Nattanai Na Songkhla (nn14), Ram Sabyrkulov (rs46), Ibtisaam Dalvi (idalvi2)

Regression Analysis of Top Three National Parks in USA Data

Abstract

This report presents an analysis and predictions of monthly visitor numbers in the top three national parks of the USA. Employing regression and spectral analysis techniques, our study aimed to elucidate patterns and factors influencing visitor trends. The methodologies applied facilitated a comprehensive understanding of the dynamics driving monthly visitation, contributing valuable insights for park management and tourism planning.

Introduction

As a group of future data scientists, our collaboration is not just founded on our professional expertise but also on a shared and deeply rooted passion and love for nature. Surprisingly, all of us were left speechless after viewing the beauty of Yellowstone, Yosemite, and Grand Canyon National Parks. This common ground inspired us to conduct a time series analysis and find helpful data insights that could help us understand and predict the dynamics of visitor trends in the top three national parks in the USA.

The primary focus of our study is to analyze and forecast the monthly visitor count for each of these top national parks over the next five months. National parks are not just scenic retreats but vital components of our ecological system, cultural heritage, and economic structure. These parks are not just natural reserves but also serve as critical economic drivers through tourism and related activities. In our studies, we will include weather and gasoline prices to help understand the dynamics of the visitors better, we believe that accurate predictions can aid in efficient park management, resource allocation, and enhancing visitor experiences.

Dataset Introduction

We utilized three distinct datasets for our analysis; the National Park dataset, the Weather Temperature dataset and the Gasoline Price dataset.

The **National Park** dataset, sourced from the National Park Service Statistics, was employed to forecast the annual visitor count for specific national parks, namely "Yellowstone," "Yosemite," and "Grand Canyon." This dataset spans the selected years from 2000 to 2022 and includes the following columns for each park: Year (covering selected years), Month (comprising twelve months for each year), and N_visitor (indicating the number of visitors per month per year).

The **Weather Temperature** dataset, serving as an independent variable in our analysis, incorporates average temperature data by state and originates from the NOAA National Centers for Environmental Information. Its columns include Date (formatted as YYYY-MM-DD), Value (indicating the average temperature in Fahrenheit per month per year), and Anomaly (representing the difference from an average or baseline temperature).

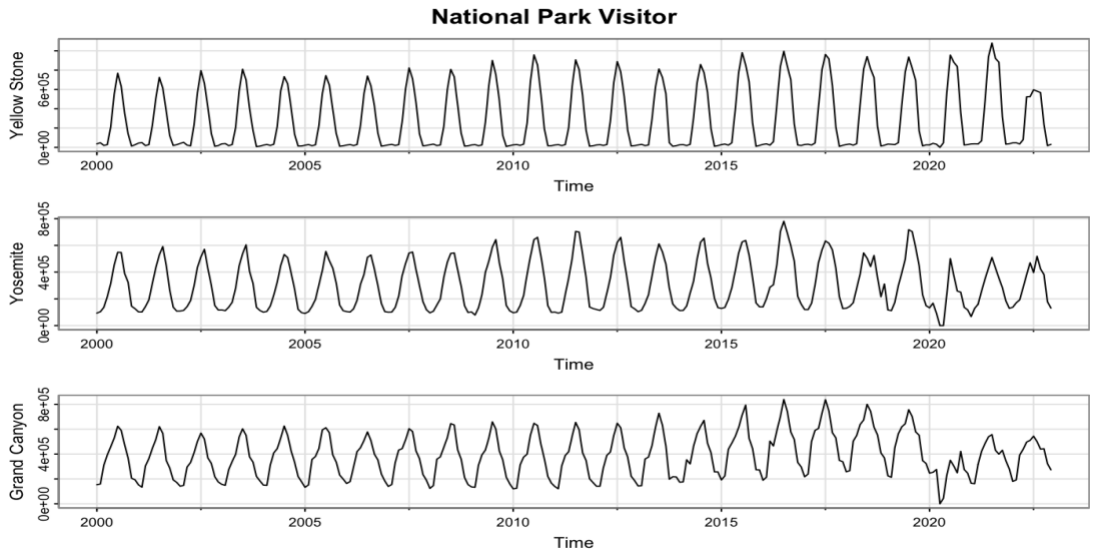
In addition to the Weather Temperature dataset, we integrated the **Gasoline Price** dataset as another independent variable in our analysis, since driving is quite required for most national parks, and we could make a hypothesis that gasoline price may affect the number of national parks visitors. The dataset includes the following columns: Month (formatted as MM-YY), Price (indicating the average price in US dollars per gallon), and Change (depicting the month-to-month price change).

