Andrea Bruckner
Predict 413: Midterm

Forecasting Mt. Rainier National Park Visitors

Introduction

Every year millions of people visit our country's national parks. The National Park System (NPS) is made of more than 400 official units, 59 of which are designated as national parks. Some of the other NPS units include national monuments, preserves, and historic sites. For this analysis, we will be focusing on Mt. Rainier National Park (MRNP), a 236,000-acre park located southeast of Seattle, Washington. Congress established MRNP on March 2, 1899, making it the nation's fifth oldest national park[i]. Last year nearly 2 million people visited the park. While weather conditions, forest fires, natural disasters, gas prices, and many other variables influence the total annual visitors to NPS sites, the number of visitors is on the rise. In 2015, more than 305 million people visited NPS sites, which is a four percent increase from the year prior[ii]. This year marks the 100th anniversary of the National Park System, and special centennial celebrations are likely to attract even more visitors this year.

Forecasting the number of visitors to national parks is important to ensure the safety of the visitors as well as the preservation of the park's natural features. MRNP is home to numerous state or federally protected species and contains various sensitive ecosystems. An increase in visitors puts a strain on the park and can lead to irreparable damage to the environment if the park is not equipped to manage the influx of visitors. More visitors also typically increase the likelihood of dangerous animal encounters. As more and more people visit national parks, animals grow accustomed to humans and begin to associate them with food. Visitors who do not store their food or dispose of their trash properly provide an easy food source for wild animals. Animals that associate humans with food can quickly become a threat to visitors. An increase in visitors calls for an increase in staff to ensure the wellbeing of the visitors and the park's natural features. Park rangers, environmental scientists, trail maintenance crews, and various other national park staff help maintain and protect the park, and they can help educate visitors on how to safely behave in the park to protect themselves and the park's environment.

In this report, I will be building two models to forecast the number of visitors to MRNP. These models could potentially be used for park planning or staffing purposes.

Data and Literature Review

The data set for this analysis comes from the Integrated Resource Management Applications Portal on the National Park Service's official website[iii]. I filtered the data to include only MRNP visitor data for all the years with monthly data available:1979 to 2015. There are 444 rows of data in total, representing each month for the 37 years of data examined. The original data set included a date variable plus nine visitor variables: Recreation Visitors, Non-Recreation Visitors, Concession Lodging, Tent Campers, RV Campers, Concession Camping, Backcountry Campers, Misc Campers, and Total Overnight Stays. I decided to focus on just the Recreation Visitors variable since it represents the visitors most typical to MRNP (unlike Non-Recreation Visitors, which include scientists and government officials), and because this variable has positive values for every month (unlike the various Campers variables that show a value of 0 for the winter months). After looking over the data, I divided the data set into two sets: a training set containing the years 1979 to 2009, and a testing set containing the years 2010 to 2015.

Prior to building my models, I completed some exploratory data analysis to familiarize myself with the data. In **Figure 1**, we can clearly see that the data is highly seasonal.

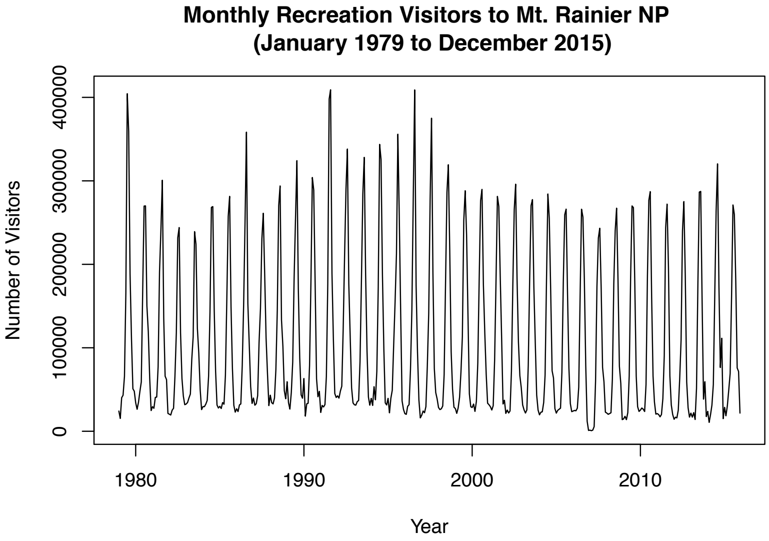


Figure 1: Plot of MRNP Recreation Visitors Data

I then created month and seasonal deviation plots to further explore the data. In **Figure 2**, the data makes a clear bell shape. We can easily distinguish the high season (July and August) from the low season (January to April and November to December). It appears that May, June, September, and October make up the shoulder season.

