Project 1: Forecasting Monthly Yellowstone Visitors

Elizabeth Combs (eac721)

04/06/2021

Introduction

The data for this project was retrieved from National Park Service (NPS) Stats, which houses the National Park Service data for monthly visitors per park in the United States: (link to site and link to data). Yellowstone National Park was chosen for this project because it has both a long history and a high visitation rate. As the first official US National Park, Yellowstone remains one of the most popular. The data is available for each month from January 1979 through February 2021.

This project focuses on the number of `Recreation Visits By Month`. Recreation visits include any entry of a person onto lands or waters administered by the NPS excluding non-reportable and non-recreation visits. Additional documentation on the park statistics gathered is available here.

Please note that a dynamic version of this report is available at this link.

Data Exploration

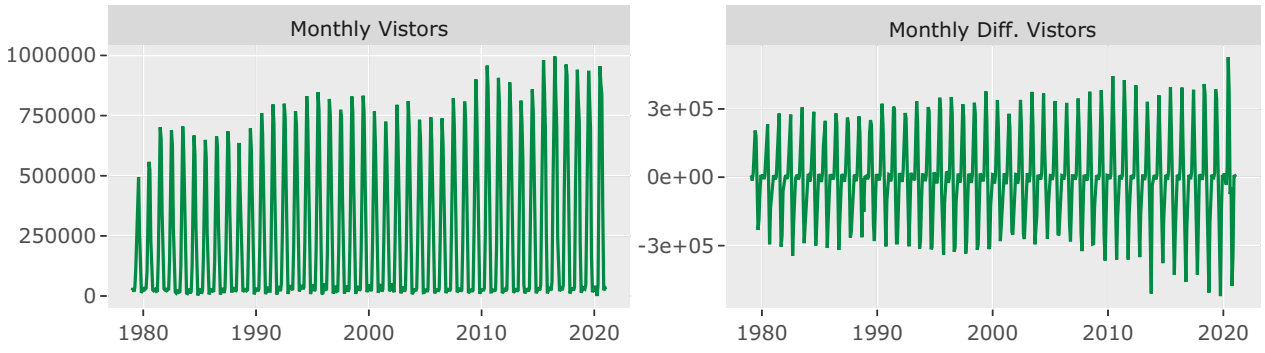
Data Transformation

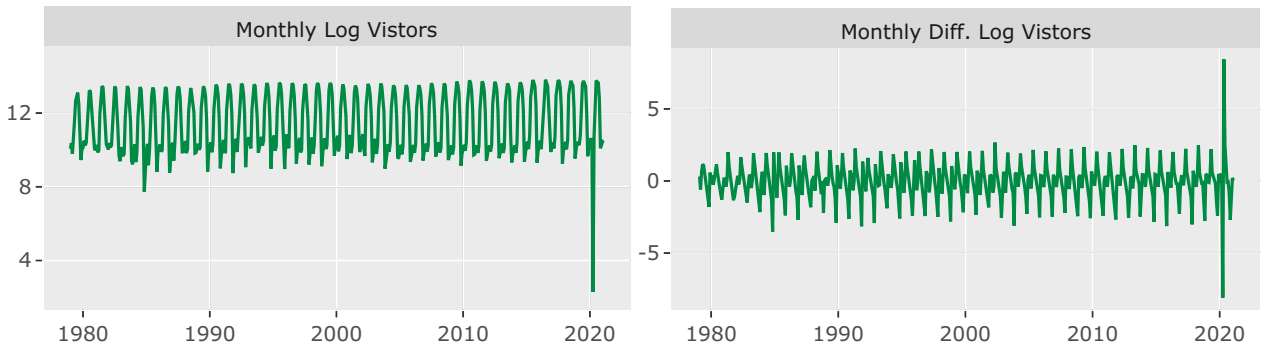
In order to conduct this study, the data was first transformed to a long instead of wide dataset using the `reshape` library in `R`. After transformation, the data is a time-series dataset with monthly data from January 1979 to February 2021, a total of 506 data points.

