National Park Traffic Exploration Tool

Team 037: *Anh Tran, Bryan Truong, Chase Diaz, Olivia Hadlaw, Priyal Patel*

1. Introduction:

US national parks are turning out to be popular destinations which can be an unwelcome surprise for visitors who weren't expecting crowds. In addition, understanding and predicting levels of park traffic is important for the US National Park Service (NPS), as it must take steps to accommodate surges in expected park traffic. The goal of this project was to develop a tool that allows visitors to explore expected levels of park traffic during various times of the year as they make travel decisions.

2. Problem Definition

Currently there is no single, comprehensive and interactive source for information on expected levels of traffic at all the national parks and visitation forecasts are not easily accessible to potential visitors so visitors must seek this information themselves; this can be time consuming and can lead to inaccurate findings.

3. Literature Survey

Future visitation for the 47 national parks in the United States is estimated to increase by 1.6 million a year through 2026 (Bergstorm 2020) and accessibility of park resources has impacted tourist satisfaction negatively (Shen 2019), along with congestion and social carrying capacity (McIntosh 2011) but it's not always easy for travelers to find information on expected traffic levels. These studies explore the problem and forecast future visitation levels and their consequences but there is no solution, such as the one this project aims to develop, that gathers and presents the information tourists need when planning.

While more people are enjoying the beauty offered by these parks, a surge in visitation would diminish the overall quality of the visitor's experience at the national park. Research has been done exploring the impacts of overcrowding and potential fixes (Timmons, 2018), as well as visitor waiting preferences upon arriving to parks (Newton et al., 2015). Visitation estimation through geotagged photographs on Flickr was also explored (Session, et al., 2016). These studies are all useful for our project, as they highlight the growing overcrowding problem and provide ways to improve the visitor experience. However, they also do not provide a readily available visualization of the data provided by the park data to allow potential visitors to easily understand and make data-driven decisions themselves when deciding when to visit.

The tool to be developed will focus on presenting travelers with traffic data at a US National Park at a time of their choosing so they can make decisions based on how crowded a park is expected to be but there are several studies that have explored other factors important to visitors. Musonera Abdou conducted a study finding that price, park-specific qualities, and month of visit were the factors with the most power in predicting visitation to national parks in Rwanda (Abdou, 2022). Marjo Neuvonon studied the relationship between park-specific characteristics and visitation at 35 national parks in Finland and found characteristics such as extent of the trail network, and population density of the region affect visitation behavior (Neuvonon, 2010). Additionally, Choe, Schuett and Sim (2017) analyzed the travel behaviors of visitors at one of Korea's largest parks and provided data on satisfaction levels of crowding that supports the reasoning for creating the project described. Each of these studies highlights factors that can be incorporated into the tool to be developed but their shortcomings are that they focus on parks outside the US whereas this tool will focus on parks managed by the US National Park Service (NPS).

As mentioned, several key points influence visitor's decisions including park type, region, size, accommodations, price, activities, weather, and time/season of the visit (Liandi, 2021). Man-Keum Kim also found that visitors have varying levels of sensitivity to wildfires depending on if the location of the park is closer to metropolitan regions (Kim, 2010). While these studies further highlight features important to visitors that may be incorporated in the solution developed for this project, their findings are not easily accessible for visitors that may find them useful. Therefore, the solution developed for this project is unique in that it will make such information easily digestible for visitors planning park visits.

While factors valued by visitors will be explored, the primary feature presented by the tool will be visitor traffic. In recent decades, various models have been employed to forecast visitation levels at national parks. Clark and Wilkins found that an autoregressive model was outperformed by one created using Google Trends search engine data (Clark, 2019) and Bangwayo-Skeete et. all (2015) similarly used an autoregressive mixed data-sampling model with time series data obtained from Google Trends to improve forecasting accuracy. Hopken et all (2020) included extra variables like web search traffic into their deep learning machine ANN to improve prediction accuracy and Wilmot and McIntosh identified the methodologies that demonstrate the highest forecast accuracy (Wilmot, 2014). Each of these studies presents insights on models for best predicting future visitation levels at national parks which will be valuable to the development of this project, but the solutions are geared towards use by the NPS for ongoing management of each park. They do not explore bringing the data to the visiting public as planning tools which is the main limitation of current practice and the innovation of this project.

