

## Are mobile device location data a substitute for travel cost surveys?

Jude Bayham<sup>a\*</sup>, Aaron J. Enriquez<sup>b</sup>, Leslie Richardson<sup>c</sup>

<sup>a</sup> *Department of Agricultural and Resource Economics, Colorado State University, Fort Collins, CO, USA*

<sup>b</sup> *U.S. Geological Survey, Fort Collins Science Center, Fort Collins, CO, USA*

<sup>c</sup> *National Park Service, Economics Program, Fort Collins, CO, USA*

*\* Corresponding author*

*E-mail: [jbayham@colostate.edu](mailto:jbayham@colostate.edu)*

### Statements and Declarations

Views and conclusions in this article are those of the authors and do not necessarily reflect the opinions or policies of the National Park Service.

### Acknowledgments

Colorado State University organized, funded, and implemented collection of the mobility data described in this information product. The visitor surveys described in this manuscript were organized, funded, and implemented by the National Park Service. Data were not collected on behalf of the U.S. Geological Survey. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Bayham thanks Kate Floersheim and Isa Naschold for research assistance. The authors thank Christian Crowley for valuable feedback on a draft manuscript.

The authors acknowledge the use of ChatGPT, Microsoft 365 Copilot, and Github Copilot for developing code to process and analyze data. However, the manuscript is entirely the product of the authors.

## **Abstract**

Mobile device location data offer a low-cost alternative for measuring visitation to outdoor recreation sites and are known to correlate with official visitation counts. Less is known about whether these data can recover recreation demand and consumer surplus comparable to survey-based methods. We compare travel cost models estimated using mobile device and survey data for 17 U.S. National Park Service sites. Results are mixed. We examine the roles of aggregation and sampling bias and use LASSO regression to assess whether sample and site characteristics explain discrepancies.

# 1. Introduction

Mobile device location data have become widely available in the past five years and are being used to study many aspects of human behavior. Environmental economists have demonstrated the potential to use these data to study recreation demand, which has long relied on expensive and time-consuming survey methods (Kubo et al. 2020; Kling et al. this issue). Mobile device location data promise a relatively low-cost source of high-resolution visitation data across broad geographies over time. Having credible welfare estimates over time for a wide range of recreation sites could unlock many research questions and provide important information about the economic value of public lands. The critical research question is whether mobile device location data are a reliable substitute for conventional travel cost surveys.

We seek to answer the substitute question by estimating travel cost models for 17 U.S. National Park Service (NPS) sites using both conventional survey data and mobile device data (See Figure 1). The NPS recently launched a visitor survey effort that collects detailed information from park visitors necessary to estimate travel cost models. We compare survey- and mobile-based welfare estimates across a range of park sites, from large and well-known scenic parks to small historic sites. We prepare the data using the same or equivalent assumptions to ensure the most appropriate comparisons. At each site, we estimate a series of models across a spectrum of aggregation. We start by using all the individual-specific information from the individual-level survey data, then progressively aggregate the data to investigate the consequences of aggregation, characteristic of mobile device data. We also estimate models with the mobile device data alone and with augmentation, leveraging information from the survey data. Together,

comparison across model specifications and sites allows us to disentangle the effects of data aggregation from other sources of difference between the consumer surplus estimates.<sup>i</sup>

<<Fig 1 about here>>

Recreation demand researchers have sought out alternative data sources to better understand recreation values for years. Social media metadata is now stored in central repositories that can be mined to collect data on visitors (Heikinheimo et al. 2017; Keeler et al. 2015; Sinclair, Ghermandi, and Sheela 2018; Wood et al. 2013). Mobility data have become widely available since the late 2010s and contain the information needed to estimate recreation demand models. These data have been used to estimate single-site travel cost models (Kubo et al. 2020) as well as more complex discrete choice models (Kling et al. this issue).

Alternative data sources must not only contain origin and destination visitation data, but the data must accurately describe the behaviors of the population of interest. In general, research has found that social media data can approximate visitation trends to recreation sites (Ghermandi 2018; Goebel et al. 2023; Heikinheimo et al. 2017; Sinclair et al. 2018; Wilkins, Wood, and Smith 2021; Wood et al. 2020). Sessions et al. (2016) show that Flickr data corresponds to visitation at U.S. national parks. Using Airsage and Safegraph data, respectively, Merrill et al. (2024) and Tsai et al. (2023) find that mobile data correlates with visitation data at some U.S. national parks, but does not correlate well in others. RRC Associates (2024) show how high-resolution data can be used to study the ways in which visitors use parks.

Most mobile device location data are now collected via smartphone applications where users permit location services to collect timestamped GPS coordinates over time. For most research applications, there is a need to ensure that the sample of devices contributing to the dataset is free

from systematic bias. Indeed, Li et al. (2024) compare Advan (our data source) mobile device counts to U.S. Census data and find minimal systematic biases. We conduct similar bias analyses on our sample and draw similar conclusions. We then compare the distribution of visit distances (origin to destination) between the survey and mobile device data to investigate systematic differences between the two samples of site visitors.

We contribute to this growing literature in three ways. First, we estimate single-site travel cost models for a set of U.S. national parks and derive welfare estimates. Our approach is most similar to Sinclair et al. (2020), who compare travel cost and welfare estimates based on geolocated photographic data and conventional survey data in German national parks. However, we use mobile device data that samples a much larger share of the population and focuses on U.S. national park sites that exhibit more variation in site types. To our knowledge, our analysis is one of the first applications of mobile device data to estimate welfare values for recreation sites within the U.S. National Park System. Second, we develop methods to overcome data limitations inherent in mobile device data. Specifically, we train regression models based on the visitor survey data and use them to impute the probability of flying and the number of people splitting expenses, both of which are key inputs to travel cost models. Third, we compute welfare estimates for NPS sites over time, demonstrating how these data can be used to study values across various timeframes and seasons.

We find that mobile device location data are not a perfect substitute for on-site surveys, in the sense that we cannot simply replace conventional survey data with mobile device data without losing information. However, mobile device data can provide valuable information to fill in gaps, both temporal and spatial, when no surveys can be completed. This can help land management agencies reduce monitoring costs by reducing the required frequency of on-site surveys and has

important implications for improved policy and management decisions. For instance, continued comparisons of mobile and survey data can support calibration factors applied to consumer surplus estimates used in natural resource damage assessments, regulatory analyses, visitor use management strategies, and other analyses that involve recreation opportunities occurring in a season or location with no available survey data. Mobile device data can also play a role in providing information for new legislation. For instance, the 2025 Expanding Public Lands Outdoor Recreation Experiences (EXPLORE) Act (Public Law 118-234) is a significant and wide-reaching piece of bipartisan legislation that seeks to better understand and improve visitor experiences on Federal public lands, largely through improved data collection and monitoring.<sup>ii</sup> Given the considerable resource constraints faced by agencies, supplementing more traditional data collection methods with mobile device data that are relatively easier to collect over longer time horizons and at dispersed/low use recreation areas can lead to improved estimates of both recreation visitation and welfare.