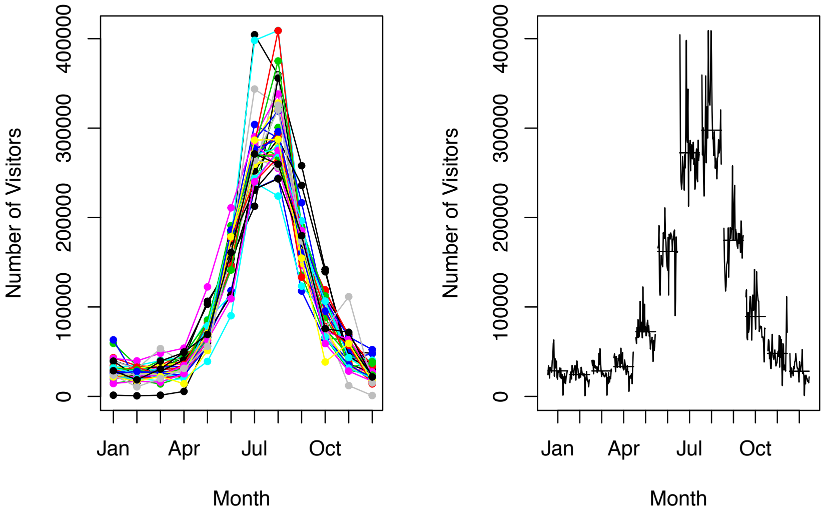


Figure 2: Month and Seasonal Deviation Plots

The seasonal deviation plot in **Figure 2** shows that the greatest variation for the number of recreation visitors to MRNP occurs during the high season, or July and August. As an avid backpacker, I interpret the years with a lower number of visitors during the high season as likely due to persistent snowpack. If the snow does not melt enough for visitors to hike on the trails during the peak months, the park will probably attract fewer visitors. Further analysis of weather data corresponding to the years of my data set might help confirm my snowpack theory, but this will have to wait for another analysis.

Because the number of visitors to MRNP is very seasonal, I decided to look into how other researchers forecasted seasonal visitor data, as well as look at sources that could provide more information about the process and importance of collecting data about visitors to national parks. I found an article in an economics journal detailing time series forecasting methods for international visitors to Thailand[iv]. Because of the long monsoon season, Thailand sees low numbers of tourists for about half the year but a surge when the weather is nice. The researchers used various forecasting methods, such as SARIMA, ARIMA, Holt-Winter-Additive, and Holt-Winter-Multiplicative, to predict the number of visitors. They also used other methods to forecast two variables: the number of international tourists and the country's economic growth rate. Their investigation of the relationship between tourists and economic growth inspired me to look into the correlation between MRNP visitor statistics and weather data. I only had a chance to look at plots of the correlation during my exploratory data analysis, but creating models using the visitor data and weather data are goals for the future.

After looking at the study in the economics journal, I looked at a study in an environmental management journal about how the number of visitors to a public land can affect the area's water quality[v]. Although the study focused on two rivers in Puerto Rico, its framework could be applied to MRNP data. Mt. Rainier is home to 27 glaciers that feed six major river systems[vi]. Investigating the impact visitors have on the park's waterways could help MRNP's natural resource and recreation managers decide whether to impose more stringent visitor caps during peak times.

In addition to the two journals mentioned above, I looked at a study conducted in 2003 by the National Park Service Social Science Program[vii]. This study described how forecasting visitor rates could help the parks plan and design visitor centers and other park facilities. It also mentioned that using GIS and looking at qualitative data of who is visiting can help each park improve its planning process.

Although the website where I retrieved my data provided some information on how the visitor data was calculated, I searched for more information on the NPS visitor-counting process. The National Parks Conservation Association provided more insight into how national parks calculate their visitor rates. In addition to automatic traffic counters, entrance tickets, camping permits, gift shop statistics, and manual counting, some parks also use planes to count the number of boats, kayaks, and other water vessels to factor these visitors into their counts[viii]. The article stressed that the numbers of visitors recorded for each national park are merely estimates since it is impossible to account for everyone who visits the park each year.

