Measuring Recreational Visitation with Crowd-Sourced Photographs

Caroline Sessions

A thesis submitted in partial fulfillment of the requirements for the degree of:

Master of Science

University of Washington

2015

Thesis Committee:

Sergey Rabotyagov

Spencer Wood

Program Authorized to Offer Degree:

School of Environmental and Forest Sciences

University of Washington

Abstract
Measuring Recreational Visitation with Crowd-Sourced Photographs

Caroline Sessions

Chair of the Supervisory Committee:
Sergey Rabotyagov, Assistant Professor
School of Environmental and Forest Sciences

In the age of big data and internet social media sites, there are myriad opportunities for researchers to leverage crowd-sourced online data for research purposes. One such opportunity is the possibility of using photos posted online to infer information about visitors to recreational areas. In this paper, I assess the validity of using data from photos posted on the website *Flickr* to infer information about visitors to National Parks in the Western United States. By comparing the photo data to statistics published by the National Park Service (NPS), I examine the relationship between the two datasets and if used properly, whether the photos can yield similar information to that provided by the Park Service.

I address two aspects of visitor use data: the number of people visiting a park and where those visitors reside. Together, researchers can use these two datasets in conducting travel cost studies to estimate the economic recreational value of a site. Using multiple regression analysis, I build a statistical model whereby I can infer the count of visitors to a park in a given month through the number of photos posted online. Overall, I find that a one percent increase in "photo-user-days" corresponds to a 0.65 percent increase in NPS visitation, holding all else constant. To evaluate the validity of using photos to infer visitor's home origins, I compare information from photo-taker's profiles to survey data provided by the NPS. I find that the photos give an accurate

big-picture view of visitor origins and moreover, provide an accurate view of the distance that visitors live from the park.

Using findings from my analyses, I estimate the travel costs incurred by in-state visitors to reach Mount Rainier National Park in 2012. Using data from the NPS, I find in-state visitors incur an average cost of $61.38 per person and a total cost of 36.31 million US dollars across all visitors. Using the photo data, I estimate the average cost incurred is $58.90 per person, with a total cost of 37.20 million US dollars (95% CI 29.32-45.01 million). This finding shows that the photo data can produce a strikingly similar economic estimate to those obtained by more traditional methods.

The ability to accurately infer visitor use information with photos has the potential to reduce costs for land managers and opens new opportunities for research on the recreational value of lands. As visitor use information is expensive and time-consuming to obtain, this method has significant potential to save time and money for the Park Service. In addition, other land managers and researchers can apply this method to evaluate visitation to remote or hard-to-measure locations that currently lack detailed visitor use data. Overall, this work represents a crucial first step in evaluating a new and exciting means to measure recreational use value.

Acknowledgements

Foremost, I am incredibly grateful to my thesis committee, all of whom provided exceptional guidance throughout this process. Sergey Rabotyagov, the chair of my committee, was especially helpful in giving me timely feedback on my work, providing excellent insights into statistical methods, and remaining a positive and encouraging influence. I want to thank Sergey for initially taking on a last-minute student and subsequently supporting me throughout my graduate career. Spencer Wood served as an invaluable resource— I am deeply grateful for his development of this idea and willingness to serve as my advisor. In addition, his passion for scientific rigor and dedication to holding me to a high standard of work pushed me to continually improve my analysis. Finally, I am especially thankful to David Layton, who provided me with excellent advice, both on the big picture of my research as well as on methodological details. His willingness to explain difficult statistical concepts was instrumental in developing my methods.

I am also incredibly grateful to the entire Natural Capital Project team. Dave Fisher, my technical "guru" was invaluable in providing me with data—his willingness to teach me programming codes and patience in helping me with data processing was invaluable to this research. I am also thankful to Katherine Wyatt who provided support, statistical advice, and encouragement throughout the process. Lastly, I want to thank Anne Guerry for making me feel a part of the Natural Capital team and for giving me the opportunity to present my research. Lastly, I want to thank the entire Natural Capital team for their financial support of this research.

Finally, I am grateful to my friends, family, and other academic mentors who have supported me through this process. Their willingness to share in my successes and give advice in times of stress sustained me through this process.

Table of Contents

Chapter 1: Background

Chapter 2: Validating Visitor Counts

Chapter 3: Validating Visitor Home Locations

Chapter 4: Application of Methods to a Travel Cost Study

Chapter 5: Synthesis and Conclusion

List of Tables and Figures

Chapter 1: Background

Section 1: Introduction

In the age of big data and internet social media sites, there are myriad opportunities for researchers to leverage crowd-sourced online data for research purposes. One such opportunity is the possibility of using photos posted online to infer information about visitors to recreational areas. Recreationalists often post photos of their trips to social media sites, such as Facebook, Instagram, and Flickr. While the photographer's main purpose of posting photos may be to share the experience with their friends and other users, their photos actually convey valuable information to the public. Many photos posted online contain a geotag of where the photo was taken and a timestamp of when it was taken; in some cases, this information can be linked with personal information about the user, which they may voluntarily post on a user profile. Together, this information can provide valuable insights to researchers on the tastes, preferences, and habits of the photo-poster. This research represents one attempt to leverage data from crowd-sourced photographs to learn more about people's tastes and preferences. Specifically, I explore the validity of using online photo data to estimate recreational visits to National Parks and in turn, use that data to estimate the economic value of selected parks.

Why Value Recreation?

Recreation on public lands has long played an important part in United States history. Since Congress created the first National Park in 1872, Americans have sought-out public lands to enjoy nature and to adventure. Currently, roughly one-third of land area in the United States is federally owned (Gorte, Hanson, and Rosenblum 2012). The use of federal land, however, has been hotly contested from its inception (Stevens and Frank 2009). As public lands have long been an important source of natural resources for energy development, there are competing perspectives on the best use of the land. Determining the best use requires balancing objectives between economic development, recreation, and conservation.

Recreation became recognized as a worthy use of federal land around the mid-twentieth century. In the Multiple-Use Sustained Yield Act of 1960, Congress stated that National Forests "shall be administered for outdoor recreation, range, timber, watershed, and wildlife and fish

purposes" (Congress 1960). Since the mid-1980's, researchers have tried to substantiate the argument for recreational lands by developing quantitative assessments of their recreational value. As a consequence, recreation gained more legitimacy as an alternative to land dedicated towards resource extraction (Stevens and Frank 2009). However, as Stevens and Frank (2009) explain in their research on conflicts on public lands, "new challenges, including recent pressures to devote portions of our public lands to renewable energy project development and the multifaceted threats presented by climate change, will continue to test the public's and government policymaker's commitment to devote public land to recreational purposes" (pg. 29).

In the face of mounting challenges, it is important that researchers and policymakers have the best information possible regarding the use and value of recreational lands. Having comprehensive knowledge of these areas will allow policymakers to more effectively allocate resources towards their best use. For recreation, this entails having a thorough understanding of how many people recreate on public lands, peoples' preferences and activities, and their economic value of that experience.

This issue is most salient at National Parks. The Park Service strives to understand visitation and recreation for two reasons. First, the National Park Service (NPS), created in 1916, has a dual mandate to "conserve the scenery and the natural and historic objects and the wild life therein *and* to provide for the enjoyment of the same in such a manner and by such means as will leave them unimpaired for the enjoyment of future generations" (Antolini 2009 pg. 862). This mandate requires that the NPS understands who enjoys National Parks, how they use parks, and what factors influence their use and enjoyment (Benson et al. 2013). Second, the NPS is required to evaluate the economic costs and benefits of proposed policies and management alternatives (Loomis 2000). In their most recent five-year strategic plan, the NPS listed one goal as to better understand how national parks contribute to economies—to do this, they plan to study the economic value of NPS activities and programs (National Park Service 2012).

Despite the importance of understanding recreational visitation to National Parks and other public lands, there is a lack of comprehensive data on the subject. According to Loomis (2000), "government agencies that supply outdoor recreation have been slow to recognize the importance of consistently collected and defensible use data" (pg. 93). Loomis later continues on the consequence of this issue: "If we don't even know how many customers we have and whether their use is increasing or decreasing, it is hard to make good recreation management

decisions. Better data collection is an important step to better funding, management, and allocation of natural resources to recreation" (Loomis 2000, pg. 94). **In this thesis, I build upon the work of others to address this paucity in data and further develop new methods of measuring recreational visitation and its economic value.** I first review current techniques for understanding NPS visitation and for measuring the recreational value of parks. I then validate a new method for examining visitation in hopes that it can then be used to gain a more comprehensive view of recreational visitation to National Parks and other public lands.

Research Objectives

The primary objective of my research is to evaluate the use of photos posted online to infer information about recreational visitation. **Specifically, I test whether Flickr photos can act as a reliable proxy for visitor use statistics obtained from the National Park Service.** National Parks serve as an appropriate venue to test the validity of using photos as a proxy for empirical visitation data for two reasons. First, the existence of visitor use statistics at National Parks gives me a dataset against which to test the photo data. More or less, it serves as an empirical "truth" against which I can measure how the photo data matches, a method known as "ground truthing". Second, the NPS could benefit from further techniques to measure recreational visitation, for the reasons outlined in the Introduction. The organization of this paper is as follows:

Chapter 1 gives background information, continuing with a historical look at past techniques to measure recreational value, a survey of existing literature on recreational visitation, and a description of my datasets and the biases introduced by each.

Chapter 2 begins the analysis by testing how the counts of visitors measured by each technique align. The primary goal is to validate the use of photos in approximating the count of visitors measured by the Park Service. This chapter goes through the methods used for statistical modeling, results of those models, and discusses the conclusions drawn.

Chapter 3 continues the analysis by testing whether the photos give an accurate view of visitors' home locations and the distances between their homes and the park. My objective in this section is to validate the use of photos in accurately inferring where visitors originate. Similar to the previous chapter, this section goes through the statistical methods, results, and discussion of implications.

Chapter 4 applies findings from Chapters 2 and 3 into travel cost studies for Mount Rainier and Great Sand Dunes National Parks. Using the photo data, I apply regression models from Chapter 2 and information on the home locations from Chapter 3, to calculate the average economic travel expenditure per in-state visitor to each park.

Section 2: Literature Review

There has been significant research regarding visitation to National Parks and the valuation of recreational benefits. In this section, I start by providing background information on techniques for valuing environmental goods. I then continue by surveying the existing literature in the field to better understand standard practices and see how my research fits within the existing scholarly and policy environments. I examine research in the following three arenas: economic valuation of recreation in National Parks, measurement of how natural and policy events impact recreational visitation, and new methods for measuring recreational visitation.

Economic Valuation of Environmental Goods and Services

My research seeks to measure the recreational use value of certain National Parks. *Recreational use value* is defined as people's willingness-to-pay (WTP) for recreational activities.[1] Most environmental goods and natural resources derive value in two ways, through use value and non-use value. *Use value* involves an individual deriving value from the good or place by actually using it. Examples include hiking, fishing, hunting, and kayaking. *Non-use* or *passive-use value*, on the other hand, is when an individual derives value from the good without going there or physically interacting with the good; examples include existence value and bequest value. As the name implies, *recreational use value* is a type of use value whereby the user derives value from the place by recreating there—the user does not need to participate in consumptive activities, but rather may simply enjoy the place.

Assessing recreational use value can be challenging. Many recreational activities are non-market goods, meaning that they are not sold in the traditional marketplace. Consequently, economists cannot use the typical price signal as an indicator of the good's economic value.

[1] Willingness-to-Pay is the maximum amount that a consumer would pay to receive a good or service.

Instead, economists look to alternate methods to evaluate a good's monetary value; there are two types of methods typically used. *Revealed preference studies* examine goods that consumers do buy in the standard marketplace and use these transactions as clues to the economic value of the good of interest. *Stated preference studies*, on the other hand, present consumers with hypothetical scenarios and elicit their willingness-to-pay for that good. While researchers use both revealed preference and stated preference studies to value recreation and recreational lands, I focus here on studies using the *travel cost method* of revealed preference.

Travel cost studies operate under the premise that the amount of money spent en route to a location is a lower-bound estimate of the amount that one values the site. Travel cost studies were first developed by Harold Hotelling in a letter that he sent to the National Park Service in 1947. In the letter, Hotelling described how one can estimate points on a demand curve by plotting how many people visited the park and how far each one travelled to get there. In a meta-analysis on travel cost studies, Smith and Kaoru (1990) explain that the essential element in Hotelling's analysis is the "recognition that people pay an implicit price for the use of a recreation site" (pg. 269). Since 1947, travel cost studies have become a standard and accepted method to measure the recreational use value of sites. While researchers originally used travel cost studies to estimate the total value of a site, economists now also use it to estimate values for specific characteristics or site attributes. The *hedonic travel cost method* uses variation in people's willingness-to-pay to visit different sites to isolate certain attributes and back-out a value for a specific characteristic. These studies acknowledge that visitors have a variety of places to choose to visit, each one offering a different package of characteristics. In an early look at this method, Brown and Mendelsohn explain that by "treating heterogeneous sites as if it was a bundle of characteristics, the site price can be decomposed into a set of implicit prices for each characteristic using the traditional hedonic method" (Brown and Mendelsohn 1984). Thus, visitors to recreational sites do not decide where to visit solely on cost of travel, but on a host of other site attributes as well. In the following section, I look at how researchers have used the travel cost method to value recreation, recreational lands, and changes to those environments.

Since Hotelling's first application of travel cost studies to valuing National Parks, researchers have continued to measure the recreational use value of wild places. However, this research has been largely dependent upon having visitor use data, something that was scarce through the 1970s; the consequence was a paucity in comprehensive studies. Starting in 1982, however, the NPS facilitated this area of research by conducting and publishing visitor use surveys. The surveys were created and administered by Visitor Service Project (VSP) at the University of Idaho. While the main purpose of the VSP surveys has been to assess visitor satisfaction (Neher et al. 2013), the surveys ask some questions that are helpful for economic valuation. For example, most surveys ask for the home zip code of the respondent as well as the number of times he has visited the park in the last year and in his lifetime.

Since the VSP started, several researchers have used the survey results to estimate recreational value. Heberling and Templeton (2009) were the first researchers to do so. In their study, they use information on the home zip codes of visitors to conduct an individual travel cost model for the Great Sand Dunes National Park. Since Heberling and Templeton's work, there have been numerous research projects using VSP data to assess the economic value of National Parks. Until recently, however, the work has largely been site-specific and disaggregated (Neher et al. 2013). In a survey of NPS valuation studies, Duffield et al. identify 27 different studies of willingness-to-pay of NPS visitors, including 128 estimates of WTP for specific parks (Neher et al. 2013). Neher et al. (2013) aim to consolidate this research in a meta-regression analysis measuring the average and total WTP for NPS recreational visitation system-wide. They use NPS visitor use data from 58 parks; overall, they estimate that the WTP in 2011 per park visit averaged $102 across parks and ranged from $67 to $288 for specific parks. From this, they estimate the total WTP among 2011 visitors for National Park visits at $28.5 billion.

Benson's work in 2013 was notable as it pushed the boundaries of how researchers use VSP data (Benson et al. 2013). Rather than measuring aggregated economic recreational value, they conducted a travel cost study of Yellowstone National Park with cluster analysis to assess how benefits vary by participants of different activities (i.e., backcountry users versus visitors going for a scenic drive). They found that the average benefit across all visitors ranged between $235 and $276 per person per trip. They also found, however, that the economic value varied greatly between types of visitors, varying from $90-$103 for "value picnickers" to $323-$714 for "creature comfort seekers" who stay in lodges and dine in restaurants.

Assessment of the Impact of Natural and Policy Events on Recreational Visitation

In addition to studying the aggregated economic value of recreation at National Parks, researchers have begun to study how different events change visitation to National Parks. Using NPS monthly visitation statistics, researchers are studying how marginal changes in visitation correlate to outside factors.

Several studies, for example, examine the marginal impact of environmental and weather events on recreational visitation. In 2013, Duffield et al. study the effect of wildfire on recreational use at Yellowstone National Park (Duffield et al. 2013). By comparing the monthly visitation counts to wildfire activity in the area in a time series model, they find that people do, in fact, have a willingness-to-pay to avoid fire-affected areas. They conclude that the presence of fires results in short-term economic losses to the area. In a similar vein, Poudyal et al. (2013a) examine the effect on impaired visibility on visitation at Great Smokey National Park. Using a polynomial distributed lag model, they estimate that improving the average visibility by 10% from the current level could lead to an increase in roughly one million recreational visits per year. In another work, Poudyal et al. (2013b) assess the effect of the economic recession on NPS visitation. They regress the number of NPS monthly visits to a park on economic variables while controlling for month, gasoline price, and population using a Generalized Method of Moments model. Overall, they find that recession was negatively associated with demand to visit national parks.

Morgan et al. (2011) conduct a similar type of analysis by analyzing daily visitation counts to a state park in southwestern Florida. They measure the impact of red tide events and the publicity of those events on visitation to Lover's Key State Park. Using time series analysis, they estimate how daily changes in park attendance track environmental and media covariates. Overall, they find that almost 400 fewer people visited the park on a day when the term "red-tide" was included in at least one local newspaper article on that day. This work is significant as it is the only research I have found that studies daily changes in visitation, rather than monthly aggregated counts.

Exploration of New Datasets on Recreational Visitation

Numerous studies discuss limitations of using NPS count and survey data (Benson et al. 2013; Heberling and Templeton 2009; Poudyal, et al. 2013). These limitations come in two forms. First, the aggregated monthly counts of visitors contain little to no detail on demographics. Poudyal et al. (2013), for example, would have liked to analyze the count of domestic and international visitors separately. Second, researchers using the VSP survey data are limited by data that may not be representative of all visitors to that park (Benson et al. 2013; Neher et al. 2013). As the surveys are only conducted during a short period, usually a week or two during the summer, visitor responses may not be representative of the visitors across the entire year. Lastly, Morgan's (2011) research of the effect of red tide highlights another limitation of NPS data—the inability to study daily changes in visitation. As the NPS only publishes monthly counts of visitation, researchers have thus far been unable to assess daily trends. In speaking with Pam Ziesler with the Visitor Use Statistics program at the NPS, she mentioned that there is a significant but unmet demand for counts of daily visitors (Ziesler 2015).

Therefore, there is a current void in recreational visitation data (specifically to National Parks, but other public lands are even more lacking in measuring visitation). As noted in Heberling (2009), surveys of National Park visitors that are appropriate for economic analysis are rare. There is a niche for data that gives a multi-season view of visitation by being collected throughout the year, gives demographic information on the visitor, and can be accessed through daily visitor counts.

Researchers at the Natural Capital Project have recently explored this niche. In 2013, Wood et al. (2013) published a paper assessing the relationship between geo-tagged Flickr photos and empirical visitation rates for recreational site. Overall, they compare photo-estimated visitation rates to survey and human-measured visitation rates at 836 sites worldwide. Through simple linear regression analysis, they find a reliable statistical relationship between the two estimates of visitation and thus, conclude that geo-tagged photos can be a reliable proxy for empirical data on visitation. The following graph shows their results, where the x-axis represents visitation based on geo-tagged photos and the y-axis shows the empirical data:

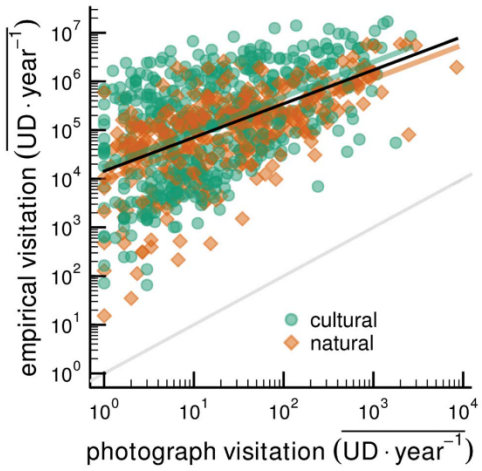

Figure 1.1: Photograph Visitation vs. Empirical Visitation (source: Wood et al. 2013)

They further explore this finding in two primary ways. By controlling for whether the site is a recreational or cultural attraction and for the income-level of the country, they conclude that neither variable has a statistically-significant impact on the relationship between photos and visitor counts. Next, and more interestingly, they examine whether the home countries of the photo takers match the home countries of the observed visitors. By comparing the home countries reported on users' Flickr profiles to the home country reported by incoming visitors surveyed at international borders, they find that the two datasets give a similar view of visitor origins. They then conclude "crowd-sourced data [are] not only useful for estimating visitation rates, but also for understanding where visitors originate." Wood et al. finish by explaining how researchers could use this new dataset in travel cost studies.

In a subsequent publication, Keeler et al. (2015) apply the use of Flickr photos to a travel cost study examining recreational demand for clean water in Minnesota and Iowa lakes. Overall, they assess how marginal changes to lake clarity would effect recreational visitation. By using Flickr photos as a proxy for the count of visitors to a lake, they estimate multiple regression models analyzing how visitation changes with travel time, lake attributes, and other factors. They find that users are willing to travel 56 minutes further, thereby incurring $22 more in travel costs, for every one-meter increase in water clarity in MN and IA lakes.

This recent work is important, as it is the first application of using Flickr photos as a proxy for survey-measured visitation rates. While Keeler's work does an excellent job at

demonstrating the utility of this method, they are limited in their ability to provide further validation to the method due to paucity in survey data. While they do regress photo visitation rates against survey-measured visitation rates, their ability to rigorously examine the relationship between photo counts and visitor counts and origins is limited. For example, they can only examine annual visits to a site and are unable validate the visitor's home locations against a baseline. In my research, I aim to provide additional statistical validation of using photos to infer visitor information, as well as test the method on a different subset of public lands.

Section 3: Data

To assess whether photos accurately predict visitation at National Parks and the distance traveled by visitors, I compare visitation rates and home location data measured by Park Service to that measured by the photos. To do this, I examine three datasets, two published by the NPS and one from Flickr. The Park Service has traditionally studied visitation through two primary methods: counting visitors and conducting a detailed survey to a sample of visitors. I use data obtained from both methods for my research. The following section describes the three datasets, then examines potential biases introduced with each dataset.

NPS Visitor Counts

Since 1904, the NPS has published Visitor Use Statistics, a monthly count of the number of people that have entered the park through each entrance. This data gives a big-picture view of overall park visitation, but with little data on whom the visitors are. To obtain these statistics, parks calculate a daily count of visitors using a variety of methods. For each park, the NPS establishes "Visitor Use Counting and Reporting Instructions" which instructs the method of counting. Many of these instructions were issued during the early 1990s, when the Park Service implemented a new visitor count reporting system (NPS 2014). Some of these instructions have since been updated, but many have not. In a Director's Order issued in 2004, the NPS described that it would "conduct periodic reviews of the public use counting instructions for each park and verify and issue the specific counting instructions to keep the data consistent and reliable" (NPS 2004). In my interview with Pam Ziesler, the NPS employee in charge of monitoring visitation, she explained that the NPS does periodic audits of the instructions and only edits them if visitor

use patterns have changed significantly (Ziesler 2015). She mentioned that since visitation has remained relatively steady over the last 20 years, there is often little need to change the instructions.

The method by which parks count visitors varies. Some parks do manual counting, whereby a ranger records the number of visitors entering through a gate. Most parks, especially larger ones, use automated counters, including ones on vehicles, doors, and trails. An automated vehicular traffic counter that counts the number of vehicles passing a specified entrance is the most common method of counting. Parks then use a per-person-vehicle multiplier to estimate the total number of people entering. The multiplier varies by park, and often also by the time of year (NPS 2014). Parks calculate the multiplier through a stratified sample of visitors over a one-year period (Ziesler 2015).

As an example of how parks collect data, Yosemite National Park operates under reporting instructions from 1994. It measures recreational visitor counts through five traffic counters at different entrances (Yosemite National Park 2004); each counter measures the number of vehicles passing through the entrance. It then reduces the count by a specified percent to account for non-reportable vehicles and non-recreational vehicles (mostly NPS staff or researchers). The reduction percent varies between 4% and 23% between entrances—the reporting instructions do not explain how NPS determined these percent. The count is then multiplied by the persons-per-vehicle (PPV) multiplier, which is 2.9. The PPV is constant across seasons and has not been updated since its issuance in 1994. For comparison, Arches uses a PPV of 2.4 during the winter months and 2.7 during the Spring and Summer (Arches National Park 1993). Yellowstone varies the PPV by entrance gate and month; the multiplier ranges between 2.2 and 3.1 (Yellowstone National Park 1995). Mount Rainier, which updated its Instructions in 2005, varies both the reduction percent of non-recreational vehicles as well as the PPV by both entrance and month (Mount Rainier National Park 1995).

In this research, I analyze data for all 38 National Parks in the Western United States, including Alaska.[2] For each park, I have compiled monthly visitation counts from 2005 through 2012. The following graph illustrates the total number of visitors in the 38 parks during that timeframe, according to the NPS:

[2] A table of the 38 parks studied and their corresponding identification codes, states, and regions is located in the Appendix.

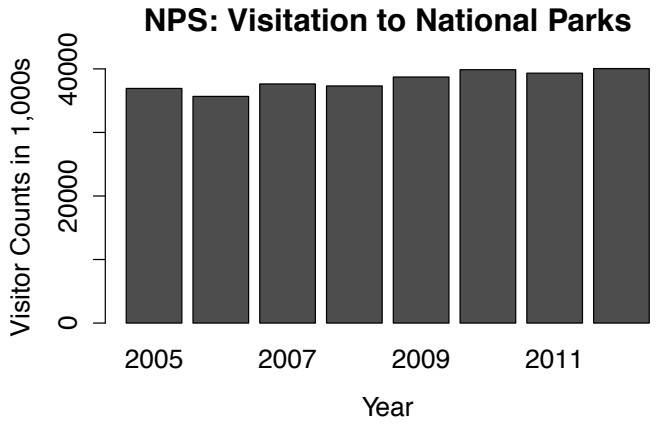

Figure 1.2: NPS Visitation to National Parks

NPS Visitor Home Locations

The second way in which the NPS assesses visitation is through detailed Visitor Service Projects (VSPs) conducted by the Park Studies Unit (PSU) at the University of Idaho[3] The surveys are in-person, written surveys approximately 15 pages in length. Following a standard protocol for all VSPs, the surveyor identifies one person in the visiting party to participate. The surveyor asks questions for approximately two minutes, then asks the visitor to fill out the paper survey in the days following his/her visit. The visitor then mails back the survey. The PSU sends out reminders and replacement questions to those participants who have not submitted their questionnaire within four weeks. Response rates are typically high, ranging from 56% to 90%. National Parks conduct VSPs with varying frequency, ranging from once every ten years to once every two to three for each park. While most parks conduct VSPs during their peak visitation time, others do it in different seasons to explore seasonal variations in visitation. Surveys ask visitors numerous questions about themselves and about their visit to the park. One question specifically asks them to indicate the home zip code of everyone in the party. If they are from outside of the United States, they are asked to list their home country.