## 2. Data

We assemble data from multiple sources to facilitate a comparison of travel cost models estimated with survey data and mobile device data for a sample of NPS sites across the United States. Survey data come from a new NPS monitoring program that administers visitor surveys at 24 randomly-selected park units each year. Mobile device location data are aggregated and provided by Advan using Safegraph Point of Interest locations via DeweyData.io. We supplement these data with socioeconomic and demographic data from the U.S. Census 5-year American Community Survey. We calculate driving distance and time using Open Source

Routing Machine (OSRM). In this section, we describe each data source and how we process it, and we present summary statistics comparing the survey and mobile device datasets.

The NPS recently launched a new visitor survey effort to collect systematic and comprehensive data from visitors across a representative sample of park sites. A stratified random sampling approach is used to determine which park sites will be included in the sample each year. First, all park units with adequate visitation data are grouped into one of four categories representing the “type” of park unit, including natural, recreational, historic urban, and historic non-urban/other. These categories are based primarily on Congressional designation, as well as the population of the urban area or metropolitan statistical area in which the park unit is located. Park units are then grouped into a “high” or “low” visitation category, based on the top 80% and bottom 20% of park visits within each of the four park unit types. Based on these classifications, every eligible unit of the National Park System is included in one of eight strata. Each year, three parks from each stratified group are randomly selected, resulting in 24 parks surveyed each year. Parks are sampled without replacement to ensure adequate coverage across the system. Our analysis focuses on the first two years of survey administration (2022 and 2023).

The NPS visitor survey has two components: a relatively short on-site intercept survey that collects information from visitors as tablet-based responses during their park visit, and a longer, post-trip follow-up survey that can be completed by the respondent online or mailed back. The on-site survey is typically administered over a 10-day sampling period and includes key questions for the travel cost modeling, such as the number of trips the respondent took to the park over the past 12 months, the number of people in the respondent’s group splitting trip expenses on the current trip, the transportation modes used by the respondent to travel to the site, and the respondent’s home ZIP code. To ensure a representative sample of respondents,

surveyors are spread throughout each park unit and placed in locations where they are most likely to intercept all types of visitors. Areas such as campgrounds and lodges are typically avoided, since they tend to draw a specific visitor segment. For busier parks, every Xth visitor group is intercepted (X varies depending on how busy the park is) and within each intercepted group, the adult respondent with the most recent birthday is asked to complete the survey. For smaller parks with low visitation, surveyors contact nearly a census of all visitors arriving during the survey period. While response rates vary across park units, on average, slightly more than 80% of intercepted visitors agree to take the intercept survey. These same respondents are asked to complete an additional post-trip follow-up survey, which includes additional questions relevant to the travel cost modeling, such as the respondent's mode of travel and household income. On average, around 30-45% of visitors respond to the follow-up survey. For more information on the NPS visitor surveys, see Otak Inc., RRC Associates, and University of Montana (2023).

We build a mobile device dataset based on monthly aggregate unique visitors to NPS sites provided by Advan via deweydata.io (Monthly Patterns).<sup>iii</sup> Advan constructs monthly aggregate visits to a point of interest (POI) based on mobile device GPS time-stamped coordinates from a sample of phones that permit certain apps to use location services. A visit occurs when a device generates a GPS signal from inside a predefined polygon, regardless of dwell time (refer to Appendix 1 for more details on how visits are measured). Consequently, transient devices that pass through a large POI may be counted as a visit. We are unable to change the way visits are calculated without access to the underlying data. Moreover, we are unable to change the POI boundary or geofence used to calculate visitation. Some boundaries include the entire site and parking lots, while others contain just an important structure like a visitors center. Advan's

Monthly Patterns product reports these visitation data as different spatial and temporal aggregates. Specifically, they report “visitor\_home\_aggregation” - the number of unique devices visiting a POI (over a month) by devices’ “home” census tracts in the United States, generating a set of monthly origin-destination pairs necessary to estimate a travel cost model (refer to Appendix 1 for how home census tract is determined).

These visit data disaggregated by census tract are subject to “differential privacy” rules, which intentionally distort the data by: 1) adding statistical (Laplace) noise to the count, 2) suppressing counts if fewer than two devices visit from a specific census tract (censoring), and 3) reporting four visits if between two and four devices visit (truncation). We describe the econometric consequences of the differential privacy in Section 3. Refer to Wan et al. (this issue) for a detailed discussion about differential privacy.

We extract visits to points of interest that best align with the official boundaries of U.S. park sites included in the NPS visitor survey effort. Prior to December 2022, Advan calculated visitation metrics for all official NPS boundaries. Advan no longer reports visitation metrics based on NPS boundaries and instead reports on specific sites within many U.S. NPS sites.<sup>iv</sup> In several cases, mobility visitation data did exist, but they were too sparse to estimate a travel cost model. We chose the set of 17 sites based on good alignment of park boundaries and sufficient visitation data from both sources. The final list of sites used in our analysis is shown in Table A1, along with the site location, survey period, and annual visitation.

We align the visitation units across the visitor survey and mobile datasets to the greatest extent possible. The NPS survey asks respondents to report the total number of visits taken to the site over the past 12 months. As a result, any repeat visits by the same visitor within that timeframe

are included in the data, regardless of whether those visits occurred in the same week or month. In contrast, the mobile device data report the number of unique visitors by census tract within the calendar month. Multiple visits by the same device within a month are counted once, whereas visits by the same device in different months are counted separately. We align the mobile device data with the visitor survey data by summing the monthly visits by census tract over the trailing 12 months prior to the survey sampling period. Under ideal conditions in which both the NPS survey and mobile device visitors represented the true visiting population, the number of trips reported by survey respondents and the summed unique monthly mobile device visitors by census tract should be comparable.

One of the challenges of using mobility data for travel cost modeling is the lack of individual socioeconomic and demographic information. To investigate the consequences of unobserved data, we estimate survey-based travel cost models using U.S. Census socioeconomic and demographic data, emulating the limitation in the mobility data. We collect both tract and ZIP code-level data from the 2022 and 2023 5-year American Community Survey (ACS) on the following variables: Total Population (Table B01001), Median Age (B01002), Per Capita Income (B19301), and Average Household Size (B25010). Advan reports visitation by home census tracts using the 2010 – 2019 official tract boundaries, which do not align with ACS data post 2020. We use relationship files to map the 2022 and 2023 ACS data back to 2010-2019 tract boundaries.<sup>v</sup> However, we use the 2010-2019 Tiger geometries to determine origin location in travel distance and time calculations.

## 2.1. Calculating Travel Cost

Travel distances and times are critical inputs to travel cost models. Similar to English et al. (2018), we consider two modes of travel: 1) driving the entire way from home to destination, and 2) flying, which involves driving to the departure airport, flying to a destination airport, and driving to the final destination. For individual  $i$ , the weighted travel cost from ZIP code  $z$  to destination  $j$  is given by

$$wtc_{ijz} = (1 - probfly_{ijz}) \cdot dtc_{ijz} + probfly_{ijz} \cdot ftc_{ijz} \quad \forall i, j, z, \quad (1)$$

in which  $dtc_{ijz}$  are the driving-only travel costs,  $ftc_{ijz}$  are travel costs for both driving and flying, and  $probfly_{ijz}$  is the probability that someone flies. For NPS survey respondents who reported their mode of transportation, weighted travel costs simplify down to either just the driving-only cost if they drove (because  $probfly_{ijz}$  in that case is equal to zero) or just the driving and flying cost if they flew (because  $probfly_{ijz}$  in that case is equal to one). When survey respondents did not answer the question, and for all the mobile visitation data, we impute the probability of flying using a logistic regression model trained on the survey data. Refer to Appendix 3 for details on the prediction model and results. In the remainder of this section, we detail the construction of travel costs for the two modes of transportation.

