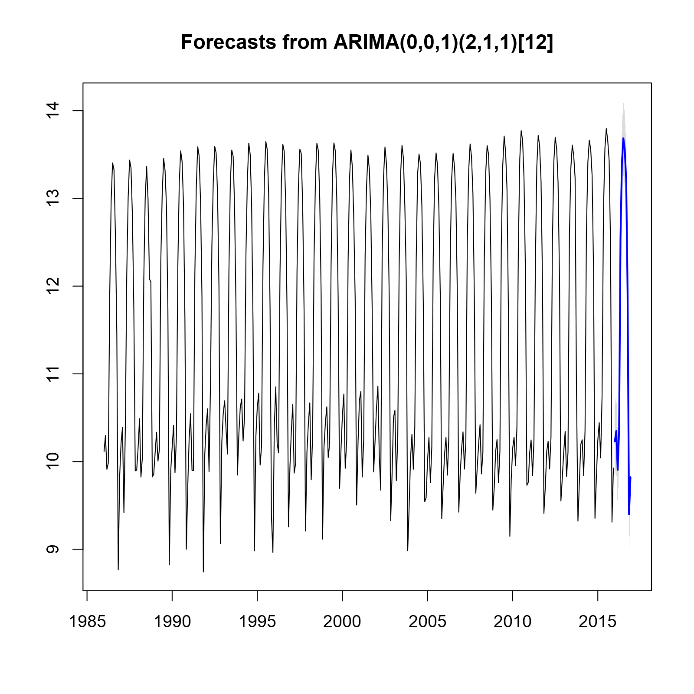
*Figure 6 Plot of residuals as well as ACF and PACF graphs for Yellowstone National Park Visits prior to transformation (left) and after log transformation (right)*

Forecast and Model Accuracy

In order to forecast, we kept the last year, 2016, out of our model fitting in order to have a set of valid values to compare to. This give us the ability to measure how well our model predicts one year out from the values that it is trained on. We ran the twelve-month forecasts on both models in order to see how each faired. When comparing the accuracy of the forecasts, we run into a bit of difficulty because the models are not using the same scale for the data, since the second model was fitted with the log transform of our series. The following graphs exhibit the results of our forecasts.



*Figure 7 Forecasts taken from the raw data prior to transformation*



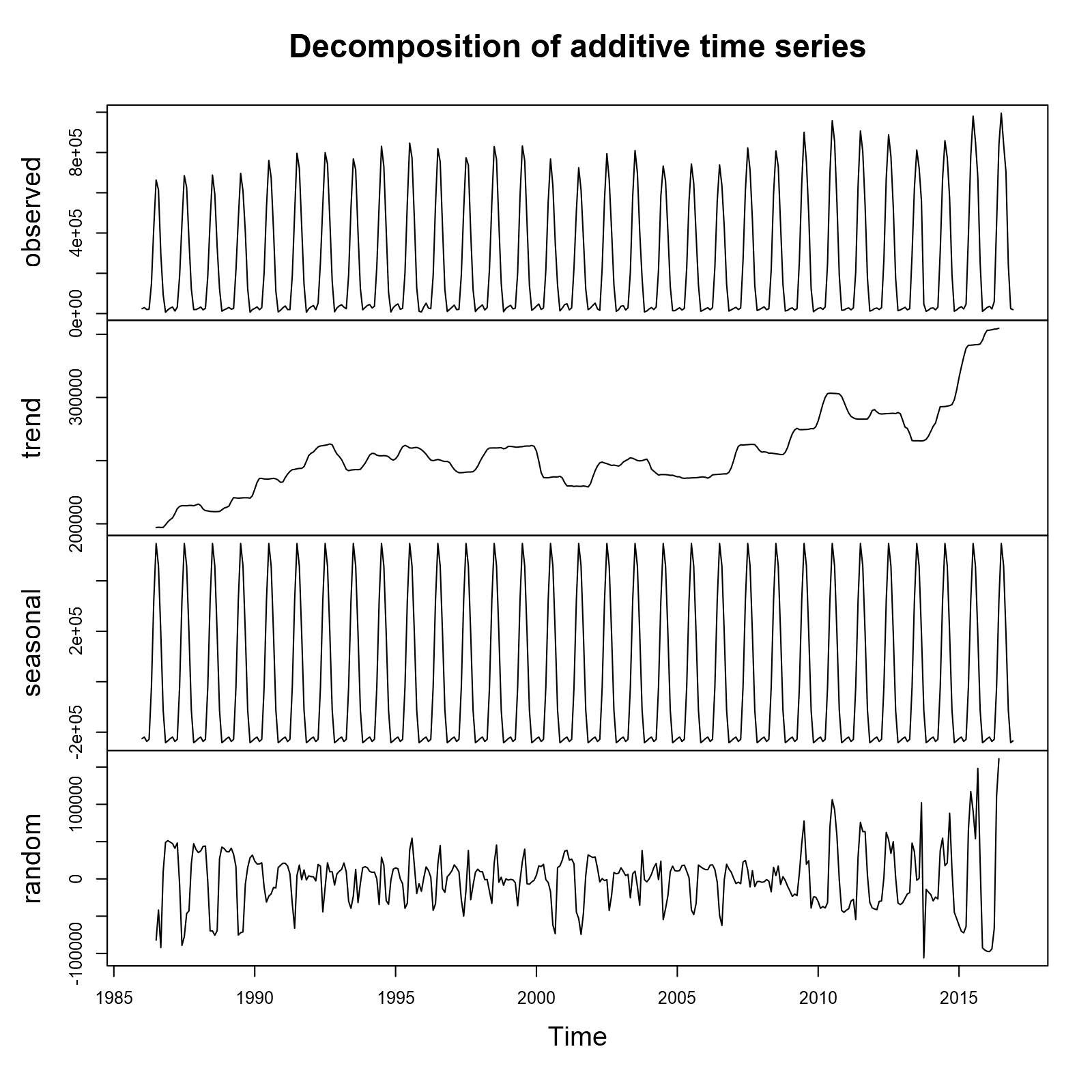
*Figure 8 Forecasts taken from the data after log transformation*