[3] The PSU has recently closed due to budget cuts, but the work will likely continue through other Universities (Collins 2015).

I have compiled a database of all VSPs conducted in National Parks in the Western U.S from 2006 to 2012. The database includes information from 16 surveys in total; the total sample size of responses is 9,228, with an average response rate of 71%. The table below gives information on the surveys that I analyze; there is significant variation in location, year, and season of the surveys.

Park Name	State	Year	Season	Dates Collected	Sample Size	Response Rate
Black Canyon of the Gunnison	CO	2010	Summer	June 26 - July 2	459	69%
Capital Reef	UT	2008	Summer	May 24- June 1	480	78%
Denali	AK	2011	Summer	July 19-25	735	71%
Denali	AK	2006	Summer	Aug 1-7	815	81%
Death Valley	CA	2010	Spring	Mar 18-24	304	82%
Death Valley	CA	2009	Fall	Nov 22-Dec 8	271	76%
Grand Teton	ID	2008	Summer	July 13-19	739	71%
Mount Rainier	WA	2012	Summer	Aug 4-10	702	63%
Rocky Mountain	CO	2011	Winter	Feb 19-27	579	73%
Rocky Mountain	CO	2010	Summer	July 18-24	755	69%
Yellowstone	WY	2006	Summer	July 23-29	903	69%
Yellowstone	WY	2012	Winter	Feb 15-21	334	77%
Yellowstone	WY	2011	Summer	July 23-29	900	68%
Yosemite	CA	2009	Summer	July 8-14	689	57%
Yosemite	CA	2008	Winter	Feb 2-10	563	60%

Table 1.1 NPS Surveys

Flickr Photos

Flickr is a website that allows users to post and publicly share their photos online (www.flickr.com). As of 2013, there were over six billion photos posted on Flickr, uploaded by over 71 million users (Wood et al. 2013). Flickr was started in 2005, gained popularity by 2008, and has continued to rise in usage since. Approximately 197 million of these photos are geo-tagged, meaning that they indicate the GPS coordinates of where they were taken; 40% of the geo-tagged photos are taken in the U.S (Wood et al. 2013). Uploaded photos automatically include a timestamp, indicating when the photo was originally taken and when it was uploaded to the website. A portion of Flickr users have a profile in which they self report their home location.

Researchers at the Natural Capital Project have developed software that identifies and counts Flickr photos taken within a geographic boundary, on a given day, by a unique user (Sharp et al. 2015). The process uses GIS to query a database and select all photos taken within one of the geographic boundaries of the 38 National Parks that I study. The software then counts the number of photos taken by a unique user on a unique day—this returns a count of Photo-User-Days (PUD) (Wood et al. 2013). This count thereby gives the number of users who take at least one photo within a geographic boundary each day across the study period. For example, if five people visit Yellowstone National Park on a certain day and each upload one photo, then the resulting PUD equals five. However, if one person visits Yellowstone but uploads five photos, PUD equals only one, as this person is a unique user. Researchers at the Natural Capital Project initially developed their methodology using Flickr photos due to the website's large database of photos and ease of extracting photo data from the application program interface. The figure below indicates the total number of PUD from 2005 to 2012 taken in all 38 parks. As apparent from the graph, there are significantly fewer photos taken in 2005 and 2006 than in later years. Also, PUD peaked in 2011, then declined slightly in 2012.

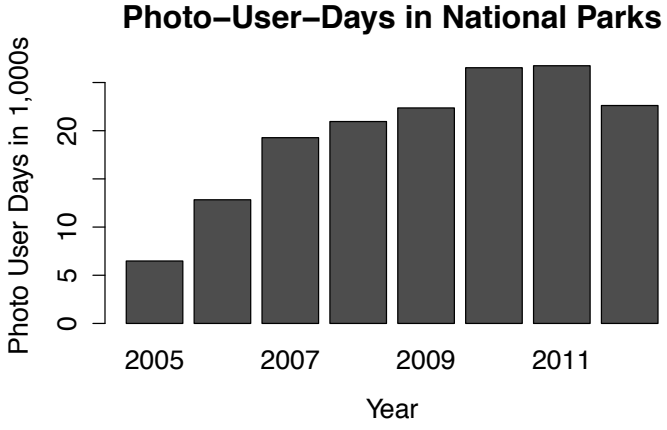

Figure 1.3: Photo-User-Days in National Parks

Biases in Datasets

Most datasets in social science research include sources of bias. Data collected from a sample of a population can rarely be applied towards a population without some degree of error. Rather than trying to eliminate all bias in a dataset, it is advisable to try to understand that bias

and then control for it. These three data sources, like most, include bias. Bias in the photo data tends to be more obvious; only a small subset of visitors to National Parks post photos on Flickr and moreover, Flickr uses likely share common characteristics. However, the NPS data may also be biased, but perhaps in different ways. In attempt to better understand potential biases, I first look at the "average" user of National Parks and Flickr. I then describe more broadly the types of biases potentially introduced with each dataset.

One might expect the stereotypical NPS visitor to be quite different from the average Flickr user—as this section shows, however, the two are more similar than one might expect. Surprisingly, the Park Service does not publish aggregated statistics on the ages or demographics of visitors; rather, they only publish this information through the individual park-level VSP surveys. Therefore, the information provided below is simply an average of a sample of surveys that I study. In that way, it is meant as an approximation rather than an absolute known. Information on Flickr users is taken from Ignite Social Media (Ignite Social Media 2012).

NPS visitors and Flickr users have a relatively similar breakdown in age. The following table shows the percent of visitors/users falling into each age category.

Age Range	Percent of NPS Visitors	Percent of Flickr Users
25 or younger	15%	15%
25-45	25%	45%
45-65	45%	35%
65+	10%	<5%

Table 1.2: Ages of NPS Visitors and Flickr Users

Both user groups have relatively few users under 25 and over 65. Flickr has more users falling between 25 and 24 than NPS, pointing to a slightly younger demographic than NPS. In addition, over half of Flickr users are female, the majority have some college education, and earn between $25,000-$75,000 a year. Due to lack of data, I was unable to estimate these values for NPS visitors.

NPS visitors and Flickr users also originate from similar countries. The ten most common countries of origin for NPS visitors and Flickr users are listed below. Countries written in bold are common between the two datasets.

- NPS visitors: **US, Canada, the United Kingdom,** Germany, Spain, France, Netherlands, Switzerland, **Australia**, and Belgium
- Flickr users: **The United Kingdom,** Singapore, **Spain**, Italy, Ireland, **Canada**, **US**, Taiwan, **Australia**, and Vietnam

These comparisons combat the assumption that NPS visitors and Flickr users come from different demographics. Surprisingly, users from both groups are relatively similar in age breakdown and in country of origin. Moving forwards in comparing the two datasets, it is important to know that while selection bias still exists, the datasets are not drawing upon completely different demographics.

However, neither dataset is a true representation of visitor counts of home locations—rather, both are approximations based on the subpopulation sampled. The following table describes the broad types of biases introduced by each method and their subsequent implication.

	Dataset	Type of Bias Introduced	Example	Implication
NPS	Counting Procedures	Measurement Error	Parks use a PPV to estimate the number of people in a vehicle. The multiplier may be inaccurate, especially for certain parks or seasons.	Introduces variability into the count of visitors.
		Sampling Error	The Parks likely determined the PPV and Percent of Reduction for non-recreational vehicles through a small sample of vehicles at an entrance. This may not be accurate for the entire population.	
	Surveys (home locations)	Undercoverage and Smaller Sample Size	As the surveys are only distributed during a short period of time, they likely miss groups of visitors. For example, perhaps Germans are more likely to travel in September due to school schedules. Conclusions may not be accurate for the entire population.	The reported home locations of visitors represent only a subset of the entire population of visitors. This sample is likely biased in the demographic that is included.
		Non Response	Respondents must mail their survey back to the NPS. There may be a pattern in who is more likely to follow through and mail it in.	
PUD	Count of PUD	Voluntary Response	Users voluntarily post their share their photos on Flickr. Flickr may be used more heavily by certain populations (e.g. 20-30 year old males) and from certain locations (e.g. in England but not France). The sample is biased towards people who own a camera and have the ability to upload photos. Visitors only may only take and post photos when visiting an especially picturesque place or during days of good weather.	Photos only represent a sample of visitors to National Parks. This sample is likely biased in the demographic that is included.
	Home Locations	Voluntary Response	Only selected users choose to share their home location.	

Table 1.3: Biases in Datasets

Moving forwards, I often treat the NPS data as the *truth*, as it is the best (or most accepted) approximation of actual visitation that we have. However, it is important to keep in mind that this dataset is also biased in its own way. Throughout my analysis, I try to further understand the biases invoked and control for them when possible.

Chapter 2: Validating Visitor Counts

Section 1: Objectives

The goal in this chapter is to validate the use of photos as a proxy for NPS visitor counts. Further objectives of this section are to explore whether there are seasons, parks, or regions for which the PUD-NPS relationship is more or less robust. The flow of this chapter is as follows: Section 2 explains the data and shows findings of an exploratory data analysis where I visualize the NPS-PUD relationship to better understand its correlation. Section 3 explains the methodology used in building regression models; it starts with model specification that parameterizes the relationship between NPS and PUD, then follows with model selection that compares the performance of different models; it concludes with constructing additional models specific for each park. In Section 4, I report the results from the chosen model, both for the pooled model analyzing all parks and for the individual park-specific models. Section 5 then discusses the significance and meaning of the results and Section 6 provides concluding remarks.

Section 2: Exploratory Data Analysis

As this chapter focuses on the number of visitors coming to a park, I use the NPS monthly statistics on visitation, as well as the monthly counts of PUD in each park. To gain a better general understanding of the data and their relationship to one another, I use graphs to visualize the relationship between the two datasets. Figure 2.1 (below) plots the number of NPS visitors against PUD to assess their relationship. As shown, there is a strong relationship between PUD and NPS at both high and low levels of visitation—though it is more consistent at higher levels. The relationship also appears relatively stable over time.

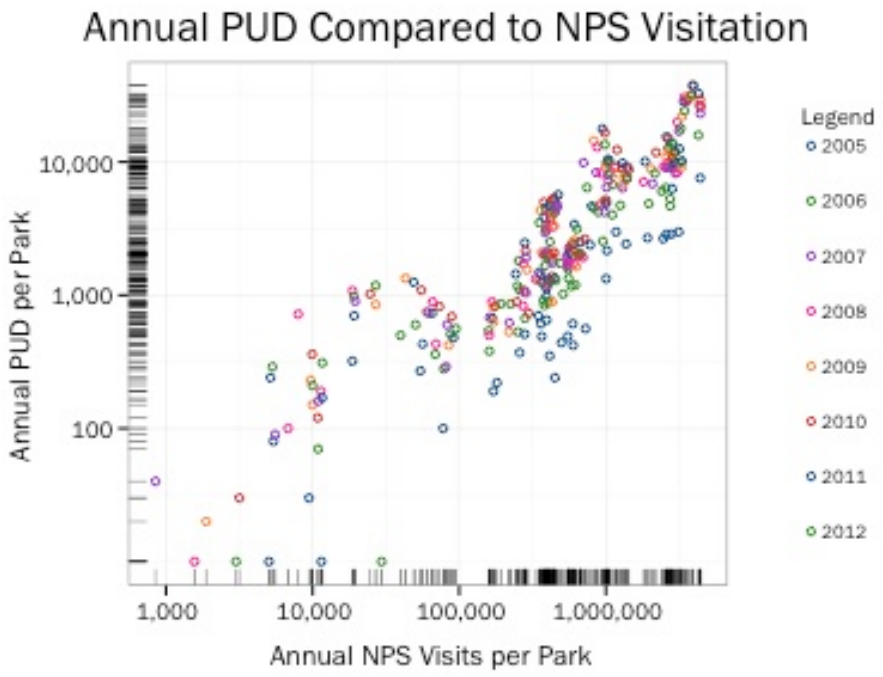

Figure 2.1: Annual PUD Compared to NPS Visitation

To examine the relationship on a finer scale, I visualize how NPS and PUD correspond to each other over the calendar year for each park. To do this, I graph the average monthly visitation rate, measured by each dataset separately. Monthly visitation rates are calculated as the 5-year average between 2008 and 2012.[4] Rather than reporting the number of visitors in absolute terms, I instead report the percent-of-annual visitors per month, thereby facilitating comparison between the two datasets.[5] The goal of this task is to see whether the two datasets give a roughly similar view of seasonal visitation; lines that track well with each other would indicate that the datasets yield similar results; diverging lines would indicate that PUD and NPS give different views of seasonal visitation. Figure 2.2 (below) shows my results, with NPS data shown in black and PUD data shown in red. Parks are color coded and grouped by region.

[4] To compute the average percent of visitors per month, I averaged the percent of visitors for each month between 2008 and 2012. I decided to limit this to the 5-year average as to exclude any biases produced by low photo counts in the beginning years.

[5] The percent of annual is computed by dividing the average number of visits per month from 2008-2012 by the average annual number of visitors from 2008-2012.

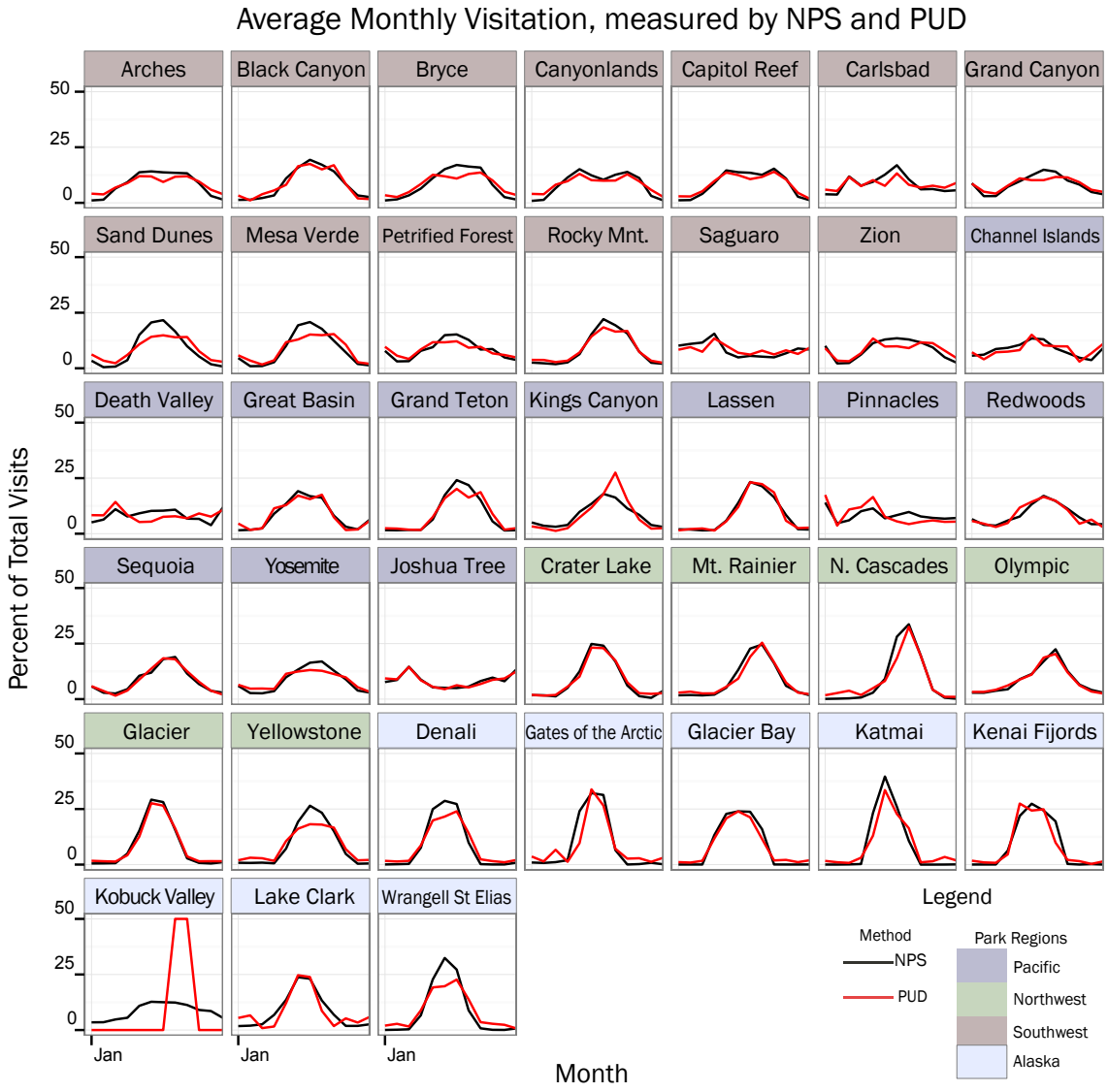

Figure 2.2: Average Monthly Visitation, measured by NPS and PUD

As apparent, the datasets align quite well for most parks. There are, however, noticeable trends of the data fitting less well. In parks with visitation that peaks during the summer, PUD often underestimates NPS visitation during those peak months; examples include Arches, Bryce, Grand Canyon, Sand Dunes, Yellowstone, Denali, and Wrangell St Elias. In addition, Kobuck Valley, which has a very low visitation rate (measured both by NPS and PUD), exhibits a poor fit between the datasets. Other parks with low visitation rates, such as Lake Clark, Katmai, and North Cascades, exhibit much better fits. Overall, these graphics are helpful as a first step in

assessing the relationship between NPS and PUD. It shows that in most cases, the two datasets visually align quite well and thus, warrant further investigation on their relationship.

Section 3: Methods

3.1. Model Specification

While the previous section assesses whether there is a consistent relationship between NPS and PUD, it does little to measure that relationship. To parameterize the relationship between NPS and PUD, I regress NPS on PUD while controlling for other factors.

Latent Assumption in Model Specification: Causation

It is worth pausing here for a brief discussion on **causation**. Many regression models imply a causal relationship between the outcome variable and explanatory variables. Indeed researchers often use regressions to assess whether x has a causal effect on y. In my models, however, this is not the case. While I use PUD as an explanatory variable, it has no causal impact on NPS. A causal relationship here is impossible, primarily because taking and posting photos occurs after one visits the park. Moreover, the two numbers are actually different measures to the same thing—visitor counts. There are two primary ways that researchers deal with atypical causation in regression models: calibration and predictive modeling.

Calibration, also known as inverse regression, is essentially a type of reverse regression. It regresses y on x, but then uses later values of y to predict future values of x (Brown and Sundberg 1987). The goal in calibration is to understand the conditional distribution of x given y (Brown 1982). This concept is both useful and not for my application. First, it is helpful in that it highlights a technique for when y is a known truth and x has an unknown distribution. For this research, I know the distribution of x (PUD), but am unsure as to its correlation with y (NPS); in this way, calibration is not helpful for my application. Further, calibration still implies that x has a causal effect on y, it just changes which is the known factor. Thus, while researching calibration techniques gives me a broader context for methods to build regression models with atypical qualities, in the end it proves not applicable to my research.

The second technique, predictive modeling, uses regression analysis to find a mathematical relationship between a target variable and various independent variables (Dickey

2012). Predictive modeling differs from causal analysis in fundamental ways. Foremost, the objective of causal analysis is determine whether an explanatory variable really affects the response variable and to measure the magnitude of that effect. In predictive modeling, on the other hand, the goal is to determine the relationship between variables in order to make accurate predictions about future values of the response variable (Allison 2014). In this way, predictive modeling invokes nothing about causality; it simply uses the independent variables that hold the most explanatory power, regardless of their causal impact on the response variable.

In my research, I use regression analysis within the purview of predictive modeling. Rather than trying to imply a causal relationship between visitation and photos, I use photos as a metric to predict and understand past, current, and future visitation rates.

Alternative Models

To build a specification with the most predictive power, I explore the performance of multiple regression models. I start with a hypothesized model (Model 1), then build eight alternate models against which to test my results. Each model regresses NPS visitation against a series of covariates, but has a unique set of covariates, functional form, and subset of observations. Table 2.1 (below) describes the specification for each model. Labels in red indicate ways in which the specification differs from Model 1.

		Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**	**Model 9**
Functional Form	Model	Negative Binomial	Poisson	Negative Binomial	Negative Binomial	Negative Binomial	Negative Binomial	Negative Binomial	Negative Binomial	Negative Binomial
Dataset Range	Year	2005-2012	2005-2012	2007-2012	2007-2012	2007-2012	2007-2012	2007-2012	2007-2012	2007-2012
	Park	All Parks	All Parks	All Parks	Excludes Alaska	Excludes Alaska	Excludes Alaska	Excludes Alaska	All Parks	Excludes Smallest
Covariates Included	PUD	Yes	Yes	Yes	Yes	Yes	No	Logged	Logged	Logged
	Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Park	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Sum-mer Month * PUD	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes

Table 2.1: Alternative Models

The following subsections address why I choose various functional forms, covariates, and subsets of observations to test and compare. Each section addresses why I build Model 1 as I do, then the justification for testing other model alternatives.

Functional form

All models use either the Negative Binomial or the Poisson models for its functional form; a decision based upon standard statistical practice. NPS is count data, characterized by observations that have a lower bound of zero and can only take integer values. It is standard to use the Poisson and Negative Binomial models to estimate count regressions (Cameron and Trivedi, 1998). One common issue with the Poisson model is that it sets the variance equal to the mean. In my data, however, the variance is actually larger than the mean. In cases such as these, using the Poisson model results in overdispersion, meaning that there is actually greater variability in the data than the model would suggest. In the presence of overdispersion, statisticians instead use the Negative Binomial model to fit count data (Cameron and Trivedi, 1998). The Negative Binomial uses the Poisson as a starting place, then adds a multiplicative random effect to allow for another free parameter. This additional parameter is drawn from the gamma distribution. To test for overdispersion in my models, I build Model 2 using the Poisson.[6]

Covariates

I start with a model that includes PUD, year, the month of visit, an interaction term between PUD and June/July/August, and the park itself as covariates. The interaction term allows summer months to have a unique relationship with PUD. I include this variable based on the visual observation that PUD underrepresents NPS during the summer months, but tracks well during the remaining months of the year. Model 5 excludes the interaction term between PUD and summer month; this decision is to test whether the inclusion of this term actually adds predictive power.

[6] I originally built another model using a linear ordinary least squares (OLS). However, this model is neither conceptually appropriate (for reasons listed prior) nor comparable to the other models based on my fit criteria. For example, I was unable to compute the log likelihood, AIC, or cross validation results for the OLS model, preventing me from comparing it to the other model specifications. Consequently, I do not report results from the linear model here.

All models include a fixed effect for park. Including the fixed effect allows for a hierarchical model in which observations within each park are correlated and non-randomly related. Including the fixed effect gives each park a unique intercept, thereby controlling for differences among parks that may alter its estimated visitation.

I also vary both the inclusion and functional form of PUD. Model 6 regresses NPS only on month, year, and park—it excludes PUD altogether. The objective of including this model is to assess the added value of the photos. In Models 7, 8, and 9, I use the log value of PUD, allowing me to better understand the multiplicative effect of photos on visitation. When using unlogged PUD as a dependent variable, I am restricted to making conclusions about the impact of one additional photo—this ignores the issue of how many photos were taken at all. Using the logged value of PUD, however, allows me to test the impact of 1% more photos, thereby incorporating the multiplicative effect of additional photos.

Testing Subsets of Observations

I test models that exclude some observations to examine whether I can improve my model fit and predictive power by eliminating certain groups of observations that may bias my results. First, I suspect that the lack of photos during the 2005 and 2006 may skew the model results. In 2005 and 2006, PUD in National Parks totaled 6,479 and 12,833 respectively. The average yearly PUD between 2007 and 2012, on the other hand, was 23,080. While controlling for year does help account for this bias, I believe that photos posted during the beginning stages of the website may have a fundamentally different relationship with visitation than in later year. Thus, I decide to build a model excluding 2005 and 2006 data.

Second, I exclude various subsets of parks. By examining the influence plot of Model 8, I question whether excluding Alaska parks would improve model fit.[7] Alaska has eight National Parks; in every year studied, five or six of these parks rank in the lowest ten annual PUD for all 38 parks. Across all parks, almost 1,000 observations have a monthly PUD below five; over half of these observations are from parks in Alaska. Further, there are 356 month-observations for which there are no photos taken; over three quarters of these observations are in Alaska parks. I hypothesize that the exceptionally small number of photos taken in Alaska parks could skew

[7] A copy of the influence plot is shown on page 31.

other results. I build Model 7, which excludes Alaska parks, to test this hypothesis. In Model 9, I also test whether excluding the ten parks with the lowest visitation rates improve model fit.

Section 3.2: Model Selection

I use five criteria to assess model fit and choose a model, including two fit statistics, two cross-validation tests, and diagnostic graphs. The following information describes each of the criterion and then reports the results of model fit.

Fit Statistics

The **Log-Likelihood** is a fit statistic used for generalized linear models. Specifically, it measures the probability of seeing your data given the estimated parameters. The higher the log likelihood, the better the model fit. Note that often in model comparison, researchers compare not the actual log likelihoods, but rather compute a **Likelihood Ratio Statistic,** the difference being that the ratio allows researchers to account for the lost degrees of freedom when choosing a critical value for significance. I compute likelihood ratios, as well as their significance, for some of the models. I am unable to compute ratios for those models that exclude a subset of parks; because the number of observations between models varies, it changes the degrees of freedom, making the ratios less meaningful for model selection.

The **Akaike's Information Criterion** (AIC) is another goodness of fit statistic that evaluates the fit of the model and its complexity. The statistic improves with model fit, but then penalizes the inclusion of additional parameters; in this way, it is discouraging of overfitting. The AIC is not meaningful on its own, but rather only in comparison to other AIC values. The lower the AIC, the better the model.

Cross-Validation Tests

Whereas fit statistics measure how well the model fits the dataset, cross-validation tests assess the ability of a model to correctly predict outcomes on outside data. Cross-validation splits the dataset; it builds a model on one part, called the "training set" then tests it on the other part, called the "test set." The tool then calculates the difference between the predicted value for the test set and the actual value in the set; the result is the prediction error. Models with better fits

minimize the prediction error. I use two techniques to perform cross-validation techniques; they differ in the number of observations that go into the test and training sets.