4. Proposed Method

4.1. Intuition

Throughout the course of this project's development, a data-driven solution was designed to support tourism decisions by presenting users with traffic data at national parks at a time of the user's choosing. There were minimal costs associated with developing this project other than the time required to build it as there have been no plans to deploy the solution to a cloud platform. This solution can impact park visitors, the NPS and local businesses near parks and is designed to benefit visitors as they will be better informed on the traffic levels to expect and plan around.

While Google does provide popular times and live visit data using the Google Location History of select users when researching specific national parks, there is no comprehensive source for all this information. Tourists must search each individual national park in order to obtain this information. If they are interested in visiting different national parks and don't necessarily prefer a specific park, they would have to search each national park on Google and make their own list of popular times for each park. Our approach helps tourists determine which national park is the optimal location for a visit based on their individual preferences during the specific time of year they choose in one simple visualization. Our approach is an innovative solution to the traditional approach's nuances because it:

- Provides visitation data for all national parks in one simple, easy-to-use visualization
- Predicts future visitation using machine learning models
- Allows users to filter forecasted visitation based on region and time of year

4.2. Approaches

An interactive data visualization was created to allow users to explore traffic trends at the park of their choosing during a chosen month/year, with dates ranging back from January 1979 to December 2021. In addition, a machine learning model was trained on the NPS visitation data to predict park traffic for 2022 and 2023 (the NPS has not yet published monthly data for 2022). Our visualization was created using Tableau and published/hosted to Tableau Public to allow users to easily interact with the park data and forecasts via web URL.

During the initial data exploration and preparation phase, we observed sharp drops in the rolling mean and standard deviation of national park visitation in 2020 and 2021 due to the COVID-19 pandemic. However, the trend returned to its pre-pandemic pattern towards the end of 2021. Based on this observation, we decided to use the NPS visitation data from 1979-2018 as our training set, while holding out the 2019 visitation data for testing against our model-generated 2019 predictions. While exploring the data, we realized that we would need to use a time series forecasting algorithm that accounts for seasonality when training our model. We found that summer months tend to have the highest visitation, with July and August having the highest visitation across all national parks. When exploring the data, we also found that eight of the national parks did not have consistent monthly data: Channel Islands National Park, Congaree National

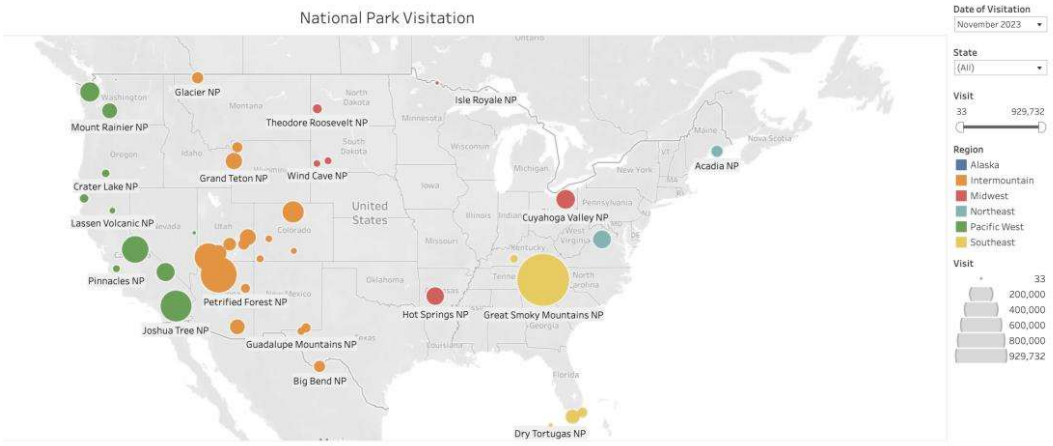
Park, Gates of the Arctic Nark and Preserve, Katmai National Park and Preserve, Kenai Fjords National Park, Kobuk Valley National Park, Lake Clark National Park and Preserve, and Wrangell-St. Elias National Park and Preserve. These national parks have been omitted from the dataset and from the forecasting features.