We estimate the cost of individual  $i$  driving from origin zone  $z$  (zipcode or census tract) to destination  $j$  as

$$dtc_{ijz} = 2 \cdot \left[ \frac{c_d \cdot dd_{jz} + c_h h_{jz}}{ns_{iz}} + \frac{1}{3} \cdot \frac{inc_{iz}}{2040} \cdot dt_{jz} \right] + F_j, \quad (2)$$

in which  $c_d$  is the average cost per mile driven,  $dd_{jz}$  is the one-way driving distance from the origin to the destination,  $c_h$  is the cost of a hotel night,  $h_{jz} = dt_{jz}/12$  is an integer number of hotel nights assuming one night in a hotel for every 12 hours of driving (which matches the assumption used by English et al. (2018)),  $ns_{iz}$  is the number of people splitting the driving cost,<sup>vi</sup>  $inc_{iz}$  is the individual's reported income or the per capita income from zone  $z$ ,  $dt_{jz}$  is the one-way driving time from home zone to destination, and  $F_j$  is the per-person fee to access the destination.

We obtain driving distances and times between all survey-based and mobile device origin-destination pairs using OSRM, specifically using the R package *osrm* (Giraud 2022).<sup>vii</sup> We use polygon centroids of home ZIP codes (ZIP code tabulation areas (ZCTAs) for the survey data) and census tracts (for the mobile data) based on U.S. Census TIGER polygons accessed via the R package *tigris* (Walker 2024). We select popular locations near the entrances of large NPS sites (e.g., visitor centers) as destinations. We acknowledge that the choice of destination can influence travel distance estimations, particularly as large parks can have multiple potential entry points. However, we use the same destination points for the survey and mobile data.

The  $c_d \cdot dd_{jz}$  term of the first sum represents driving costs. We use the American Automobile Association's (AAA's) weighted average operating cost per mile, which considers fuel and maintenance costs per mile as averaged over all major vehicle types. The AAA cost estimates were 27.7 cents per mile in 2022 and 25.8 cents per mile in 2023 (AAA 2022, 2023). The  $c_h h_{jz}$  term of the first sum represents hotel costs. For the cost of a hotel night, we use estimates from the American Hotel & Lodging Association (AHLA), which detail average daily rates for hotels

in 2022 (\$149) and 2023 (\$155) in their 2024 State of the Industry Report (Kilic, Cashour, and Carrier 2024).

The second term of the sum represents the opportunity costs of time spent driving. We calculate wage rate by dividing individual- or zone-level income by 2040 hours (51 40-hour work weeks), which provides an estimate of the number of hours worked per year. We value time spent traveling by multiplying the wage rate by one-third, as is standard in travel cost literature (Champ, Boyle, and Brown 2017). We multiply the wage rate by the time spent driving one-way, which is directly provided by OSRM. The first two terms of the sum are multiplied by two, such that costs are on a roundtrip basis.

The third and final term in the sum represents NPS site fees. Some of the sites have per-person entry fees, which we use directly for  $F_j$ . Other sites have per-vehicle fees, which we divide by the number of people splitting expenses.

Travel costs for those who both drive and fly are more difficult to measure. We decompose such routes into three legs. First, a person drives from their home to an origin airport. Second, a person flies from an origin airport to a destination airport near the recreation site. Third, a person rents a car and drives from the destination airport to the recreation site. We estimate the cost of individual  $i$  flying from ZIP code  $z$  to destination  $j$  as

$$ftc_{ijz} = 2 \cdot \left[ \frac{c_d(dd_{jz}^{leg1} + dd_{jz}^{leg3}) + r_j}{ns_{iz}} + a_{ijz} + \frac{1}{3} \cdot \frac{inc_{iz}}{2040} \cdot (dt_{jz}^{leg1} + ft_{jz}^{leg2} + dt_{jz}^{leg3}) \right] + F_j \quad (3)$$

in which  $dd_{jz}^{leg1}$  and  $dd_{jz}^{leg3}$  are the one-way driving distances for legs 1 and 3, respectively,  $r_j$  is the cost of a rental car,  $a_{ijz}$  is the airline fare from origin airport to destination airport,  $dt_{jz}^{leg1}$

and  $dt_{jz}^{leg3}$  are the driving times for legs 1 and 3, respectively,  $ft_{jz}^{leg2}$  is the flying time for leg 2, and all other terms are defined as in the equation for driving-only travel costs.

We treat an individual's choice of origin and destination airport as a cost-minimization problem, in which the individual considers the ten nearest origin airports to their home ZIP code centroid and the ten nearest destination airports to the recreation site. We narrowed down the list of all U.S. airports to those that appear every year from 2010 to 2022 in the Bureau of Transportation Statistics' airport rankings for originating domestic passengers (Bureau of Transportation Statistics 2025a). We assigned spatial coordinates to those airports using a Bureau of Transportation Statistics dataset on aviation facilities (Bureau of Transportation Statistics 2025b). This in turn allowed us to identify the closest airports to the home ZIP code centroids and destination sites, using the R package *sf*. Overall, we assume that individuals pick the origin-destination airport combo that entails the lowest total travel cost across all three legs (out of the 100 possible combinations across the 10 nearest destination airports and 10 nearest destination airports).

To approximate the cost of a rental car,  $r_j$ , we obtained average weekly rental car prices at the 15 largest U.S. airports from Nerd Wallet (French 2025). We converted from weekly to daily prices. We then assigned each U.S. airport a daily rental car price by using the daily rental car price from the closest of the 15 largest U.S. airports. The spatial matching was again done using the R package *sf*. The NPS visitor surveys contain a question asking respondents how many days they planned to spend in the local area near the recreation site. We calculate the median number of local days across all respondents surveyed at a given recreation site, which we multiply by the daily rental car price for all airports near the recreation site.

To calculate airline fares,  $a_{jz}$ , we obtained average fares for contiguous U.S. city-pair markets averaging at least 10 passengers per day, as provided by the U.S. Department of Transportation in their Consumer Airfare Report (U.S. Department of Transportation 2025). We assigned every city in the dataset to the NPS Region in which it is located, covering the Pacific West Region, the Intermountain Region, the Midwest Region, the Southeast Region, and the Northeast Region. We then predict average airfare for each city-pair as a function of distance between the cities, year, and region-region combination. We report the linear regression results in Appendix 4. Using the results, we predict the airline fare for all possible airport-airport combinations.