Take Logs

To build time series models, we need to decide whether we must take logs. Logs can help turn exponential growth into linear growth in macroeconomic variables. It also helps eliminate level-dependent volatility. We can review four plots to make this decision:

1. Original Time Series
2. Differenced Original Time Series
3. Log Time Series
4. Differenced Log Time Series





According to the figures above, we can see high visitation levels in summer months and low visitation levels in winter months. This phenomenon is likely due seasonal effects which could be dependent on temperatures in the park, availability of summer holidays, etc.

The first two figures also indicate that there is likely level-dependent volatility, so the first step is to take the log to stabilize it. In this case, this could be a macro-economic series dependent on population growth, economic conditions and other factors. After taking logs, we are assured that the series is not exponentially increasing (second row plots).

Note that in order to compute the log of the series, any values that were zero (which occurred only recently in the series due to nation-wide COVID lock downs in April 2020) were transformed to value 10. Other ways to transform the data include removing the data from the analysis completely (replacing with NA value) and replacing the missing value with the previous year's value or another small value. Replacing a small value for the zero was chosen because it maintains the integrity of the actualized values while simplifying the transformations made.

Seasonality Adjustment

Following the level-change (we used natural log), the data still has contains the seasonal effect. For simplicity, we can remove the seasonal component by subtracting out the seasonal averages (monthly). This value is referred to as the Log Seasonally Adjusted Visitors.

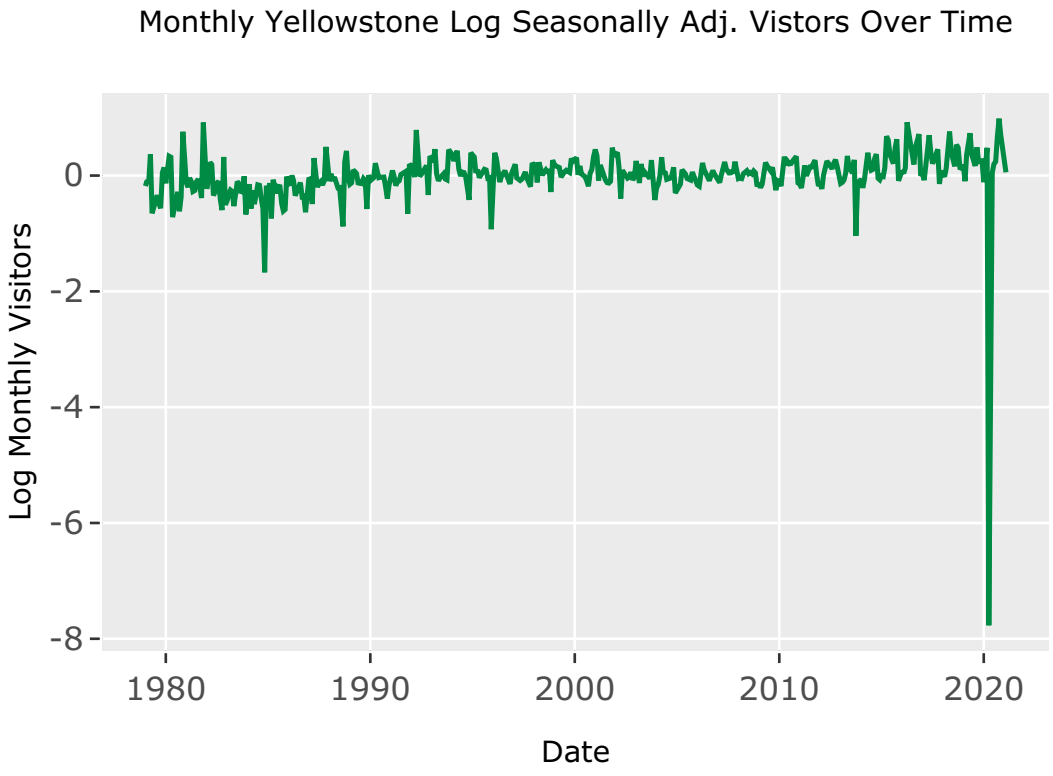
While this simple method helps reduce the impact of the seasonal differences, there may be better ways to account for it either in data transformations or within the modeling (i.e. using SARIMA). The Discussion & Next Steps section discusses this further.