We trained two different models: a Holt-Winters and a SARIMA (Seasonal Autoregressive Integrated Moving Average) model.

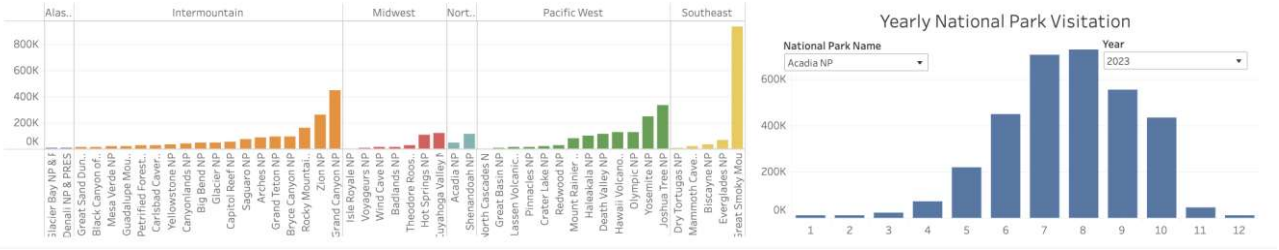
The Holt-Winters model was built in R, and the SARIMA model was built within a Jupyter Python Notebook. In the SARIMA model, trend and seasonal elements were tuned as hyperparameters for each national park and were then used for forecasting. A Holt Winters model was created using multiplicative seasonality and the optimal alpha, beta and gamma hyperparameters were identified for each park. We evaluated both models against the 2019 data using several metrics which are elaborated upon in section 5.2 Experiments and Observations. Mean Absolute Percentage Error (MAPE), Mean Absolute Deviation (MAD), Mean Forecast Error, and Mean Absolute Error (MAE) were all metrics considered for evaluating the performance of our models. We decided to use MAPE as our primary metric for evaluation.

After evaluating the predictions from both models, we opted to use the SARIMA model to forecast the visitation for all national parks covered by the visualization. The SARIMA class of models is applicable for the NPS visitation dataset, as it can capture trends and seasonality (as opposed to ARIMA models, which do not consider seasonality patterns). Holt-Winters models also take seasonality trends into account but utilize exponential smoothing (which places more emphasis on recent historical data than on older data). When comparing the performance between the Holt Winters model and the SARIMA model, the SARIMA model had lower mean absolute percentage error per month than the Holt Winters model.

To create our Tableau visualization, we utilized the Tableau Professional Desktop application and connected our csv file with the latitude and longitudinal values and the actual/forecasted visitation data as the main data source. The latitude and longitudinal values served as the geographic roles to create an interactive map chart. The coordinate data was downloaded from Wikipedia and merged with the park data. We plotted each of the national parks on the map as a dot and determined the size of the dots based on the sum of the visits of that national park in a particular month and year. The color of each dot was determined by the region of the national park. We then created additional views for the user to visualize the data in a more intuitive way. The first chart we added grouped each national park by region and displayed the visitation count in ascending order so that the user could see which national park had the highest visitation numbers in a particular month and year. The other chart we incorporated allows the user to select a particular national park from the dropdown and based on the year selected, the user can see the yearly trend of visitation over each month. For the visualization as a whole, we offer a variety of customization options. We allow users not only to select the month and year of visitation but also the state that they prefer (if needed). We also implemented a visit filter which would allow the user to determine the minimum and maximum values for the number of visitors they would prefer when making a travel decision. This would allow users to filter out national parks with high visitation or vice versa, based on personal preference. Overall, we attempted to provide users with an interactive and visually appealing interface that also allowed customization based on user preferences.