To obtain time spent flying,  $ft_{jz}^{leg2}$ , we divide the geodesic distance between origin and destination (as approximated using the R package *sf* package) by the average speed of a commercial passenger aircraft (528 miles per hour<sup>viii</sup>). Like English et al. (2018), we add two hours to approximate time spent in airports.

## 2.2. Datasets

We construct a total of three datasets designed to probe limitations in the mobile device data. First, we build a dataset based on the NPS visitor survey individual trip counts, reported income and demographic information, as well as travel mode and the number of people splitting trip expenses (NPS Individual). When respondents do not report travel mode and the number of people splitting expenses, we impute missing data using the models trained on individual survey data (refer to Appendix 3). Second, we limit the NPS survey data further by summing trips by all individuals from each ZIP code, using 5-year ACS aggregate estimates of income and demographics, as well as imputed travel mode and number of people splitting expenses (NPS Zonal). These data effectively transform the NPS survey data into a zonal structure similar to the

mobility data. The aggregated zonal structure of this data enables us to study the consequences of aggregation independent of the other sources of variation in the mobile data. Third, we build a dataset based on mobile device visits by census tract, 5-year ACS estimates of income and demographics, and imputed travel model and number of people splitting expenses (Mobile). We summarize the datasets in Table 2.

We construct the mobile datasets to temporally align with the NPS survey to facilitate the most appropriate comparison of the datasets. We also construct Mobile datasets in each year to investigate changes in recreation demand over time. We use demographics and travel costs associated with 2023, such that the only source of variation is changes in the distribution of travel distances.

### 2.3. Summary Statistics

We present summary statistics in both a table and a dot-whisker plot. Trip and Travel Cost are presented in a table because the variation across NPS sites is not clear in a plot. The demographic variables (Age, Income, Household size) are easier to compare visually. We focus on the comparison between the NPS Zonal and Mobile datasets, as the structure and processing of the data are most similar. Full summary tables are available in an online appendix.

Table A6 compares the average trip count per person across the NPS Zonal and Mobile data. The mean trip counts differ across NPS sites and between data sources. Lake Meredith (LAMR) and Cuyahoga Valley (CUVA) have the highest mean visits in the NPS data, while Great Smoky Mountain (GRSM) and CUVA have the highest mean in the mobile data. The percent difference in visits between the two data sources averages 117%. At nearly all sites, the mobile visits exceed the survey data, suggesting that the mobile data may capture more visitation.

The travel costs show more agreement between the NPS and mobile data sources, with an average percent difference of 57%. Some of the large scenic parks have the highest travel cost in the NPS survey data, including Everglades, Capitol Reef, and Great Sand Dunes (all above \$900). While we similarly find high travel costs for Capitol Reef and Grand Teton in the mobile data, the average travel cost at Everglades is \$162. The lower average travel cost to Everglades suggests that the Mobile data is capturing many more local visitors compared to the NPS survey. The average percent difference between average travel cost is 64%.

Figure A1 compares the demographic characteristics between the survey and mobile data. We compare aggregate measures of age, income, and household size from the Census ACS for both data sources. Differences arise for two reasons. First, the geography differs by data source. The Mobile data use census tracts, whereas the NPS data use ZIP codes. Second, each data source may capture different samples of the population. While there are some differences across sites, the two data sources tend to agree. Age tends to be slightly higher in the Mobile data, whereas Income is slightly higher in the NPS data.

### 3. Empirical Model

We specify a series of travel cost models that facilitate the comparison of recreation values between the NPS survey and mobile device data. At its most granular, the NPS survey data contain individual-level trip counts, demographics, and travel cost inputs. However, the mobility data are reported as the aggregate flows from a census tract to a point of interest (POI). We estimate three types of models corresponding to the datasets described in section 2.2: 1) Negative Binomial TCM using individual-level visits NPS survey data (NPS Individual), 2) Negative

Binomial Zonal TCM using aggregated visits NPS survey data (NPS Zonal), and 3) Truncated and Censored Negative Binomial Zonal TCM using the mobility data (Mobile).

We specify a typical Negative Binomial TCM for individual-level visits (denoted  $i$ ) and aggregate visits (denoted  $z$ ) for each NPS site,  $j$ :

$$\log(\lambda_{ijz}) = \beta_{j0} + \beta_{j1}wtc_{ijz} + \beta_{j2}age_{ijz} + \beta_{j3}income_{ijz} + \beta_{j4}hhsiz_{ijz} + \beta_{j5}pop_{jz} \quad (4)$$

in which  $wtc_{ijz}$  is the weighted travel cost,  $age_{ijz}$  is age of respondent  $i$  or median age of zone  $z$ ,  $income_{ijz}$  is the income of respondent  $i$  or per-capita income of zone  $z$ , and  $hhsiz_{ijz}$  is the household size of respondent  $i$  or median household size of zone  $z$ . In the zonal version of the model we include the population of zone  $z$  to absorb variation in trip counts due to differences in population (Hellerstein 1991). Note that we estimate a zonal model using the aggregated data, so the  $i$  subscript does not apply. The NPS models are estimated via `nbstrat` in Stata to accommodate endogenous stratification introduced by the intercept survey (Hilbe and Martinez-Espineira 2005).

We estimate a zonal travel cost model to accommodate the aggregate mobile device visitation data. However, the mobile data are subject to truncation and censoring. The differential privacy applied by Advan does not report data if fewer than two devices visit the KUW from a single census tract (truncation at one), and reports four visits if between two and four devices visit the KUW from a single census tract (censoring at four). We account for this by estimating a Negative Binomial regression truncated at one and censored at four, with log likelihood,

$$L(\lambda_{zj}, \theta_j | y_{jz}) = \sum_{y_{jz} > 4} \left[ \log f(y_{jz} | \lambda_{jz}, \theta_j) - \log(1 - F(1 | \lambda_{jz}, \theta_j)) \right] \\ + \sum_{y_{jz} = 4} \left[ \log(F(4 | \lambda_{jz}, \theta_j) - F(1 | \lambda_{jz}, \theta_j)) - \log(1 - F(1 | \lambda_{jz}, \theta_j)) \right] \quad (5)$$

in which  $\log(\lambda_{jz})$  is the usual log link function with travel cost and demographics defined in equation 4 at the zone level, and  $y_{jz}$  is the total number of trips from zone  $z$  to site  $j$ .  $f()$  and  $F()$  are the Negative Binomial PDF and CDF, respectively. The term  $\log(1 - F(1 | \lambda_z))$  accommodates truncation by conditioning the contribution of the  $z$ -th observation to the log-likelihood function on the probability that an observation greater than one is observed. The term  $\log(F(4 | \lambda_z) - F(1 | \lambda_z))$  accommodates censoring by accounting for the probability of observing a 2, 3, or 4 when we observe a 4 in the data. Hypothesis testing is based on robust standard errors.

We calculate consumer surplus as  $-1/\beta_{tc}$  in both the survey and mobile negative binomial models since both use log link functions. Confidence intervals are based on the delta method for the survey models and Krinsky-Robb simulation method in the mobile models (Krinsky and Robb 1986).