K-Fold Cross Validation partitions data into k sections; it then uses k-1 sections for the training set and the remaining 1 section for the test set. It then loops through such that every section is used as the training and test set. I used k=10, which is a standard accepted number in statistical research (Kohavi et al., 1995).

Quite similar to K-Fold testing, **Leave One Out Cross-Validation** (LOOCV) partitions the dataset into training and testing groups. This time, however, k, or the number of folds, is equal to n. Thus, the training set is built on n-1 observations and the test set includes 1 observation. The output from this test is an adjusted estimate of the prediction error.

Results from each of the four statistics are summarized in Table 2.2, found below.

		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Functional Form	Model	Negative Binomial	Poisson	Negative Binomial	Negative Binomial	Negative Binomial	Negative Binomial	Negative Binomial	Negative Binomial	Negative Binomial
Dataset Range	Year	2005-2012	2005-2012	2007-2012	2007-2012	2007-2012	2007-2012	2007-2012	2007-2012	2007-2012
	Park	All Parks	All Parks	All Parks	Excludes Alaska	Excludes Alaska	Excludes Alaska	Excludes Alaska	All Parks	Excludes Smallest
Covariates Included	PUD	Yes	Yes	Yes	Yes	Yes	No	Logged	Logged	Logged
	Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Park	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Summer Month* PUD	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Fit Statistics	Log Likelihood	-40,378	-2.4E+07	-30,324	-24,701	-24,717	-24,886	**-24,434**	-30,064	-29,494
	Likelihood Ratio Statistic	Baseline	n/a	20,108	n/a	n/a	n/a	n/a	20,629	n/a
	Sig. of Likelihood Ratio	n/a	n/a	0	n/a	n/a	n/a	n/a	0	n/a
	AIC	80,863	4.8E+07	60,755	49,493	49,522	49,859	**48,958**	60,234	59,096
	K Fold Prediction Error	2.1E+10	**2.0E+09**	3.5E+10	1.6E+10	1.7E+10	7.5E+9	3.1E+09	4.7E+09	4.5E+09
	LOO Prediction Error	2.0E+10	**2.0E+09**	3.5E+10	1.7E+10	1.6E+10	7.4E+09	3.1E+09	4.6E+09	4.6E+09

Note: Labels in bold indicate the best result.

Table 2.2: Results of Alternative Models

Examining the log likelihoods and AICs, it appears that Model 7, which uses years 2007-2012, excludes Alaskan parks, and uses a logged PUD value, has the best values. Moreover, the statistics show that Model 2, which uses the Poisson as its functional form, is significantly worse than models using Negative Binomial. In addition, Model 5, which excludes the interaction term between PUD and summer month, is slightly worse than the Model 4, which includes it. This suggests that while the interaction term does improve model fit, it only does so slightly. Lastly, comparing Models 4 and 6, which only differ in their inclusion of PUD, shows that Model 6, which excludes PUD, performs only marginally worse than Model 4. This suggests that the explanatory power of month and year alone are actually quite strong. As it turns out, the Likelihood Ratio Statistics were not very helpful for comparing models to Model 1, as many of the other specifications excluded subsets of variables and therefore were inappropriate for computing the ratios.

The results from the cross-validation tests corroborate that Model 7 has the best fit. While the Poisson model has the lowest prediction errors values of $2.0E^{09}$, it is typical that Poisson models underestimate standard errors and thus show overconfidence (Cameron and Trivedi, 1998). For this reason, I do not trust the low delta values. Of the other models, Model 7 has the lowest prediction errors, equaling $3.1E^{09}$. This shows that this model has the best predictive power, meaning that it had the smallest errors between the values it predicted and the true observed values.

Assessment of Diagnostic Graphs

Overall, fit statistics and cross-validation give a snapshot of model fit; diagnostic graphs, however, often give a more comprehensive evaluation of fit. While Model 7 appears most appropriate from its fit statistics, it is good practice to visualize different models. Specifically, I use influence plots and average-versus-predicted plots to assess whether Model 7, which excludes Alaskan parks, performs visibly better than Model 8, which includes them.

Influence Plots are a convenient way to assess how the model fits specific data points and helps to identify outliers that have a significant effect on the model. Figure 2.3 (below) shows the influence plots for Models 7 and 8, which differ only by the inclusion or exclusion of Alaskan parks. The x-axis shows the standardized hat values, an assessment of how much influence an observation has on the model—the higher the value, the greater effect. The y-axis

shows the standardized residual, measuring how much error is associated with that observation in comparison to other observations; the further from 0, the worse fit. When assessing influence plots, one looks for observations that are both high in hat values and extreme in residuals—these points have poor fit and are influential.

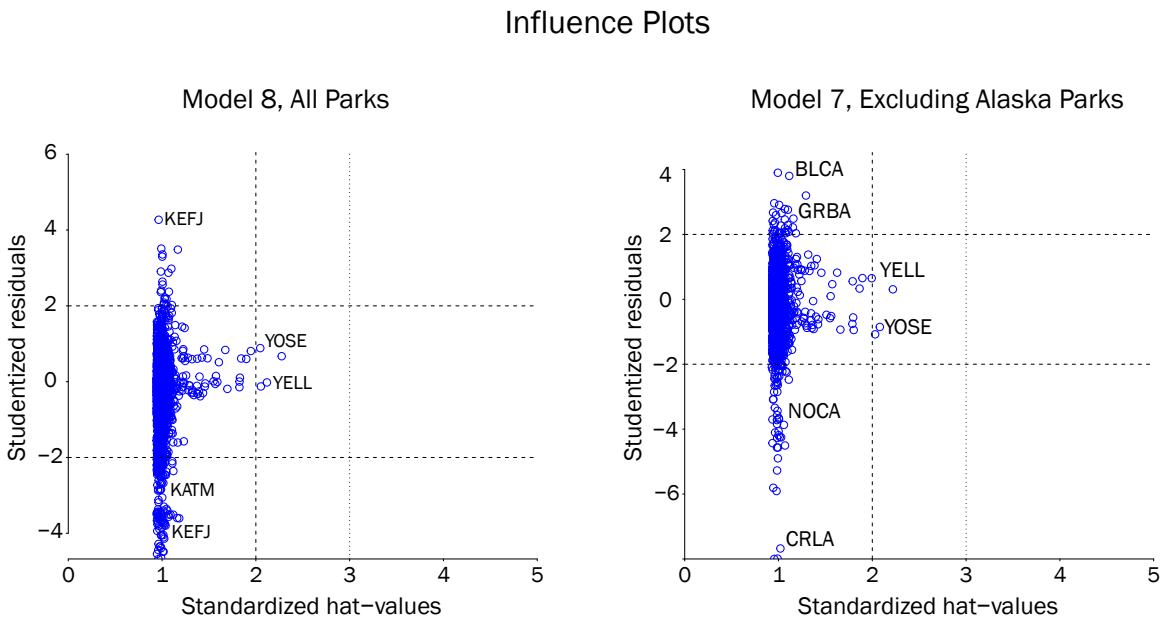

Figure 2.3: Influence Plots

Examining the plots above, it appears that Model 8 has numerous observations with poor fit; many of which are observations from Alaskan parks. However, none of these points have much influence on the model, likely due to their small visitation counts. Therefore, while the model does not fit these parks as nicely, there is little harm in still including them. However, in the plot showing Model 7, there are significantly fewer observations with poor model fit. In both plots and models, there are also several observations, mostly from Yellowstone and Yosemite, which have more influence on the model than other observations—however, these points have small residuals, showing good model fit.

Average versus Predicted Plots are another way to assess the performance of a model. They use the model to predict the outcome variable, then compare the predicted value to that of the actual value. Figure 2.4 (below) examines the performance of Models 7 and 8. The color of the points represents the Park. The x-axis shows the monthly visitation count predicted by the

model; the y-axis shows the actual observed visitation count (NPS). With a perfect model, all points would fall exactly on the one-to-one line.

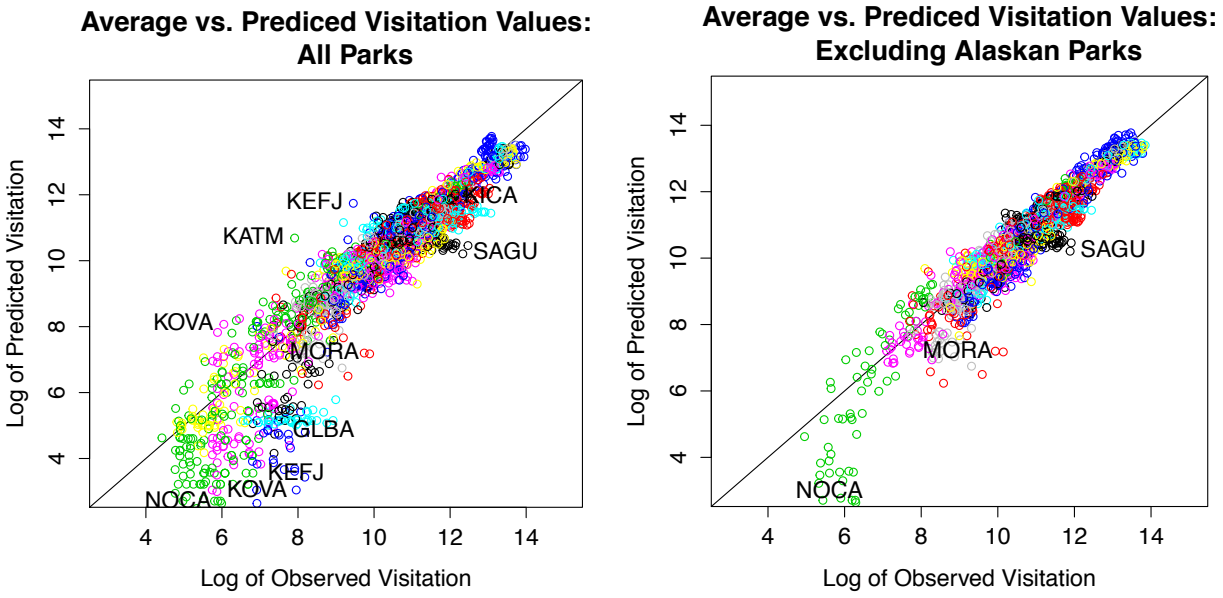

Figure 2.4: Average versus Predicted Plots

Overall, it appears that parks with lower visitation numbers have a worse fit than parks with larger visitation numbers. This is likely due to larger parks having more influence on the model than smaller ones, and thus forcing the model to conform to their trends better. Many of the parks with poor fit in the Model 8 graph are located in Alaska.

Given all of these results, I choose to proceed with Model 7, which uses a Negative Binomial specification with data from 2007-2012 and excludes Alaskan parks. This decision results in a total sample size of 2160 observations across 30 parks. The model specification is as follows:

$$NPS_{ij} \sim \text{Negative Binomial } (\exp(\mathbf{x}_{ij}\,\beta), \theta)$$

Where **X** is a vector of covariates including: the logged value of PUD, year, the month of visit, an interaction term between PUD and June/July/August and a fixed effect for the park.

Park-Specific Models

To answer park-specific questions, I build additional models that apply the same specification as the pooled model to individual parks. Note that in the original model, I include park as a fixed effect, thereby allowing each park to have its own intercept—the slope of the regression line, however, is the same for each park. By building individual models for each park, it allows parks to have both a unique intercept and slope, thereby improving model accuracy. Each model regresses number of NPS-measured visits to a park in month *j* using a negative binomial regression.

$$NPS_j \sim \text{Negative Binomial } (\exp(\mathbf{x}_j \beta), \theta)$$

Where **X** is a vector of covariates including: the logged value of PUD, year, the month of visit, an interaction term between PUD and June/July/August.

Section 4: Results

Model Results

The chosen regression models the number of observed visits to park *i* in month *j* using a negative binomial regression:

$$NPS_{ij} \sim \text{Negative Binomial } (\exp(\mathbf{x}_{ij} \beta), \theta)$$

Where **X** is a vector of covariates including: the logged value of PUD, year, the month of visit, an interaction term between PUD and June/July/August, and a fixed effect for the park.[8] April is treated as the reference case for month and Arches is the reference case for park. The model analyzes data between 2007 and 2012 for National Parks in the western contiguous United States (i.e., excluding Alaska). Table 2.2 (below) shows the resulting coefficients and standard errors:

[8] In the model, I use log(PUD+1) to escape having to take the log of 0.

Variables			Park Fixed Effects		
Variable	**Coefficients**	**Standard Errors**	**Variable**	**Coefficients**	**Standard Errors**
Intercept	41.256	12.140	Arches (reference case)	n/a	n/a
PUD	0.649**	0.020	Black Canyon	0.065	0.095
Year	-0.016**	0.006	Bryce	0.234**	0.080
January	-0.589*	0.051	Canyonlands	-0.319**	0.081
February	-0.482*	0.051	Capitol Reef	0.546**	0.085
March	-0.169**	0.050	Carlsbad Caverns	0.655**	0.089
April (reference case)	n/a	n/a	Channel Islands	0.396**	0.088
May	0.219*	0.051	Crater Lake	-0.405**	0.082
June	0.440*	0.054	Death Valley	-0.163**	0.080
July	0.524*	0.056	Glacier	0.409**	0.081
August	0.484*	0.057	Great Basin	-0.518**	0.099
September	0.381*	0.053	Grand Canyon	0.893**	0.083
October	0.149*	0.051	Great Sand Dunes	-0.412**	0.086
November	-0.288*	0.051	Grand Teton	0.713**	0.080
December	-0.511*	0.051	Joshua Tree	0.760**	0.080
PUD* Summer Month	0.001**	0.000	Kings Canyon	0.740**	0.086
* Significant at the 0.1 level			Lassen Volcanoes	0.055	0.088
** Significant at the 0.05 level			Mesa Verde	0.332**	0.086
			Mount Rainier	-0.013	0.080
			North Cascades	-2.791**	0.094
			Olympic	1.257**	0.080
			Petrified Forest	0.707**	0.085
			Pinnacles	0.351**	0.091
			Redwoods	-0.022	0.082
			Rocky Mountain	0.977**	0.080
			Saguaro	0.846**	0.083
			Sequoia	0.660**	0.081
			Yellowstone	0.127	0.082
			Yosemite	0.519**	0.084
			Zion	0.989**	0.080

Table 2.2: Results from Pooled Model

From this table, we see that across parks, a one-percent increase in PUD is correlated with a 0.65 percent increase in observed visitation. However, during summer months, this relationship actually decreases by .001 percent, as shown by the coefficient on the PUD-Summer interaction term. As expected, the coefficients on month are positive during the summer months and negative during the winter, indicating that the NPS-PUD relationship increases from May

through October, and then decreases during the winter months, all else constant. PUD is significant at the 0.05 level, as is year, most months, and the interaction term between summer month and PUD. While many parks appear significant, the significance of fixed effects are not actually meaningful, as the only measures the relative difference from Arches National Park.

Model Fit (revisited)

The purpose of the initial section on model fit was to assess which model was best suited for my data. The objective of the following section, which revisits diagnostic graphs, is to evaluate the performance of the chosen model. To do so, I conduct an out-of-sample test whereby I rebuild the model using only 2009/2010 data, then predict visitation for 2011/2012 PUD levels. Figure 2.5 (below) illustrates the actual versus predicted values. Again, many of the points fall close to the one-to-one line, with the notable outlier indicated.

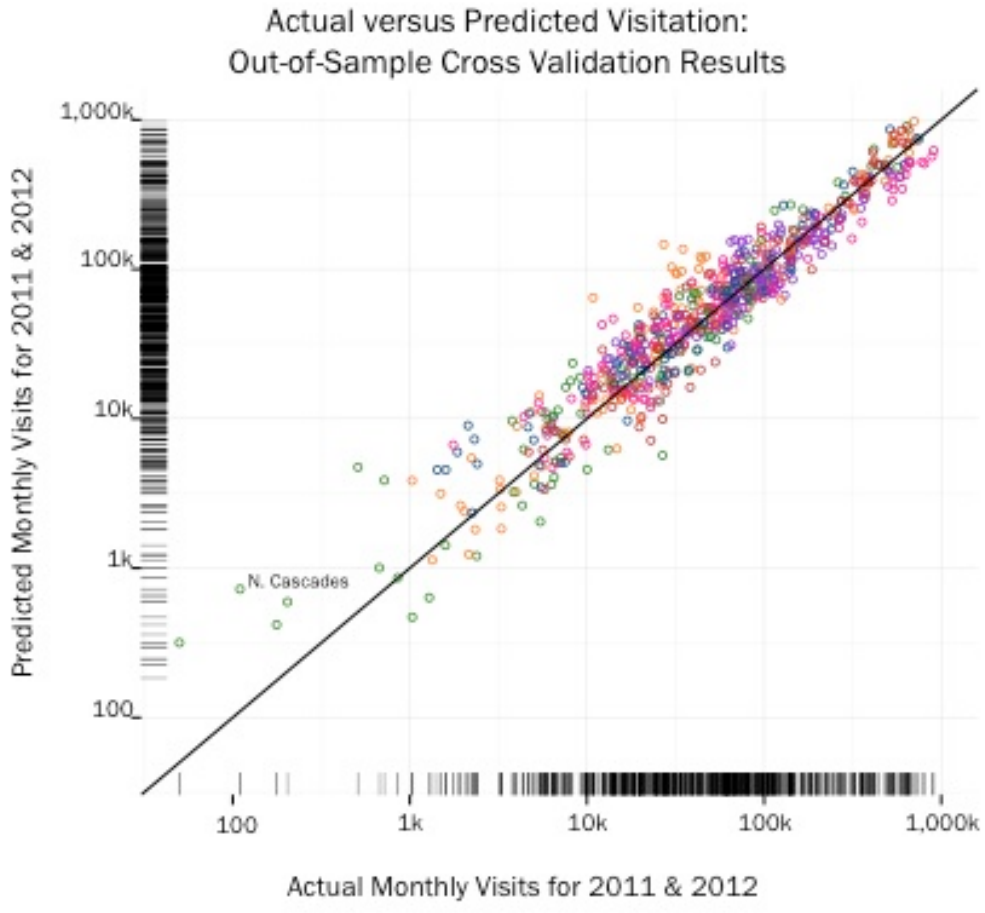

Figure 2.5: Actual versus Predicted Visitation, Out-of-Sample Cross Validation Test

As illustrated, many of the predicted points are quite close to the actual values. The noticeable exception is North Cascades, which exhibits worse model fit.

Results of Park-Specific Regressions

To build models that are more accurate for individual parks, I apply the same specification to individual parks as I did for the pooled model. Table 2.3 (below) shows the resulting coefficients and standard errors for each model:

Park	PUD	Year	PUD*Summer	Intercept
MORA	0.5601**	0.0193	-0.0023	-30.8007
	(0.2036)	(0.0305)	(0.0024)	(61.3163)
NOCA	0.6057**	0.0446	-0.0242	-85.0176
	(0.2062)	(0.0561)	(0.0377)	(112.6665)
CRLA	0.2475**	-0.0075**	-0.0032**	24.1178**
	(0.004)	(4.00E-04)	(1.00E-04)	(0.7259)
ROMO	0.1977**	0.027**	-5.00E-04	-43.7615**
	(0.0741)	(0.0077)	(9.00E-04)	(15.3105)
CHIS	0.1464**	-0.0953**	-0.0016	48.2288**
	(0.0641)	(0.0164)	(0.0147)	(12.4995)
JOTR	0.1352**	0.0082	-0.0011	-4.8973
	(0.0584)	(0.006)	(0.0024)	(12.0605)
YOSE	0.1346**	0.0121**	2.00E-04	-12.6614
	(0.0606)	(0.0059)	(3.00E-04)	(11.7887)
LAVO	0.1231**	-0.0103	-0.0025	29.4406
	(0.0673)	(0.0127)	(0.0043)	(25.4345)
SEQU	0.0954**	0.0089	-0.0019	-7.1826
	(0.0431)	(0.0062)	(0.0014)	(12.3955)
CARE	0.0864**	0.0318**	-9.00E-04	-53.2192**
	(0.0363)	(0.0073)	(0.0042)	(14.6643)
KICA	0.0641**	-0.0098	7.00E-04	29.8922**
	(0.0325)	(0.007)	(0.0027)	(13.987)
ARCH	0.098*	0.0362**	0.0015	-61.67881**
	(0.0592)	(0.0054)	(0.0012)	(10.82)
GRTE	0.0596*	-0.0187**	8.00E-04*	48.2288**
	(0.0351)	(0.0062)	(4.00E-04)	(12.4995)
PEFO	-0.0828*	0.0312*	0.0029	-51.6321**
	(0.0473)	(0.0066)	(0.0043)	(13.2631)
BLCA	-0.105	-0.0653**	8.00E-04	140.2149**
	(.101)	(.0262)	(.0274)	(52.6021)
BRCA	-0.1259	0.0573**	0.0033*	-103.4008**
	(.0825)	(.0093)	(.0018)	(18.6706)
CANY	0.0583	0.0219**	-0.0012**	-33.4055*
	(0.0603)	(0.0093)	(0.0023)	(18.6072)
CAVE	-0.0183	-0.0121*	-0.0033	34.6802**
	(0.0299)	(0.0069)	(0.007)	(13.8597)
DEVA	0.137	0.0518**	-2.00E-04	-93.3211**
	(0.1587)	(0.0197)	(0.0033)	(39.1172)
GLAC	-0.074	-0.0165	0.0013	43.4072**

	(0.0805)	(0.0107)	(9.00E-04)	(21.4397)
GRBA	-0.0144 (0.0425)	0.0562** (0.0109)	-0.0026 (0.0106)	-104.2514** (21.9789)
GRCA	-0.0177 (0.043)	-0.0055 (0.0038)	5.00E-04 (3.00E-04)	23.9622** (7.5518)
GRSA	-0.0454 (0.0316)	0.01 (0.0064)	-0.0011 (0.0036)	-10.7431 (12.947)
MEVE	-0.0052 (0.039)	-0.0028 (0.0074)	0.0048 (0.0041)	15.783 (14.8701)
OLYM	0.0911 (0.0927)	-0.0366** (0.0097)	5.00E-04 (0.0012)	84.9496** (19.256)
PINN	-0.0951 (0.1605)	0.1531** (0.0235)	0.0193 (0.0484)	-296.9537** 47.1273 ()
REDW	-0.085 0.0935	-0.0172 (0.0143)	0.0043 (0.0053)	44.8921 (28.6806)
SAGU	0.0378 (0.0456)	-0.0136** (0.0079)	-0.0046 (0.0037)	38.3939** (15.8638)
YELL	0.0541 (0.0564)	0.0023 (0.0085)	6.00E-04* (3.00E-04)	5.4839 (16.9999)
ZION	0.0316 (0.0347)	0.016** (0.0046)	-3.00E-04 (0.001)	-19.7163** (9.2025)

* Significant at the 0.10 level
** Significant at the 0.05 level

Table 2.3: Results from Park-Specific Models

As apparent, PUD is not always a significant predictor of NPS visitation; it is significant at the .05 level in ten of the 30 parks (at the .10 level, it is significant in 14 parks).[9] Interestingly, PUD is a significant predictor in many of the parks that exhibited poor fit in the model built upon all parks, such as North Cascades, Joshua Tree, and Channel Islands. Also, in Petrified Forest, PUD was both significant and negative (indicating that all else constant, more photos correlate with a fall in visitation), however, the interaction term between PUD and summer month is positive but not significant.

Using the models above, I predict the number of visitors at two parks in 2012 to serve as examples. For Mount Rainier, the regression predicts total visits at 1,142,200 (95% CI 900,093 to 1,384,304); the NPS reports total visitation for that year at 1,049,178, a value within my predicted range. For Great Sand Dunes National Park, I predict 2012 visitation at 285,334 (95%

[9] To test whether PUD has added benefit in the models, I compare two models for each park: Model 1 has only year and month as explanatory variables; Model 2 uses logged PUD, the interaction term between PUD and summer month, year, and month as the explanatory variables. I then compute Likelihood Ratios between the two models for each park. The results from this analysis are located in Table 2A in the Appendix.

CI 271,828 to 298,840). The NPS reported 254,674 visits, 17,154 people, or 6.7% below my lower-bound estimate.