We believe our visualization tool provides a solution to the problem we sought out to solve: a single, comprehensive and interactive source that allows users to view and interpret historical visitation figures for national parks in the United States as well as forecast future visitation. Our tool is accessible to anyone with internet access and is easy to use and interpret. By default, the visual will open with a map of the United States, with the selected time period of visitation set to the month/year of the latest available prediction (November 2023). "Bubbles" representing the national parks across the United States are placed in their precise latitude/longitude location on the map and are sized to represent the amount of expected or actual visitation relative to all other parks. These bubbles are also colored by region, a legend is provided on the right-hand side of the map for easy interpretation. The legend also allows you to change the date of visitation, filter the map by State, or filter by visitation. By hovering your mouse over any of the bubbles, you will see the name of the national park, its region, and either the actual visitation or expected visitation depending on if the Date of Visitation is set to a past month or future month. Selecting/clicking on a particular bubble will allow you to exclude the national park from the visualization, as well as to "solo" the national park (removing the other national parks' bubbles from the visualization).



Below the main visual are two smaller bar-charts. These additional visualizations provide the user with further insight into visitation from a regional perspective, as well as a yearly perspective. The bar-chart on the left-hand side gives the user an additional scale for interpreting how much visitation one park may have compared to another, it may be difficult by only looking at the size of the bubbles to understand one park may have twice as much visitation that another. So, we provided this chart additionally with the intent that a user will likely see multiple parks in the main visual that interest them and then they can "drill-down" and easily compare visitation using this chart. The bar-chart on the right-hand side provides the monthly visitation for any national park and year as an additional drill-down tool.

Our solution is innovative because there is currently no single, comprehensive and interactive visual tool that allows users to interact and drill-down into national park visitation. For someone that is planning to visit a national park but does not yet know which specific park to visit, it is an extremely tedious process. For each park under consideration, the prospective visitor would need to research what the expected visitation is for different months in the year through different platforms and then compare the findings.

While Google Maps does offer "Popular Times" functionality, this only allows users to look at a single national park at a time and does not provide a comprehensive view of the popularity of all the national parks simultaneously. In addition, the "Popular Times" widget only shows an hourly visualization of the popularity by day of week. It also lacks specific numbers, making it difficult to quantify the "business/popularity" of a particular park. This is extremely time-consuming and can be unreliable/incomplete, as many travel decisions and reports of crowded parks are passed through word-of-mouth, or do not cite specific data. Since our data comes from National Park Services, we consider our reported data to be a reliable source for visitation data. Our tool relays this data into a user-friendly interface with additional forecasting that allows the user to answer these kinds of travel questions both quickly and easily.

Our original approach considered using JavaScript with D3.js to build our solution. However, after considering the scope and time constraints of the project, we elected to use Tableau. While all team members were exposed to D3.js over the course of the semester and the assigned activities, none of us considered ourselves particularly proficient with D3.js. On the other hand, team members were already proficient with Tableau and felt comfortable leveraging Tableau to deliver the interactive visualization. In addition, leveraging Tableau over D3.js also made end user testing easier. Had we pursued the original D3.js route, additional barriers would be added for the solution to be publicly accessible. We could have hosted the solution on a cloud provider, which, in addition to financial cost, would bring about a significant amount of additional workload for the team (creating the account, setting up billing, establishing the resource group/provisioning resources, containerizing the web app, etc.). We also could have zipped the project, with references to the D3.js CDN and other used libraries (or bundling them locally in the zipped project), but for users to test our project, they would have had to download the project files and view the HTML locally before being able to answer the questionnaire. Tableau Public allows for free hosting and is responsive to multiple orientations and screen sizes, allowing respondents to visit the questionnaire and interact with the tool from mobile devices and tablets. Implementing responsive layouts for a D3.js visualization would be a significant undertaking. For this project, with our intention of being able to move quickly and with agility, Tableau (and the ability to publicly host on Tableau Public for free) was the best option available.

