Journal of Outdoor Recreation and Tourism xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Outdoor Recreation and Tourism

journal homepage: www.elsevier.com/locate/jort

Research Article

Using social media user profiles to identify visitor demographics and origins in Yellowstone national park

Yun Liang [a], Junjun Yin [b], Soyoung Park [c], Bing Pan [a,*], Guangqing Chi [d], Zachary Miller [e]

[a] Department of Recreation, Park, and Tourism Management, Pennsylvania State University, United States
[b] Social Science Research Institute, Pennsylvania State University, United States
[c] Department of Marketing, College of Business, Florida Atlantic University, United States
[d] Department of Agricultural Economics, Sociology, and Education, Pennsylvania State University, United States
[e] National Park Service, Intermountain Regional Office, U.S. Department of Interior, United States

ARTICLE INFO

Keywords:
Social media
User profiles
Visitor demographics
Representative issue
National park

ABSTRACT

Despite the growing body of studies on mining visitor perceptions and attitudes of national park visitors using social media data, few research investigated user demographics and its representative issues. This study assessed visitor demographics, including gender, age, racial groups, and origins of visitors in a U.S. national park through their Twitter user profiles, and compared the results to a traditional visitor survey. The results showed similar percentages of gender groups between Twitter user profiles and the traditional survey. However, significant differences existed across all age groups and all racial groups between the two data sources. Compared to the survey, the visitors identified from social media data were younger and from more diverse race groups. The lists of the top 10 states and countries of residency of visitors from the two data sources overlapped but had different orders. The findings indicated that social media data could only be a complementary data source due to its representative issues. The results allow researchers to explore social media users' demographics by advanced social data analytics. However, this study suggests that analyzing Twitter profile information, such as self-reported names and profile photos, requires special attention from researchers even if the data were publicly available. The authors recommend that future research should attend to the representative and private issues of social media data.

Management implications:

- Social media user profiles can be utilized for predicting users' demographics, such as gender, age, and racial groups.
- Social media data can only be a conplementary data source to understand visitor demographics in future research.
- The ethical issues of social media data, including private domain and machine learning algorithms, need further discussion.

1. Introduction

According to the National Park ServiceOffice of Communications (2020), National Park System in the U.S. has experienced significantly increase in visitation over the last decade and attracted 327.5 million visitations in 2019 alone. The growth in visitor volumes was especially prominent in Arches, Zion, Glacier, and Yellowstone National Parks. However, Americans do not have equal access to national parks. According to Resource Systems Group and Wyoming Survey and Analysis Center (2019), 71.4% of the people who visited national parks in the last two years were non-Hispanic Whites, whereas the percentage was 11.8% for Hispanics and 6.0% for Blacks. Similarly, Xiao et al. (2022) indicated that Hispanics and Blacks were less likely to visit national parks than non-Hispanic Whites among the study participants. Therefore,

* Corresponding author.
E-mail addresses: yjl5451@psu.edu (Y. Liang), jyin@psu.edu (J. Yin), soyoungpark@fau.edu (S. Park), bingpan@psu.edu (B. Pan), gfc5047@psu.edu (G. Chi), zach_miller@nps.gov (Z. Miller).

https://doi.org/10.1016/j.jort.2023.100620
Received 31 December 2021; Received in revised form 7 February 2023; Accepted 15 February 2023

information about visitor demographics is critical for park managers to assess the inequality in Americans' access to U.S. national parks (Byrne et al., 2009; Park et al., 2021; Tarrant & Cordell, 1999).

Traditional approaches to collecting visitor demographics and origins in U.S. national parks rely on survey-based visitor use studies (Ednie et al., 2020; Rice et al., 2020). For example, Yellowstone National Park (YNP) conducted two studies in the summers of 2016 and 2018 to understand visitor demographic (e.g., gender, age, race, educational level, etc.) and country/state of residence (National Park Service, 2019a; National Park Service, 2019b). However, as a government agency, National Park Service (NPS) faces limited and potentially depleting labor and financial resources to conduct large-scale visitor use studies. Furthermore, the survey approach has several known limitations, such as a short data collection window, slow data collection process, and limited spatial coverage when intercepting visitors (Cessford & Muhar, 2003; Di Minin et al., 2015; Hadwen et al., 2007). The response rates of minority groups could also be impacted by administrative bias and language barriers (Gstaettner et al., 2020). In addition, conducting visitor surveys is challenging during the COVID-19 pandemic due to the requirements of social distancing and safety concerns for in-person interactions.