A final source I looked at when starting this project pertains specifically to the visitors who apply for a permit to hike and backpack on the Wonderland Trail, a 93-mile-long trail around Mt. Rainier[ix]. Typically visitors can apply for a permit in the last two weeks of March each year to hike or backpack on this trail, or they can try to secure a permit as a walk-in. In recent years, the number of permit requests has exceeded the number of available permits by 100 percent. (The park limits the number of people who can use the trail in order to protect the environment and ensure the safety of the hikers and backpackers.) This year, there was glitch in the permitting reservation system, so all permits are only available on a first-come, first-serve basis. Given that this is the centennial of the National Park System and permits for the Wonderland Trail are available only to walk-ins, the number of monthly visitors for 2016 might be hard to accurately predict. More people might visit the park hoping to secure a Wonderland Trail permit but end up having to settle on visiting the front-country sites of the park instead. I will not be predicting 2016 visitor statistics, but the centennial and permitting system glitch are both factors I would have to consider if I did.

Overall, the studies and articles I reviewed helped me form ideas on which kinds of models to create, as well as provide me with ideas on how estimating the number of visitors to MRNP can be useful to the park, if sometimes challenging.

Forecasting Models

For this analysis, I created two models: a seasonal naïve model and a time series multiple linear regression model. I chose these two types of models for various reasons. Because my data are very seasonal, creating a seasonal naïve model was an obvious appropriate modeling technique. I am also very familiar with linear regression models and was curious how this type of model might perform given the seasonality of the data.

For each of these models, I forecasted six time periods: the years 2010 to 2015. These are the years I included in my test set. Also, for the multiple linear regression model, I forecasted the number of visitors using the trend and season components of the data.

In **Figures 3 and 4**, we can see the six-year forecasts for each model in blue against the actual visitor data in red.