Statistical Methods

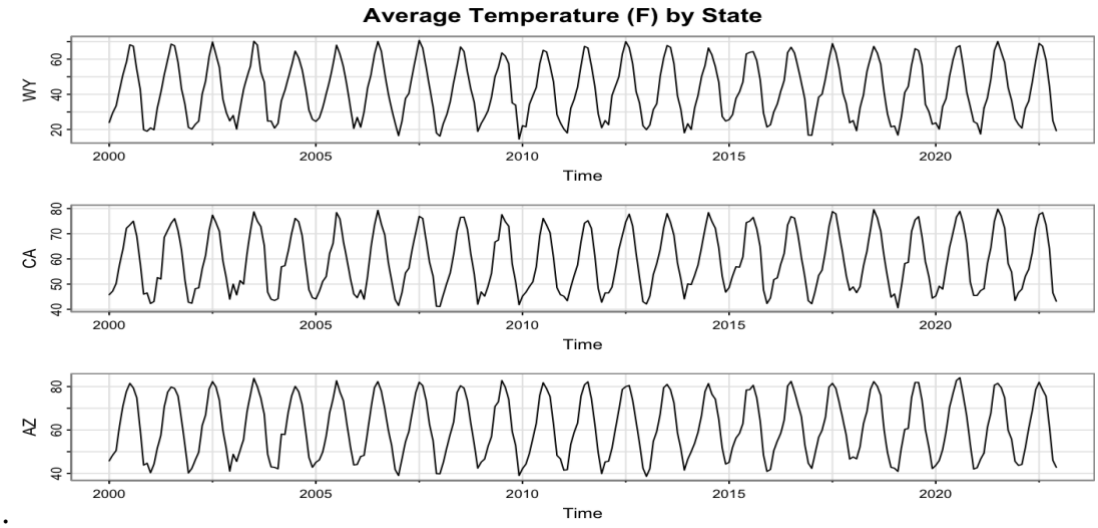
We first prepared our data by normalizing the year data by a difference of 2011 and split the data between 2000-2020 as a train set, and 2021-2022 as a test set.

Time Series Plot

According to the national park visitor plot, we found that all of these national parks show a repeating number of visitor trends across the year (peak during summer and dip in winter).