5. Experiments/Evaluation

5.1. Testbed

The experiments and evaluation were divided into those that test the predictive ability of the visitation models developed and those that test the visualization and user interface. The questions we wanted to answer were:

1. Is the forecast we are showing of good accuracy?
2. Is the visualization easy to interpret?
3. Does the visualization make a user want to visit one park versus another?
4. Does the visualization make a user want to visit a park during one month versus another?
5. Do users see this tool as beneficial to their planning process?

5.2. Experiments and Observations

The predictive ability was tested by splitting the data into a training dataset and a testing dataset. The SARIMA and Holt Winters forecasting models were developed using cross validation across the training dataset and once they were finalized, the predictions were tested against the testing dataset. The tests were a comparison of the predicted traffic for a national park during 2019 to the actual traffic at that park and time. As mentioned previously, we observed sharp drops in the rolling mean and standard deviation of national park visitation in 2020 and 2021 due to the COVID-19 pandemic so the NPS visitation data from 1979-2018 was used as our training set, while holding out the 2019 visitation data for testing against our model-generated 2019 predictions.

Mean Absolute Percentage Error (MAPE), Mean Absolute Deviation (MAD), Mean Forecast Error, and Mean Absolute Error (MAE) were all metrics considered for evaluating the performance of our models. We decided to use MAPE as our primary metric for evaluation. The overall mean MAPE of all national park forecasts in 2019 from our Holt Winters model was 21%, which is above the 20% threshold that we used to define "good" or "acceptable" predictive accuracy. Meanwhile, the overall mean MAPE of all national park forecasts in 2019 from our SARIMA model was 17%, which is below the 20% threshold that we used to define "good" or "acceptable" predictive accuracy. After evaluating the predictions from both models, we opted to use the SARIMA model since it can capture trends and seasonality and had lower mean average error than the Holt Winters model. While we always would prefer a lower measure of error (we initially defined "success" as a MAPE below 10%), we are nevertheless pleased with our results, as we do not want to overfit our model to the 2019 data. The MAPE of 17% indicates that our model does indeed provide meaningful predictions.

The visualization and user interface were evaluated by having users complete short questionnaires after interacting with the tool. The questionnaires addressed whether the visualization had impacted and benefited their travel decision-making. Questionnaire respondents were individuals from our personal networks who identify as being interested in traveling to a national park in the future. The questionnaire asked if they would make use of the tool when making travel decisions about the national parks. Second, the questionnaire asked the visitation numbers made them want to visit one park over another. Third, it asked if the tool, with its existing features, had impacted which month they choose for traveling. Lastly, the questionnaire was concluded with a section for suggestions to make the visualization more useful. With 21 total responses the results were as follows:

- 95.2% of respondents said they would use the visualization to make national park travel decisions.
- 85.7% of respondents said the visitation numbers made them want to visit one national park over another
- 90.5% of respondents said the visitation numbers made them want to visit a park during one month over another

The respondents mentioned that they would like it if there was information about average temperature, special features, and events across each month at each national park. They also suggested that a ratio of people expected vs the overall capacity of the park would make the tool even more helpful. Overall, we believe that our tool was successful because more than 60% of respondents reported that they would use the tool to make national park travel decisions.

6. Conclusion and Discussion

The goal of our tool is to inform users of expected visitation numbers and popularity at national parks for a time of their choosing. Equipped with this knowledge, our hope is that users will be able to make informed decisions on which park to visit, as well as when to visit, allowing them to make the most of their experience. Based on the results of the experiments described, the project can be considered successful, though it could benefit from further development and exploration as there are limitations to the project. Visitation/forecasted visitation is visualized, which gives prospective visitors insight into the number of visitors, but to understand how visitation numbers may impact the visitor experience, additional factors such as average temperature/weather, capacity of the park, and visitor reviews for the prospective time of year would give greater insight into travel decisions. These features are potential future extensions that could be added in future iterations. Throughout this process, all team members have contributed a similar amount of effort.