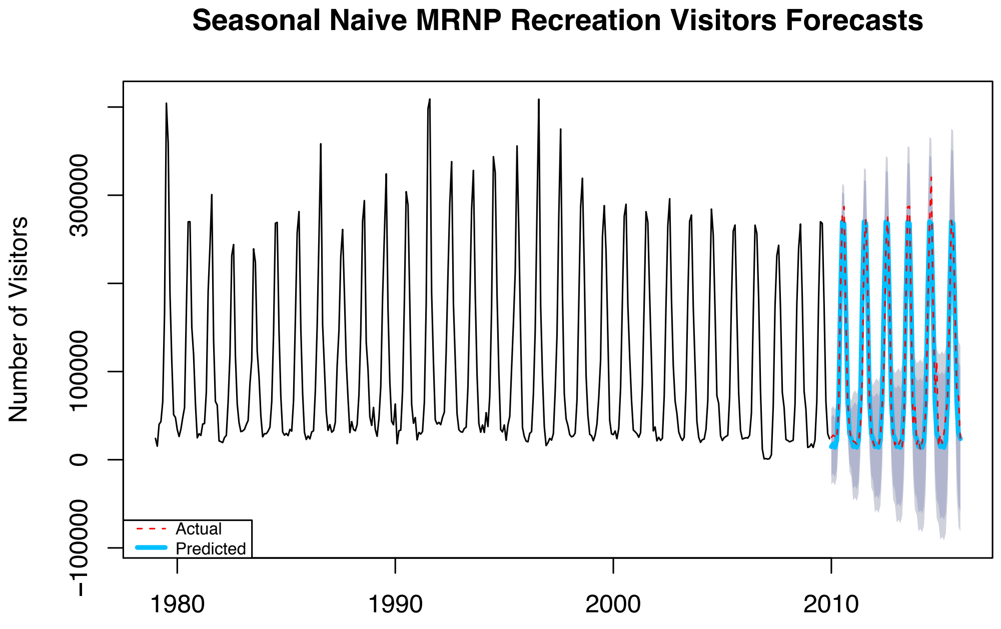


Figure 3: Seasonal Naive Forecasts

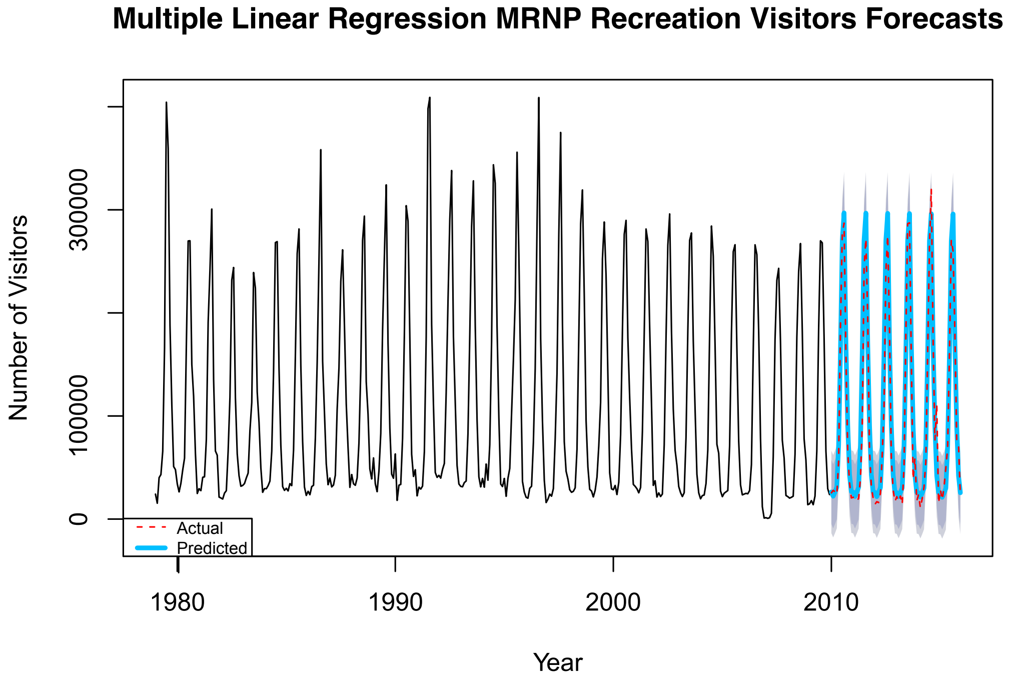


Figure 4: Multiple Linear Regression Forecasts

The confidence intervals for the seasonal naïve model in **Figure 3** appear to extend further than the confidence intervals for the multiple linear regression model in **Figure 4**. To confirm this impression, I looked at the numerical output for the confidence intervals. In **Tables 1 and 2** we can see the 80 percent and 90 percent confidence intervals for the 1-year forecasts for each model, as well as the point forecasts. Both models have a wide range of values for the confidence intervals for each month, and both show negative values for the lower bounds of the confidence intervals for the colder months. Though the confidence intervals for the remaining years I forecasted (2011 to 2015) are not printed in the tables due to space constraints, I observed that the range of the confidence intervals increased with each year for both models. This suggests that neither model is very good at predicting the number of visitors too far into the future.

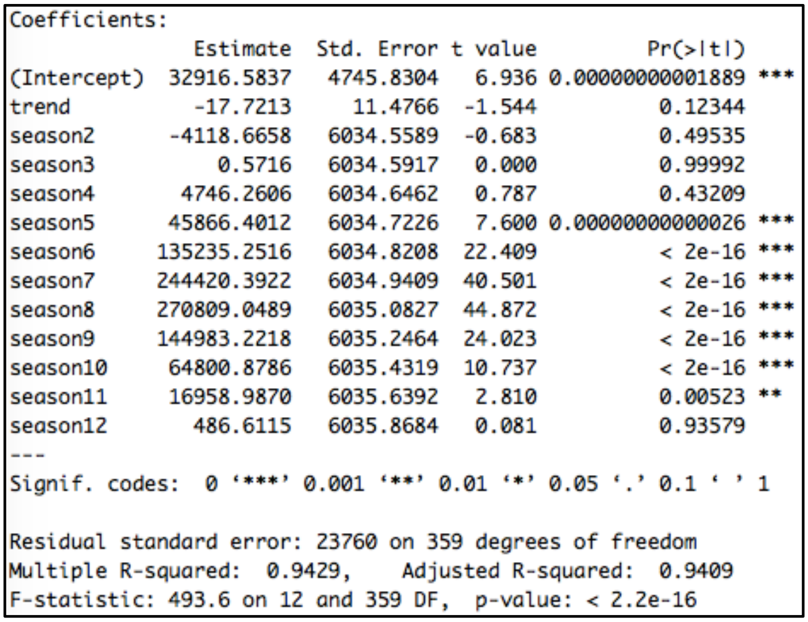
Seasonal Naïve Model					
Date	Point Forecast	Lo 80	Hi 80	Lo 90	Hi 90
Jan 2010	14766	-18736.0105	48268.01	-28233.365	57765.36
Feb 2010	17919	-15583.0105	51421.01	-25080.365	60918.36
Mar 2010	13917	-19585.0105	47419.01	-29082.365	56916.36
Apr 2010	22254	-11248.0105	55756.01	-20745.365	65253.36
May 2010	66161	32658.9895	99663.01	23161.635	109160.36
Jun 2010	184078	150575.9895	217580.01	141078.635	227077.36
Jul 2010	269951	236448.9895	303453.01	226951.635	312950.36
Aug 2010	267980	234477.9895	301482.01	224980.635	310979.36
Sep 2010	175431	141928.9895	208933.01	132431.635	218430.36
Oct 2010	65988	32485.9895	99490.01	22988.635	108987.36
Nov 2010	29473	-4029.0105	62975.01	-13526.365	72472.36
Dec 2010	23736	-9766.0105	57238.01	-19263.365	66735.36