## 4. Results

Our objective is to compare per-trip consumer surplus estimates across our subset of parks to determine whether mobility data are a reliable substitute for survey data. In addition, we seek to explain why differences exist between the data sources. We estimate three primary travel cost models for each of the 17 sites for which we have sufficient data, the models converged, and the

travel cost coefficients were statistically different from zero. Figure 2 contains the consumer surplus estimates and 95% confidence intervals of the three models comparing different levels of aggregation and information: NPS Individual labels estimates of the model based on individual-level NPS survey data, NPS Zonal labels estimates of the model based on aggregated NPS data, Mobile labels estimates of the model based on the mobile device data (only reported in aggregate form). We include the data underlying Figure 2 in Appendix 8.

<<Fig 2 about here>>

We explore the possibility that the CS estimates based on the survey and mobile data differ because of aggregation. The NPS survey collects individual-level data. We compare CS estimates based on the individual responses (NPS Individual) to those from the artificially aggregated NPS data (NPS Zonal). We find that the average of the zonal CS estimates (\$228) are almost 10% lower than the average of the individual estimates (\$252). Moreover, the Pearson correlation coefficient between the two series of CS estimates is 0.73. The aggregation alone leads to some discrepancy between the CS estimates.

We focus on the comparison between CS estimates from the NPS Zonal model and the Mobile model (Figure 3). The average of the mobile CS estimates (\$214) is about 5% less than the zonal CS estimates (\$228). However, the Pearson correlation coefficient is only 0.41, suggesting discrepancies in individual parks. The estimates for Cape Hatteras National Seashore and Great Sand Dunes National Park are within 5% of each other, whereas Tuskegee Airmen National Historic Site and Lake Meredith National Recreation Area differ by nearly 90%. We also compare the rank of each NPS site within its respective model. We find that the average absolute difference in park rank between the NPS Zonal and Mobile models is 4.1 (median 3). While the

majority of sites have similar rank, Gauley River National Recreation Area is ranked number 1 in the NPS Zonal model and number 15 in the Mobile model.

<<Fig 3 about here>>

#### 4.1. Why CS estimates differ

We investigate why the CS estimates may differ between the NPS Zonal and Mobile data. CS estimates may differ for several reasons. In large NPS sites with multiple entry points, it may be difficult to sample a representative set of visitors via intercept surveys. The mobile device data is only as accurate as the geofences used to construct the visitation measure. Sites near population centers may erroneously capture foot traffic passing nearby. For example, Gateway Arch National Park includes a walking path over Interstate 44 in St Louis. A geofence that includes the interstate may simply sample those driving on the interstate. We make sure to use a POI that excludes the overpass. Lastly, we explore the distribution of visitor travel distances.

Discrepancies in the CS estimates, despite similar travel distances, may indicate that assuming census demographics for visitors is inappropriate.

##### **Least Absolute Shrinkage and Selection Operator (LASSO)**

We use a LASSO regression to estimate the site and sample characteristics associated with the difference between CS estimates from the survey and mobile datasets. LASSO regressions allow us to include many regressors with very few observations (17) because the penalty will shrink coefficients that do not explain variation (or are redundant to other covariates). We estimate two LASSO regressions using different but related dependent variables:  $NPS\ Zonal - Mobile$  and the  $abs(NPS\ Zonal - Mobile)/NPS\ Zonal$ . The set of independent variables includes the following NPS site characteristics: the average visits per person, annual visitation (total annual visits 5-year

average), number of visitor centers, number of campgrounds, large site (binary), high profile site (binary), historic site (binary), Eastern United States Western United States, site total acreage, site total acres of water, percent of local visitors (permanent or seasonal residents of the local area surrounding the park), percent of visitors whose primary trip purpose was visiting the site, difference in mean travel distance between NPS Zonal and Mobile data, difference in mean median income between NPS Zonal and Mobile data, and difference in number of observations between NPS Zonal and Mobile data. The site characteristics data were obtained from various sources, including the NPS visitor surveys, NPS' Integrated Resource Management Applications (IRMA) portal, NPS' Hydrographic and Impairment Statistics (HIS) database, and an internal NPS Application Programming Interface (API). All of the covariates are scaled (z-score), so the magnitudes are comparable. However, we caution against overinterpretation of the coefficients as LASSO willingly trades bias for better prediction.

We find that NPS site type, size, and nature of visitors correlate with differences between the NPS survey CS estimates and those from the mobile data (Table 4). The CS estimates from the mobile data exceed the survey-based CS estimates at historic sites (e.g., Mount Rushmore and Tuskegee Airmen), while the opposite is true at sites with a higher percentage of primary purpose visitors. When we seek to explain the absolute value of the difference as a percent of the survey-based CS estimates, we find that larger, more visible parks correlate with smaller differences, while a larger percentage of local visitors and larger income discrepancies correlate with larger CS differences. Together, these results suggest that the CS estimates are more likely to agree at larger parks when there are more sites where visitors congregate, possibly increasing the intercept rate and measurement at a mobile data Point of Interest. On the other hand, the CS estimates are more likely to disagree when there is a larger share of local visitors, indicating that

the structure of the aggregate mobile data makes it difficult to measure frequent local visitors. The income discrepancy between the data also points to a difference in who is being sampled.

### **Travel Distance**

We explore the extent to which differences in the distribution of travel distances explain the differences in estimates. Differences in the distribution of travel distances suggest that the survey and mobile datasets are capturing different visitors and may arise for several reasons. The surveys are generally administered at busy times of the year, but may not capture visitors who tend to travel at off-peak times. On the other hand, if the point of interest geofence does not align with the site boundary, the mobile data may not reflect the true visitors.

Figure A2 depicts the empirical cumulative distribution function (ECDF) of each data source (Mobile and NPS) at each site. At most sites, the NPS ECDF lies to the right of the Mobile ECDF, suggesting that people traveling further are more represented in the NPS surveys.<sup>ix</sup> If travel costs were purely a function of driving distance, this difference in distributions would lead to higher inverse demand curves and, thus, larger consumer surplus estimates. While this difference in travel distances may explain some of the discrepancies in CS estimates (e.g., Great Smoky Mountains NP), it does not fit in all cases. For instance, the ECDF of Cuyahoga Valley National Park travel distances suggests the NPS sample traveled further; however, the NPS CS estimates are lower than the Mobile estimates. This observation suggests that travel mode and the opportunity cost of time mediate the relationship between travel distance and CS.

## **4.2. Consumer surplus over time**

One of the advantages of the mobility data is that it is continuously collected over time, allowing us to estimate the zonal travel cost model for any window of time. We estimate models for a

subset of NPS sites for the years 2018 – 2022 and present the results in Figure 4 (Cape Hatteras National Seashore (CAHA), Grand Teton National Park (GRTE), Mount Rushmore National Memorial (MORU), and Rocky Mountain National Park (ROMO); plots for all other sites are available in Appendix 8). Each plot shows the daily CS estimate for visits over the course of the year prior to Aug 1 of the date listed. We use the 2023 ACS data for all years, so the only source of variation in a park’s travel cost from one year to another is the travel distance and the opportunity cost of time.