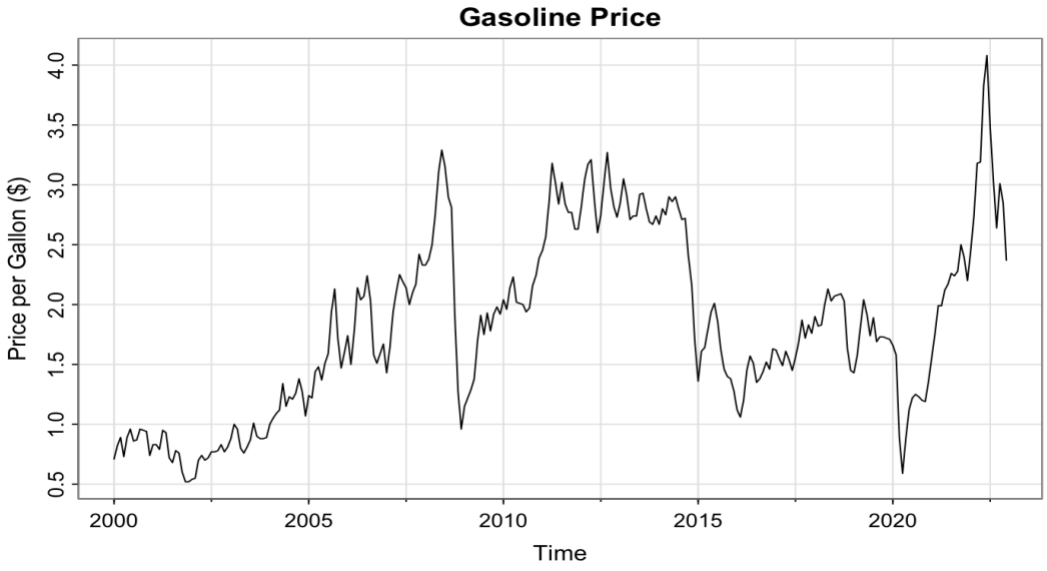


Fortunately, average temperatures also show a similar trend to the number of visitors, which is a good indicator for making a regression model as can be seen in the average temperature by state plot.



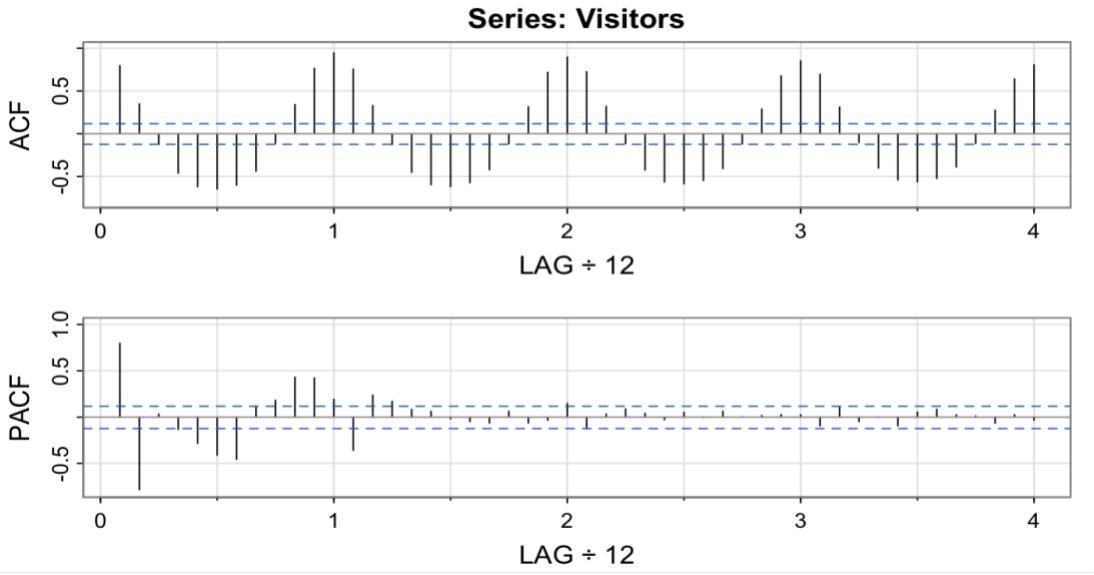
.

According to the ‘Gasoline Price’ plot, it seems to show no obvious trend that looks like the trend of national parks visitors. However, it may be related and significant for predicting the number of national visitors.



ACF/PACF

From our ACF and PACF plot we observed a seasonal **lag of 12 months.** Therefore, we made the decision to perform regression analysis on both with lag and without lag models.