Table 1: Seasonal Naive Model Confidence Intervals

Multiple Linear Regression Model					
Date	Point Forecast	Lo 80	Hi 80	Lo 90	Hi 90
Jan 2010	26306.55	-4813.683	57426.78	-13665.483	66278.58
Feb 2010	22170.16	-8950.0701	53290.39	-17801.87	62142.19
Mar 2010	26271.68	-4848.5539	57391.91	-13700.354	66243.71
Apr 2010	30999.64	-120.5862	62119.87	-8972.386	70971.67
May 2010	72102.06	40981.8331	103222.29	32130.033	112074.09
Jun 2010	161453.19	130332.9622	192573.42	121481.162	201425.22
Jul 2010	270620.61	239500.3815	301740.84	230648.582	310592.64
Aug 2010	296991.55	265871.317	328111.78	257019.517	336963.58
Sep 2010	171148	140027.7686	202268.23	131175.969	211120.03
Oct 2010	90947.93	59827.7041	122068.16	50975.904	130919.96
Nov 2010	43088.32	11968.0912	74208.55	3116.291	83060.35
Dec 2010	26598.22	-4522.0056	57718.45	-13373.806	66570.25

Table 2: Multiple Linear Regression Model Confidence Intervals

Although predicted values in the plots in **Figures 3 and 4** looked similar, the point forecasts in **Tables 1 and 2** confirm that the multiple linear regression model predicts a higher number of recreation visitors overall, while the seasonal naïve model's predictions match the last year of the training data set, 2009. Looking at accuracy metrics and residuals will help us determine which of these models more accurately forecasts the number of recreation visitors to MRNP, but before we examine those metrics, it is important to look at the coefficients and adjusted R-squared value for the multiple linear regression model.



```
Coefficients:
               Estimate  Std. Error t value         Pr(>|t|)    
(Intercept)  32916.5837   4745.8304   6.936 0.00000000001889 ***
trend          -17.7213     11.4766  -1.544          0.12344    
season2      -4118.6658   6034.5589  -0.683          0.49535    
season3          0.5716   6034.5917   0.000          0.99992    
season4       4746.2606   6034.6462   0.787          0.43209    
season5      45866.4012   6034.7226   7.600 0.00000000000026 ***
season6     135235.2516   6034.8208  22.409          < 2e-16 ***
season7     244420.3922   6034.9409  40.501          < 2e-16 ***
season8     270809.0489   6035.0827  44.872          < 2e-16 ***
season9     144983.2218   6035.2464  24.023          < 2e-16 ***
season10     64800.8786   6035.4319  10.737          < 2e-16 ***
season11     16958.9870   6035.6392   2.810          0.00523 ** 
season12       486.6115   6035.8684   0.081          0.93579    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23760 on 359 degrees of freedom
Multiple R-squared:  0.9429,	Adjusted R-squared:  0.9409 
F-statistic: 493.6 on 12 and 359 DF,  p-value: < 2.2e-16
```

Figure 5: Multiple Linear Regression Results

In **Figure 5**, we can see that the coefficients increase until August (represented by season8), and then decrease. This echoes the pattern we saw in **Figure 2**: the number of visitors increased as the months get warmer, peaked in August, and then decreased as the months get colder. February (represented by season2) is negative, indicating that fewer people visit MRNP during that month. January is not represented in **Figure 5**, but if it were, it would likely show a very low or negative value as well.