Comparing the two models accuracy of prediction for the last year of our time series cannot be done directly, since one is on log scale and the other is not; when attempting to convert the log transformed prediction back to the original scale, we could not obtain comparable predictions and thus it did not make sense to include those pursuits. We instead observe the quality of the forecast through the Theil’s U coefficient; this gives us an idea of which model is best for this particular forecast. The model fitted the original time series data yields a Theil’s U of  0.2975, while the model fitted on the log transformed data gives 0.3615. From this we can gather that the model fit on the original time series data provides us with a better forecast, since its U value is closer to zero [1]. Thus, the ARIMA (1,0,0) X (0,1,1) model fit on the original time series data is a more realistic fit for the data for the 2016, and thus should be used if we wanted to make a statement about future years.

Outlier Detection, Data Manipulation, Model Refitting, and New Forecasts

On June 30,1988, a small forest fire broke out in Crown Butte region of Yellowstone National Park. The fire started when a lightning bolt struck a tree and spread rapidly across 1,800 acres in a span of 3 days. During that summer, the fires burned about 683,000 of the parks 2.2 million acres. The fires spread so much because Yellowstone had a park policy to do nothing to take away the fire. Usually Yellowstone has a fire at least once a year, but this fire was different. Most of the fires would go out on its own in a couple of days but this summer was a dry and windy summer. The fire did not go out until November when it started to snow, and the fire has reached about 1.2 million acres. This fire caused the decrease in visitors for the summer of 1988, our time series’ major outlier.