Regression Analysis

Our 2 regression models with and without lag-12 for each parks

- visitors ~ year + month + temp + gasoline_price
- visitors ~ year + month + temp + gasoline_price + lag-12

Comparison:

National Parks	Model	R2	RMSE	AIC	BIC
Yellowstone	lag	0.9815	178267.2	5843.666	5902.837
	w/o lag	0.968	136159.9	6241.677	6298.085
Yosemite	lag	0.8974	68378.27	5996.685	6055.856
	w/o lag	0.8989	85304.81	6261.561	6317.969
Grand Canyon	lag	0.8193	90129.46	6116.406	6175.577
	w/o lag	0.8084	95837.91	6405.864	6462.271

We observed that the RMSE, AIC and BIC for models with lag-12 performs better. Hence, we selected this model to perform our forecasting.

Forecasting Result (Jan - May 2023)

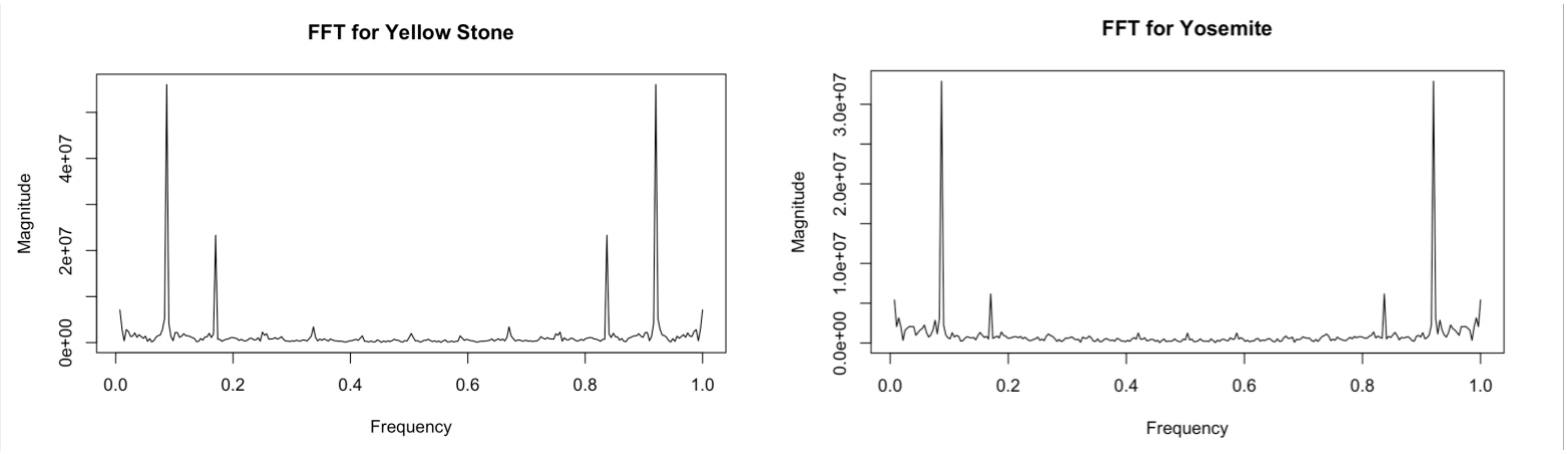
National Parks	Value	JAN	FEB	MAR	APR	MAY
Yellowstone	forecast	56,540	59,450	38,741	79,912	475,826
	real	45,709	45,717	30,044	69,247	454,262
Yosemite	forecast	149,965	159,811	191,209	265,351	378,542
	real	107,256	107,012	25,005	205,802	322,308
Grand Canyon	forecast	200,278	207,000	394,630	433,321	486,099
	real	134,361	151,395	326,916	457,189	492,037