The results above also show a very high adjusted R-squared value, 0.9409. This value means that 94.09 percent of the variation in the number of annual recreation visitors to MRNP is accounted for by this model.

Model Selection

Looking at plots of the models, such as **Figures 3 and 4**, is not enough to determine which model performs better. Examining various metrics, such as RMSE and MAE, as well as the models' residuals, can give us a better indication of which model is better at forecasting the number of visitors to MRNP.

In **Tables 3 and 4**, we can see various accuracy metrics for each model. The metrics for each model indicate that the models are both best at predicting the number of visitors for a 1-year forecasting period. The RMSE and MAE values are lowest for this time period, indicating the models' predictions contain fewer errors.

Seasonal Naïve Model								
Data Set	**Forecast Period**	**ME**	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**	**ACF1**
Training Set	N/A	-1014.025	26141.737	16410.947	-31.023	48.349	1.000	0.243
Test Set	1 Year	3341.667	9756.444	7999.833	9.516	16.663	0.487	-0.062
Test Set	2 Years	-3055.208	22272.110	12015.290	-0.780	20.525	0.732	0.414
Test Set	3 Years	-4883.361	22778.830	12090.140	-3.825	18.818	0.737	0.364
Test Set	4 Years	-3727.146	21357.430	12515.940	-4.602	21.371	0.763	0.290
Test Set	5 Years	-1104.967	23254.820	13765.470	-3.018	23.121	0.839	0.206
Test Set	6 Years	267.764	22417.300	13590.240	0.453	22.979	0.828	0.205

Table 3: Seasonal Naive Model Metrics

Multiple Linear Regression Model								
Data Set	**Forecast Period**	**ME**	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**	**ACF1**
Training Set	N/A	0.000	23339.220	14722.705	-32.184	45.338	0.308	0.392
Test Set	1 Year	1022.287	12586.730	9120.371	-4.840	16.223	0.191	N/A
Test Set	2 Years	-10202.540	21375.240	15356.130	-22.371	26.476	0.936	0.280
Test Set	3 Years	-11924.370	21654.160	16172.160	-26.716	29.890	0.985	0.215
Test Set	4 Years	-10661.820	21194.120	15972.540	-27.671	31.888	0.973	0.148
Test Set	5 Years	-7933.316	21474.340	15482.930	-25.080	30.985	0.943	0.099
Test Set	6 Years	-6454.257	20617.760	14684.130	-20.202	28.365	0.895	0.074

Table 4: Multiple Linear Regression Model Metrics

Determining which of these two models is better at predicting the number of recreation visitors to MRNP largely depends on how many forecasting periods we look at and which metrics we consider. For instance, if we want a model that can predict the number of people who visit MRNP in 1 year, the seasonal naïve model is better since its metrics, namely RMSE and MAE, are lower than the metrics for the multiple linear regression model. However, if we want to forecast a longer period of time, it is not as clear which model is better. The RMSE values for a 6-year forecast are overall smaller for the multiple linear regression model than for the seasonal naïve model. However, the MAE values are smaller for the seasonal naïve model. Looking at the metrics can lead to conflicting conclusions. Even if we consider all of the accuracy metrics presented in the tables above, deciding which model is better is a toss-up because about half of the metrics suggest the seasonal naïve model is better, while the other half suggest the multiple linear regression model is better.

Looking at residual plots can help us also get a better sense of which model might be better. In this case, however, the plots in **Figures 6 and 7** do not definitively steer us to one model over the other. The residuals for both models appear random, have a mean zero, and generally do not indicate any cause for concern.