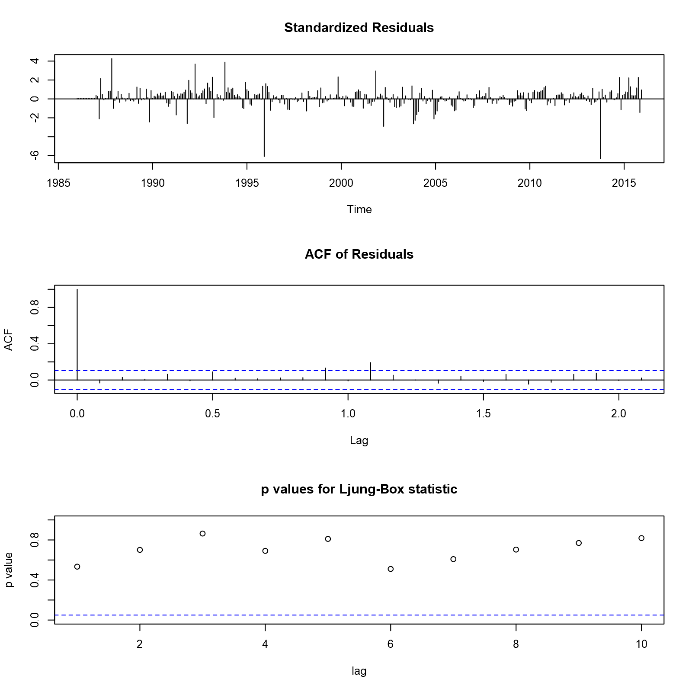
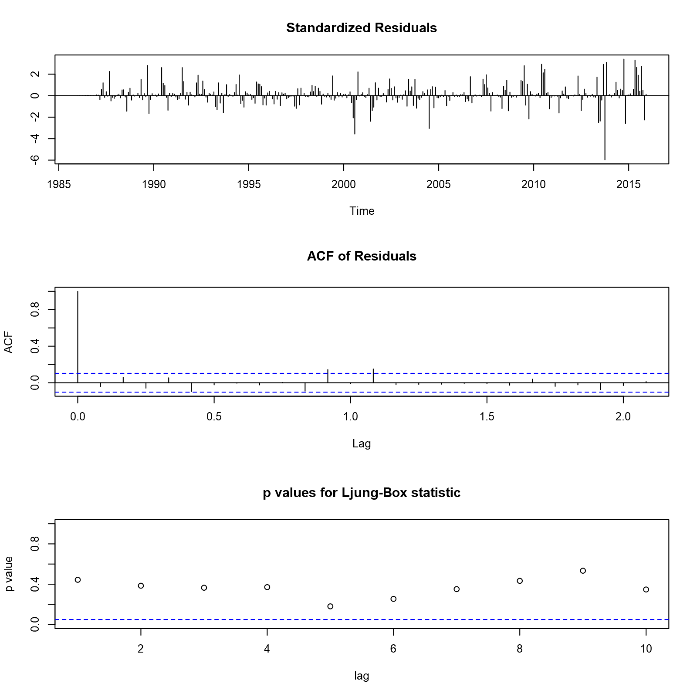
In order to remedy this outlier, we replaced the data from 1988 with a five-year average made using the years 1986 to 1990. With this manipulation of our data, we run through a similar analysis of residuals that we performed before and refit models to this new data.



*Figure 9 Decomposition of Time Series after removing 1988 outlier*

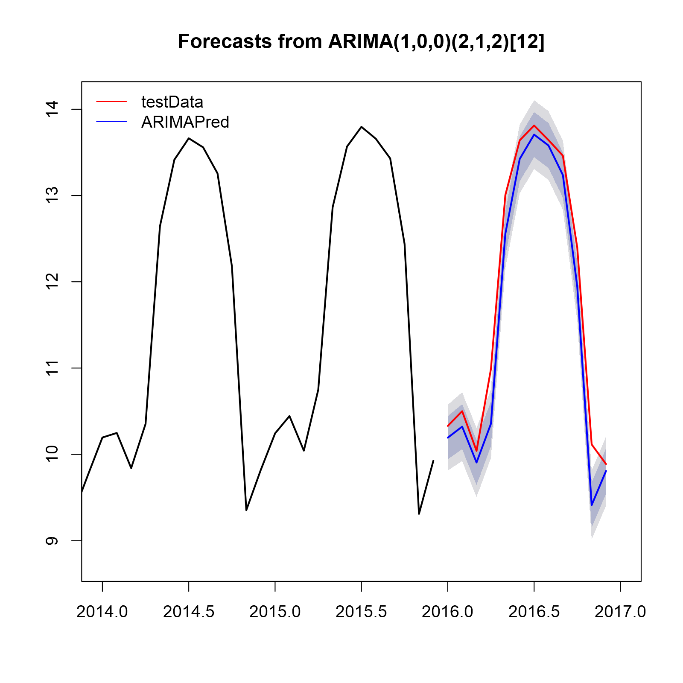
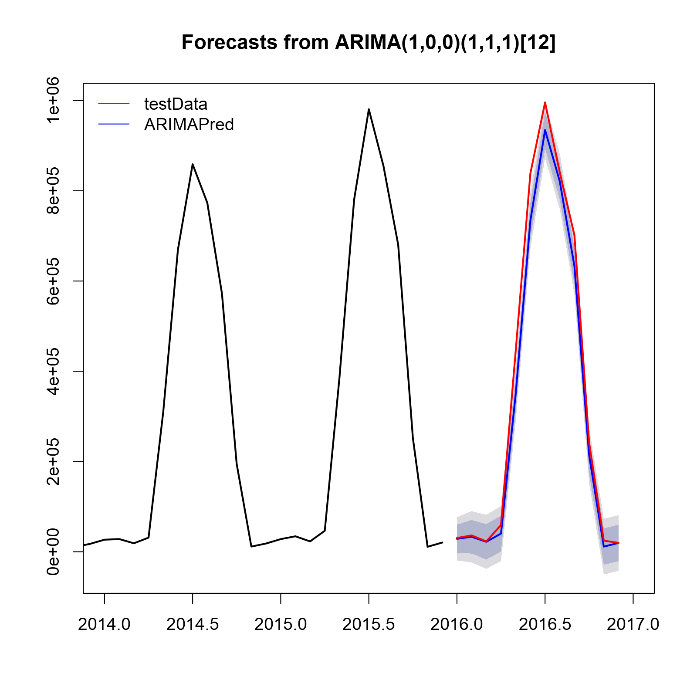
The decomposition above shows our time series after removing the outlier before taking a log transform, we see that the residuals are mostly random and do not show the 1988 outlier any more. Although the residuals seem to increase in variance toward the end of our series, this is an improvement over the previous, and we will continue on looking refitting models to this data and a log transformed version of it.  It should be noted that the decomposition of the log transformed data set without outliers, that the increased variance in residuals is lessened, which is one of the reasons we still pursue the log transformation. Also, both series passed the Augmented Dickey-Fuller Test and can be considered stationary.