To assess the performance of each model for predicting visitation at its respective park, I construct average vs. predicted plots, similar to the actual versus predicted plot built on the pooled model. Figure 2.6 (below) shows the results:

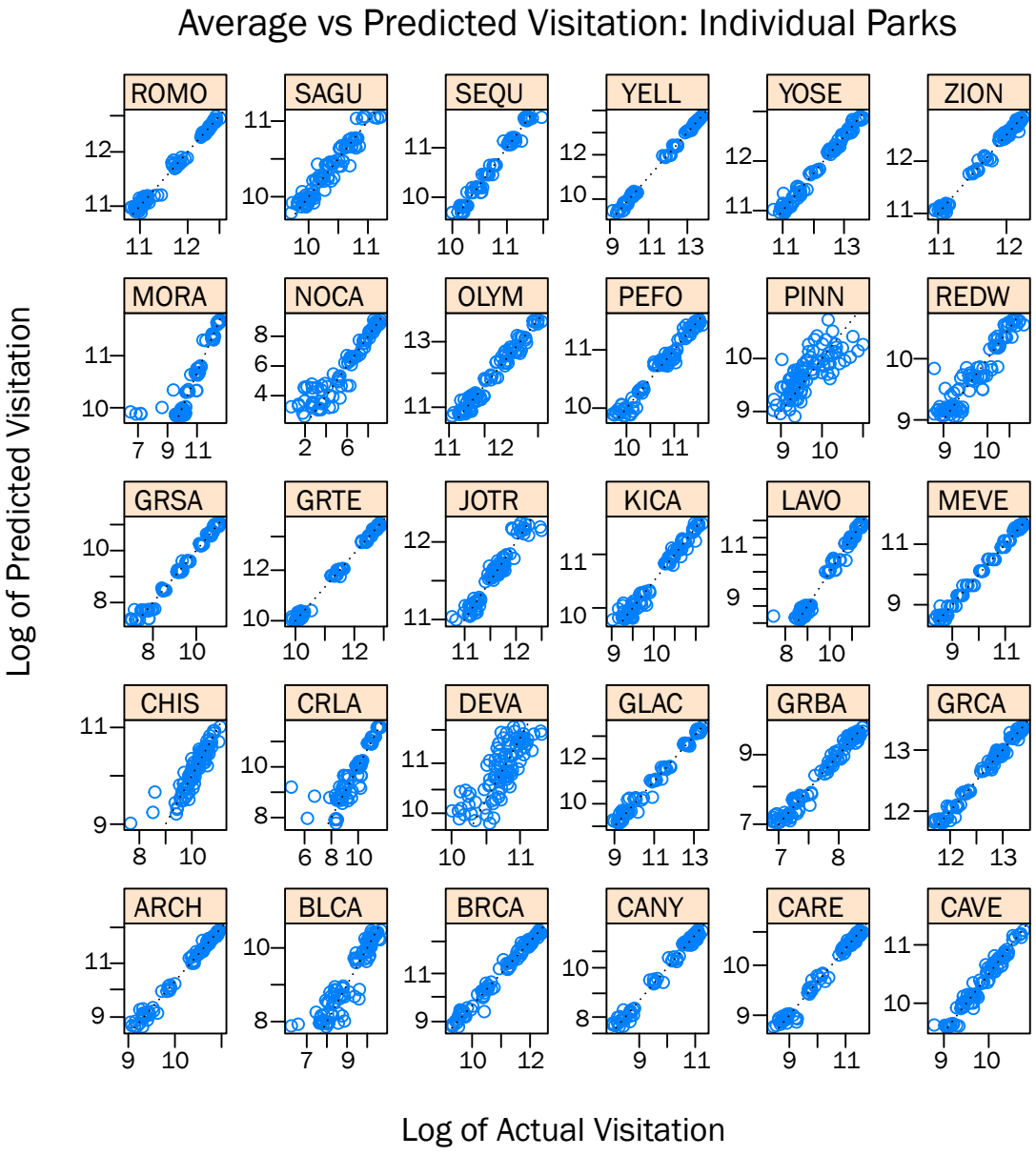

Figure 2.6: Average versus Predicted Visitation Values from Park-Specific Models

Many of predicted values fall quite close to the actual visitation numbers, showing good model fit. In addition, most models are accurate at both low and high visitation counts, examples include Yellowstone (YELL), Yosemite (YOSE), and Zion (ZION). However, six models are more accurate at higher levels of visitation than they are at lower levels (examples are Mount Rainier, MORA, and North Cascades, NOCA). Pinnacles (PINN) and Death Valley (DEVA) show the worst results, with the predicted values lying further from their actual points.

Section 5: Discussion

Obtaining accurate and consistent counts of visitors is one of the most important components of measuring the recreational value of lands. However, its time-intensive and expensive nature can make it a burdensome task on land managers (Ziesler 2015). Using crowd-sourced photo data as a proxy for some empirically measured visitation counts can give a reliably similar view of visitation while reducing the cost and burden on land managers. The previous analysis showed that NPS and PUD have a consistent statistical relationship that can be parameterized, and thus that photos can act as a reliable proxy for empirically measured visitation rates. In this section, I first discuss the performance of the pooled and individual models in predicting visitation. Following that, I discuss the resulting implications for the Park Service and for other land managers.

Model Performance: Ability to Infer Visitation Rates from Photos

Overall, my findings show that it is possible to build a predictive model that uses photos to infer the number of visitors to a park. The pooled model for parks shows that there is a consistent and reliable relationship between photos and measured park visitation. While the model shows good fit at both high and mid levels of visitation, fit worsens at lower visitation rates. North Cascades, which has poor fit in the pooled model, has quite few visitors (averaging 20,000 per year), most of who visit during the summer months. It also has very few photos, averaging 100 per year, with many months having 0 visitors. Similarly, many Alaskan parks, which I chose to exclude from the pooled model altogether, have low visitation rates and photo counts. My hypothesis is that in a model built on all parks, many of which have much larger visitation and photo rates (average of ~85,000 visitors and ~500 photos), the smaller parks have

less statistical power to influence the relationship. The consequence is that the larger parks dominate the regression, causing a less accurate model for the smaller parks.

In the park-specific analysis, I build models that have high accuracy in their predicted values. Many of the models produced estimates falling within the range of the observed visitation counts. Interestingly, however, PUD *is not* always a statistically significant predictor in the park-specific models. Rather than interpret this as a signal that PUD has little additive value, I believe it is rather that month sometimes is an especially strong predictor and may drown out the effect of PUD. In the parks that did have a significant PUD value, there was greater variation in month visitation between different years (i.e., June 2008 was very different from June 2010). Thus, when month was less strong of a predictor, PUD could make a greater impact. This signals that PUD has little additive power in the parks, indicating that the photos may not be very helpful in predicting visitation at these parks. Also interesting, PUD is a significant predictor in many of the parks that exhibited poor fit in the pooled model, such as North Cascades, Joshua Tree, and Channel Islands. I believe that this is related to the unpredictable pattern of visitation at these parks. Because visitation does not follow the standard seasonal trend of visitation peaking in the summer, month is less strong of a predictor, leaving more room for PUD to have explanatory power. Overall, these models show that one can build an accurate model for most parks that predicts visitation based on PUD and other temporal factors. Even for those parks that did not show good fit in the pooled model, like North Cascades and Sequoia, the individual-park model shows an accurate predictive model.

Importantly, I found that while NPS and PUD have a consistent statistical relationship for a given park, the nature of that relationship varied among parks. Comparing the park-specific regressions, the coefficient on PUD varied among the models, thereby indicating that the number of visits per photo changed between parks. This highlights the importance of having some amount of empirical count data against which to calibrate the photos to determine the NPS-PUD relationship best suited for that park. Without data to calibrate, the photos can still show trends in park visitation, but cannot be used to infer total visitor counts.

Overall, the pooled model and the park-specific ones serve difference purposes. The pooled model is useful it allows me assess that there is a statistical relationship between NPS and PUD in the parks studied, though the specific parameterization and accuracy varies among parks. Moreover, it gives the Park Service a tool with which they can measure the impact of discrete

events across all parks. However, this predictive model has little further pragmatic utility for the Park Service—seldom do I imagine the NPS wanting to predict visitation together for all parks in the West. Instead, the individual park models have more pragmatic value, as they allow the NPS to accurately measure visitation at a specific park and track how this changes with events. For example, by building a model just on Mount Rainier National Park, the Park Service can track how visitation correlates with sunshine, or in other words, how many more visitors come to the park when sun is forecasted rather than rain or clouds.

Applications for the NPS

The Park Service can use photo data in two primary ways. First, it can use the regression models to infer the number of visitors to their park based on the number of photos and the month. For example, the Figure 2.7 (below) illustrates the expected number of visitors based on different values of PUD for Yosemite and the Grand Canyon. The blue lines illustrate expected visitation for June, while the red lines shows visitation for January.

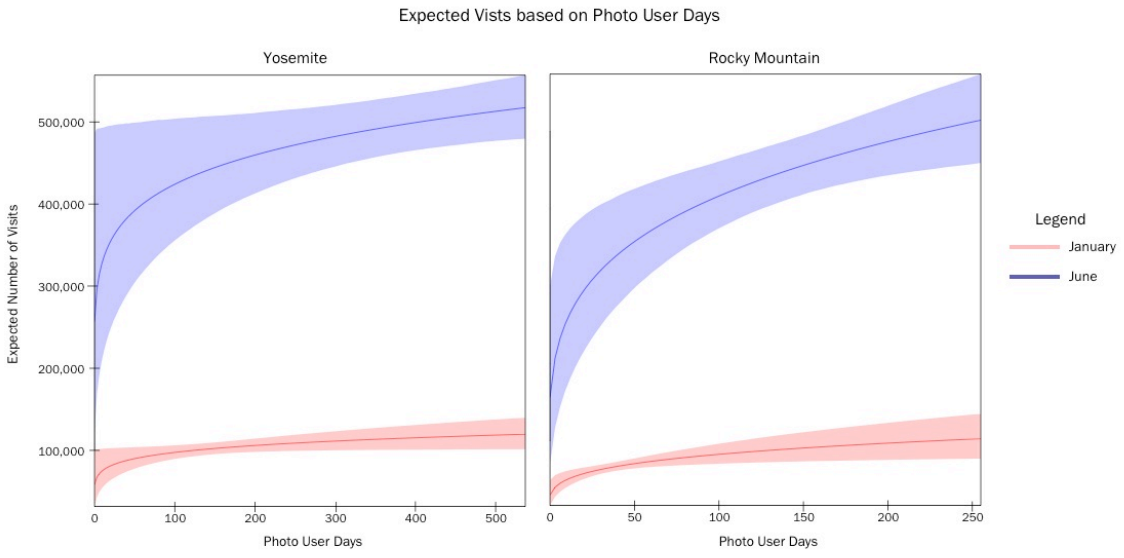

Figure 2.7: Expected Visits to Yosemite and the Grand Canyon based on PUD

Findings such as these will be most helpful to augment the NPS's current visitation counts. The NPS currently collects visitor count data for all entrances at all parks for every month; these data

are then compiled by one NPS employee (Ziesler 2015). Rather than collecting all of this data, the NPS could instead only collect a portion to use in conjunction with the photo data. For example, if the NPS collected a stratified sample of visitors, they could use this as training data against which to calibrate the photos. They could then use the calibrated photo data and regression output to predict the missing observed-visitation values. In addition, they could use the data from the stratified sample to continue monitoring the NPS-PUD relationship and watch for changes over time. The ability to use photo data in place of some NPS data would allow the NPS to collect data less frequently, which could be especially beneficial for remote or hard-to-measure locations. As the photo data gives an accurate view of visitation, this process would leave the Park Service with the same information that they currently have, but for a fraction of the time, effort, and cost.

Second, the NPS can use the photo data to provide a finer-scale look at the counts of visitors. Currently, the Visitor Use Statistics Program only collects monthly visitation totals from parks. According to Pamela Ziesler, the officer in the program, only a portion of parks retain numbers on weekly or daily counts due to the high cost and burden of doing so (Ziesler 2015). However, her office repeatedly fields requests for this data. For parks with a large enough sample of photos, the NPS could use the photo data to infer information about daily counts of visitors. The ability to analyze daily visitor counts is a large value-added and could be applied to a myriad of research fields. For example, the NPS and researchers could use this data to correlate visitation with weather, landslides or other large natural events, road closures, weekends and holidays, or other events of interest. They could also leverage the temporal nature of the photos to study trends or events on time-scales more narrow than the monthly level.

Applications for Other Land Managers

For assessing recreational value at other lands, land managers can apply this technique to places that lack consistent visitation data. Again, managers could conduct a stratified sample to obtain baseline visitor counts and calibrate the photo data. They could then use PUD to track visitation over time to detect further trends. This is an enormous value added and opens the possibility of assessing recreational value at many different types of lands, including remote ones.

For example, the Forest Service (FS) runs the National Visitor Use Monitoring program, which measures recreational visits to Forest Service land. Due to the large number of lands that

the FS manages, they sample each National Forest once every 5 years. The program started in 2003; they have recently finished the second round of surveys (USDA Forest Service 2012). Thus, the FS has an estimate of annual visits to an area, but they only have two estimates for each area. In their Summary Report, they are quick to state that this data is a "snapshot" of visitation and "does not yet provide any true trend measures" (USDA Forest Service 2012). The Forest Service could use this data as baseline and training data for the photos. They could then track visitation with the photos to detect trends. Continuing with their 5-year measurement cycle would allow them to periodically check the calibration of the photos.

This technique could also be especially helpful to monitor recreation in lands currently not studied or under federal management. For example, many backcountry ski enthusiasts visit sites that have little tracking monitoring of visitor use. Researchers could use the photos to estimate visitation rates at various popular backcountry ski areas to better understand what places are most popular and start managing the lands to better cater to user's needs (and potentially charge a user fee to extract people's consumer surplus).

Lastly, with further study, researchers could also refine the regression equation and decrease the need for calibration numbers. With more information, for example, one could build a model for recreational lands featuring x attributes and located y distance away from population center of z size. This would allow managers to predict visitation at areas that were perhaps never surveyed.

It is worth noting here that this methodology is also not limited to photos posted on Flickr. I analyzed photos on Flickr due to the large number of photos posted and ability to extract data from the website. However, researchers with access to other photo-sharing websites, such as *Instagram* and *Facebook*, could apply the same methodology to these datasets.

Section 6: Conclusion

In this chapter, I built a predictive model that uses photos online to infer the number of monthly visitors to a park. By building a pooled model using 30 parks together, I developed a model that had relatively accurate predicted values at higher levels of visitation, but less accurate values at lower levels. To remedy this problem, I built models specific to each park, the vast majority of which had good predictive power. Importantly, I found that while NPS and PUD

have a consistent statistical relationship for a given park, the nature of that relationship varied among parks—meaning that the number of visits per photo changed between parks. This highlights the importance of having some amount of empirical count data against which to calibrate the photos. The ability to use photos to infer the number of visitors to a site allows the NPS to reduce their reliance on measuring visitor counts, as well as opens the possibility measuring and tracking visitation at remote locations.

Chapter 3: Validating Visitor Home Locations

Section 1: Introduction and Objectives

The purpose of this chapter is to assess the validity of using photos to make conclusions about the home locations of National Park visitors. The National Park Service spends considerable time and money in conducting Visitor Use Surveys. While the main objective of the surveys is to assess visitor satisfaction, one output of interest is visitor's home locations. The Park Service uses this information to better understand trends in visitation and to provide relevant information to visitors on how to reach the park (Collins 2015). In addition, researchers use the home locations obtained by the surveys to estimate visitor's travel distances to a park, which they in turn use as an inputs to travel cost studies (Heberling and Templeton 2009; Neher et al. 2013). I hypothesize that the NPS can obtain the same information on visitor's home locations from photos posted online, thereby alleviating one need for the surveys. Obtaining home locations from the photos rather than surveys would provide a low-cost, less time consuming approach to access visitor information.

By comparing the home location information provided by the Flickr photos and NPS VSP surveys, I explore how the proportion of visitors from various locations differs between the datasets. Further, I assess whether the datasets indicate similar results on where visitors originate from and therefore, the distance they travel to reach the park. Specifically, my objective is to answer the following questions:

- Do both datasets indicate that proportion of visitors from each county, state, and country are statistically similar? Moreover, what is the relationship between the proportions reported by NPS and by PUD and do any trends appear among parks, years, or seasons?
- Given that I can calculate the distance between a visitor's home and the park visited, is the distribution of distances statistically different between datasets?
- As the NPS surveys and Flickr photos are both samples from the population of visitors, are they statistically likely to have been drawn from the same population?

If the photo data matches that provided by the surveys, the photos could serve as a proxy for the survey data of home locations, thereby providing a method of measurement that is less expensive and easier to obtain.

Throughout this analysis, I treat the survey-obtained home locations of visitors as an empirical truth against which I test the accuracy of the photo-derived home locations. While it is necessary to have a truth or benchmark against which to test the photo data, it is inaccurate to assume that the NPS survey is a true representation of all visitors to the park. As explained in Chapter 1, the NPS conducts the surveys over a 2-week period in the park and at various attractions within each park. In this way, the surveys are a subsample of all visitors, making the information obtained from them a snapshot of information rather than a holistic representation of the population. For example, if surveys are conducted during two weeks in August, the survey may show a majority of out-of-state visitors, many of who visit during summer vacation but not other times of the year. Additionally, as the surveys only last two weeks, they may capture a one-time event that is not necessarily common. For example, it could sample a tour group of German visitors; while the group may make up 5% of all visitors sampled, the German visitors may not make up 5% of all visitors to the park in that year. Overall, the NPS sample is likely not representative of all visitors to the park; however, the lack of further data and need for an empirical truth necessitates that I use the survey data as a baseline against which to test the photos.

Section 2: Data

Analyzing the home locations of visitors requires additional data processing beyond what I described in Chapter 1. This section explains the system for obtaining home location information from the NPS surveys and from the photos, including matching data between the surveys and photos, steps for processing the data, and the resulting dataset.

Matching Home Locations between NPS and Photo Data

The NPS conducts most surveys over a two-week period at various locations in the park. To compare the results from the surveys to those from the photos, I select only the photos taken within a park during a similar date range. For example, the NPS conducted the Mount Rainier survey from August 4-10, 2012. To best try to approximate this sample, I would ideally examine photos taken during this same week at Mount Rainier. However, due to the small number of photos taken within a one-week duration at many parks, I chose to expand my photo sample to include photos taken within the same three-month span. For surveys conducted in August, for example, I analyze photos taken in July, August, and September. For Rainier, this leads to an NPS sample size of 1,842 and a PUD sample size of 217. I find that using the 3-month time span for photo best balances maximizing the sample size while still staying accurate to the season.

Extracting Home Locations from NPS Survey Data

I analyze 16 NPS surveys conducted between 2005 and 2012 in National Parks in the Western US. Each survey follows specified guidelines and asks similar questions, the details of which I described in Chapter 1. Relevant to this chapter, surveys ask respondents to list the home zip code of each visitor in their party (or their home country for international visitors). From these responses, I compile a dataset where one observation is one visitor's home zip code; the month and year of the survey are also noted. I then assign the visitor's home county, state, and country to each observation. I gathered this information by linking the visitor's zip code to a spatial database of US and international boundaries. I also assign a rough estimate of the distance between the visitor's home location and the park. From the visitor's zip code, I use a database of populated places to assign latitude and longitude coordinates to each location. These coordinates correspond to the spatial center of the zip code. Using GIS, I then calculate the Euclidean distance between the origin's coordinates and those of the geographic center of the National Park. The end output is the Euclidean distance, measured in kilometers, between the visitor's home location and the park.

Creating a Dataset of Photo-Taker's Home Locations

I use information from Flickr and GIS to map the home location of visitors who post photos. A portion of Flickr users posts a Flickr profile in which they voluntarily indicate their home location. Flickr profiles give a blank space in which users can write-in their home location; for example, a user might report "Seattle, WA." The detail in which users report their home varies; however most people indicate their city and state, or city and country if they are outside the United States. I use a database of populated places to map the home location indicated by each photo user to latitude and longitude coordinates. The database assigns the coordinates of the center of the self-reported home location (i.e., the latitude and longitude for the center of Seattle, Washington). Then, using additional spatial databases of US and international boundaries, I assign each latitude longitude coordinate to a specific county, state, and country. International locations are only assigned a country. The final output is a dataset indicating the county, state, and country of every "photo visitor"; each record is indexed by visitor, park, month, and year. I then use the same latitude/longitude coordinate of the reported home location to calculate the distance of the park from the user's home. Using the same method as done with NPS, I calculate the Euclidean distance between the home location and the center of the park.

Section 3: Methods

I use two metrics of testing to assess whether the photos and survey data give a similar view of visitors' home locations. The first metric takes a high level look at the types of visitors to a park (in-state, out-of-state, and international visitors) and compares the proportion of each type measured by each dataset. This analysis uses the actual home zip codes to categorize and compare visitor proportions. The second metric is a finer resolution analysis that compares the distance between a visitor's home location and the park visited. Analyzing distances, rather than actual home locations, zeros in on the information most useful for travel cost studies. It tests whether the photos and surveys give a similar picture of the distribution of visitors' distance between homes and parks and assesses whether the two techniques may yield similar results in a travel cost study. Even though comparing distances is less precise than comparing the homes locations (it compares a radius in which the visitor could live rather than a specific site), this analysis is a finer resolution as it compares each observation without binning it into a category. I

compare distances between homes and parks for all parks pooled together, as well as for each individual survey period. Overall, these metrics should form a robust analysis that examines the accuracy and precision of the photos in inferring visitor's home locations. The following subsections describe the methodological techniques used for each metric.

Comparing Types of Visitors through Home Locations

I compare visitor's home locations using two different categorizations. The first categorization describes visitors as in-state, out-of-state, or international, then compares the proportion of visitors in each category measured by each dataset. I compare these proportions for each survey. The second categorization takes a more in-depth view by dividing people by their home county. I compare the proportion of visitors from each county for Mount Rainier and Rocky Mountain National Parks, two parks that have a high proportion of in-state visitors. Lastly, an additional analysis comparing the home states of domestic visitors and home countries of international visitors to Mount Rainier is located in the Appendix. The purpose of these analyses are to use the comparisons to assess whether photos and surveys give a similar view of from where people are traveling to reach the park.

I compare proportions through two primary methods: visual comparison and statistical tests of equal proportions. Visual comparison allows for a quick assessment of how well the datasets align. It also allows the reader to see the raw numbers and thereby make his own judgment on the degree to which the datasets align or differ. Using statistical analysis, on the other hand, provides a quantitative and more objective measure of the similarity of the datasets. Together, these two techniques form a well-rounded analysis of the accuracy of using Flickr photos to infer the home locations of visitors.

Comparing Distance between Visitors' Home Locations and the Park

Comparing how each dataset reports the distribution of distance between visitors' homes and the park visited serves as a finer evaluation of using photos in a travel cost method. This analysis presupposes that the purpose of researching the home locations of photo users is to conduct an analysis of visitors' willingness-to-pay to reach a recreational site. Taking this assumption, it skips the interim step of analyzing the home locations and validates the quantity of interest—the distance that people travel to reach the park. This analysis also assumes that the

distance between a visitor's home location and the park can be meaningfully translated into the distance traveled—a process that I explore in Chapter 4.

I use several methods to analyze how the two measurement techniques report the distance between visitors' home location and the park. First, I gain a holistic view of the datasets by comparing descriptive statistics and the cumulative distribution functions for all parks and survey periods pooled together. Next, I dive into the specific survey periods and compare the distances specific for each of the 16 surveys. I do this by visually comparing the kernel density plots, conducting a Mann-Whitney U-Test, and conducting a Kolmogorov-Smirnov test. Kernel density estimation is a non-parametric technique to estimate the density function of x given y (Cameron and Trivedi 1998). In my case, it shows the distribution of distances traveled by visitors. By graphing the density function for both photos and surveys, I visually compare whether the distributions align.

Both the Mann-Whitney U-Test and the Kolmogorov-Smirnov test (KS test) are nonparametric techniques used to test whether two samples are independent. The Mann-Whitney U-Test is appropriate for data that may not fit a normal distribution; however, it assumes equal variances between datasets. The technique pools the samples and rank-orders the observations. It then tests whether observations from the samples are clustered together or randomly distributed throughout the set; a random distribution of observations implies that the samples are not different, while a clustering implies that the samples are different (Corder and Foreman 2009). The KS test evaluates whether two independent and random samples are drawn from the same population. It allows data to take a non-normal distribution and allows for unequal variances between the two samples. The technique compares the cumulative distribution functions of each sample. It finds the maximum vertical difference between the distributions, then uses that number as the D statistic (Kirkman 1996). This statistic, thereafter, is used to determine whether the distributions are statistically similar or different (Corder and Foreman 2009).

Together, these tests evaluate whether the samples of visitors obtained from the surveys and the samples from the photos are statistically different. If different, the samples likely show a different picture of how far visitors travel to reach the park. If statistically similar, however, then the samples likely draw from the sample population, indicating that either could accurately represent the distance visitors' may travel to reach the park.

I conduct each test twice; once using all observations of distance between visitor homes and the park, then again on a subset of observations that excludes international visitors. My hypothesis is that the photos may oversample international visitors, as I imagine that visits are more exotic and exciting for international visitors and that Flickr users may be more likely to post photos for "exciting" trips. Wood et al (2013), found that Sri Lankan and Indian tourists upload fewer photos at nearby Nepalese recreation sites than predicted from the overall trend; they conclude that local visitors may be less likely to take or share photographs of nearby sites. Should this trend appear it my data, it would create an upward bias of the photo distance distribution. Therefore, analyzing only domestic visitors may improve how well the NPS and photo distributions match each other. I also experimented with weighting photo distances to improve fit between photos and surveys, though this analysis proved ineffective and is consequently not discussed here.

Section 3: Results

Comparing Types of Visitors through Home Locations

I find that the surveys and photos indicate a similar proportion of in-state visitors, yet sometimes different proportion of out-of-state and international visitors. Figure 3.1 (below) illustrates the proportion of visitors from each location for each survey period, measured by each dataset. The vertical line in each column represents the proportion of visitors reported by the NPS. The location of the dot shows the number of percentage points by which the photo reporting differed. The solid dots represent surveys for which the difference was less than 10%; the hollow dots represent differences greater than 10%. Of the 16 survey periods tested, 12 of the in-state photo-estimated proportions fall within 10% of the survey-estimated proportion. Seven of the 16 out-of-state photo-estimated proportions are within 10% of the NPS estimate, and six of the international proportions are within 10%. Two of all the observations fall more than 25% away from each other. Overall, the surveys and photos align relatively well, especially for representing the proportion of in-state visitors. While the photos are often not widely inaccurate for out-of-state and international visitors, they do appear to underestimate the proportion of out-of-state visitors and overestimate international visitors. Moreover, the photos tend to be more accurate during the summer months. Of the five surveys conducted during the winter, one survey

has no visitor categories falling within 10% of the survey estimate, three surveys have one of the three visitor categories falling within 10%, and one survey has two of the three categories falling within 10%. Lastly, there appears to be little consistency among parks or years for when the survey and photo data align and when they do not.