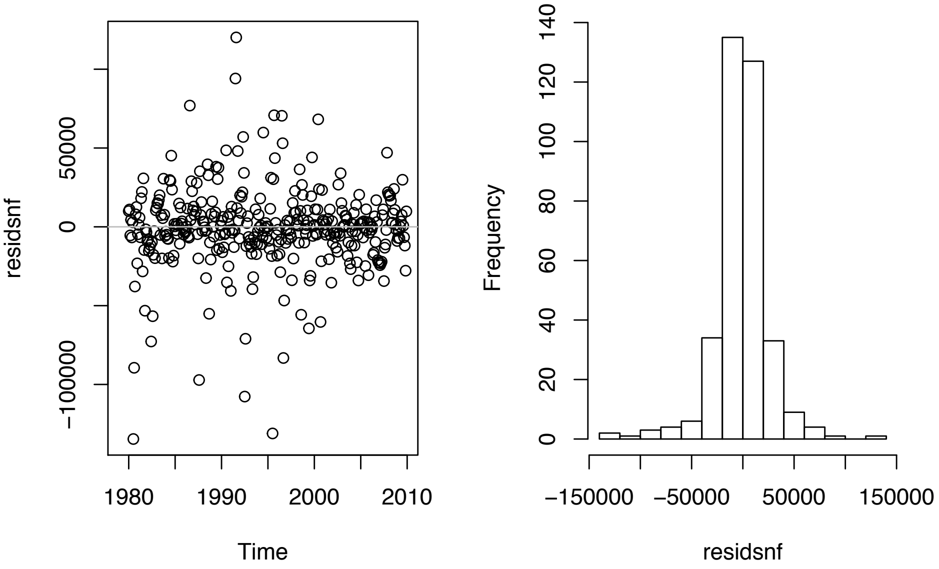


Figure 6: Seasonal Naive Model Residuals

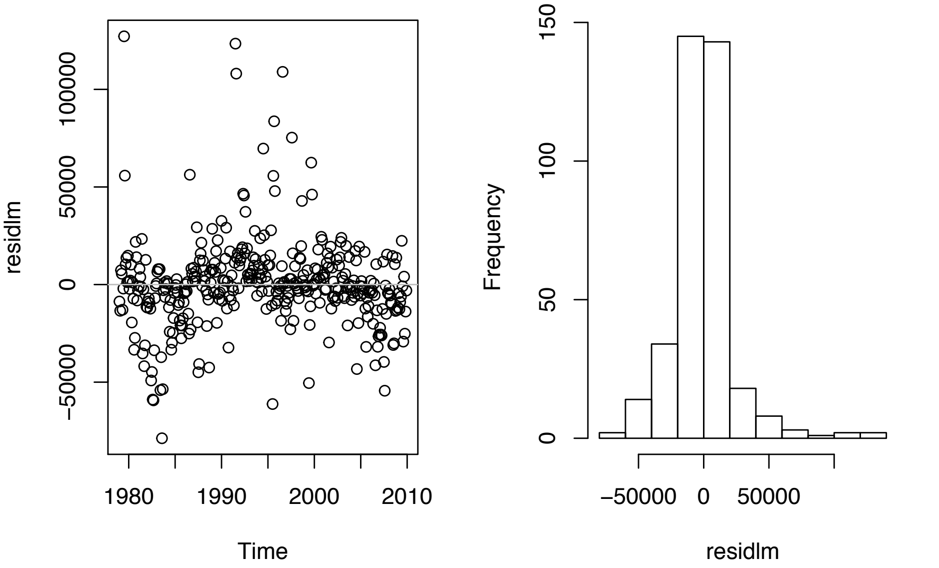


Figure 7: Multiple Linear Regression Model Residuals

If I had to choose just one model to deploy, I would lean more toward the multiple linear regression model. It scores better than the seasonal naïve model according to the RMSE metric, and its confidence intervals have smaller ranges than the seasonal naïve model's confidence intervals. I also like the fact that this model is relying on trend and season for its predictions, rather than just the last year's data. As **Figure 1** depicted, the number of visitors to MRNP vary each year, sometimes significantly. Being able to factor in this variation for forecasting the number of visitors is more appealing than relying on just the previous year's data.

Limitations and Lessons Learned

While working with this data, I discovered several limitations. As I mentioned in my literature review, the visitor statistics published by the National Park Service are educated estimates. They do not reflect the actual number of visitors per year, as that value is very difficult to measure. There is nothing I can do about this limitation other than acknowledge it. Another limitation is not having data available regarding the number of visitors per day because the National Park Service publishes only monthly and yearly data. Being able to work with daily data might reveal more patterns and allow me to match up some of the various factors that probably influence the number of visitors to the park, such as forest fire data, weather data, and fuel price data. While I did try to work with weather data (as evident in my R code), I ultimately abandoned it, the reasons which I will discuss shortly. Having daily visitor data would open up more possibilities for discovery.