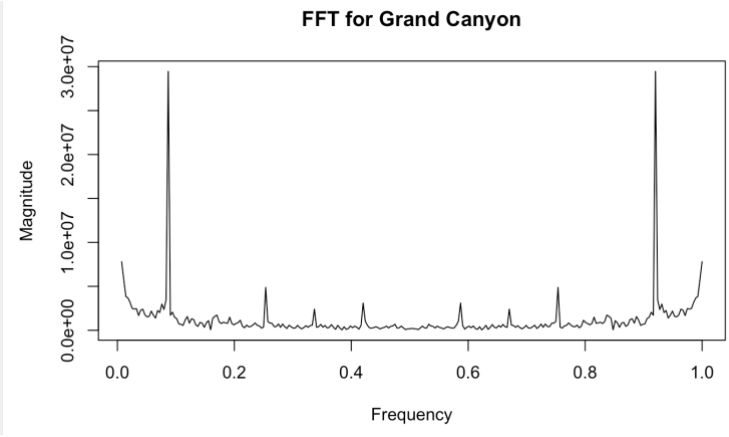
As our data contained values up to 2022, we performed forecasting for the next 5 months. results we observe that our model is close to the real values but still slightly over predicts the number of visitors for each national park.

Spectral Analysis

Given that we have monthly visitor data from 2000 to 2022 for three National Parks, *Spectral Analysis* can be used to identify cyclic patterns or any dominant frequencies in visitation patterns.

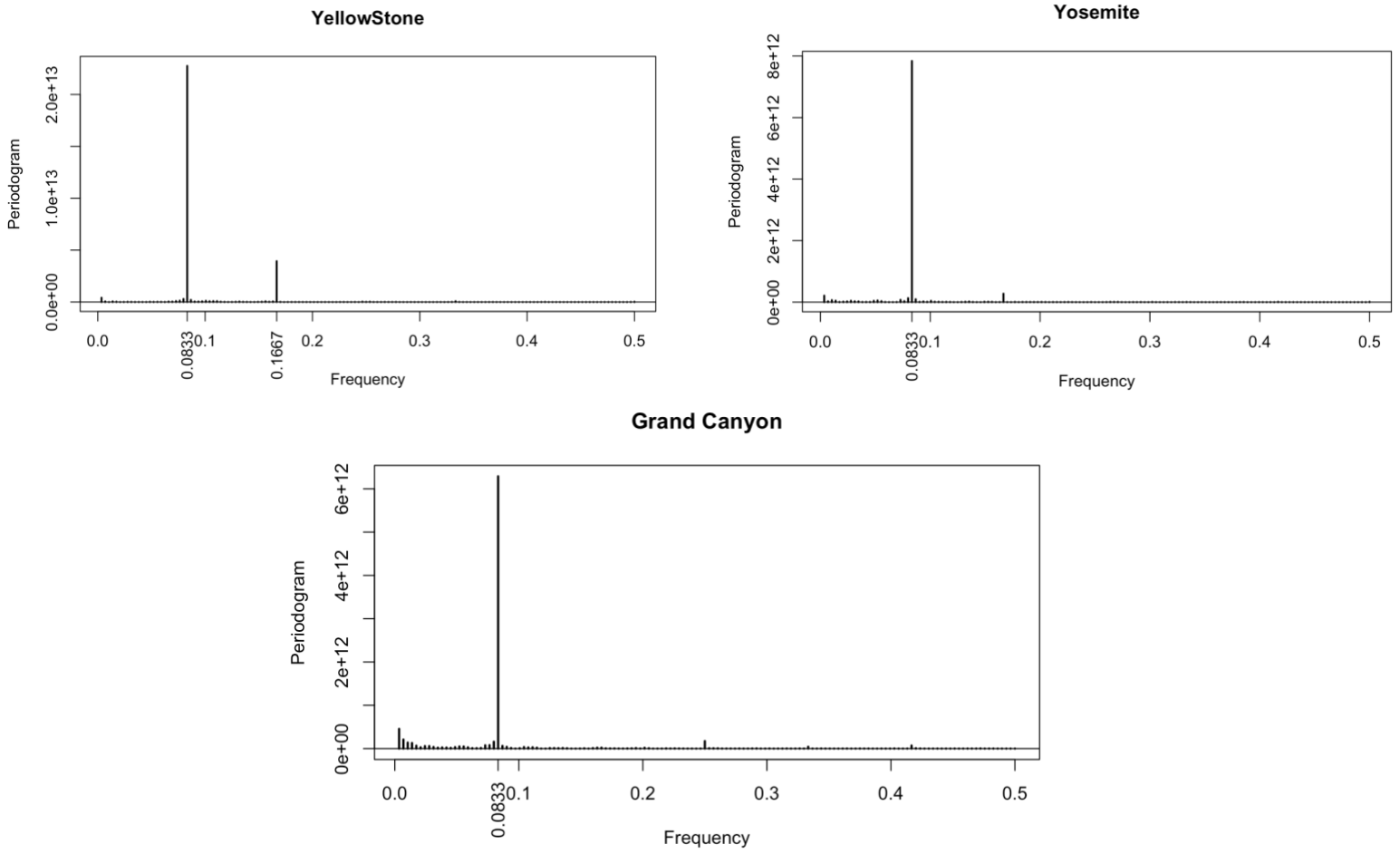
First we applied the Fast Fourier Transform (FFT) to the time-series data of visitor counts for each park. The plots for each park with x-axis represents the frequency, and the y-axis represents the magnitude of each frequency component. The peaks in the plot tell us the presence of a strong frequency component within the data. In our case, we see the dominant annual pattern of visitors due to one high peak in the year.