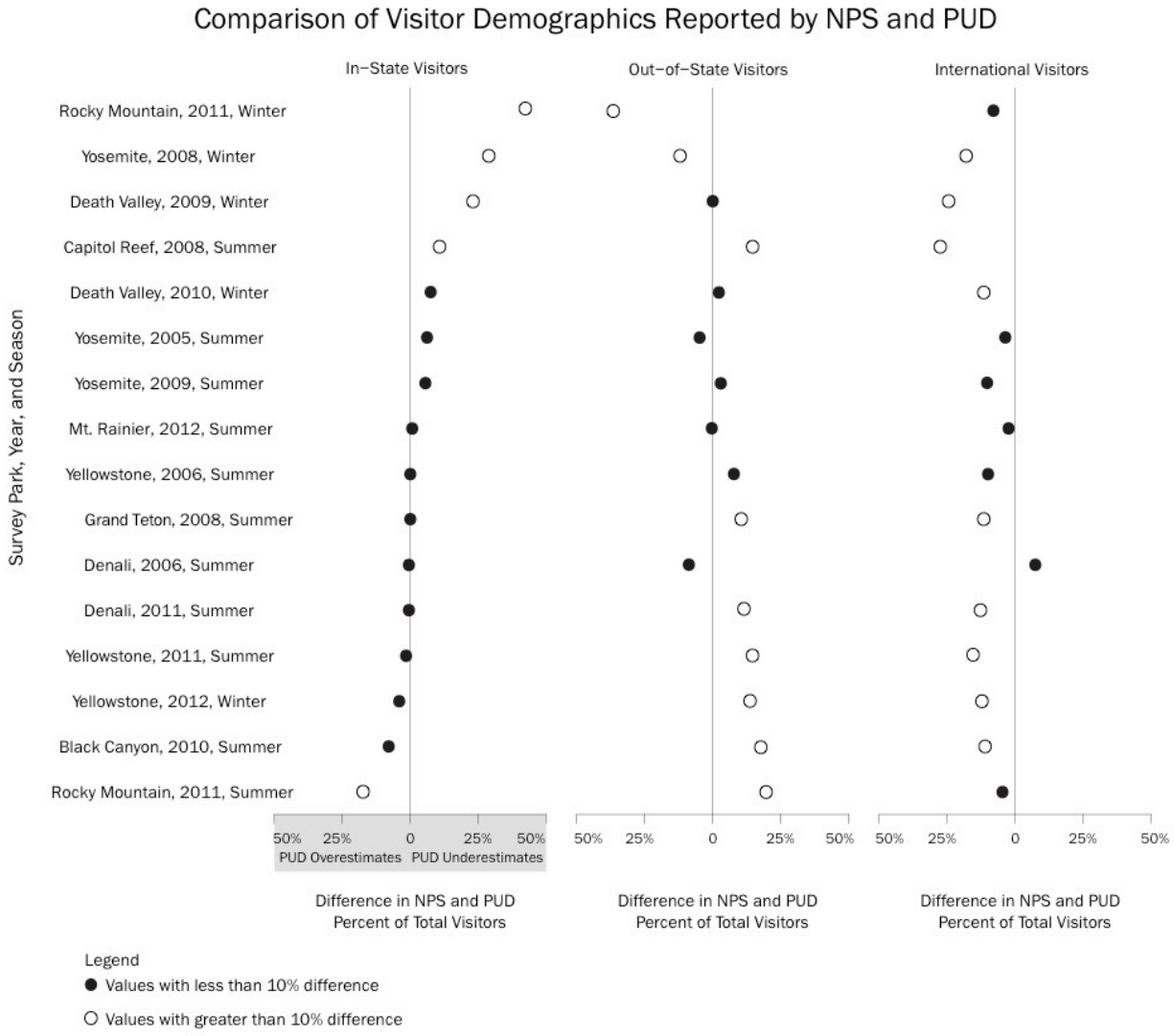

Figure 3.1: Comparison of Visitor Homes

To quantitatively assess the relationship between the proportion of visitors in each category indicated by photos and by NPS, I conduct a test of equal proportions between the two datasets. The test assesses whether the two proportions indicated by each dataset are statistically different from each other. Table 3.1 (below) reports the results for each survey period, indicating those periods that are found *not* to be statistically different at the .05 level. Of the 16 survey periods, 11 are statistically *not different* for the proportion of in-state visitors, eight for out-of-

state visitors, and only five for international. This analysis corroborates my findings from the visual comparison—that photos and surveys align well for in-state visitors, okay for out-of-state, and less well for international visitors.

Park	Year	Season	In-State		Out of State		International	
			P Value	Result	P Value	Result	P Value	Result
BLCA	2010	Summer	0.685	No Difference	0.186	No Difference	0.162	No Difference
YOSE	2006	Summer	n/a*	No Difference	0.191	No Difference	0.191	No Difference
CARE	2012	Summer	0.729	No Difference	0.980	No Difference	0.345	No Difference
DENA	2005	Summer	0.119	No Difference	0.313	No Difference	0.448	No Difference
GRTE	2008	Summer	0.262	No Difference	0.179	No Difference	0.002	
YELL	2010	Winter	0.097	No Difference	0.554	No Difference	0.000	
MORA	2011	Summer	n/a*	No Difference	0.003		0.003	
DEVA	2008	Summer	0.900	No Difference	0.001		0.000	
YELL	2006	Summer	0.527	No Difference	0.001		0.000	
YELL	2011	Summer	0.398	No Difference	0.000		0.000	
YOSE	2012	Winter	0.441	No Difference	0.003		0.005	
DENA	2009	Winter	0.000		0.975	No Difference	0.000	
DEVA	2009	Summer	0.026		0.191	No Difference	0.000	
ROMO	2011	Winter	0.000		0.000		0.064	No Difference
ROMO	2010	Summer	0.000		0.000		0.000	
YOSE	2008	Winter	0.000		0.000		0.000	

* Cells with n/a values had 0 in-state visitors measured by both NPS and PUD

Table 3.1: Results of Test of Equal Proportions

For a more detailed look at in-state visitors, I also compare the proportions of visitors from each county, measured by each dataset. I compare these proportions for Mount Rainier and Rocky Mountain National Parks, as these parks have a high proportion of in-state visitors and are located in states with relatively few counties.[10] Table 3.2 (below) shows the proportion of visitors from each county, reported by each dataset.

[10] For Rocky Mountain NP, I pooled the 2010 Summer and 2011 Winter together for a larger sample size.

Washington Visitors to Mount Rainier			Colorado Visitors to Rocky Mountain		
County	**NPS Proportion**	**PUD Proportion**	**County**	**NPS Proportion**	**PUD Proportion**
King	43.74%	72.50%	Larimer	28.44%	15.73%
Pierce	23.03%	16.67%	Boulder	14.94%	25.84%
Thurston	9.06%	1.67%	Denver	9.59%	29.21%
Snohomish	6.26%	0.83%	Jefferson	9.46%	6.74%
Benton	2.89%	0.00%	Weld	8.81%	3.37%
Yakima	2.79%	1.67%	Arapahoe	6.46%	7.87%
Clark	2.02%	0.83%	El Paso	5.35%	1.12%
Kitsap	2.02%	0.83%	Adams	5.15%	4.49%
Cowlitz	1.25%	0.00%	Douglas	3.72%	1.12%
Spokane	1.06%	0.83%	Grand	3.07%	0.00%
Mason	0.87%	0.00%	Broomfield	1.04%	0.00%
Whatcom	0.77%	0.00%	Pueblo	0.52%	0.00%
Jefferson	0.67%	0.00%	Lincoln	0.46%	0.00%
Lewis	0.58%	0.00%	Elbert	0.39%	0.00%
Chelan	0.48%	3.33%	Eagle	0.33%	0.00%
Skagit	0.48%	0.83%	Teller	0.33%	0.00%
Whitman	0.48%	0.00%	Clear Creek	0.26%	0.00%
Franklin	0.39%	0.00%	Summit	0.26%	0.00%
Kittitas	0.39%	0.00%	Logan	0.20%	0.00%
Klickitat	0.29%	0.00%	Mesa	0.20%	0.00%
Clallam	0.19%	0.00%	Morgan	0.20%	0.00%
Walla Walla	0.19%	0.00%	Alamosa	0.13%	0.00%
Island	0.10%	0.00%	Garfield	0.13%	0.00%
			Jackson	0.13%	0.00%
			Moffat	0.13%	0.00%
			Park	0.13%	4.49%
			Gilpin	0.07%	0.00%
			Gunnison	0.07%	0.00%
			Routt	0.07%	0.00%

Table 3.2: Proportion of Visitors from each County, Mount Rainier and Rocky Mountain National Parks

Again, the overall trends of photo and NPS visitors are similar, but the specific proportions differed by up to 30%. The photos and surveys align well for determining the three

or four counties in which the most visitors resided; however, this information could likely be inferred simply from the county populations. For both Mount Rainier and Rocky Mountain, the three counties most represented in the NPS surveys match the three counties most represented in the photo data. However, the photos are less accurate in conveying the specific proportion of visitors from a specific county. The difference between the NPS-measured and photo-measured proportion of visitors from each county ranges from 0 to 30%, with an average deviation of 2.64%. Examining visitors to Mount Rainier, for example, photos report that King County residents make up 73% of Washington visitors, while the NPS report only having 44%—a difference of 29%. Similarly for Rocky Mountain, photos indicate 30% of visitors coming from Denver County, but the NPS report only 10%.

Comparing Distance between Visitors' Home Locations and the Park

I assess whether the photos and surveys give a similar indication of the distance between visitors' home and the park visited on two levels: a pooled level whereby I analyze all parks, years, and seasons together and disaggregated level in which I analyze each survey period individually. In the pooled analysis, I find that the distances represented by each dataset yield similar statistics. Figure 3.3 (below) shows descriptive statistics for each dataset. As shown, the descriptive statistics for all visitors reported by the two datasets are strikingly close, though the standard deviation is quite different. For all visitors, the mean distance between home locations and the park is 2,275 km measured by NPS (± 2,620km) and 2,787 km measured by the photos (± 3,325). For domestic visitors only, the mean distance is 1,590km according to NPS and 1,308km by the photos; a difference of only 282km (standard deviations of 1,583 and 1,288km respectively). In conducting a test of equal variances, I find that in both cases, there is a statistically-significant difference between the variance of survey-measured distance and photo-measured distance.

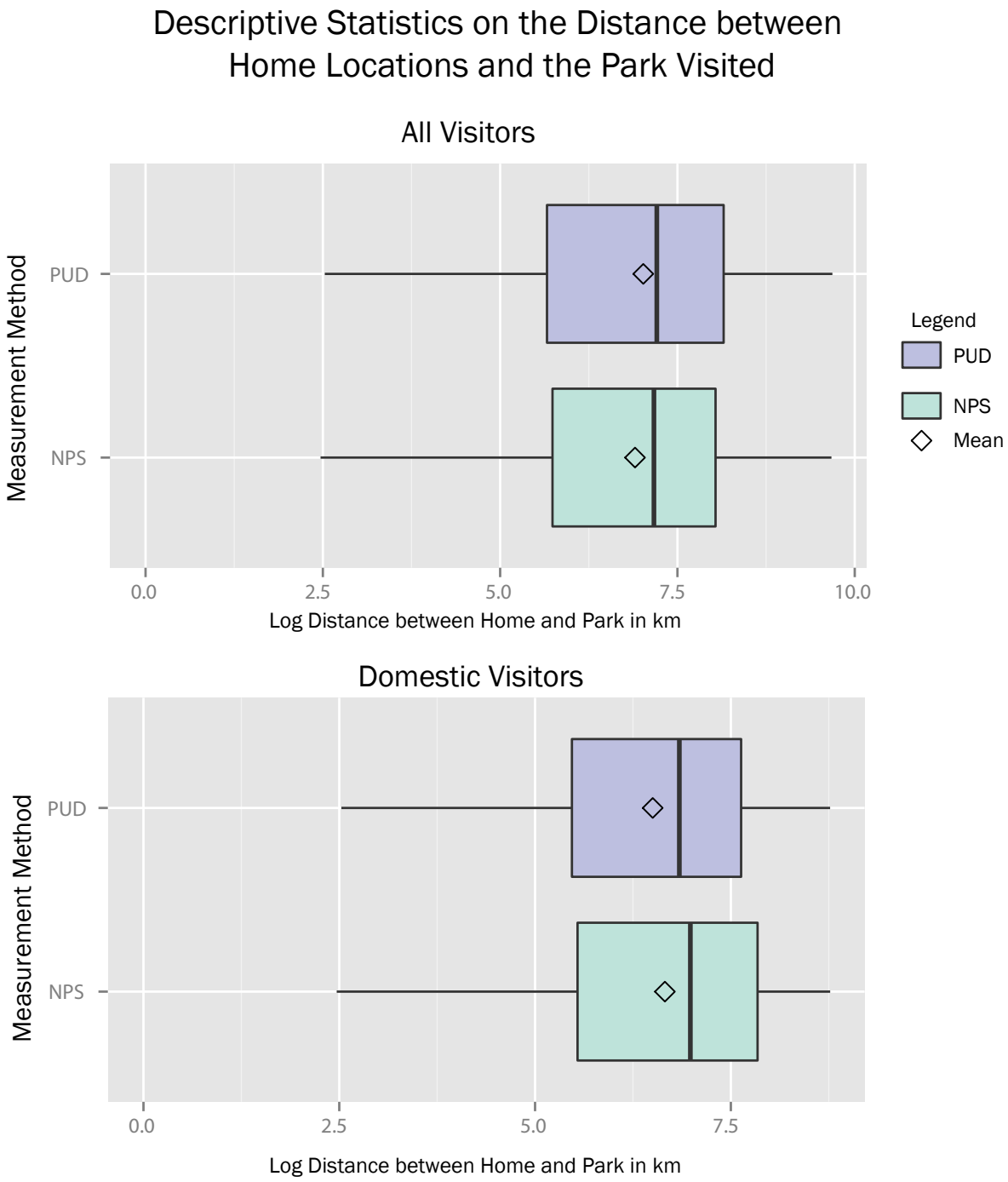

Figure 3.3: Descriptive Statistics for Distance between Visitor Homes and the Park

The cumulative distribution functions (CDFs) for the two datasets also show congruency. Figure 3.4 compares the CDFs for all visitors and for domestic visitors, with the black line

indicating distances measured by the surveys and the red line showing distances according to the photos. Both graphs show similar CDFs between the survey and photo data. In the CDF for all visitors, the photos match the surveys quite well, except visitors travelling between 5,000km and 7,500km. While the surveys show more visitors in this range, the photos show almost none. Interestingly, of the visitors coming from 5-7,500 km away, 88% are visiting Denali National Park. Denali observations make up 16% of the total NPS survey observations; however, they only make up 5% of all photo observations. Thus, this divergence between the NPS CDF and the PUD CDF shows that the NPS surveys have a greater sampling of Denali visitors relative to other park visitors, compared to the photos. The CDF for domestic visitors shows alignment between the two datasets, but this time with the photos slightly overestimating the number of visitors coming from mid-range distances.

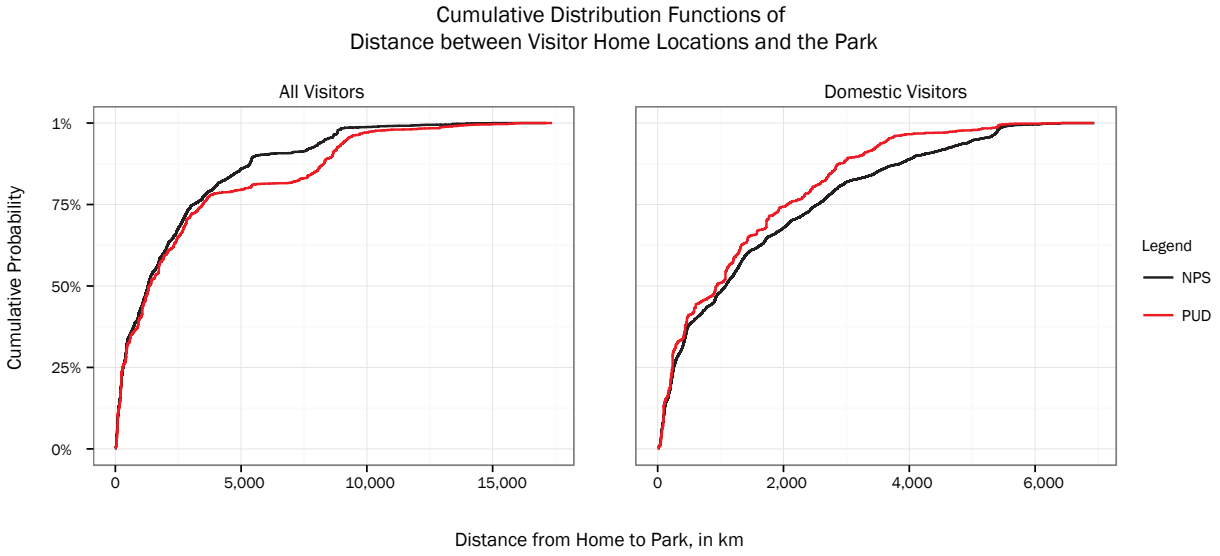

Figure 3.4: CDFs of Distance between Visitor Homes and the Park

In addition to analyzing the distances of visitors traveling to all parks together, I compare how the distance is represented by the surveys and photos for each specific survey period. I use three metrics to compare the distances: visual comparison of the kernel density plots, a statistical test of the sampling, and another statistical test of the similarity of the CDFs. This analysis allows me to test whether the surveys and photos align better for certain parks, years, or seasons. Figure 3.5 (below) shows the kernel density plots for each survey conducted; the blue line shows

data from the NPS surveys while the pink line represents data from the photos. The circles at the base of each graph is a rug plot, with each circle representing one observation at that distance.

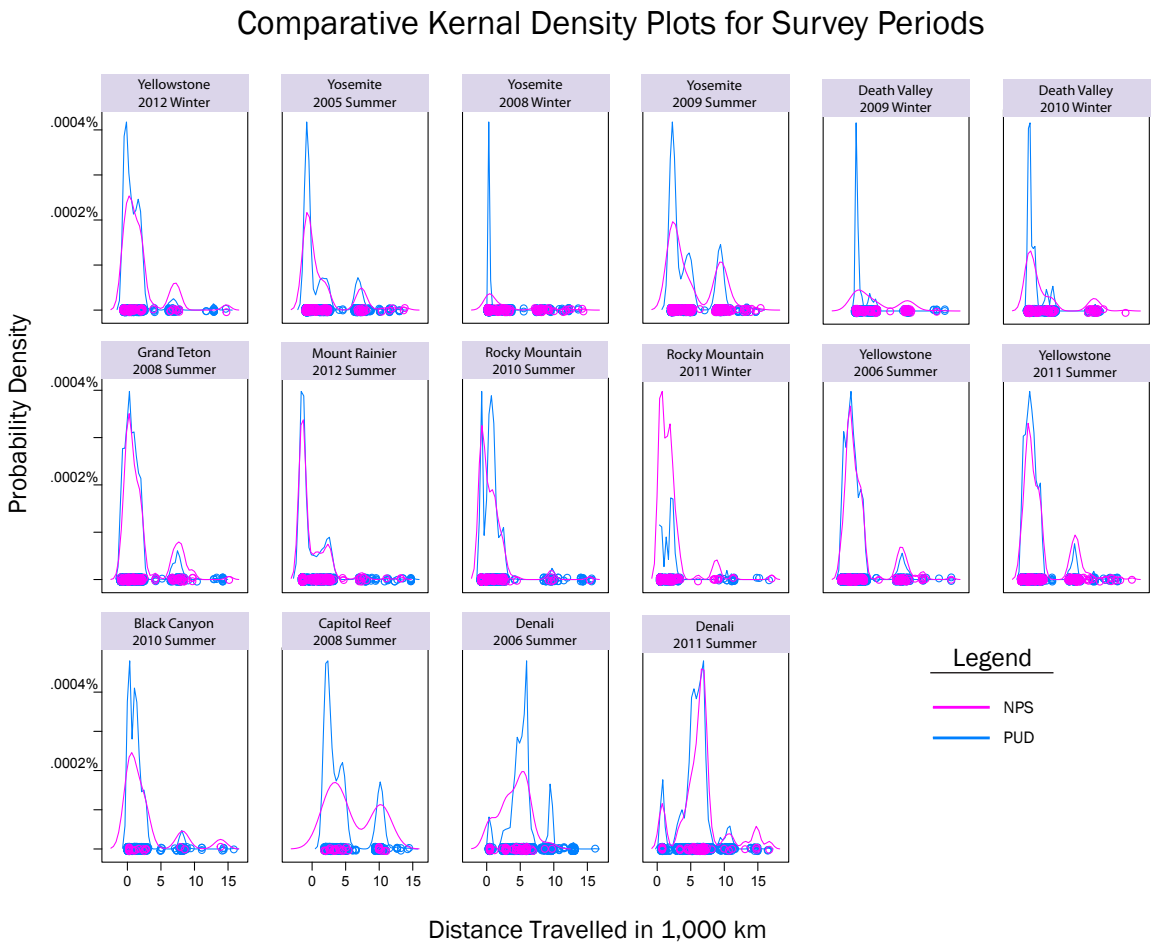

Figure 3.5: Kernel Density Plots for Distance between Visitor Homes and the Park

As shown, the densities align quite well for the majority of survey periods. Notably, many surveys show two peaks in visitors, one from a close distance and one from a midrange— in most cases, the photos pick up on this trend and closely mirror the shape. However, in numerous cases, such as Capitol Reef 2008 and Death Valley 2009 Winter, the photos under-represent closer visitation and over-represent further visitation.

While comparing the kernel density plots is largely descriptive, the following two statistical tests serve as a more quantitative and objective measure of similarity. Table 3.3 (below) show the results from the Mann-Whitney U-Test and the KS test. For both tests, a smaller W statistic and D statistic indicate that the photo and survey samples are more statistically different from each other; an insignificant p-value reports that there is no difference among the samples, indicating that the two samples are similar or were taken from the same population.[11] In my case, this means that the photos and surveys indicate an equal sample distribution of distances between home locations and the park.

Survey			Mann-Whitney U-Test			KS Test		
Park	**Year**	**Season**	**W Stat.**	**P Value**	**Conclusion**	**D Stat.**	**Two-Tailed P Value**	**Conclusion**
BLCA	2010	Summer	8980	0.286	No Difference	0.227	0.644	No Difference
YOSE	2005	Summer	152271	0.090	No Difference	0.114	0.124	No Difference
MORA	2012	Summer	198541	0.874	No Difference	0.143	0.001	
CARE	2008	Summer	10332	0.013		0.282	0.081	No Difference
DENA	2011	Summer	49647	0.017		0.220	0.014	
GRTE	2008	Summer	25112	0.000		0.597	0.000	
YELL	2012	Winter	20564	0.039		0.222	0.029	
DEVA	2010	Winter	46408	0.002		0.241	0.000	
YELL	2006	Summer	245523	0.001		0.141	0.002	
YELL	2011	Summer	343843	0.000		0.169	0.000	
YOSE	2009	Summer	362070	0.000		0.120	0.000	
DENA	2006	Summer	79050	0.000		0.294	0.001	
DEVA	2009	Winter	20458	0.000		0.350	0.000	
ROMO	2010	Summer	158169	0.014		0.242	0.000	
ROMO	2011	Winter	39648	0.000		0.435	0.000	
YOSE	2008	Winter	63794	0.000		0.313	0.000	

Table 3.3: Results from Statistical Tests for Distance between Visitor Homes and the Park, All Visitors

Three of the 16 surveys show no difference at the 0.05 level according to both the Mann-Whitney U-Test and the KS Test. Interestingly, the two tests differ in their assessment of two survey periods. Overall though, these tests show that in 13 of the 16 survey periods, the photos

[11] Significance is defined at a 95% confidence level.

and surveys give a statistically different view of the distance between visitor's home locations and the park visited.

I also conduct the tests on a subset of observations, analyzing only the distance between visitors' home location and the park for domestic visitors. The results are located in Table 3.4. The CDF curves are more similar with the exclusion of international visitors. Analyzing only the distance between visitor's home locations and the park for domestic visitors, 11 of the 16 survey periods show no difference with the Mann-Whitney U-Test, at the 0.05 significance level; six of the 16 show no difference with the KS Test.

Survey			Mann-Whitney U-Test Domestic Visitors Only			KS Test Domestic Visitors Only		
Park	**Year**	**Season**	**W Stat.**	**P Value**	**Conclusion**	**D Stat.**	**Two-Tailed P Value**	**Conclusion**
BLCA	2010	Summer	7889	0.767	No Difference	0.2406903	0.716	No Difference
YOSE	2005	Summer	102654.5	0.212	No Difference	0.1319866	0.097	No Difference
CARE	2008	Summer	4606	0.150	No Difference	0.276273	0.564	No Difference
DENA	2011	Summer	40009	0.206	No Difference	0.1891387	0.132	No Difference
GRTE	2008	Summer	94955.5	0.198	No Difference	0.132225	0.112	No Difference
YELL	2012	Winter	17874	0.417	No Difference	0.1517227	0.580	No Difference
MORA	2012	Summer	178368	0.880	No Difference	0.1538943	0.001	
DEVA	2010	Winter	40480	0.465	No Difference	0.1685288	0.012	
YELL	2006	Summer	188669.5	0.079	No Difference	0.1469149	0.005	
YELL	2011	Summer	247428	0.196	No Difference	0.1097767	0.040	
YOSE	2009	Summer	195956.5	0.814	No Difference	0.1305514	0.002	
DENA	2006	Summer	57797	0.008		0.2909672	0.002	
DEVA	2009	Winter	17446.5	0.002		0.212194	0.014	
ROMO	2010	Summer	207709	0.002		0.2352922	0.000	
ROMO	2011	Winter	19744	0.000		0.3255705	0.000	
YOSE	2008	Winter	47041	0.000		0.2465848	0.000	

Table 3.4: Results from Statistical Tests for Distance between Visitor Homes and the Park, Domestic Visitors

Section 4: Discussion

In this analysis, I have compared how the photos and NPS surveys each represent visitors' home locations and distance traveled. Both datasets, the photos and the surveys, sample a fraction of visitors—the question that I test is whether these samples are statistically similar or different and whether they are likely to have been drawn from the same population. Were the photos and survey data to yield similar results on where visitors originate from, it would indicate that the photos could act as a reliable proxy for the survey data. The following subsections discuss my results from each part of the analysis, then continues with implications for using the data and potential applications.

Comparing Types of Visitors through Home Locations

Overall I find that while the surveys and photos yield similar results on the general makeup and origin of visitors, the two methods do not consistently match in the specific proportion of visitors from each group. First, the surveys and photos yield a similar view of the proportion of in-state visitors to a park, and to a lesser degree, on the county from which in-state visitors travel. This shows that the photos' ability to infer in-state visitors versus out-of-state is fairly accurate, as is its ability to infer the counties that draw the most visitors. This finding is relatively surprising, as one would expect fewer in-state and nearby visitors to post photos, as the trip may be less exciting and more commonplace for them. Second, the photos are less accurate and consistent in reporting the proportion of out-of-state and international visitors. As a general trend, the photos tend to underestimate the number of out-of-state visitors and overestimate international ones, though in many cases the magnitude of error is between 5 to 20%. Again, this is not entirely surprising, as one would expect more international visitors to post photos, as the trip was likely more exotic for them.

However, the NPS surveys may have biases that impact the proportion of visitors from each location that it reports. For example, the NPS conducts surveys in English and does not direct surveys to members of large tour groups, causing the surveys to potentially under-sample international visitors who come on an arranged tour. In addition, the surveys rely upon people filling out the survey then mailing it back after their trip. As returning the survey to the NPS requires extra effort on the part of the visitor, visitors who are local and more invested in the park may be more likely to mail it back; international visitors who have fewer emotional ties to

the NPS and who are traveling and on-the-go are plausibly less likely to mail back the survey. These biases point to the potential that the misalignment between the photo and survey proportion of international visitors is not due to the photos over sampling international visitors, but may be caused by the surveys under sampling them.