Because some areas of MRNP are not accessible in certain months due to heavy snow or rainfall, I thought that looking at weather data might help explain some of the yearly fluctuations in the number of visitors. I gathered monthly data from the National Oceanic and Atmospheric Administration for the same time period as my NPS data (January 1979 to December 2015). I chose the Paradise Ranger Station in MRNP as the source for the weather data. Seven dates from my weather data set were missing, so I imputed values for the missing dates using the means from the same month the year before and the year after the missing dates. I created flags for the imputed values. I also ended up correcting one existing value as well since it was stored as -9999 for the amount of snowfall. Given that the date was in August, I corrected the value to 0 since no August dates in the entire data set had any snow reported. I then merged the weather data set with the MRNP visitors data set. In spite of all of my data cleaning, I ended up not using the weather data in my models because I could not figure out how to call specific variables, such as rain, once the data was in time series format. I later learned that I should have used the "fpp" package as it is more flexible than the "forecast" package for this kind of analysis. Although I did not end up using the weather data in my models, I learned a lot from cleaning it.

Conclusions

The two models I created both did a reasonably good job at predicting the number of visitors to MRNP for the time periods selected. As I mentioned in the Model Selection section, the accuracy metrics and the residual plots for each model did not clearly favor one model over the other. I ultimately chose the multiple linear regression model as the best model because it relied on trend and season for its predictions rather than just the last year of data. It also had smaller confidence intervals and a better RMSE, my preferred metric. However, the seasonal naïve model performed nearly equally well and could easily be considered the best model if I had prioritized different metrics.

While the models I created both performed well, I am curious how they would fare predicting 2016 visitor statistics. I think the centennial celebrations and the permitting system glitch at MRNP would make accurate predictions very difficult. I am also curious how incorporating weather data into my models would affect predictions. I think dividing the data into high, low, and shoulder seasons could also yield interesting results. Theses are all paths I hope to explore with my data in the future in the future.

References

[i] National Park Service. (n.d.). Frequently Asked Questions. Retrieved from https://www.nps.gov/mora/faqs.htm

[ii] National Park Service. (2016, January 27). America's National Parks: Record Number of Visitors in 2015. Retrieved from https://www.nps.gov/aboutus/news/release.htm?id=1775

[iii] National Park Service. (n.d.). NPS Stats. Retrieved from https://irma.nps.gov/Stats/SSRSReports/Park%20Specific%20Reports/Summary%20of%20Visitor%20Use%20By%20Month%20and%20Year%20(1979%20-%20Last%20Calendar%20Year)?Park=MORA

[iv] Chaitip, P., Chaiboonsri, C., & Mukhjang, R. (2008). Time Series Models for Forecasting International Visitor Arrivals to Thailand. *International Conference on Applied Economics*. Retrieved from https://www.researchgate.net/file.PostFileLoader.html?id=57319a145b49520ed668dbed&assetKey=AS%3A360110745505792%401462868500296

[v] Santiago, L. E., Gonzalez-Caban, A., & Loomis, J. (2008, January 28). A Model for Predicting Daily Peak Visitation and Implications for Recreation Management and Water Quality: Evidence from Two Rivers in Puerto Rico. *Springer Science+Business Media*. Retrieved from https://warnercnr.colostate.edu/docs/ess/biocomplexity/Luis_et_al_Env_Mgmt.pdf

[vi] National Park Service. (n.d.). Glaciers. Retrieved from https://www.nps.gov/mora/learn/nature/glaciers.htm

[vii] Gramann, J. H. (2003, August). *Visitation Forecasting and Predicting Use of NPS Parks and Visitor Centers: Focus Group Report*. Retrieved from https://www.nature.nps.gov/socialscience/docs/archive/NPS_Forecasting_Report.pdf

[viii] Brulliard, N. (2016, March 2). *Who Counts? A Closer Look at Parks' Record Visitation Numbers*. Retrieved from https://www.npca.org/articles/1146-who-counts-a-closer-look-at-parks-record-visitation-numbers

[ix] Yuasa, M. (2016, March 24). *Reservation-system failure at Mount Rainier a game changer for hikers, climbers*. Retrieved from http://www.seattletimes.com/life/outdoors/online-registration-failure-for-wonderland-trail-back-country-permits-could-result-in-long-lines-at-ranger-stations/