Afterwards, we conducted the periodogram analysis and clearly saw that all three national parks show the main frequency of **0.0833** that confirms the annual cycle in visitor numbers.

However, while the annual pattern is consistent across all parks the relative strength of parks varies, it can be explained by each park's unique environmental and recreational offerings, as well as their geographic locations.



Overall, we see that Spectral Analysis confirmed our assumption that each park experiences an increase in visitors numbers during the summer season and holiday days. Spectral analysis with combination of the regression analysis vital for park management team that can be used in efficient allocation distribution, staff employment, construction planning and optimization of visitor experience.

Our predictive results are helpful to ensure that the park can accommodate the influx of visitors without compromising the visitor experience or environmental sustainability, and during off-peak times, the park may offer discounts or special events to attract visitors and balance the annual visitation distribution.

Discussion

Our analysis findings and the forecast model for each national park shows pretty accurate results that support our study in predicting the visitors numbers by months. It brings us closer to an easier strategic decision in terms of park financial distributions. A clear picture of expected number of visitors provides park management a chance to plan ahead for construction road works, improvement of park activities and human resources. Moreover, it helps to plan marketing campaigns for each park for each month, special events and activities, and an efficient discount program for each month. It allowed us to discover significant and insignificant variables in predicting the dynamics of attendance.

Our model significance indicates that temperature seems to not significantly affect the number of visitors as it has p-value greater than 0.05 for all three national parks, this may be resulted from the high correlation between the temperature and visitors numbers, whose trend is mostly captured by the time series variable, like year and month.

On the other hand, gasoline prices are significant for only Grand Canyon national park. While other national parks like Yosemite and Yellowstone are more likely to be a sole destination for the trip that could take many days for visitation, Grand Canyon seems to be an on the way visiting destination during the road trip for people who come to California or Las Vegas. Additionally, the gasoline price in California is also relatively higher than other states, thus gasoline price may be one of the major considerations that affect the decision of people who want to travel to the Grand Canyon for the road trip or sightseeing.

As a group of nature enthusiasts, we are happy with our outcome, and willing to continue working on our research to bringing more insights by studying the relationship of the hotel cost and flight cost to these three national parks, so the local business would be able to accommodate their resource in best way to gain more benefit form our study.

Limitations

We wanted to include flight data and employee data however, the data sets were expensive to purchase. The absence of flight data limits our ability to explore correlation between travel patterns and park visitations. Additionally, employee data would have helped us understand staffing levels and its impact on park management.

Resources

1. National Report website (https://irma.nps.gov/Stats/Reports/National)
2. NOAA National Centers for Environmental Information website (https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/statewide/time-series)
3. Index Mundi website (https://www.indexmundi.com/commodities/?commodity=gasoline&months=300)