7. References

Abdou, Musonera, et.al "Factors Affecting the Visitation to National Parks Using Machine Learning Techniques: the Case of National Parks in Rwanda", *EBSCO Host*, African Journal of Hospitality, Tourism and Leisure 11.2, 2022, https://web.s.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=5926645f-9232-4d53-8958-fee0cb4d8c44%40redis

Bangwayo-Skeete, P.F. and Skeete, R.W. , "Can Google data improve the forecasting performance of tourist arrivals?", Mixed-data Sampling approach Tourism Management, Vol. 46, pp. 454-464, 2015.

Bergstrom, John C., et al. "What does the future hold for US National Park visitation? estimation and assessment of demand determinants and new projections.", Ag Econ Search, Journal of Agricultural and Resource Economics 45.1, 2020, https://ageconsearch.umn.edu/record/298433

Choe, Schuett, M. A., & Sim, K.-W. (2017). An analysis of first-time and repeat visitors to Korean national parks from 2007 and 2013. *Journal of Mountain Science*, *14*(12), 2527–2539. https://doi.org/10.1007/s11629-017-4387-y

Clark, Wilkins, E. J., Dagan, D. T., Powell, R., Sharp, R. L., & Hillis, V. (2019). Bringing forecasting into the future: Using Google to predict visitation in U.S. national parks. *Journal of Environmental Management*, *243*, 88–94. https://doi.org/10.1016/j.jenvman.2019.05.006

Höpken, W., Eberle, T., Fuchs, M. and Lexhagen, M., "Improving tourist arrival prediction: a big data and artificial neural network approach"., 60, 998-1017, 2020F

Kim, Man-Keum, et.al "Wildfire, National Park Visitation, and Changes in Regional Economic Activity", *Science Direct*, Journal of Outdoor Recreation and Tourism 26, 2019, https://www.sciencedirect.com/science/article/pii/S2213078019300209

Liandi Slabberta, Elizabeth AnnDu Preeza (2021), "Where did all the visitor research go? A systematic review of application areas in national parks", Journal of Hospitality and Tourism Management, Vol 49, p. 12-24.

McIntosh, Christopher R., et al. "An empirical study of the influences of recreational park visitation: the case of US National Park Service sites.", Sage Journals, Tourism Economics 17.2, 2011, https://journals.sagepub.com/doi/abs/10.5367/te.2011.0036

Neuvonon, Marjo, et.al "Visits to National Parks: Effects of park characteristics and spatial demand", *Science Direct*, Journal for Nature Conservation 18.3, 2010, https://www.sciencedirect.com/science/article/pii/S1617138109000752

Newton, Jennifer N., et al. "If I Can Find a Parking Spot: A Stated Choice Approach to Grand Teton National Park Visitors' Transportation Preferences." Science Direct, Journal of Outdoor Recreation and Tourism, 2020, https://www.sciencedirect.com/science/article/pii/S2213078018300227.

Sessions, Carrie, et al. "Measuring Recreational Visitation at U.S. National Parks with Crowd-Sourced Photographs." Science Direct, Journal of Environmental Management, 1 Sept. 2016, https://www.sciencedirect.com/science/article/pii/S0301479716306685.

Shen, Xing-ju, et al. "US national parks accessibility and visitation.", *Springer*, Journal of Mountain Science 16.12, 2019, https://link.springer.com/article/10.1007/s11629-019-5379-x

Timmons, Abby L. “Too Much of a Good Thing: Overcrowding at America's National Parks.” Gale Academic Onefile, Notre Dame Law Review, 2018, https://go.gale.com/ps/i.do?p = AONE&u=gainstoftech&id=GALE%7CA572943409&v=2.1&it=r.

Wilmot, & McIntosh, C. R. (2014). Forecasting Recreational Visitation at US National Parks. *Tourism Analysis*, *19*(2), 129–137. https://doi.org/10.3727/108354214X13963557455487