Show 12 entries Search:

	Month	Monthly Avg. Log Visitors
1	JAN	10.25
2	FEB	10.46
3	MAR	9.91
4	APR	10.07
5	MAY	12.25
6	JUN	13.21
7	JUL	13.57
8	AUG	13.45
9	SEP	12.96
10	OCT	11.81
11	NOV	9.4
12	DEC	9.89

Showing 1 to 12 of 12 entries Previous 1 Next

Once we subtract the Log Monthly Average, we can view the seasonally adjusted time-series in the plot below. While the seasonal effect is significantly lowered using this method, there does still seem to be some pattern related to the time of year. We can also see this phenomenon with the slightly cyclical nature of the autocorrelations and partial autocorrelations in the following plots.

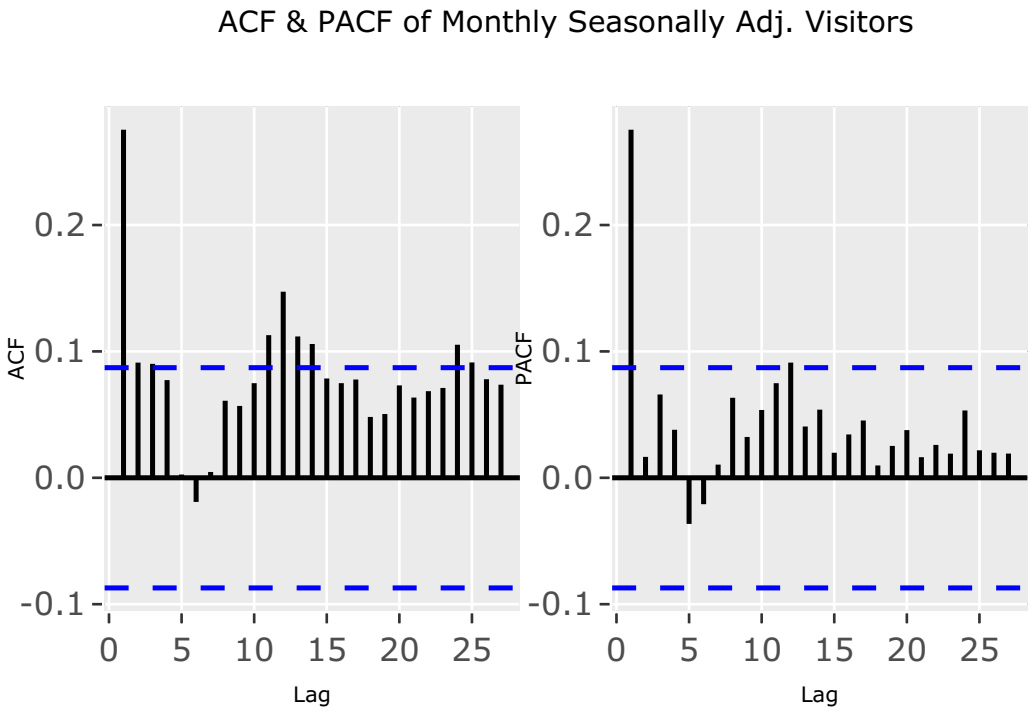


Take Differences

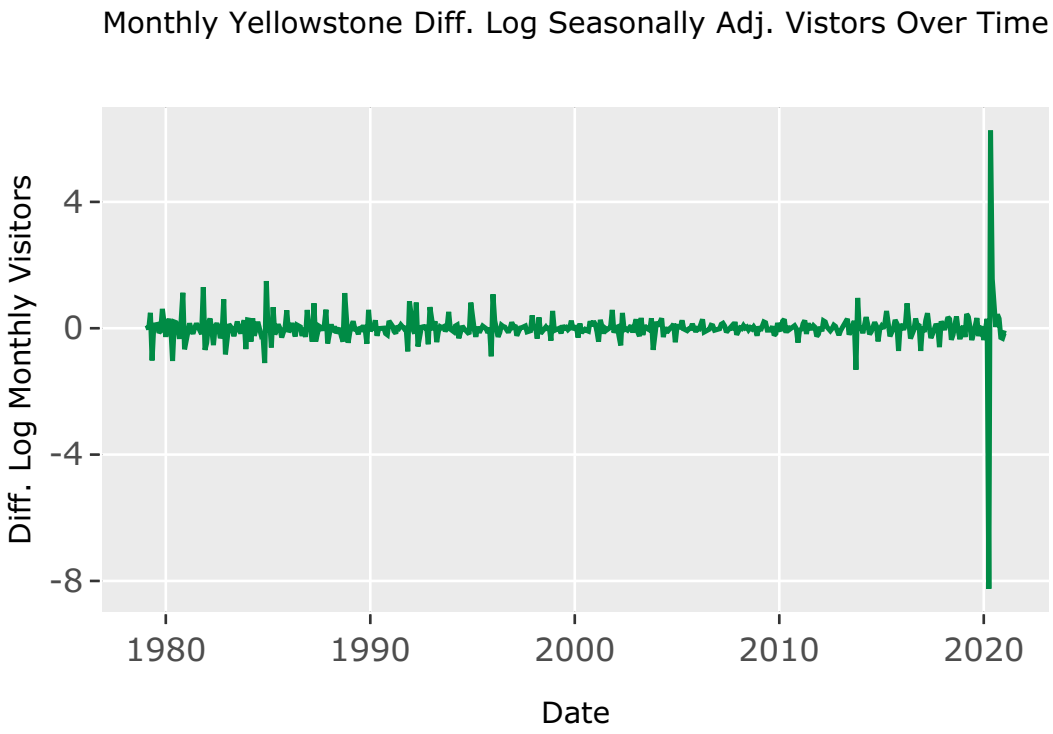
Now that we have transformed our data through logs and seasonal averages, we can start the model selection process to select d in the ARIMA(p,d,q). We can use the ACF plot to assess whether we think our series is

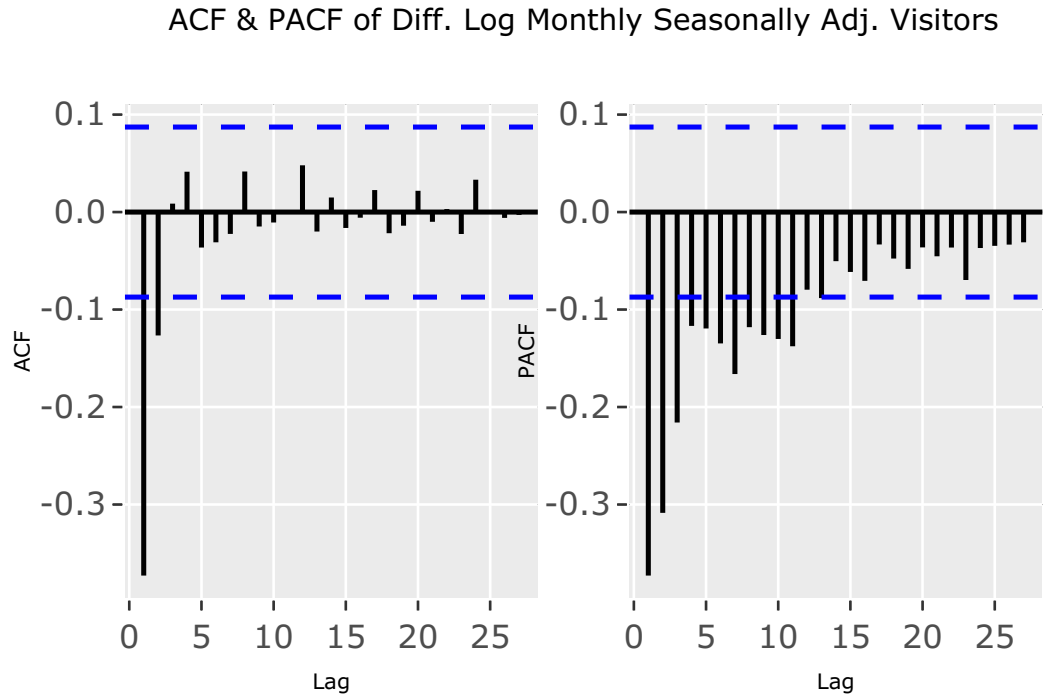
stationary (mean-reverting). We would like to difference our data only as many times as necessary to make it stationary but want to make sure not to over-difference it.