Overall, these findings show that the photos give an accurate indication of the general makeup of visitors, but may have less accuracy in reporting the specific percentage of visitors coming from a place. This conclusion may be acceptable given the survey collection method; as the surveys are not a representative sample of the visitor population, the exact proportion reported by the surveys is likely not a precise estimate of the true proportion. Therefore, the finding that the photos are relatively in-line with the survey estimates show that they both possibly center on the true number. Further study is necessary to fully explore the bias in each dataset and whether they do truly circle around the true proportion.

That said, there are several challenges with using photos to infer the proportion of visitors in each category. One problem is that there is not a consistent trend or pattern in when the survey and photo proportions align and when they do not; that is, there are not specific parks or years that have more error than others. Consequently, it becomes difficult to fully understand the bias in the photo data or to build a regression model to calibrate it. Also, it is worrisome that the winter survey periods show less congruency than those during the summer months. This could point to a bias in the photo data that aligns with a bias in summer surveys. For example, the summer surveys may sample proportionally more international visitors than come throughout other parts of the year; the photos may also be oversampling international visitors. The last challenge is that in this analysis, the three samples (in-state, out-of-state, and international visitors) are not independent. Therefore, an overestimation in one group leads to an underestimation in another. This makes it such that an error in estimating one group visitors has to decrease the accuracy of estimating another group.

Comparing Distance between Visitors' Home Locations and the Park

Comparing the distance between visitor's homes and the park visited between the Flickr and NPS surveys, I find that the two approaches yield quite similar results when analyzing all parks together, then mixed results for the individual survey periods. Both the descriptive statistics and the CDFs of distances for all parks align quite well for all visitors and for domestic

visitors only. This shows that the photos can give relatively accurate information on the distances visitors travel to reach parks when analyzing multiple parks, years, or seasons together. However, when analyzing each survey period individually, I find that the distances are similar between datasets for some parks, but not for others. Visually, most kernel density plots align well. However, the statistical tests often show a difference in results between the two sampling techniques; this disparity is most pronounced when analyzing all visitors and improves greatly with analyzing only domestic visitors. Depending on the test and subset of visitors, between three and 11 of the survey periods are found to show no statistical difference between the samples—indicating that again, some survey periods show little difference between the photos and surveys, while others do.

The KS test, which tests for statistical differences between CDFs, often signals that the surveys and photos yield different results; *no difference* was found in three of the 11 survey periods for all visitors and in six of the 11 periods for domestic visitors. Interestingly, the KS test may be too stringent for my purposes. Figure 3.6 shows the comparative CDFs for each survey period. The surveys labeled in green were found have no statistical difference between the NPS CDF and the photo CDF according to the KS test; surveys labeled in a grey box were found to be statistically different. Looking at the CDFs, those in green do not look to have considerably more alignment than some of the survey periods in grey. For example, Capitol Reef, shown to have no difference, looks more diverging than Yellowstone 2011. Examining Mount Rainier, the CDFs align quite well for most of the distribution; it is only between ~ 2000-2400 km where the lines separate. According to the KS test, however, the data are statistically different and therefore not taken from the same sample. Consequently, I believe that visual comparisons and the Mann-Whitney U-Test may be more telling for my purposes. However, it is also possible the results from the Mann-Whitney U-Test are potentially overconfident in that the test assumes equal variances among the samples, an assumption that my data does not fulfill.

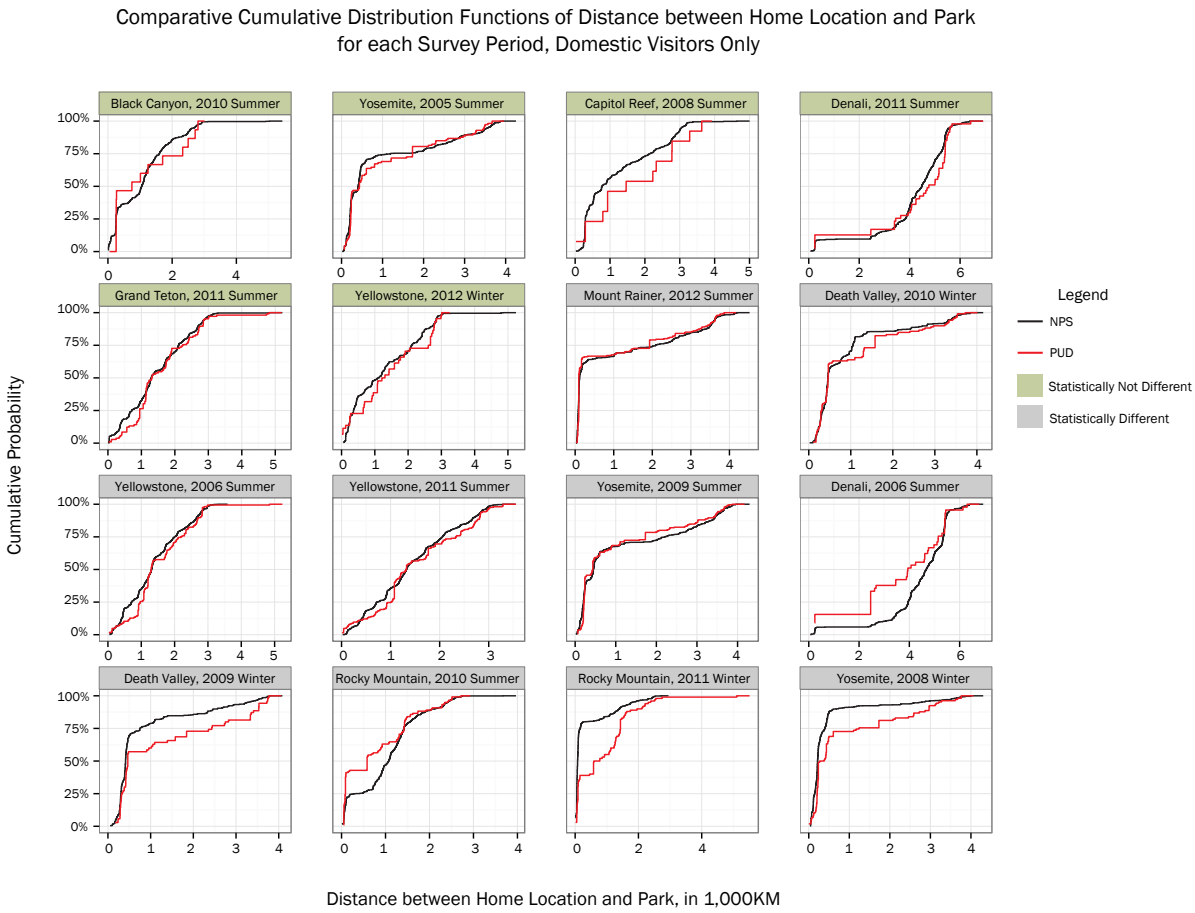

Figure 3.6: Comparative CDFs of Distance between Visitor Homes and the Park

Also telling is a comparison of the CDFs for one park across multiple years. Examining the CDFs for Yellowstone in the Summer, the NPS distribution looks appears relatively different between 2006 and 2011— in each case, however, the photo distribution matches with relative accuracy. This highlights that there is variability in the data collected by the NPS, indicating trends that may or may appear in the full population.

Conclusions and Implications

These results show that the photos are relatively accurate in showing where visitors come from and their distance away from the park. This accuracy increases when analyzing the proportion of in-state visitors, the distance between home locations and the park for domestic visitors, and when analyzing multiple parks and time periods together. Due to variation in results,

however, the photos may be less accurate in signaling the exact proportion or number of visitors from a specific location, though this is difficult to assess with limited survey data.

Once again, the methodological challenge comes in that neither are random samples of the population; each measurement technique involves sampling biases that result in a biased representation of the population. As the surveys and photos are both small and nonrandom samples of the larger population of visitors, it is likely that neither shows a true representation of visitor demographics. Rather, each technique or sample likely gives one perspective of visitor origins. In this section, I have thus far treated the survey data as an *empirical truth*; in reality, however, it is heavily biased and likely not a true representation of visitation throughout the year. That said, the survey data is currently the best estimate of visitor's home locations that we have against which to test the photo data. Thus, the fact that the photo and survey data do not align perfectly does not invalidate the photo data—rather, it indicates that the two techniques have similar, but different representations of the population. Further research and testing will elucidate the biases and errors in both estimation techniques, thereby helping us understand when the photos are likely accurate and when less so. Having more NPS surveys, conducted in different parks and through various times of the year, will yield a larger dataset against which to test the photo data. In turn, researchers will be able to identify when and where the datasets align more closely and potentially, allow for calibration to adjust photo data for more accuracy.

Moving forwards with the current data, however, I find that researchers can use the photo data to infer visitor home locations with accuracy in certain circumstances. Foremost, researchers can use the photos for the parks and time periods for which I tested and found positive results. In addition, researchers can accurately infer the distance between visitors' homes and the park when analyzing multiple parks, years, and seasons together. Barring either of these circumstances, I believe that researchers can use this technique in applications for which they are more concerned with the general trend in visitation and visitor origins than they are with precise estimates of visitor's homes or travel distance.

Potential Applications of Data

In the above circumstances, the NPS and researchers can use the photo data to draw conclusions about seasonal trends in visitor origins. As the NPS surveys are only conducted during a two-week period, the current surveys may less accurately predict how visitor origins

change throughout the year. One strength of the photo data is that it allows the NPS and researchers to understand seasonal changes in visitation. As an illustration, Figure 3.7 (below) shows the average proportion of instate, out-of-state, and international visitors to a subset of parks between 2008 and 2012.

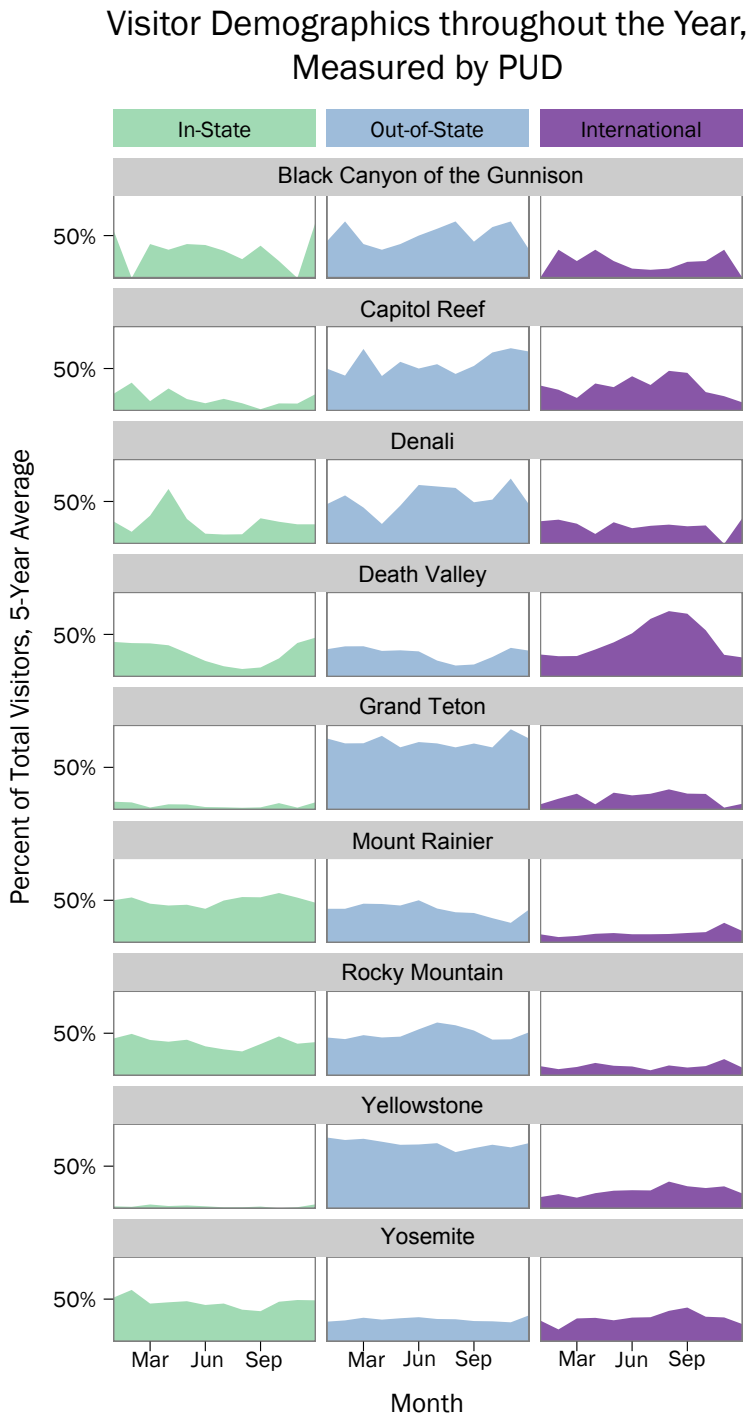

Figure 3.7: Visitor Origins throughout the Year, Measured by PUD

The Park Service can also use the photo data to conduct more detailed analyses on visitors' home locations. Figure 3.8 shows the proportion of visitors from each location visiting Mount Rainier in 2012. The NPS could use similar figures to compare visitors' origins during different seasons, years, or parks.

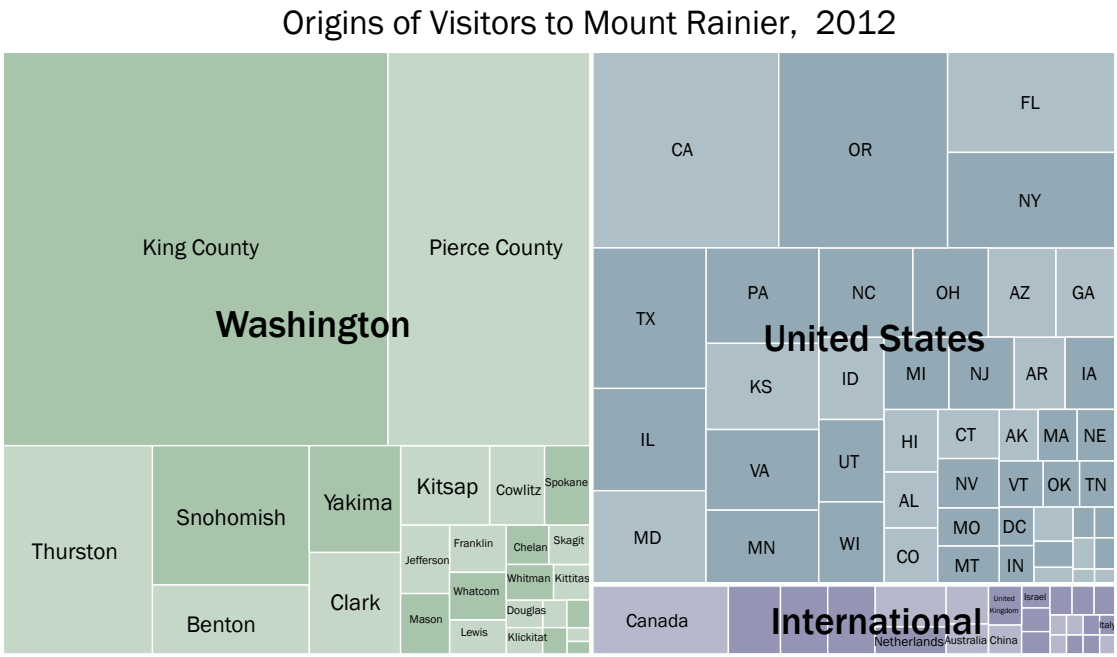

Figure 3.8: Visitor Origins to Mount Rainier, 2012

Knowing where visitors travel from during certain times of the year will better equip the NPS to provide relevant and helpful information regarding transportation options. Using these two graphs, the NPS can tailor their programs to visitor groups that come during that season. For example, if the NPS knew that most foreign visitors come during the late summer, they can plan more interpretative talks in foreign languages during this time and plan fewer during other months. In addition, if a park knew that the majority of winter visitors come from nearby counties, they could publish travel information pertinent for these specific groups. In speaking with Rachel Collins at the NPS, she mentioned that Parks want to encourage visitors to use transportation options other than their personal car to reach the Park (Collins 2015).

In addition, the NPS can combine photo home location data with spatial analysis to study which sites within a park attract certain demographics. As each photo has a geotag of where it was taken, researchers can combine this data with the home location of the user to understand the

origins of visitors to specific sites or attractions. This would allow the NPS to provide relevant information at the site for the most predominant user group.

The photo data also provides park managers with effective and lower-cost methods for conducting travel cost studies. As previously explained, travel cost studies rest on knowing the distance traveled by visitors to reach a site or attraction. The home locations indicated on Flickr users' profiles provide enough information to assume their travel distance. By using the Network Analysis tool in ArcGIS (as in Keeler et al. 2015), or open-source tools like PyRoute and openflights, one could calculate the route that visitors would likely have driven to reach the park. It is important to note here that many assumptions are necessary to use this data as an input to travel cost studies, the details of which I will describe in Chapter 4. However, as I have shown in this analysis, the distribution of distances between visitors' home and the park for domestic visitors, represented by the photos, aligns relatively well with that given by the surveys. In using this data as an input to travel cost studies, I expect that researchers will get a slightly different estimate than by using survey data, but potentially comparable in scale.

Section 6: Conclusion

In this chapter, I assess whether the photos and NPS surveys give a similar view of visitor home locations. By comparing the home locations and distance between visitor's homes and the park, reported by each method, I evaluate whether the two sampling methods give similar results and draw from the same population. I find that the photos and surveys align relatively well and give a similar view of visitor homes, especially for in-state visitors and when analyzing multiple parks, years, or seasons together. However, due to lack of consistency in my results and lack of statistical significance, I find that the two methods do not give the *same* picture of visitation. Overall though, the photos and surveys match in giving an accurate big-picture view of visitor's home locations and the distance their homes lie from the park. That said, neither the photos nor the surveys are true representations of the population visiting National Parks—each dataset samples different subsets of the population, leading to biases within each dataset. Rather than manipulating the photo data to match that provided by the surveys through statistical calibration, I find it preferable to understand the potential biases in each dataset and then use each method while acknowledging its potential biases.

Moving forwards, the photo data can aid the NPS and researchers by inferring seasonal changes in visitor origins and in providing a key input necessary for conducting basic travel cost studies. Researchers should recognize that while the photos present a valid view of visitor origins, it is also a biased one. At the same time, however, the NPS surveys also present a biased view of demographics; further investigation is needed to assess whether one dataset is more valid or representative than the other. It is important that researchers acknowledge the biases and errors in both datasets before making conclusive claims. Researchers will have to decide whether the photos align well enough with the surveys for their research purposes. In many cases, knowing the general trend in visitor demographics and rough home locations of visitors will be accurate enough. In other cases, however, such as detailed travel cost studies, researchers may want a larger or more representative sample of visitors. What I find in this chapter, however, is that the photos are relatively accurate in the demographic information they provide and therefore could be helpful if applied in the right circumstances.

Chapter 4: Application of Methods to a Travel Cost Study

Section 1: Objectives

In this section, I apply my findings from Chapters 2 and 3 to conduct travel cost studies for Mount Rainier and Great Sand Dunes National Parks. The objective of this section is not to complete a thorough and robust travel cost study that yields precise estimates of the recreational value of National Parks. Rather, it is to show how estimates of the cost incurred by visitors computed using information from photographs differs from cost estimates derived from more traditional methods. When possible, I compute the cost estimations using both NPS survey data and photo data, then compare the results. I also compare my findings to those published by other researchers. **The important finding of this analysis will not be the final estimate of economic cost incurred by visitors, but rather how well the photo-derived estimates mirror the survey-derived ones.** Should the photo data yield similar cost estimates to those from the survey methods, it would give additional validity towards using photo data as inputs to travel cost studies and other economic analyses. Moreover, my results should help point to potential inaccuracies and biases in both the survey and photo approaches—further understanding these biases will aid in finding ways to improve both methods.

The decision to model travel costs of visitors to Mount Rainier and Great Sand Dunes was strategic. First, I wanted to analyze parks for which there were already published estimates on their economic value, thereby giving me a baseline against which I could compare my estimates. Heberling et al. (2009) were the first researchers to use NPS survey data in a travel cost study; they conducted their analysis on Great Sand Dunes National Park. Neher et al. (2013) continue their work by using NPS survey data in travel cost studies for 58 National Parks, including both Great Sand Dunes and Mount Rainier. As such, Great Sand Dunes is the only park for which there are two estimates of its recreational value, thereby giving me an additional estimate against which to compare my results. One challenge, however, is that the NPS conducted the survey for Great Sand Dunes in 2002, three years before Flickr's inception. Consequently, I was unable to compare the home locations of visitors derived from the photos against a benchmark produced by the NPS. Therefore, I wanted to analyze another park that I analyzed in Chapter 3 and ideally, that performed well in my analyses. This priority led to choosing Mount Rainier. My findings in Chapters 2 and 3 showed that the survey and photo data

aligned well for this park; I was able to use photos to accurately predict the number of visitors to the park, as well as to infer the distance between visitor's home locations and the park. In addition, living in Seattle, WA, I have a personal connection to the park, having visited it often for skiing and climbing trips. These several priorities led me to estimate economic values for Mount Rainier and Great Sand Dunes National Parks.

Section 2: Methods

As my primary objective is to assess how photo-derived estimates of economic value compare to survey-derived ones, I conduct travel cost studies using each technique and compare the results. For Mount Rainier, this entails conducting the travel cost study twice, once using only NPS data and again using Flickr data. For Great Sand Dunes, however, I am unable to model a study using NPS data, as I deemed the 2002 survey too outdated for comparison. As such, I only compute the photo-derived economic value for Great Sand Dunes. **Therefore, my primary comparison in this analysis is between the survey-derived and photo-derived estimates of the economic cost incurred by visitors to Mount Rainier in 2012.** My secondary comparison is between the estimates from my analysis for Mount Rainier and Great Sand Dunes to other published estimates of their recreational value (my analysis including estimate from both surveys and photos). To compute the cost incurred by visitors to the parks, I calculate both a total cost estimation (Equation 1) and the average cost per visitor (Equation 2).

Rather than computing the value for all visitors, however, I only analyze the value for those visitors coming from in-state. This decision allows me to assuage two (often problematic) assumptions in travel cost studies. First, travel cost studies assume that people visit a site on a single-purpose trip (Ward and Beal 2000); that is, visiting the park is the primary reason of the person's trip. This assumption allows researchers to attribute the entire travel costs to the person's value for that place, without having to consider the visitor's value for other destinations along the way. By analyzing only in-state visitors, I increase the likelihood that visitors are on single-purpose trip. Out-of-state and international visitors may be more likely to visit other locations on their trip for which they have value. Second, travel cost studies often assume that visitors are on a one-day trip (Ward and Beal 2000); this assumption allows researchers to

attribute round-trip travel costs to one recreational visit.[12] As in-state visitors are more likely to visit Rainier for a day trip than out-of-state or international visitors, the choice to limit my dataset makes this assumption more realistic. An added benefit of analyzing only in-state visitors is that from Chapter 3, the photos are more accurate in predicting the proportion of in-state visitors than they are in predicting the proportion of out-of-state and international visitors.

Researchers typically use two methods to estimate economic values from travel cost studies. Many studies use the *demand estimation*, whereby the number of trips taken during a specified period is a function of the travel cost and other variables (Ward and Beal 2000). Other studies instead use *cost estimation*, whereby the researchers aggregate all travel costs incurred by visitors. The primary difference is the interpretation of the final value; the first method, which derives the demand function for visits, includes visitors' consumer surplus for visiting the site and therefore represents a visitor's *total willingness-to-pay* to visit the site. Cost estimation, on the other hand, is a representation of the amount visitors spend in time and money to reach the park; it therefore does not include consumer surplus and is a portion of *total willingness-to-pay.* Researchers can use the photo data for either type of travel cost study. By extracting the number of trips taken by a visitor in a year (that is, count the number of photos taken in a park by a specific user), researchers can regress travel costs on this count to obtain the total WTP. Conversely, researchers can aggregate all travel costs by estimating the number of visitors and the costs incurred by each one to calculate the total amount expended to visit the park.

As I collected data on the total number of visitors and their home locations, I am well situated to follow the second approach of aggregating travel costs across visitors. My models take the following forms:

$$\text{Total Cost Incurred by Visitors} = \Sigma\,(\text{Travel Cost } i \times \text{No. of vistiors } i)$$

(Equation 1)

And

$$\text{Average Cost Incurred per Visitor} = \frac{\Sigma\,(\text{Travel Cost } i \times \text{No. of vistiors } i)}{\text{Total No. of Visitors}}$$

(Equation 2)

[12] Researchers have found ways to work around these two assumptions with advanced statistical techniques.

where *i* represents the distance traveled, ranging from 0 to 325 km, measured and aggregated in ten kilometer increments.[13] This process involves four primary steps: first, computing the number of visitors traveling from each distance away; second, computing the cost of travel for each distance; third, multiplying the number of visitors from that distance by the corresponding travel costs, and; forth, summing these numbers for the total value, and averaging these numbers for the per visitor average. The following two sections describe the methods for the first two calculations. I conduct the same process for Mount Rainier and Great Sand Dunes.

Estimating the Number of Visitors from Each Distance

This step entails finding the total number of visitors to each park in 2012, then finding the inferred number of visitors coming from each distance away. To obtain the total number of visitors according to the NPS data, I aggregate the monthly totals supplied by the NPS. For the Flickr data, I use the regression created in Chapter 2 to scale up the number of photos. By plugging in the monthly PUD and other corresponding variables into my original Mount Rainier-specific regression, I infer the total number of visitors to the park during each month of 2012. The regression equation for Mount Rainier is as follows:

$$NPS_{ij} \sim \text{Negative Binomial } (\beta_0 + \beta_1(PUD) + \beta_2(Year) + \beta_3(Month) + \beta_4(PUD*Summer\ month))$$

Or

$$NPS_{ij} \sim \text{Negative Binomial } (-30.80 + .5601(PUD) + .0193(Year) + \beta_3(Month) - .0023(PUD*Summer\ month))$$[14]

Summed over all months, the result is the estimated number of visitors to the park in 2012, including a 95% confidence interval (CI).