<<Fig 4 about here>>

The results show different trends in daily CS over time across the four sites. The CS estimates at Cape Hatteras fall in the year 2020, but then rebound to their near 2019 levels. However, Grand Teton estimates peak in 2020 at nearly \$330, then fall below \$300 by 2022. Other high-profile sites (i.e., Mount Rushmore and Rocky Mountain National Park) show slight increases peaking in 2020 before plateauing or slightly declining. The variation in CS/trip may reflect the strong increase in US National Park attendance during the COVID-19 pandemic and the subsequent response to congestion.<sup>x</sup> However, the trends we document are not causal evidence of any single explanation.

Following the closure during the COVID-19 lockdowns, ROMO implemented a timed-entry system to limit crowding at the park.<sup>xi</sup> The timed-entry system requires visitors to make advanced reservations to enter the park during peak season. It is not clear how the timed-entry system affects per-trip CS a priori. It may reduce visitation, which could decrease CS; however, it may shift the composition of visitors to those traveling from further away who are more likely to make reservations in advance. We find no evidence that the timed entry system reduced CS

from 2019-2022. Further analysis is required to answer this question and is outside the scope of this paper.

## 5. Discussion and Conclusions

Are mobility data a substitute for travel cost surveys? Our findings suggest that *our mobility data* are not a perfect substitute for travel cost surveys, despite our efforts to consistently process the data and calculate travel costs. While the consumer surplus estimates based on the mobility data capture some heterogeneity across our sample of sites, the mobility estimates do not always align with estimates based on the NPS survey data. Our findings are similar to Tsai et al. (2023), who show that mobility visitation data match well at some NPS sites and not others. We investigate several possible reasons for this discrepancy, including the aggregation of Advan mobility data (along with truncation and censoring), NPS site-specific attributes (i.e., popularity and primary purpose), and sampling and differences in the travel distance distributions. No single factor was found to explain the discrepancies between the survey-based and mobility-based CS estimates.

Implicit in our comparison of results is that the NPS survey data are the “gold standard” that we should measure the mobility data against. However, NPS surveys are administered during a few weeks of the year during peak visitation times. This sampling procedure maximizes sample size given a limited budget but misses visitors who visit outside of that sampling window. If these missed visitors systematically differ by travel distance, income, or travel mode, then extrapolating from a sampling window to the entire year may be inaccurate. We show that mobility data can help supplement survey data by providing information about the periods not sampled. This is likely an area where mobile device data can play a critical role. Exploration of this potential use could be beneficial, given the constraints land managers face in collecting

conventional survey data over multiple seasons and time horizons. Perhaps the survey and mobile data could be used in a hybrid-individual-zonal model as proposed in Loomis et al. (2009).

We find that the Advan mobility is representative of the U.S. population, an important prerequisite of any data purporting to measure visitation. We regress census tract device counts on census population and demographics to assess any systematic biases correlated with observables and find no evidence of large biases. We do find that in census tracts where we observed NPS visitors, device counts were higher where the median age was higher and lower where household size was larger. Higher device counts in tracts with higher median age may reflect the ubiquity of smartphones and overturn the notion that older populations are not represented in mobility data. Lower device counts in areas with larger household sizes are likely a consequence of child privacy laws that prohibit the collection of private data on known underage users.

Literature leveraging mobility data are expanding rapidly, and we build on the area of work focused on understanding outdoor recreation. We contribute to the literature by estimating consumer surplus on a sample of U.S. NPS sites using both conventional survey data and mobility data. We investigate where the estimates align and where they do not, and we explore several possible reasons why. We illustrate how researchers can use survey information to “impute” missing information on the mobile device visitors using survey information.

The data we use in this study are from Advan and are subject to several limitations. First, the data are reported as aggregate visitation metrics by origin census tract as opposed to individual-level “breadcrumb” data that are capable of re-creating time diaries. Consequently, we are unable to study multipurpose trips – an important area of the study that alternative mobile data sources

are capable of quantifying. Multi-purpose trips have always created empirical challenges in the travel cost literature. If a visit to an NPS site is part of a larger “bundle” of experiences, then fully attributing costs accrued to visit our study site is inappropriate. The aggregated Advan data suffer from the same limitations as other secondary data such as backcountry permits or camp reservations; we do not observe the entire good. While the Advan data do contain information on other chain stores visited on the same day as the NPS site, we do not know whether these were incidental visits to a restaurant or destinations. Moreover, we do not observe visits across multiple days. Because we do not account for multipurpose trips in the survey data either, our comparison methods suffer from the same limitations and may overestimate the consumer surplus of visitors whose visit is not the sole purpose of the trip.

Advan applies a differential privacy policy that introduces censoring and truncation. We use a negative binomial model that accommodates this censoring and truncation. However, we do not know the consequences of the differential privacy policy on our results. Wan et al. (this issue) study this very issue.

Another limitation of aggregated mobile device data is that aspects of the data processing pipeline are subject to change. In December of 2022, Advan discontinued the use of NPS official site boundaries as the geofence used to construct visitation metrics and now uses POIs of specific, smaller locations within the park. For instance, Advan now reports visitation on several popular hiking trailheads within Arches National Park, but not the entire park itself. Any changes to the methods of visit attribution affect the fidelity of visit measurement and may create challenges for conducting analyses over time. Advan’s change in geofences effectively cut our visit sample to November 2022, which affected our comparison to NPS data collected in 2023. Recall that we compare the 12 months preceding the survey date. We include an indicator for

this affected set of POIs in our LASSO analysis and find no evidence that it explains differences in the CS estimates. In general, researchers using aggregated mobile device data must constantly track any changes that data providers make and think through how each change affects their analysis.

Another limitation of the Advan mobility data is the lack of income and demographic information on the visitors. We use census tract aggregates as the best estimate of information about visitors from an origin zone. However, studies have shown that visitors to NPS sites may not be accurately described by aggregate demographics. We emulate this restriction by aggregating the NPS survey data to the ZIP code-level and using ACS values for demographics. We find that the per trip NPS Zonal CS estimates are still reasonably correlated with the NPS Individual CS estimates, but they are about 10% less than the individual model estimates. This disparity suggests that estimates from models relying on aggregate data may need calibration, but may be capable of tracking changes over time.

Visitation to U.S. national parks remains near an all-time high, suggesting that people derive significant value from this national resource. The NPS depends on reliable estimates of consumer surplus to inform a wide range of management decisions. Further, such estimates can help meet the goals of new legislation such as the EXPLORE Act, which seeks to improve recreation opportunities on Federal lands and waters, largely through improved data collection.

Collecting conventional survey data on public lands is costly, time-consuming, and requires complex approval processes. We show that mobility data can supplement traditional travel cost survey methods. In particular, mobility data can highlight the limitations of sampling over a few weeks out of the year and can help identify any systematic biases in those who choose

to respond to surveys. However, our findings indicate that the aggregate mobility data from Advan alone, in their current form, are not a reliable substitute for surveys.

R code is available on Github at [https://github.com/jbayham/tc\\_nps/tree/main](https://github.com/jbayham/tc_nps/tree/main) and archived at 10.5281/zenodo.18383266.