Like before, by choosing models with the lowest AIC values we found that a ARIMA(1,0,0) X (1,1,1) fit best for the non-log transformed data and a ARIMA(1,0,0) X (2,1,2) fit best for the log transformed data. When looking at our fitted models’ residuals below (left is non log and right is log transformed), we see that the original data only has a larger outlier at 2014 now, but the log transformed data as a couple outliers. This leads us to believe that removing the outlier also normalized our data and made the log transform less beneficial. Additionally, similar to before we see that the box statistics are significant still.



*Figure 10 Time series diagnostics of raw data prior to transformation without the 1988 outlier (left) and after log transformation(right)*

Forecasting using the same methodology yielded the following graphs. When checking the Theil’s U values, we again see that the non-log transformed model has a lower U value of 0.2662 than the log transformed model’s U value of 0.3558. This tells us that the ARIMA(1,0,0) X (1,1,1) is best for forecasting than all previous models. Since our best model is also our simplest model, we believe that we found the best fit to our data and that it is appropriate to remove the 1988 wildfire outlier to improve our model’s forecasting power.



*Figure 11 Forecasts of Yellowstone National Park visits prior to transformation after taking out the 1988 outlier (left) and after taking the log transformation (right)*

Conclusion

We first presented two stationary time models, the first an ARIMA (1,0,0) X (0,1,1) made with the original time series data, the second an ARIMA (0,0,1) X (2,1,1) made with the log transformed series data. Both of the models were the obtained by minimizing the AIC values and were fitted to their respective series excluding the last year to obtain a measure of accuracy for the models. We found that  ARIMA (1,0,0) X (0,1,1) model fit on the original time series data was more accurate in predicting the last year of our time series and recommend it if forecast of future attendance is required. Additionally, we presented in-depth analysis of the models’ residuals and outliers to justify our selection. After manipulating our data to remove the wildfire outlier in 1988, we refit our model and re-ran the forecasts to obtain better Theil’s U values. This implies that the model fit on the original time series data after the 1988 outlier was removed gave us the best results for forecasting 2016 park attendance than all previous models. Thus, our recommended model for this dataset is the ARIMA(1,0,0) X (1,1,1) with the modification of the fire year average to replace the 1988 visitor attendance because of the large wildfire.

References

1. Bliemel, F. (1973). Theils Forecast Accuracy Coefficient: A Clarification. Journal of Marketing Research, 10(4), 444. doi:10.2307/3149394
2. Fire. (n.d.). Retrieved from https://www.nps.gov/yell/learn/nature/fire.htm
3. Park Facts. (n.d.). Retrieved from <https://www.nps.gov/yell/planyourvisit/parkfacts.htmPrabhakaran>,
4. S. (n.d.). Time Series Analysis. Retrieved from <http://r-statistics.co/Time-Series-Analysis-With-R.html>
5. Snehal. (2018, May 08). YellowStone National Park. Retrieved from <https://www.kaggle.com/snehal1405/yellow-stone-national-park>
6. Time series analysis - accuracy in forecasting. (n.d.). Retrieved from <https://rpubs.com/RatherBit/90267>