Next, I used the cumulative distribution functions (CDFs) of distance between visitors' homes and the park from Chapter 3 to estimate the proportion of visitors coming from each

[13] I aggregate in 10km bins so that I can estimate the total number of people coming to the park from that distance.

[14] I do not report the Beta values on month here, as each of the twelve months has a different value.

location. Figure 4.1 (below) shows the CDFs of all visitors to Rainier obtained from NPS and PUD data and the CDF of all visitors to Great Sand Dunes according to photo data.

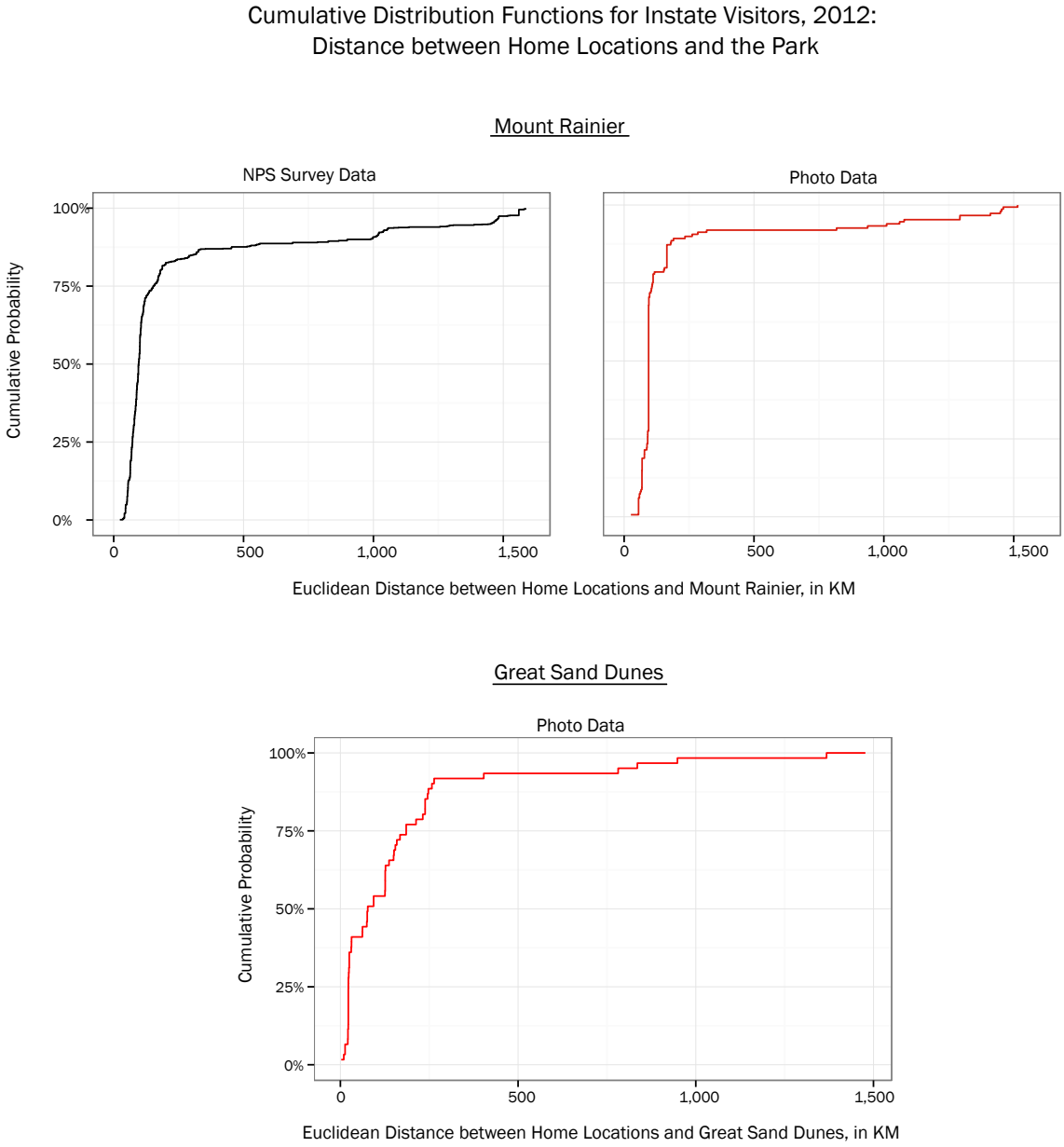

Figure 4.1: CDFs of Distance between Visitor Homes and the Park, Mount Rainier and Great Sand Dunes National Parks

Categorizing the distances into 10km bins, I used the CDFs to compute the proportion of the sample coming from each distance away. For example, according to the Mount Rainier PUD distance CDF, 0.0461 percent of all visitors to Rainier traveled between 50 and 60 km to reach

the park. To obtain the number of visitors traveling from this distance, I multiplied the percent (0.0461) by the total number of expected visitors. The result is a table indicating the total number of visitors coming from each distance away, according to NPS data and to PUD data. Estimates from PUD data include a point estimate, lower estimate, and upper estimate.

Estimating the Travel Costs

For each category or bin of distance between visitors' home location and the park, I compute the travel cost incurred by visitor. I use the median distance of each bin to compute costs. Overall, estimate the cost per visitor with the following model:

$$\text{Cost per Visitor} = (\text{travel distance} \times \text{cost of travel}) / \text{number of people in vehicle}) + (\text{travel time} \times \text{value of time}) + \text{per person entrance fee}$$

The following bullets explain the variables I use as inputs to the model. A summary of all variables and their computations are located in the Appendix.

- **Travel distance:** From Chapter 3, I have the Euclidean distance between the visitor's home location and the park for each visitor, computed from their zip code and the center of the park. To convert this into a more meaningful metric of travel distance by road, I multiply the Euclidean distance by a circuitry factor of 1.15, as done in Heberling and Templeton (2009). The circuitry factor is a multiplier created by the US Forest Service that converts a straight-line distance to a probable distance traveled over road (Hellerstein 1993).[15] I then multiply the travel distance by two to obtain the round-trip distance. I also convert the distance from kilometers to miles.

[15] 1.15 is the average circuitry factor for all 50 states. While the manual states that there are state-specific circuitry factors, I have been unable to find the one for Washington. The manual also states that the factor is "surprisingly accurate." They explain, "Comparison of distances computed using ZipFip to those published in road atlases typically differ only by a few percentage points, especially for longer trips." However, "Problems do arise in short trips, especially in mountainous regions, where specific journeys might necessitate a very indirect route (Hellerstein 1993)." Considering my analysis of only in-state visitors, it is likely that the circuitry factor of 1.15 is an underestimate. In other publications, such as (Heberling and Templeton 2009; Keeler et al. 2015), the authors compute a distance using GIS or Google Maps. Either of these techniques would be viable options for a more in-depth analysis.

- **Cost of travel:** In line with Donovan and Champ (2009) and Heberling and Templeton (2009), I use the private vehicle mileage deduction rate to estimate the cost of travel. As I study only in-state visitors, I use the 2012 state reimbursements rates rather than the federal IRS rate. This is $0.565 per mile for Washington and $0.52 per mile for Colorado (State of Colorado 2015; State of Washington 2015).
- **Number of people in vehicle:** As travel costs are incurred by all people in the vehicle together rather than individually, I divide the cost of travel by the assumed number of passengers in the vehicle. As explained in Chapter 1, each park uses a Persons-Per-Vehicle (PPV) statistic to estimate the number of visitors per vehicle. Mount Rainier National Park uses a PPV of 2.8 in the winter months and 3.0 during the summer months (Mount Rainier National Park 1995). I use the average rate of 2.9 people per vehicle. Great Sand Dunes varies the PPV among months; the average rate across months is 2.66 (Great Sand Dunes National Park 2002).
- **Travel time:** Following Bhat and Bergstrom (1996), I assume a constant rate of travel of 50 miles per hour.
- **Value of time:** Though still debated, it is common for travel cost models to incorporate the value of time into the travel cost (Ward and Beal 2000). That is, studies account for the opportunity cost born by the traveller of not being able to participate in other fun and valuable activities while in transit. Most studies value time at between one third and one half the wage rate (Ward and Beal 2000); in line with recommendations in Wolff (2014), I use one half the wage rate. Following Heberling et al. (2009) and Neher et al. (2013), I obtain the wage rate by using the average income. While Neher et al. use the average income in each zip code, I use the average value across each state in 2012, obtained from the Bureau of Labor Statistics (Bureau of Labor Statistics 2014).[16]
- **Entrance fee:** Each park charges a different entrance fee. The fee for Great Sand Dunes is $3 per person (Great Sand Dunes National Park n.d.). Mount Rainier, however, has two options for entrance fees to enter the park (Mount Rainier National Park n.d.). First, they offer an annual pass for $30 that is valid for the entire car to enter the park as many times as they wish throughout the year. Second, they offer a single pass for $15 that allows all passengers in the vehicle to enter. For my analysis, I assume that one third of in-state-

[16] To obtain the average hourly wage, I divide the weekly wage by 40 hours.

visitor vehicles have an annual pass and that they visit the park three times per year, making the per-trip cost $10. I assume that two thirds of vehicles entering buy a single pass. Again, I use the PPV rate as a proxy for the number of passengers in the vehicle. I then divide the vehicle entrance fee by the PPV to obtain the estimate of the entrance fee per visitor.

Section 3: Results

Number of Visitors from Each Travel Distance

Table 4.1 (below) shows the results for the number of visitors coming to the parks from each distance away; it shows the percent of visitors from each distance according to the CDF, then the predicted number of visitors calculated by multiplying the percentage by the total number of 2012 visitors.[17] Total 2012 visitation to Mount Rainier, according to NPS data, was 1,049,178; using the regression model, I find photo-predicted visitation in 2012 to be 1,142,200, (95% CI 900,093 - 1,384,304). For Great Sand Dunes, the NPS reported 2012 an annual visitation of 254,674 people; using the photo data, I estimate visitation at 285,334 (95% CI 271,828 - 298,840). Examining the CDFs of the distances between home locations and the park, I find that in-state visitors comprise 56.4 % of total visitors to Mount Rainier according to NPS data and 55.3% according to the photos. In-state visitors comprise 34.4% of all visitors to Great Sand Dunes, according to the photos.

	Mount Rainier				Great Sand Dunes	
	NPS Data		PUD Data		PUD Data	
Median Euclidean Distance (in km)	Total Percent	Visitors	Total Percent	Visitors	Total Percent	Visitors
15	0.05%	570	0.46%	5,264	1.64%	4,678
35	0.27%	2,849	0.00%	-	0.00%	-
45	3.10%	32,484	0.00%	-	0.00%	-
55	5.54%	58,129	4.61%	52,636	0.00%	-
65	7.22%	75,796	7.83%	89,481	0.00%	-
75	6.52%	68,387	1.84%	21,054	0.00%	-
85	7.22%	75,796	1.38%	15,791	0.00%	-

[17] Note that I only report the point estimate for the PUD number of visitors, though I have computed the lower-bound and upper-bound estimates.

95	7.50%	78,646	33.18%	378,979	1.64%	4,678
105	8.58%	90,044	2.30%	26,318	0.00%	-
115	2.99%	31,344	2.30%	26,318	0.00%	-
125	1.41%	14,817	0.00%	-	0.00%	-
135	1.03%	10,828	0.00%	-	3.28%	9,355
145	0.60%	6,269	0.00%	-	0.00%	-
155	0.65%	6,839	0.46%	5,264	0.00%	-
165	0.11%	1,140	0.00%	-	0.00%	-
175	0.65%	6,839	0.46%	5,264	0.00%	-
185	0.87%	9,118	0.00%	-	0.00%	-
195	0.38%	3,989	0.00%	-	0.00%	-
205	0.27%	2,849	0.00%	-	1.64%	4,678
215	0.16%	1,710	0.00%	-	3.28%	9,355
225	0.11%	1,140	0.00%	-	16.39%	46,776
235	0.05%	570	0.00%	-	0.00%	-
245	0.11%	1,140	0.00%	-	6.56%	18,710
255	0.11%	1,140	0.00%	-	0.00%	-
265	0.00%	-	0.00%	-	0.00%	-
275	0.00%	-	0.00%	-	0.00%	-
285	0.00%	-	0.00%	-	0.00%	-
295	0.11%	1,140	0.00%	-	0.00%	-
305	0.05%	570	0.00%	-	0.00%	-
315	0.38%	3,989	0.46%	5,264	0.00%	-
325	0.33%	3,419	0.00%	-	0.00%	-
TOTAL	**56.38%**	**591,552**	**55.30%**	**631,631**	**34.43%**	**98,230**

Table 4.1: Number of Visitors from Each Travel Distance, In-State Visitors Only

Travel Costs

Using the parameters described in the methods section, I compute the total per-person travel cost for each distance category, found in Figure 4.2 (below). The full computations are located in the Appendix.

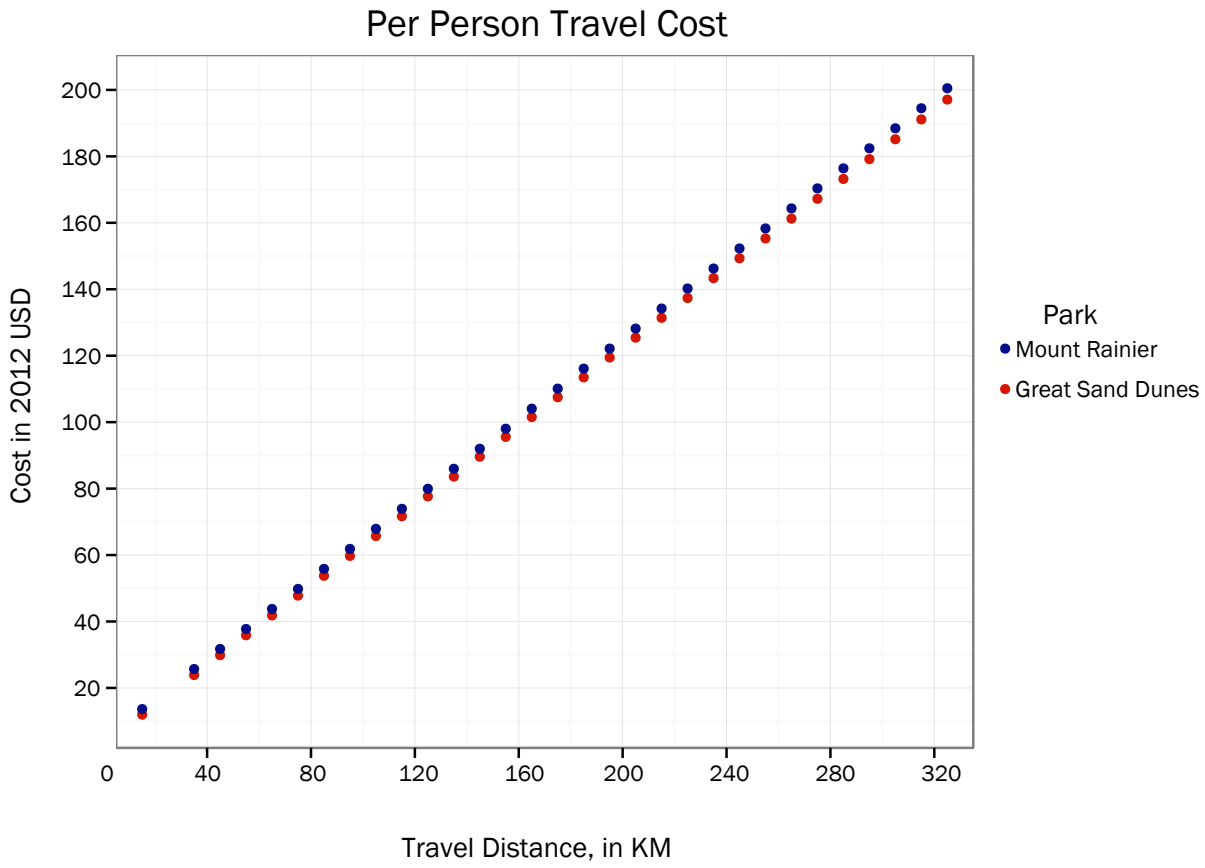

Figure 4.2: Per Person Travel Costs, Mount Rainier and Great Sand Dunes

Cost Estimation

By summing the product of the number of visitors from each distance by the amount of money they incurred to reach the park, I obtain the total cost incurred by in-state visitors to the park for Mount Rainier and Great Sand Dunes National Parks in 2012. Using the NPS data, I find the total incurred cost of Mount Rainier is $36.31 million, with an average per visitor cost of $61.38.[18] Using the PUD data, I find the total cost to equal $37.20 million, (95% CI $29.32-$45.01 million); the average cost was $58.90. For Great Sand Dunes National Park, the photo data yields an estimate of the total cost incurred to reach the park for in-state visitors of $44.69 million dollars (95% CI $44.11- $45.26 million), with an average cost of $123.72 per in-state visitor. I was unable to compute an NPS estimate due to lack of recent NPS survey data for this park.

[18] Measured in 2012 USD.

Section 4: Discussion

These results show that the photo-derived cost estimation for Mount Rainier is strikingly close to that obtained by the survey data. Comparing the survey and photo-derived estimates for Mount Rainier, I find difference of $890,000, equating to a per-person difference of $2.48. This finding has several important implications. Foremost, the small difference in estimates of the incurred costs shows that the photo method produces similar results to the NPS method, making the photo data a viable proxy for much of the survey data. While a survey sample of visitors is necessary to calibrate the photo counts, this analysis shows that ongoing sampling and detailed surveys may be less necessary for economic valuation. Moreover, this analysis gives additional validity towards the home locations and distance of travel information provided by the photos. In Chapter 3, I had mixed results on the accuracy of the photos in determining visitor's home locations and the distance between their home and park; for Mount Rainier, for example, the proportions of visitors from each county were different between the survey and photos, and only one of the two statistical tests on the distances reported that the samples were statistically similar. The analysis here, however, shows that though the home location and distance data was not consistently statistically similar, the datasets still led to similar cost estimates. In this section, I further explore the assumptions and limitations of this analysis, compare my estimates of economic values to other published estimates, and explain further applications of this method.

Assumptions and Limitations

This analysis makes numerous assumptions about the nature of a visit to the Park, some of which are typical in travel cost studies and some of which are inherent to the photo method. First and foremost, I assume that in-state visitors travel to Mount Rainier for a single-purpose day trip. This implies that the cost a visitor spends to reach the park is entirely attributable to his value for the park; that is, he does not also visit other sites on the way which he values. In addition, it assumes that each photo-visitor incurs round-trip travel, meaning that people are visiting for the day only and not staying overnight. This is presumably unrealistic and therefore leads to an overestimate of the economic value. However, analyzing only in-state visitors, rather than out-of-state and international ones, decreases the likelihood that visitors are staying more than one day.

Additional assumptions may have influenced the travel cost estimates. My specification of the travel cost assumes that all visitors are driving to the park and moreover, that they are driving from their home location. With few options for public transportation to Mount Rainier and Great Sand Dunes available, I believe this to be a valid assumption. Further, in order to estimate travel time, I assume that people travel at an average rate of 50 miles per hour, an assumption that is necessary for computation, but difficult to assess the accuracy of. Mentioned above, I also assume the average number of people in a vehicle, according to the NPS PPV statistics. Together, these assumptions mean that my calculated travel costs are likely approximations of the truth, rather than a precise estimate. I also make assumptions on the type of entrance pass that visitors buy. Without further information from the NPS, it is difficult to accurately guess the proportion of visitors buying each type of pass. Again, this assumption decreases the accuracy of my cost estimation.

Together, these assumptions limit my confidence in my final cost estimate, as different decisions on the parameterization would change the estimate (for example, using one third of the wage rate rather than one half would decrease the estimated economic value). While I believe these assumptions are necessary and rooted in best practice and the best information available, they in no doubt influence the final estimate. However, the purpose of my analysis is not to obtain an accurate estimate of the economic value of Mount Rainier and Great Sand Dunes National Parks—rather, it is to compute values using the two data sources, then to compare the values. Therefore, what is important are not the details of my assumptions, but that I have been consistent in my assumptions through both analyses.

An important limitation of my analysis is that I conducted this analysis on only in-state visitors. My analysis in Chapter 3 found that the photo data becomes less accurate for more distant visitors, such as international ones; doing a travel cost study with the photo data including international visitors would likely provide an estimate of economic value that is more divergent from the NPS one. My findings in Chapter 3 also reported mixed results in the ability of photos to accurately infer home location information on out-of-state visitors; travel cost studies using photos from all domestic visitors may or may not differ significantly from ones using the survey data. However, conducting a travel cost study on out-of-state and international visitors also bring to question the appropriateness of underlying assumptions, namely that visitors travel for a single-purpose trip and incur all travel costs for only one visit. As my analysis is only for in-state

visitors, the incurred costs are clearly an underestimate of the costs incurred by all visitors to the park.

Comparison to Other Economic Estimates

Interestingly, there have been two travel cost studies using NPS survey data against which I can compare my estimates.[19] In 2009, Heberling et al. conducted the first travel cost study using the VSP data for Great Sand Dunes National Park. Using survey data from 2002, they found that US visitors have a recreational benefit of $89 per visitor per year, or a $54 value per visitor per day (measured in 2002 dollars, equivalent to $113 and $68 in 2012 dollars). Neher et al. (2013) built upon Heberling's methods by using NPS survey data to complete travel cost studies in 58 National Parks. They found Great Sand Dunes to have an average visitor WTP of $108.38 per person per trip in 2011, equivalent to $110.62 in 2012 USD and Mount Rainier visitors to have a WTP of $190.23 per trip in 2000, equivalent to $253.63 in 2012 USD (± $32.70 2012 USD). In comparison, I find the photo-derived cost estimate for Great Sand Dunes at $123.72 per visit and for Mount Rainier at $58.90 per visit.

There are several reasons why I believe my travel cost estimates to differ from the previous estimates. First and foremost, our metrics of estimating economic value differ. Their use of *demand estimation* regresses the number of trips taken by a user in a year on travel costs. This technique computes a *total willingness-to-pay* for visiting the park, which includes consumer surplus. My technique of *cost estimation*, on the other hand, values the incurred costs of visitors, thereby excluding consumer surplus. This difference in specification should cause the other authors to find higher economic values than the technique I follow. Second, my estimate using photo data is highly sensitive to the assumptions I make about the value of time and travel costs. In their analyses, Heberling et al. (2009) and Neher et al. (2013) make different assumptions that alter their results—for example, they map driving distance rather than using a circuitry factor and decide not to include the opportunity cost of time traveled. Third, I analyze a smaller sample of visitors to obtain the distribution of distance traveled by visitors. For Great Sand Dunes, I analyze 23 visitors, whereas Heberling et al. (2009) analyze the home locations of

[19] The data used is technically called Visitor Service Project, or VSP surveys, as they are conducted by a group at the University of Idaho rather than the NPS. For ease of understanding, I still call this NPS survey data. Regardless, it is still the same data that I analyze for my NPS estimates.

364 in-state visitors. The conclusions I am able to draw about distance travelled by visitors may be less accurate or precise than those drawn from a larger sample.

The final reason why my estimates differ from those previously published lies in that I only analyze in-state visitors, whereas the other two studies calculate travel costs for all visitors. As in-state visitors incur lower travel costs than more distant visitors, I would expect an analysis of only in-state visitors to yield a lower average economic value than one encompassing all visitors. This corresponds to my finding for Mount Rainier, where my estimated costs were roughly one fourth of that found by Neher et al. Looking more closely at Rainier visitation, nearly half of visitors to Mount Rainier are from out-of-state or international locations, many of whom travel far distances to reach the park. As these distant visitors likely have a much higher costs of travel than in-state visitors, I would expect their estimates to raise the total. Therefore, while my estimate of in-state visitors' costs is significantly lower than Neher's WTP of all visitors, I find it plausible that in-state visitors' costs are much lower than other visitors'. Interestingly, my hypothesis of a travel cost study for instate visitors producing a lower economic value than one for all visitors may not play out at Great Sand Dunes, where I find a higher WTP for instate visitors than the other estimates. One possible explanation is that out-of-state and international visitors actually have a lower WTP than instate visitors for this park. As Great Sand Dunes is not a particularly popular or well-known park, many distant visitors may be visiting in route to another location, making it a multipurpose trip and decreasing the travel costs attributable to the park. As the other two studies both control for multipurpose trips in their regressions, it is possible that these visitors actually have a lower WTP than in-state visitors, who are more likely to be on a single-purpose trip.

The point of comparing my findings to other studies is not to add validity to the estimates that I find. Rather, it is explore whether the estimates of economic value are potentially in-line with one another and possible reasons why they are or are not. **Therefore, the central takeaway is not the precise economic values that I compute, but rather that an estimate derived from photo data is comparable to estimates calculated from more standard and accepted methods.**

Further Applications in Travel Cost Studies

Overall, this analysis shows that researchers can use data from the photos to derive an estimate of the economic value of a recreational site. In the case of Mount Rainier, the estimate derived from photo data falls within the same range as that derived from NPS data—this is an encouraging sign, as it shows that potentially, researchers can get similar estimates of costs incurred by visitors using a non-traditional method. To further corroborate the photo method and explore biases and inaccuracies in estimating the recreational value of sites, researchers should continue to conduct travel cost studies using different types of data and different specifications.

One specific area for further research is using photo data in demand estimation types of travel cost study. While I chose to calculate travel costs by aggregating costs incurred by visitors, researchers should not be restricted only to this method. In demand estimation, researchers model the number of trips taken by a visitor in a given timeframe against travel costs and other variables. By tracking a user's photo-user-days over a year, researchers could deduce how many times each user visited a site. One could also identify photos taken within a few days of each other and eliminate those photos, as to isolate only unique trips taken to the Park. This method would allow researchers to use the standard travel cost methodology and account for multiple day-trips.

Section 5: Conclusion

In this chapter, I present an illustrative application of photo data to parameterize a travel cost model. By comparing estimates of the costs incurred by in-state visitors to Mount Rainier National Park derived from photo data and survey data, I find that the photos yield a strikingly similar view of the economic value. The total travel cost estimated using NPS surveys falls within the range of values estimated using Flickr photographs from the same National Park. The photo method requires little data collection by the Park Service—only a few monthly counts collected over time are necessary to calibrate PUD with the number of visitors. The remainder of the parameters, namely the home locations of visitors, can accurately be inferred from the photos, thereby reducing the need for NPS surveys.