## 6. References

- AAA. 2022. Your Driving Costs. AAA. <https://newsroom.aaa.com/wp-content/uploads/2022/08/2022-YDC-Costs-Break-Out-by-Category.pdf>.
- AAA. 2023. Your Driving Costs. AAA. <https://newsroom.aaa.com/wp-content/uploads/2022/08/2022-YDC-Costs-Break-Out-by-Category.pdf>.
- Bureau of Transportation Statistics. 2025a. “Annual Airport Rankings.” <https://www.bts.gov/topics/annual-airport-rankings>.
- Bureau of Transportation Statistics. 2025b. “Aviation Facilities.” <https://geodata.bts.gov/datasets/usdot::aviation-facilities/about>.
- Champ, Patricia A., Kevin J. Boyle, and Thomas C. Brown, eds. 2017. A Primer on Nonmarket Valuation. Vol. 13. The Economics of Non-Market Goods and Resources. Dordrecht: Springer Netherlands.
- English, Eric, Roger H. von Haefen, Joseph Herriges, Christopher Leggett, Frank Lupi, Kenneth McConnell, Michael Welsh, Adam Domanski, and Norman Meade. 2018. “Estimating the Value of Lost Recreation Days from the Deepwater Horizon Oil Spill.” *Journal of Environmental Economics and Management* 91:26–45. doi:10.1016/j.jeem.2018.06.010.
- French, Sally. 2025. “Car Rental Pricing Statistics.” <https://www.nerdwallet.com/article/travel/car-rental-pricing-statistics>.
- Ghermandi, Andrea. 2018. “Integrating Social Media Analysis and Revealed Preference Methods to Value the Recreation Services of Ecologically Engineered Wetlands.” *Ecosystem Services* 31:351–57.
- Giraud, Timothée. 2022. “Osm: Interface Between R and the OpenStreetMap-Based Routing Service OSRM.” *Journal of Open Source Software* 7(78):4574. doi:10.21105/joss.04574.
- Goebel, Russell, Austin Schmaltz, Beth Ann Brackett, Spencer A. Wood, and Kimihiro Noguchi. 2023. “Modeling and Forecasting Percent Changes in National Park Visitation Using Social Media.” *Journal of Forecasting*. doi:10.1002/for.2965.
- Heikinheimo, Vuokko, Enrico Di Minin, Henrikki Tenkanen, Anna Hausmann, Joel Erkkonen, and Tuuli Toivonen. 2017. “User-Generated Geographic Information for Visitor Monitoring in a National Park: A Comparison of Social Media Data and Visitor Survey.” *ISPRS International Journal of Geo-Information* 6(3):85. doi:10.3390/ijgi6030085.

- Hellerstein, Daniel M. 1991. "Using Count Data Models in Travel Cost Analysis with Aggregate Data." *American Journal of Agricultural Economics* 73(3):860–66. doi:10.2307/1242838.
- Hilbe, Joseph, and Roberto Martinez-Espineira. 2005. "NBSTRAT: Stata Module to Estimate Negative Binomial with Endogenous Stratification." *Statistical Software Components*. <https://ideas.repec.org//c/boc/bocode/s456414.html>.
- Keeler, Bonnie L., Spencer A. Wood, Stephen Polasky, Catherine Kling, Christopher T. Filstrup, and John A. Downing. 2015. "Recreational Demand for Clean Water: Evidence from Geotagged Photographs by Visitors to Lakes." *Frontiers in Ecology and the Environment* 13(2):76–81.
- Kilic, Gizem, Curt Cashour, and Matt Carrier. 2024. 2024 STATE OF THE INDUSTRY REPORT. American Hotel and Lodging Association. [https://www.ahla.com/sites/default/files/SOTI.2024.Final\\_Draft\\_v4.pdf#page=3.06](https://www.ahla.com/sites/default/files/SOTI.2024.Final_Draft_v4.pdf#page=3.06).
- Krinsky, Itzhak, and A. Leslie Robb. 1986. "On Approximating the Statistical Properties of Elasticities." *The Review of Economics and Statistics* 68(4):715–19. doi:10.2307/1924536.
- Kubo, Takahiro, Shinya Uryu, Hiroya Yamano, Takahiro Tsuge, Takehisa Yamakita, and Yoshihisa Shirayama. 2020. "Mobile Phone Network Data Reveal Nationwide Economic Value of Coastal Tourism under Climate Change." *Tourism Management* 77:104010. doi:10.1016/j.tourman.2019.104010.
- Li, Zhenlong, Huan Ning, Fengrui Jing, and M. Naser Lessani. 2024. "Understanding the Bias of Mobile Location Data across Spatial Scales and over Time: A Comprehensive Analysis of SafeGraph Data in the United States." *PLOS ONE* 19(1):1–23. doi:10.1371/journal.pone.0294430.
- Loomis, John, Omer Tadjion, Philip Watson, Josh Wilson, Stephen Davies, and Dawn Thilmany. 2009. "A Hybrid Individual—Zonal Travel Cost Model for Estimating the Consumer Surplus of Golfing in Colorado." *Journal of Sports Economics* 10(2):155–67.
- Merrill, Nathaniel H., Samantha G. Winder, Dieta R. Hanson, Spencer A. Wood, and Eric M. White. 2024. "A National Model for Estimating US Public Land Visitation." Working Paper. Working Paper.
- Otak Inc., RRC Associates, and University of Montana. 2023. 2022 Socioeconomic Monitoring of National Park Service Visitors: Report on 2022 Data Collection. Natural Resource Report NPS/NRSS/EQD/NRR—2023/2550. Fort Collins, CO: National Park Service.

RRC Associates. 2024. 2022-2023 Richmond National Battlefield Park and Maggie L. Walker National Historic Site Visitor Movement Study. Final Report. National Park Service.

Scholz, Fritz, and Angie Zhu. 2023. “kSamples: K-Sample Rank Tests and Their Combinations.”

Sessions, Carrie, Spencer A. Wood, Sergey Rabotyagov, and David M. Fisher. 2016. “Measuring Recreational Visitation at U.S. National Parks with Crowd-Sourced Photographs.” *Journal of Environmental Management* 183:703–11. doi:10.1016/j.jenvman.2016.09.018.

Sinclair, Michael, Andrea Ghermandi, and Albert M. Sheela. 2018. “A Crowdsourced Valuation of Recreational Ecosystem Services Using Social Media Data: An Application to a Tropical Wetland in India.” *Science of The Total Environment* 642:356–65. doi:10.1016/j.scitotenv.2018.06.056.

Sinclair, Michael, Marius Mayer, Manuel Woltering, and Andrea Ghermandi. 2020. “Valuing Nature-Based Recreation Using a Crowdsourced Travel Cost Method: A Comparison to Onsite Survey Data and Value Transfer.” *Ecosystem Services* 45:101165.

Tsai, Wei-Lun, Nathaniel H. Merrill, Anne C. Neale, and Madeline Grupper. 2023. “Using Cellular Device Location Data to Estimate Visitation to Public Lands: Comparing Device Location Data to U.S. National Park Service’s Visitor Use Statistics.” *PLOS ONE* 18(11):e0289922. doi:10.1371/journal.pone.0289922.