We can start with a value of d=0, which means we have not taken any differences yet. We are looking to see where the ACF and PACF plots cut off or die down to see if we can possibly identify an autoregressive (AR(p)) *or* moving average (MA(q)) model. To identify an ARIMA(p,d,q) model we will need to rely on another model selection tactic that uses AICc (see the next section).



These graphs do not yield a very clear cut off for either AR or MA models. It looks like we could potentially use AR(1) *or* MA(1) models, but we will move forward with selecting an ARIMA(p,d,q) model in the next section. We still need to define d in this step, so we can test d=1 to see if we observe over-differencing yet. Overdifferencing signals could be an autocorrelation plot with the first autocorrelation close to -0.5.





In this case, we tested both d=0 and d=1 (above). We choose d=0 so that we avoid over-differencing, which was observed in the ACF plot of the differenced series (left plot above) since the first autocorrelation is fairly close to -0.5.

Model Selection

Once we have selected a value for d (d=0 in this case), then we can use the AICc to select values of p, q, and determine whether we need a constant in our model. We will select candidate models based on the ACF

plots above and the candidate model with the lowest AICc will be the selected model.

Show 32 entries Search:

	p	d	q	constant	aicc
14	1	0	2	FALSE	595.5
20	2	0	1	FALSE	595.96
22	2	0	2	FALSE	597.46
16	1	0	3	FALSE	597.47
13	1	0	2	TRUE	597.51
19	2	0	1	TRUE	597.99
24	2	0	3	FALSE	598.97
15	1	0	3	TRUE	599.48
21	2	0	2	TRUE	599.48
23	2	0	3	TRUE	600.99
10	1	0	0	FALSE	602.82
26	3	0	0	FALSE	604.54
12	1	0	1	FALSE	604.57
18	2	0	0	FALSE	604.71
9	1	0	0	TRUE	604.85
4	0	0	1	FALSE	605.72
32	3	0	3	FALSE	605.74
6	0	0	2	FALSE	605.96
28	3	0	1	FALSE	606.25
8	0	0	3	FALSE	606.54
30	3	0	2	FALSE	606.56
25	3	0	0	TRUE	606.58
11	1	0	1	TRUE	606.6
17	2	0	0	TRUE	606.74
31	3	0	3	TRUE	607.58
3	0	0	1	TRUE	607.74
5	0	0	2	TRUE	607.99
27	3	0	1	TRUE	608.3
7	0	0	3	TRUE	608.58
29	3	0	2	TRUE	608.61
2	0	0	0	FALSE	640.66
1	0	0	0	TRUE	642.67

Showing 1 to 32 of 32 entries Previous 1 Next

The table above contains the AICc results of the ARIMA model selection process with d=0. We will select the model with the lowest (minimum) AICc which will give us the values for p, q, and whether we need a constant in our model. If we sort the table above by increasing AICc the first model is ARIMA(1, 0, 2) without a constant (min. AICc=595.5). *Note that normally with d=0, we should assume that the model needs a constant. However, in this case we have made seasonality adjustments to the data so we would like to allow*

the average value to be 0.