Overall, this analysis shows that researchers can potentially use the photo data in economic analyses with accuracy. To further corroborate this finding, researchers should conduct more travel cost studies using the photo data and compare it to findings using other approaches.

This analysis represents a crucial first step in comparing values derived from photos to values obtained from more accepted methods.

Chapter 5: Synthesis and Conclusion

This research provides additional validity to the practice of using online, crowd-sourced photographs to infer visitor use information to National Parks in the Western United States. By comparing information from photographs on the website Flickr to data collected by the National Park Service, I find that the photos yield accurate information on the number and rates of visitors to a site and on the visitors' home locations. Together, these inputs make up the most important components of travel cost studies used to measure the recreational value of lands.

Through multiple regression analysis and robust statistical testing in Chapter 2, I build predictive models that use the number of photos posted online to accurately infer the number of monthly visitors to a site. By building models specific for each National Park, I produce estimates of visitation that fall within the range observed using more traditional methods. The ability to infer visitation counts with photos online has the potential to save the National Park Service (NPS) money and resources by decreasing their reliance on current count methods. Moreover, it opens the possibility of measuring recreational use at remote and currently hard-to-measure locations.

In Chapter 3 I find that the NPS and researchers can use photo data to accurately infer where most visitors live. The photo data yields accurate information on the proportion of visitors coming from in-state, as well as on the distribution of domestic visitors' distance from the park. This finding will allow the NPS to monitor visitor origins throughout the year, providing them with a more detailed view of visitor demographics than they currently have.

Researchers can use these findings in tandem to conduct travel cost studies with the photo data. In testing this method in Chapter 4, I find that photo-estimated and survey-estimated values of the costs incurred by in-state visitors to be within 2.4% of each other. This shows that in an economic study, the photo data can provide an estimate close to one derived from empirical sources. Overall, these findings represent a new and exciting opportunity in economic and recreational valuation. By leveraging online, crowd-source data, researchers can infer similar information to that provided by more traditional methods. This method has the potential to significantly reduce the reliance on survey-collected data, thereby saving land managers' time and money in data collection and opening new lands for economic study.

Appendix

Chapter 1

Table 1: National Parks Studied

Park	Park Code	State	Region
Arches	ARCH	UT	Southwest
Black Canyon of the Gunnison	BLCA	CO	Southwest
Bryce Canyon	BRCA	UT	Southwest
Canyonlands	CANY	UT	Southwest
Capitol Reef	CARE	UT	Southwest
Carlsbad Caverns	CAVE	NM	Southwest
Channel Islands	CHIS	CA	Pacific
Crate Lake	CRLA	OR	Northwest
Denali	DENA	AK	Alaska
Death Valley	DEVA	CA	Pacific
Gates of the Arctic	GAAR	AK	Alaska
Glacier	GLAC	MT	Rockies
Glacier Bay	GLBA	AK	Alaska
Great Basin	GRBA	NV	Pacific
Grand Canyon	GRCA	AZ	Southwest
Great Sand Dunes	GRSA	CO	Southwest
Grand Teton	GRTE	ID	Rockies
Joshua Tree	JOTR	CA	Pacific
Katmai	KATM	AK	Alaska
Kenai Fijords	KEFJ	AK	Alaska
Kings Canyon	KICA	CA	Pacific
Kobuck Valley	KOVA	AK	Alaska
Lake Clark	LACL	AK	Alaska
Lassen Volcanos	LAVO	CA	Pacific
Mesa Verde	MEVE	CO	Southwest
Mount Rainier	MORA	WA	Northwest
North Cascades	NOCA	WA	Northwest
Olympic	OLYM	WA	Northwest
Petrified Forest	PEFO	AZ	Southwest
Pinnacles	PINN	CA	Pacific
Redwoods	REDW	CA	Pacific
Rocky Mountain	ROMO	CO	Rockies
Saguaro	SAGU	AZ	Southwest

Sequoia	SEQU	CA	Pacific
Wrangell St. Elias	WRST	AK	Alaska
Yellowstone	YELL	WY	Rockies
Yosemite	YOSE	CA	Pacific
Zion	ZION	UT	Southwest

Chapter 2

Table 2A: Likelihood Ratios for Park-Specific Regressions

Park	Likelihood Ratio		Park	Likelihood Ratio		Park	Likelihood Ratio
CRLA	3841.508**		CHIS	5.005*		SAGU	1.569
ARCH	7.675**		SEQU	4.997*		MEVE	1.405
GRTE	7.272**		KICA	4.392		BLCA	1.15
NOCA	7.199**		BRCA	3.842		REDW	1.082
ROMO	7.106**		LAVO	3.249		CANY	0.937
YOSE	6.881**		PEFO	3.007		CAVE	0.921
MORA	6.536**		GRSA	2.619		DEVA	0.868
JOTR	6.109**		GLAC	2.49		ZION	0.823
CARE	5.843*		GRCA	2.41		GRBA	0.394
YELL	5.099*		OLYM	1.753		PINN	0.355

Chapter 3

Figure 3A: Home States of Domestic Visitors to Mount Rainier National Park, Summer 2012

State	Percent of Visitors: NPS	Percent of Visitors: PUD		State	Percent of Visitors: NPS	Percent of Visitors: PUD
Washington	59.63%	59.20%		**Nevada**	0.46%	0.50%
California	5.23%	4.48%		**Missouri**	0.40%	0.00%
Oregon	4.72%	7.46%		**Montana**	0.40%	0.00%
Florida	2.59%	1.49%		**Colorado**	0.35%	0.50%
Texas	2.42%	1.49%		**Nebraska**	0.35%	0.00%
New York	1.90%	4.48%		**Vermont**	0.35%	0.00%
Illinois	1.73%	1.99%		**Alaska**	0.29%	0.00%
Pennsylvania	1.73%	1.00%		**Massachusetts**	0.29%	0.50%
Maryland	1.67%	1.00%		**Oklahoma**	0.29%	0.00%

Virginia	1.50%	0.00%		**Indiana**	0.23%	0.00%
Minnesota	1.38%	0.50%		**Maine**	0.23%	0.00%
North Carolina	1.27%	0.50%		**Kentucky**	0.12%	0.00%
Ohio	1.04%	1.00%		**Louisiana**	0.12%	0.00%
Arizona	0.86%	0.00%		**Mississippi**	0.12%	0.00%
Idaho	0.86%	0.50%		**Tennessee**	0.12%	0.50%
Utah	0.86%	0.00%		**West Virginia**	0.12%	0.50%
Georgia	0.81%	0.50%		**District of Columbia**	0.06%	0.50%
Wisconsin	0.81%	0.50%		**New Mexico**	0.06%	0.00%
Kansas	0.69%	6.47%		**South Dakota**	0.06%	.00%
New Jersey	0.69%	1.00%		**Delaware**	0.00%	0.50%
Iowa	0.63%	1.00%		**International**	0.00%	0.00%
Michigan	0.63%	0.50%		**New Hampshire**	0.00%	0.00%
Arkansas	0.58%	0.00%		**North Dakota**	0.00%	0.00%
Hawaii	0.52%	0.00%		**Rhode Island**	0.00%	0.50%
Alabama	0.46%	0.50%		**South Carolina**	0.00%	0.00%
Connecticut	0.46%	0.50%		**Wyoming**	0.00%	0.00%

Figure 3B: Home Countries of International Visitors to Mount Rainier National Park, Summer 2012

Country	Percent of Visitors: NPS	Percent of Visitors: PUD		Country	Percent of Visitors: NPS	Percent of Visitors: PUD
Canada	25.26%	14.29%		Sudan	1.05%	0.00%
Germany	10.53%	7.14%		Ukraine	1.05%	0.00%
India	10.53%	0.00%		Argentina	0.00%	0.00%
France	9.47%	0.00%		Austria	0.00%	0.00%
Japan	9.47%	7.14%		Azerbaijan	0.00%	0.00%
Netherlands	5.26%	7.14%		Bangladesh	0.00%	0.00%
United Kingdom	4.21%	21.43%		Belarus	0.00%	0.00%
China	3.16%	0.00%		Belgium	0.00%	7.14%
Finland	3.16%	0.00%		Bermuda	0.00%	0.00%
Australia	2.11%	14.29%		Chile	0.00%	0.00%
Israel	2.11%	0.00%		Colombia	0.00%	0.00%
Kazakhstan	2.11%	0.00%		Costa Rica	0.00%	7.14%
Malaysia	2.11%	0.00%		Croatia	0.00%	0.00%
Russia	2.11%	0.00%		Czech Rep.	0.00%	0.00%
Slovakia	2.11%	0.00%		Dem. Rep.	0.00%	0.00%

				Congo		
Sweden	2.11%	0.00%		Denmark	0.00%	0.00%
Brazil	1.05%	7.14%		Dominican Rep.	0.00%	7.14%
Italy	1.05%	0.00%				

Chapter 4

Table 4A: Computations for calculating the per person travel cost visitor, Mount Rainier National Park

	Travel Distance				Cost of Travel				Value of Time					Cost of Time	Entrance Fee	TOTAL
Dist. Range	Median Dist. (KM)	Median Dist.: miles, euclidean	Circuity Factor	Median Dist.: miles, roads	Reimbursement Rate (per mile)	Mileage Cost	PPV	Mileage Cost per Person	Av. Speed mph	Travel Time (hours)	Av. Weekly Wage	Wage Rate	Cost of travel time	Total travel cost per person (RT)	Av. Entry fee Cost per person	Total Cost per Person
0-30	15	9.32	1.15	10.72	0.57	6.06	2.90	2.09	50	0.19	1044	26.1	2.43	9.04	4.60	13.64
30-40	35	21.75	1.15	25.01	0.57	14.13	2.90	4.87	50	0.43	1044	26.1	5.68	21.10	4.60	25.70
40-50	45	27.96	1.15	32.16	0.57	18.17	2.90	6.26	50	0.56	1044	26.1	7.30	27.13	4.60	31.72
50-60	55	34.18	1.15	39.30	0.57	22.21	2.90	7.66	50	0.68	1044	26.1	8.92	33.15	4.60	37.75
60-70	65	40.39	1.15	46.45	0.57	26.24	2.90	9.05	50	0.81	1044	26.1	10.54	39.18	4.60	43.78
70-80	75	46.60	1.15	53.59	0.57	30.28	2.90	10.44	50	0.93	1044	26.1	12.16	45.21	4.60	49.81
80-90	85	52.82	1.15	60.74	0.57	34.32	2.90	11.83	50	1.06	1044	26.1	13.79	51.24	4.60	55.84
90-100	95	59.03	1.15	67.88	0.57	38.35	2.90	13.23	50	1.18	1044	26.1	15.41	57.27	4.60	61.86
100-110	105	65.24	1.15	75.03	0.57	42.39	2.90	14.62	50	1.30	1044	26.1	17.03	63.29	4.60	67.89
110-120	115	71.46	1.15	82.18	0.57	46.43	2.90	16.01	50	1.43	1044	26.1	18.65	69.32	4.60	73.92
120-130	125	77.67	1.15	89.32	0.57	50.47	2.90	17.40	50	1.55	1044	26.1	20.27	75.35	4.60	79.95
130-140	135	83.89	1.15	96.47	0.57	54.50	2.90	18.79	50	1.68	1044	26.1	21.89	81.38	4.60	85.97
140-150	145	90.10	1.15	103.61	0.57	58.54	2.90	20.19	50	1.80	1044	26.1	23.52	87.41	4.60	92.00
150-160	155	96.31	1.15	110.76	0.57	62.58	2.90	21.58	50	1.93	1044	26.1	25.14	93.43	4.60	98.03
160-170	165	102.53	1.15	117.91	0.57	66.62	2.90	22.97	50	2.05	1044	26.1	26.76	99.46	4.60	104.06
170-180	175	108.74	1.15	125.05	0.57	70.65	2.90	24.36	50	2.17	1044	26.1	28.38	105.49	4.60	110.09
180-190	185	114.95	1.15	132.20	0.57	74.69	2.90	25.76	50	2.30	1044	26.1	30.00	111.52	4.60	116.11
190-200	195	121.17	1.15	139.34	0.57	78.73	2.90	27.15	50	2.42	1044	26.1	31.62	117.54	4.60	122.14
200-210	205	127.38	1.15	146.49	0.57	82.77	2.90	28.54	50	2.55	1044	26.1	33.25	123.57	4.60	128.17
210-220	215	133.59	1.15	153.63	0.57	86.80	2.90	29.93	50	2.67	1044	26.1	34.87	129.60	4.60	134.20

220-230	225	139.81	1.15	160.78	0.57	90.84	2.90	31.32	50	2.80	1044	26.1	36.49	135.63	4.60	140.23
230-240	235	146.02	1.15	167.93	0.57	94.88	2.90	32.72	50	2.92	1044	26.1	38.11	141.66	4.60	146.25
240-250	245	152.24	1.15	175.07	0.57	98.92	2.90	34.11	50	3.04	1044	26.1	39.73	147.68	4.60	152.28
250-260	255	158.45	1.15	182.22	0.57	102.95	2.90	35.50	50	3.17	1044	26.1	41.36	153.71	4.60	158.31
260-270	265	164.66	1.15	189.36	0.57	106.99	2.90	36.89	50	3.29	1044	26.1	42.98	159.74	4.60	164.34
270-280	275	170.88	1.15	196.51	0.57	111.03	2.90	38.29	50	3.42	1044	26.1	44.60	165.77	4.60	170.37
280-290	285	177.09	1.15	203.65	0.57	115.06	2.90	39.68	50	3.54	1044	26.1	46.22	171.80	4.60	176.39
290-300	295	183.30	1.15	210.80	0.57	119.10	2.90	41.07	50	3.67	1044	26.1	47.84	177.82	4.60	182.42
300-310	305	189.52	1.15	217.95	0.57	123.14	2.90	42.46	50	3.79	1044	26.1	49.46	183.85	4.60	188.45
310-320	315	195.73	1.15	225.09	0.57	127.18	2.90	43.85	50	3.91	1044	26.1	51.09	189.88	4.60	194.48
320-330	325	201.95	1.15	232.24	0.57	131.21	2.90	45.25	50	4.04	1044	26.1	52.71	195.91	4.60	200.51

Table 4B: Computations for calculating the per person cost of visitor, Great Sand Dunes National Park

	Travel Distance				**Cost of Travel**				**Value of Time**					**Cost of Time**	**Entrance Fee**	**TOTAL**
Dist. Range	Median Dist. (KM)	Median Dist.: miles, euclidean	Circuity Factor	Median Dist.:miles, roads	Reimbursement Rate (per mile)	Mileage Cost	PPV	Mileage Cost per Person	Av. Speed (mph)	Travel Time (hours)	Average Weekly Wage	Hourly Wage Rate	Cost of Travel Time	Total travel cost per person (RT)	Av. Entry fee per person	Total Cost per Person
0-30	15	9.32	1.15	10.72	0.52	5.57	2.66	2.10	50	0.19	1023	25.58	2.38	8.96	3	11.96
30-40	35	21.75	1.15	25.01	0.52	13.01	2.66	4.89	50	0.43	1023	25.58	5.56	20.90	3	23.90
40-50	45	27.96	1.15	32.16	0.52	16.72	2.66	6.29	50	0.56	1023	25.58	7.15	26.87	3	29.87
50-60	55	34.18	1.15	39.30	0.52	20.44	2.66	7.68	50	0.68	1023	25.58	8.74	32.85	3	35.85
60-70	65	40.39	1.15	46.45	0.52	24.15	2.66	9.08	50	0.81	1023	25.58	10.33	38.82	3	41.82
70-80	75	46.60	1.15	53.59	0.52	27.87	2.66	10.48	50	0.93	1023	25.58	11.92	44.79	3	47.79
80-90	85	52.82	1.15	60.74	0.52	31.58	2.66	11.87	50	1.06	1023	25.58	13.51	50.76	3	53.76
90-100	95	59.03	1.15	67.88	0.52	35.30	2.66	13.27	50	1.18	1023	25.58	15.10	56.74	3	59.74

100-110	105	65.24	1.15	75.03	0.52	39.02	2.66	14.67	50	1.30	1023	25.58	16.69	62.71	3	65.71
110-120	115	71.46	1.15	82.18	0.52	42.73	2.66	16.06	50	1.43	1023	25.58	18.28	68.68	3	71.68
120-130	125	77.67	1.15	89.32	0.52	46.45	2.66	17.46	50	1.55	1023	25.58	19.86	74.65	3	77.65
130-140	135	83.89	1.15	96.47	0.52	50.16	2.66	18.86	50	1.68	1023	25.58	21.45	80.62	3	83.62
140-150	145	90.10	1.15	103.61	0.52	53.88	2.66	20.26	50	1.80	1023	25.58	23.04	86.60	3	89.60
150-160	155	96.31	1.15	110.76	0.52	57.59	2.66	21.65	50	1.93	1023	25.58	24.63	92.57	3	95.57
160-170	165	102.53	1.15	117.91	0.52	61.31	2.66	23.05	50	2.05	1023	25.58	26.22	98.54	3	101.54
170-180	175	108.74	1.15	125.05	0.52	65.03	2.66	24.45	50	2.17	1023	25.58	27.81	104.51	3	107.51
180-190	185	114.95	1.15	132.20	0.52	68.74	2.66	25.84	50	2.30	1023	25.58	29.40	110.48	3	113.48
190-200	195	121.17	1.15	139.34	0.52	72.46	2.66	27.24	50	2.42	1023	25.58	30.99	116.46	3	119.46
200-210	205	127.38	1.15	146.49	0.52	76.17	2.66	28.64	50	2.55	1023	25.58	32.58	122.43	3	125.43
210-220	215	133.59	1.15	153.63	0.52	79.89	2.66	30.03	50	2.67	1023	25.58	34.17	128.40	3	131.40
220-230	225	139.81	1.15	160.78	0.52	83.61	2.66	31.43	50	2.80	1023	25.58	35.76	134.37	3	137.37
230-240	235	146.02	1.15	167.93	0.52	87.32	2.66	32.83	50	2.92	1023	25.58	37.35	140.35	3	143.35
240-250	245	152.24	1.15	175.07	0.52	91.04	2.66	34.22	50	3.04	1023	25.58	38.93	146.32	3	149.32
250-260	255	158.45	1.15	182.22	0.52	94.75	2.66	35.62	50	3.17	1023	25.58	40.52	152.29	3	155.29
260-270	265	164.66	1.15	189.36	0.52	98.47	2.66	37.02	50	3.29	1023	25.58	42.11	158.26	3	161.26
270-280	275	170.88	1.15	196.51	0.52	102.18	2.66	38.42	50	3.42	1023	25.58	43.70	164.23	3	167.23
280-290	285	177.09	1.15	203.65	0.52	105.90	2.66	39.81	50	3.54	1023	25.58	45.29	170.21	3	173.21
290-300	295	183.30	1.15	210.80	0.52	109.62	2.66	41.21	50	3.67	1023	25.58	46.88	176.18	3	179.18
300-310	305	189.52	1.15	217.95	0.52	113.33	2.66	42.61	50	3.79	1023	25.58	48.47	182.15	3	185.15
310-320	315	195.73	1.15	225.09	0.52	117.05	2.66	44.00	50	3.91	1023	25.58	50.06	188.12	3	191.12
320-330	325	201.95	1.15	232.24	0.52	120.76	2.66	45.40	50	4.04	1023	25.58	51.65	194.09	3	197.09

Works Cited

Allison, Paul (2014). "Prediction vs. Causation in Regression Analysis." Available from: http://www.statisticalhorizons.com/prediction-vs-causation-in-regression-analysis (Accessed April 10, 2015).

Antolini, Denise (2009). "National Park Law in the U.S.: Conservation , Conflict , and Centennial Values." *William and Mary Environmental Law and Policy Review* 33(3): 851–921.

Arches National Park (1993). *Public Use Counting and Reporting Instructions*.

Benson, Charles, P. Watson, G. Taylor, P. Cook, and S. Hollenhorst. (2013). "Who Visits a National Park and What Do They Get out of It?: A Joint Visitor Cluster Analysis and Travel Cost Model for Yellowstone National Park." *Environmental Management* 52(4): 917-928.

Bhat, Gajanan and John C. Bergstrom (1996). "Integration of Geographical Information Systems Based Spatial Analysis in Recreation Demand Analysis." *University of Georgia, Department of Agricultural and Applied Economics* No. 16649.

Brown, Gardner Jr and Robert Mendelsohn (1984). "The Hedonic Travel Cost Method." *The Review of Economics and Statistics* 66: 427–33.

Brown, Philip (1982). "Multivariate Calibration." Journal of the Royal Statistical Society. Series B (Methological), 287- 321.

Brown, Philip and Rolf Sundberg (1987). "Confidence and Conflict in Multivariate Calibration.: *Journal of the Royal Statistical Society. Series B (Methodological)*, 46-57.

Bureau of Labor Statistics (2014). "County Employment and Wages in Washington – Third Quarter 2013". Bureau of Labor Statistics, Accessed 4/30/15.

Cameron, Colin and Pravin Trivedi (1998). *Regression Analysis of Count Data*. 2nd edition, Econometric Society Monograph No.53, Cambridge University Press, 1998 (566 pages.)

Collins, Rachel (2015). Personal Communication, Conducted March 3, 2015.

Corder, Gregory and Dale Foreman (2009). *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. John Wiley & Sons.

Dickey, David (2012). "Introduction to Predictive Modeling with Examples." *SAS Global Forum 2012*.

Donovan, Geoffrey and Patricia Champ (2009). "The Economic Benefits of Elk Viewing at the Jewell Meadows Wildlife Area in Oregon." *Human Dimensions of Wildlife* 14: 51–60.

Duffield, John, Chris Neher, David Patterson, and Aaron Deskins (2013). "Effects of Wildfire on National Park Visitation and the Regional Economy: A Natural Experiment in the Northern Rockies." *International Journal of Wildland Fire*: 22(8), 1155-1166.

Gorte, Ross, Laura Hanson, and Marc Rosenblum (2012). "Federal Land Ownership: Overview and Data" *Congressional Research Service,* 42346.

Great Sand Dunes National Park (2002). "Public Use Counting and Reporting Instructions."

Great Sand Dunes National Park (n.d.). "Fees and Passes." Available from: http://www.nps.gov/grsa/planyourvisit/fees.htm (Accessed May 1, 2015).

Heberling, Matthew and Joshua Templeton (2009). "Estimating the Economic Value of National Parks with Count Data Models Using On-Site, Secondary Data: The Case of the Great Sand Dunes National Park and Preserve." *Environmental Management*, 43(4), 619-627.

Hellerstein, D (1993). "ZIPFIP: A ZIP and FIPS Database : Users Manual." *Department of Agriculture: Economic Research Service.*

Ignite Social Media (2012). "Social Network Analysis Report." Available from: http://www.ignitesocialmedia.com/social-media-stats/2012-social-network-analysis-report/ (Accessed April 5, 2015).

Keeler, Bonnie (2015). "Studying Lake Visitation Using Flickr." *Frontiers in Ecology and Environment* 10(1890/140124).

Kirkman, T.W (1996). "Statistics to Use: Kolmogorov-Smirnov Test." *College of Saint Benedict & Saint John's University*. http://www.physics.csbsju.edu/stats/ (Accessed April 8, 2015).

Loomis, John (2000). "Counting on Recreation Use Data: A Call for Long-Term Monitoring." *Journal of Leisure Research*: *32*(1), 93.

Morgan, K., Larkin and Adams (2011). "Empirical Analysis of Media Versus Environmental Impacts on Park Attendance. *Tourism Management*,*32*(4), 852-859.

Mount Rainier National Park (1995). *Public Use Counting and Reporting Instructions*.

Mount Rainier National Park (n.d.) "Fees and Passes." Available from: http://www.nps.gov/mora/planyourvisit/fees.htm (Accessed May 1, 2015).

National Park Service (2004). "Director's Order #82." *US National Park Service.*

National Park Service (2012). "A Call to Action." *US National Park Service.*

National Park Service (2014). "Visitor Use Statistics." *US National Park Service.* Available from: http://www.nature.nps.gov/socialscience/stats.cfm (Accessed January 7, 2015).

Neher, Christopher, John Duffield, and David Patterson (2013). "Valuation of National Park System Visitation: The Efficient Use of Count Data Models, Meta-Analysis, and Secondary Visitor Survey Data." *Environmental Management*. 52(3), 683-698.

Poudyal, Neelam, Bamadev Paudel, and Gary Green (2013a). "Estimating the Impact of Impaired Visibility on the Demand for Visits to National Parks." *Tourism Economics*, 19(2), 433-452.

Poudyal, Neelam, Bamadev Paudel, and Michael Tarrant (2013b). "A Time Series Analysis of the Impact of Recession on National Park Visitation in the United States." *Tourism Management*, *35*: 181-189.

Sharp, et al. (2015). "Invest User Guide" *Natural Capital Project*. Available from www.naturalcapitalproject.org.

Smith, V. and Yoshiaki Kaoru (1990). "What Have We Learned since Hotelling's Letter?" *Economics Letters* 32(3): 267–72.

State of Colorado (2015). "Mileage Reimbursement Rate History." *Office of the State Controller*. Available from: https://www.colorado.gov/pacific/osc/mileage-reimbursement-rate-history (Accessed April 23, 2015).

State of Washington (2015). "Travel." *Office of Financial Management*. Available form: http://www.ofm.wa.gov/resources/travel.asp (Accessed April 23, 2014).

Stevens, J. and Richard Frank (2009). "Current Policy and Legal Issues Affecting Recreational Use of Public Lands in the American West." *Resources for the Future:* RFF DP 09-23.

United States Congress (1960). *Multiple-Use Sustained-Yield Act*. United States Congress

USDA Forest Service (2012). "National Visitor Use Monitoring Results: National Summary Report". *United States Department of Agriculture.*

Ward, F. and Diana Beal (2000). *Valuing Nature with Travel Cost Models: A Manual.* Cheltenham: Edward Elgar.

Wolff, Hendrik (2014). "Value of Time: Speeding Behavior and Gasoline Prices." *Journal of Environmental Economics and Management* 67: 71–88.

Wood, Spencer, Anne Guerry, Jessica Silver, and Martin Lacayo (2013). "Using Social Media to Quantify Nature-Based Tourism and Recreation." *Scientific reports* 3: 2976.

Yellowstone National Park (1995). "Public Use Counting and Reporting Instructions."

Yosemite National Park (2004). "Public Use Counting and Reporting Instructions."

Ziesler, Pamela (2015). Personal Communication, Conducted February 27, 2015.