Several studies have explored the use of social media data as an alternative data source to collect information of visitors in national parks or protected areas worldwide, which is an easy and low-cost access to massive amounts of information of a large group of individuals (Di Minin et al., 2015; Mangachena & Pickering, 2021; Pickering et al., 2020; Teles da Mota & Pickering, 2020; Tenkanen et al., 2017). Previous studies in the field of park and tourism have utilized social media data for visitation counts, spatial distribution, origins, and experiences of visitors in a park context in Korea, South Africa, Finland (Chun et al., 2020; Mangachena & Pickering, 2021; Sinclair et al., 2020; Teles da Mota & Pickering, 2020; Tenkanen et al., 2017; Wilkins et al., 2021). For example, Pickering et al. (2020) extracted textual materials from a social media platform to assess how tourists view and value the highest mountain in Australia.

However, few studies have specifically examined or assessed user demographics from social media platforms (Park, 2020). If social media users are not representative of the general population, any findings solely relied on social media data will be biased and skewed. Therefore, it is necessary to explore the similarities and differences in the visitor demographics between the survey data and social media data. This study aims to address this research gap by estimating visitor demographics and origins using Twitter user profiles and compare them to a traditional visitor use survey in a U.S. national park context.

2. Literature review

2.1. NPS visitation research

U.S. National Park Service (NPS) preserves natural and cultural resources and provides recreational and educational opportunities for the current and future generations (National Park Service, 2016). Although the visitation to national parks in 2016 has increased by 156% compared to the visitation of 50 years ago (National Park Service, 2021), the demographic compositions of park visitors have not changed significantly and do not reflect the diversity of the American population (Krymkowski et al., 2014; Weber & Sultana, 2013).

Many studies explored national park visitation patterns (Byrne et al., 2009; Floyd, 1999; Xiao et al., 2022). Literature showed consistent findings that the dominant visitors in national parks are non-Hispanic Whites. Racial and ethnic minority groups are less likely to visit national parks (Byrne et al., 2009; Xiao et al., 2022), even in many different study areas (e.g., national parks, national recreation areas, etc.) and samples. Age is another factor to influence national park visitation. For example, younger participants (18–24 years old) were more likely to be non-visitors (Xiao et al., 2022).

Pettebone and Meldrum (2018) advocated that it is essential to have a comprehensive socioeconomic research program for NPS to fully understand visitor characteristics and support decision-making of park managers. Currently, traditional approaches to estimating visitors' demographic composition rely on two data sources: 1) visitor use studies from individual park units that were designed for specific management questions and park units. The majority visitor use studies were conducted by the most famous national parks, such as Yellowstone (National Park Service, 2019a; National Park Service, 2019b) and Yosemite (National Park Service, 2021). Newer and smaller national park units are limited by financial sources and staff to afford this type of study. 2) the Comprehensive Survey of American Public (CSAP) was conducted in 2000, 2008, and 2018. Although CSAP studies provide a broad generalization about visitors and non-visitors to NPS, they lack detailed information about visitation to individual parks or the types of national park units.

2.2. Social media in park and tourism research

Social media data have been applied in park and tourism research (Sinclair et al., 2020; Teles da Mota & Pickering, 2020; Toivonen et al., 2019; Wilkins et al., 2021). According to Teles da Mota and Pickering (2020) and Wilkins et al. (2021), the most popular social media platform in park-related research is Flickr; it is an image-sharing platform which provides an Application Programming Interface (API) to obtain user-generated content (UGC) and contains much nature-related content. Additionally, data from Panoramio, Instagram, Twitter, and Weibo were widely utilized in the park field (Teles da Mota & Pickering, 2020; Wilkins et al., 2021). Different types of data, such as geotags, timestamps, text, and videos, can be retrieved from social media platforms. Social media users' profiles and social networks could also be retrieved for further analysis.

Past studies utilized social media data to explore spatial and temporal distributions of visitors (Sinclair et al., 2020; Tenkanen et al., 2017; Zhang et al., 2021), park popularity (Tenkanen et al., 2017), visitors' preferences (Hausmann et al., 2018), visitors' unwanted behaviors (Liang et al., 2020), and cultural ecosystem service (Cardoso et al., 2022). Additionally, Tenkanen et al. (2017) and Hausmann et al. (2018) validated social media data with traditional survey data and official count statistics regarding temporal patterns of visitors, visitor demographics, and visitor preferences. Furthermore, Tenkanen et al. (2017) and Hausmann et al. (2018) compared data from different social media platforms.