U.S. Department of Transportation. 2025. “Consumer Airfare Report: Table 6 - Contiguous State City-Pair Markets That Average At Least 10 Passengers Per Day.” [https://data.transportation.gov/Aviation/Consumer-Airfare-Report-Table-6-Contiguous-State-C/yj5y-b2ir/about\\_data](https://data.transportation.gov/Aviation/Consumer-Airfare-Report-Table-6-Contiguous-State-C/yj5y-b2ir/about_data).

Walker, Kyle. 2024. Tigris: Load Census TIGER/Line Shapefiles.

Wilkins, Emily J., Spencer A. Wood, and Jordan W. Smith. 2021. “Uses and Limitations of Social Media to Inform Visitor Use Management in Parks and Protected Areas: A Systematic Review.” *Environmental Management* 67(1):120–32. doi:10.1007/s00267-020-01373-7.

Wood, Spencer A., Anne D. Guerry, Jessica M. Silver, and Martin Lacayo. 2013. “Using Social Media to Quantify Nature-Based Tourism and Recreation.” *Scientific Reports* 3(1):2976.

Wood, Spencer A., Samantha G. Winder, Emilia H. Lia, Eric M. White, Christian S. L. Crowley, and Adam A. Milnor. 2020. “Next-Generation Visitation Models Using

Social Media to Estimate Recreation on Public Lands.” *Scientific Reports*  
10(1):15419. doi:10.1038/s41598-020-70829-x.

## 7. Tables

Table 1. Summary of datasets used in analysis. NPS = National Park Service, ACS = American Community Survey

Dataset	Visits	Income and demographics	Travel Mode and Expense Splitting
NPS Individual	NPS Individual	NPS Data*	NPS Data*
NPS Zonal	NPS Aggregate	ACS	Imputed
Mobile	Mobile (Aggregate)	ACS	Imputed

\*We impute income and demographics, as well as travel mode and expense splitting when NPS respondents do not report.

Table 2. Trip and Travel Cost average by National Park Service (NPS) site in the NPS Zonal and Mobile data. The % difference column compares the absolute difference in means as a percentage of the average of the means. NPS site abbreviations are defined in Table A1.

	Trips (per person per year)			Travel Cost (\$)		
	NPS Zonal	Mobile	% Diff	NPS Zonal	Mobile	% Diff
AZRU	3.41	12.5	114%	715.9	517.1	32%
BADL	1.40	9	146%	791.4	501	45%
CAHA	11.42	31	92%	415.0	268.9	43%
CARE	1.40	6.1	125%	957.9	863.2	10%
CATO	11.56	25	74%	152.2	213.9	34%
CUGA	29.95	83.9	95%	319.4	151.7	71%
CUVA	70.50	207.6	99%	279.7	30.2	161%
DINO	1.74	8.8	134%	754.2	451.4	50%
EVER	2.65	102.6	190%	908.9	162.5	139%
FOLA	1.20	6	133%	734.1	615.3	18%
GARI	4.88	22.2	128%	439.8	90	132%
GRBA	7.06	4.4	46%	296.5	528.4	56%
GRSA	3.91	6.5	50%	980.2	584.2	51%
GRSM	1.60	289.5	198%	704.5	99.1	151%
GRTE	1.43	14.5	164%	747.4	918.9	21%
GUMO	14.69	16.6	12%	324.2	476	38%
ISRO	1.35	11.4	158%	830.3	299.6	94%
JEFF	1.39	7.2	135%	489.0	128.6	117%
LAMR	80.33	54.6	38%	199.5	294	38%
MORU	1.19	15	171%	731.2	548.7	29%
PINN	2.11	5.4	88%	475.3	259.3	59%
ROMO	5.21	19.9	117%	761.9	466.6	48%
TUAI	1.30	29.8	183%	401.4	265.4	41%

## 8. Figures

*Figure 1 Map of National Park Service (NPS) and Mobile sites. NPS site abbreviations are defined in Table A1*

*Figure 2. Consumer surplus (CS) per trip estimates from each model: Mobile uses models trained on National Park Service (NPS) data to predict flight probability and number of people splitting expenses, NPS Zonal uses aggregates trip counts to the ZIP code and census aggregates for controls, NPS Individual uses individual information when reported but imputes missing data.*

*Figure 3. Plot of per trip consumer surplus (CS) estimates from National Park Service (NPS) and mobile device data with 45 degree reference line. Tuskegee Airmen NHS (TUAI) and Gauley River National Recreation Area (GARI) are omitted for clarity. NPS site abbreviations are defined in Table A1.*

*Figure 4. Advan mobility data per-trip CS estimates by year in 2022 dollars for a subset of National Park Service (NPS) sites for the years 2018 - 2022: Cape Hatteras National Seashore (CAHA), Grand Teton National Park (GRTE), Mount Rushmore National Memorial (MORU), and Rocky Mountain National Park (ROMO). Note that the vertical axes scales differ across sites. The point of interest boundaries used to estimate visitation changed for many parks in December 2022.*

---

<sup>i</sup> We use data collected through the NPS visitor surveys to study the strengths and limitations of mobility data in a pure research context. To facilitate fair comparisons between the survey data and mobile data, we focus primarily on travel cost models that intentionally restrict the survey data (e.g., ignore available information). As a result, the welfare estimates reported in this paper are not necessarily based on travel cost model specifications that result in the most accurate, policy-relevant estimates of consumer surplus.

<sup>ii</sup> <https://www.congress.gov/bill/118th-congress/house-bill/6492/text>

<sup>iii</sup> These data were made available by Advan Research (<https://advanresearch.com/>) via the Dewey Data platform. (<https://www.deweydata.io/>)

<sup>iv</sup> <https://docs.deweydata.io/docs/faqs-advan-research#what-is-the-coverage-of-national-parks>

<sup>v</sup> <https://www.census.gov/geographies/reference-files/time-series/geo/relationship-files.2020.html>

<sup>vi</sup> We develop a similar strategy to impute the number of people splitting expenses for survey respondents that do not answer the question and for the mobile data. Model details and results are in Appendix 3.

<sup>vii</sup> We also tested using the Google Maps API and the ESRI Routing Server and found the resulting distance and time estimates to be comparable. Unlike OSRM, those alternatives entail a cost. Moreover, the bulk download and caching of driving distance and time is a violation of the Google Maps API terms of service.

<sup>viii</sup> <https://aviex.goflexair.com/flight-school-training-faq/commercial-plane-speeds>

<sup>ix</sup> We estimate Kolmogorov-Smirnoff and Anderson-Darling tests to test whether the NPS and Mobile samples are from the same distribution. In every case, we reject the null hypothesis that our samples arise from the same unspecified distribution (p-values all less than 0.001). Tests are implemented in R using `ks.test()` in base R and `ad.test()` from `kSamples` (Scholz and Zhu 2023).

---

<sup>x</sup> <https://www.nps.gov/subjects/socialscience/visitor-use-statistics-dashboard.htm>

<sup>xi</sup> Refer to <https://home.nps.gov/romo/pilot-timed-entry-permit-systems.htm> for more information on the timed-entry system.