Modeling

The model we select based on the AICc criterion is ARIMA(1, 0, 2) without a constant. Now, we can fit the model to obtain the coefficients based on our data.

```
## Series: ys.data$visitors_adj
## ARIMA(1,0,2) with zero mean
##
## Coefficients:
##          ar1      ma1      ma2
##       0.9880  -0.7418  -0.2104
## s.e.  0.0123   0.0451   0.0434
##
## sigma^2 estimated as 0.1879:  log likelihood=-293.71
## AIC=595.42   AICc=595.5   BIC=612.33
##
## Training set error measures:
##                       ME      RMSE       MAE      MPE     MAPE     MASE
## Training set 0.003082728 0.4321955 0.1830168 115.5256 168.5945 0.851466
##                      ACF1
## Training set 0.002835995
```

The model is $x_t = 0.9880x_{t-1} + \varepsilon_t - 0.7418\varepsilon_{t-1} - 0.2104\varepsilon_{t-2}$, where x_t is the log seasonally adjusted visitors of Yellowstone National Park.

Model Diagnostics

The Ljung-Box test will compute hypothesis tests to see if our model is adequately representing our data. Since our data is monthly, we use yearly quantities (12, 24, 36, 48) to compute the cumulative fit.

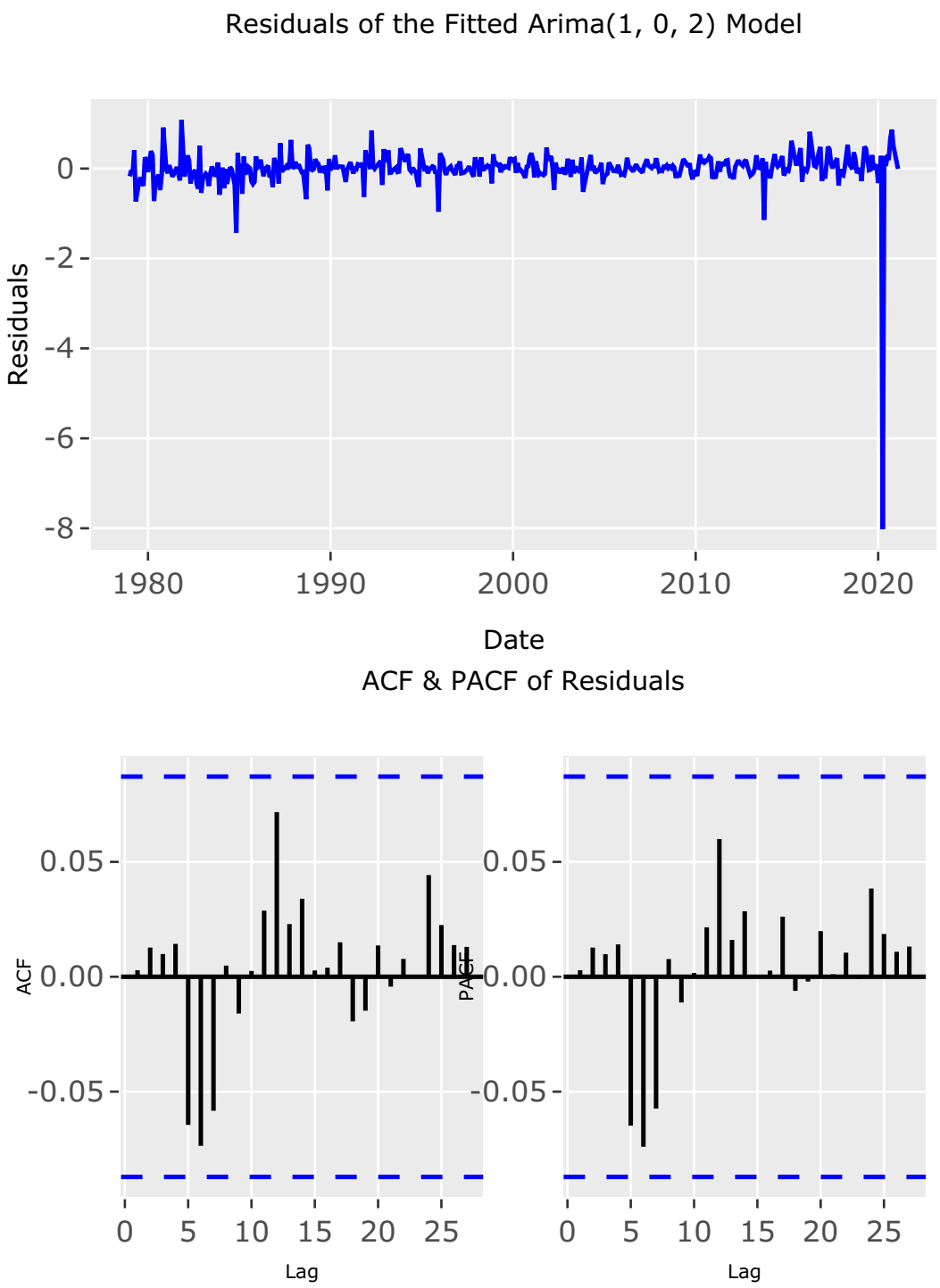
Show 4 entries Search:

	statistic	parameter	p.value	method
1	10.16	12	0.602	Box-Ljung test
2	12.66	24	0.9713	Box-Ljung test
3	15.28	36	0.999	Box-Ljung test
4	18.76	48	1	Box-Ljung test

Showing 1 to 4 of 4 entries Previous 1 Next

Since we do not observe any low p-values (p-value < 0.05) in the table above, we observe that our model is adequate according to these criteria.

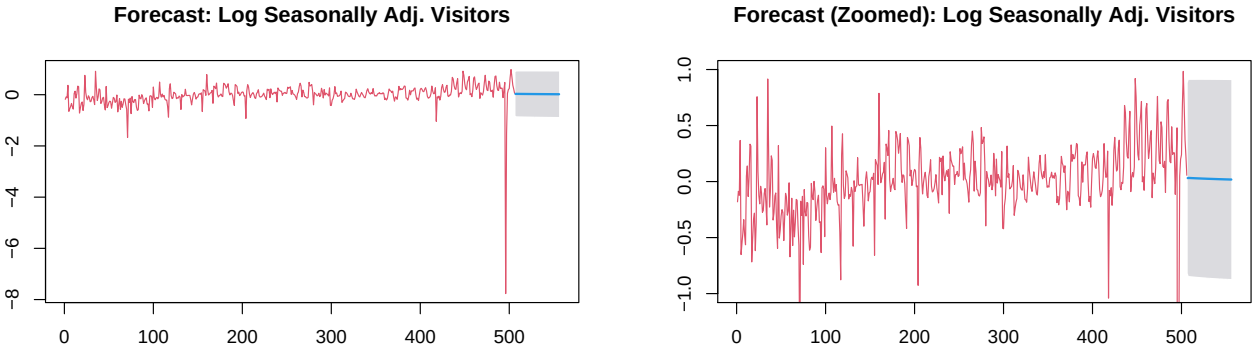
Another diagnostic check is to review the residuals of our model visually and compute the ACF and PACF using the residuals.



In each case, it appears that there is not a pattern in the residuals and no autocorrelations are significant. Thus, our model fits our data relatively well. However, given the simplistic seasonal transformation we have performed it is worth nothing that the residuals do seem to show a seasonal pattern with rises and falls over the course of each year.

Forecast

We can visualize the point forecast (blue line) along with the 95% confidence interval from lead time 1 to 50 in the graphs below (gray shaded region). Given that our model does not have a constant, the forecasts well into the future are reverting to zero. As the lead time increases, we would expect our forecasts to decrease in accuracy due to the cone of uncertainty phenomenon. In the right plot (zoomed), we can more clearly see that as the lead time increases, the forecast is reverting to zero quite quickly.



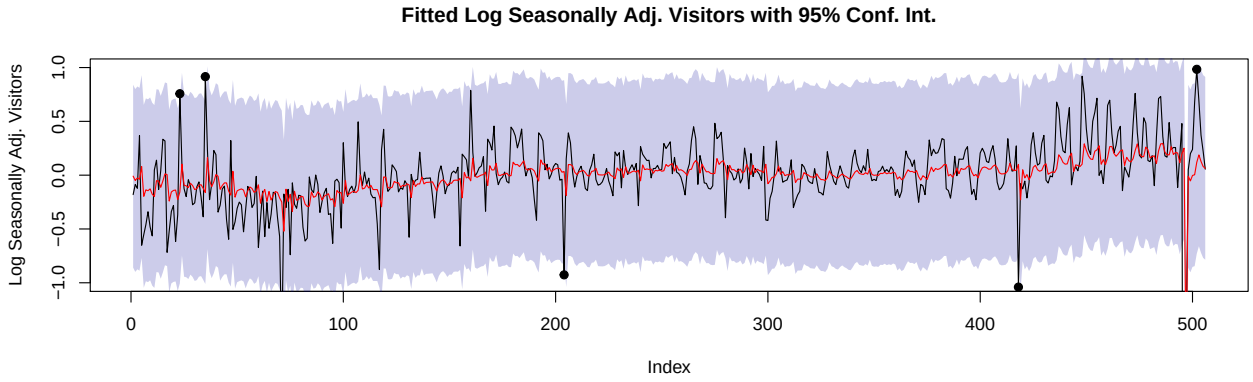
While it may be difficult to see for our forecast interval with lead time 50, both the upper and lower intervals grow in magnitude as h increases as expected based on the cone of uncertainty:

```
## [1] "Lo 95 Forecast First (1): -0.819 and Last (50) Value: -0.87"
```

```
## [1] "Hi 95 Forecast First (1): 0.881 and Last (50) Value: 0.906"
```

The upper and lower forecast intervals above grow in absolute value as the lead time increases. The lower bound decreases from approx. -0.819 to -0.870, and the upper bound increases from approx. 0.881 to 0.906.

We can further assess the forecast by reviewing the 95% confidence interval of the fitted values compared to the actualized data in the graph below. This way, we can assess whether the forecasts seem reasonable and/or excessively wide.



The forecast intervals may seem wide looking at the plots. However, given the volatility of the series (both in spring 2020 due to COVID-19 and more generally), it is a difficult series to predict, especially considering the month to month fluctuations. Because we are forecasting relatively far into the future (h=50), the cone of uncertainty comes into play. Since lead time, h, is relatively large, we expect the forecasts to get worse over time and the intervals need to be larger to accommodate our lack of information far into the future. Given the noisiness in the data, our forecast generally bets on reversion to the mean (0), especially for longer lead times.

Furthermore, the black line in the plot above signifies actual values, while the red line represents our forecast from the ARIMA model. The gray shaded region is the 95% confidence interval, and the black dots indicate when the actualized values do not fit in the interval. Looking at the graph above, the fitted interval misses the actualized data in April 2020, but also in multiple other occasions. Thus, it is likely necessary for the

forecast interval to be so wide to ensure that the actual stays in the forecast 95% of the time, especially given the current data volatility / uncertainty with the COVID-19 pandemic and the month to month variations. For instance, should another lock-down occur in the future, we would want our model to be robust for it. (The Discussion section below further examines this question.)

Discussion & Next Steps

While the outlier of the COVID-19 pandemic impacts our forecast, the national lockdown that resulted in zero visitors happened, and our model has tried to capture this moment in the historical series. While many macroeconomic indicators experienced a shock in spring 2020 and have taken time to recover, park visitation has had a different pattern. While there is an outlier in April 2020, parks traffic across the country generally *increased* during the pandemic as more people went outside this summer compared to previous years. It remains to be seen whether the higher visitation will be sustained in a post-pandemic world. We should incorporate this outlier into the model rather than ignore it since potential future shocks could result from other phenomena and we want our model to be robust to them.

Since our forecasts are made in log seasonally adjusted visitors, if we wanted to predict the actual visitors we could transform the forecasts back to non-seasonally adjusted by adding back the average to each forecast that corresponds to the month of the forecast and taking the exponential of the forecast to transform from log to visitor levels.

In future analyses, we could also try using other types of models; for example, non-linear models or seasonal ARIMA methods. Using seasonal differences instead of seasonal averages could better capture the visitation during modeling rather than using this simple data transformation. Finally, we can also assess whether volatility is changing in our dataset, which may require an ARCH model.