Although social media generates rich information in terms of textual and visual content, the data quality is still a challenge (Toivonen et al., 2019). First, social media users vary among different population groups and geographic regions. Although social media platforms are popular among all age groups, young adults (18–29 years old) are more likely to use social media and share their experiences online (Pew Research Center, 2021; Toivonen et al., 2019). Additionally, females are more likely to use social media platforms (Pew Research Center, 2019). From a geographical perspective, certain social media platforms are extremely popular in specific countries. In western countries, Facebook, Instagram, and Twitter are the most popular, while in China, Weibo is more commonly used (Toivonen et al., 2019). Secondly, computational acquisition approaches could lead to uncertainty in retrieving data (Brooker et al., 2016). Official APIs provided by social media platforms can only access limited datasets and metadata for researchers (Joseph et al., 2014; Toivonen et al., 2019), and little is known about platform APIs' sampling algorithms (Joseph et al., 2014).

2.2.1. Twitter in park and tourism research

On one of the most popular social media platforms, Twitter users can post tweets (280-character messages), and share and comment on other users' tweets. Researchers can mine different types of information, such as geolocations, timestamps, short textual materials, and user profiles,

from the platform. Facebook and Instagram are the other two most common social media platforms in U.S. (Pew Research Center, 2021). However, Facebook has extremely restrictive data access policies and the Instagram API only allows users to retrieve recent posts. Although only 23% of U.S. adults ever use Twitter (Pew Research Center, 2021), Twitter provides a most publicly-accessible data API that allows users to collect user profiles and the tweet data streams.

According to Pew Research Center (2019), the age group of 18–29 years old, accounting for 29% of total Twitter users, is the primary age group. Female and male users, account for 50% of total Twitter users separately. White users account for 60% of total Twitter users, followed by Hispanic (17%) and African American users (11%).

Tourism research have adopted Twitter data to investigate tourist behaviors, perceptions, and attitudes towards destinations (Chua et al., 2016; Lu & Zheng, 2021; Nadeau et al., 2021; Park et al., 2016, 2020). Park (2020) utilized the descriptions from Twitter user profiles to identify demographics (age and gender) and origins of message audiences for anti-orphanage tourism campaigns. However, in the park-related research, limited studies selected Twitter as the data source since its function is regarded as a communication platform rather than an outdoor recreation experience sharing space (Pinckney et al., 2018; Toivonen et al., 2019). Studies using Twitter data have focused on assessing spatial and temporal visitation patterns (Hamstead et al., 2018; Tenkanen et al., 2017). Fisher et al. (2019) identified preferred visitor attractions using geotagged Twitter data. Although various types of data from Twitter have been applied in park and tourism research, few studies have investigated the underlying Twitter user demographics as compared to actual visitor population.

2.3. Identify user demographics by social media profiles

Social media user profiles have been utilized for obtaining their demographics, such as gender, age, and race, which can be detected by supervised or unsupervised machine learning methods (Cesare, Grant, Nguyen, Lee, & Nsoesie, 2018; McCormick et al., 2017; Yin et al., 2018; Park, 2020).

A review by Cesare et al. (2018) suggested that user profiles, user posts, or the combination of the two can be utilized for identifying social media users' genders (Liu & Ruths, 2013; Yin et al., 2018). For example, Zagheni et al. (2014) and An and Weber (2016) utilized profile photos to investigate the genders of social media users. Yin et al. (2018) employed self-reported profile names and the images of Twitter users to predict their genders. Although users are not required to provide an authentic first and last name in their profiles, according to Mislove et al. (2011), 64.2% of Twitter users reported at least their first names in the profiles. Cesare et al. (2018) found that gender is the easiest characteristic to accurately predict, and the average accuracy of identifying users' gender was 83%.

Researchers employed profile descriptions and photos to predict social media users' age, age category, and life stage (An & Weber, 2016; Jung et al., 2017; McCormick et al., 2017). In addition, when combining content from users' posts, age prediction could reach a higher accuracy (76%). The predicted ages of social media users are heavily skewed toward the bracket between 30 - 40 years old, which are consistent with the primary population of social media users (Cesare et al., 2018).

User posts, profiles, and network metadata were utilized to infer users' race/ethnicity (Bergsma et al., 2013; Jung et al., 2017; Oktay et al., 2014). Similar to age prediction, previous studies indicated that the textual content could provide valuable insights for estimating the race/ethnicity of users. Based on a review by Cesare et al. (2018), the average accuracy of race prediction was 82%.

Although, in the computer science field, many studies identified users' demographics by textual and visual content from social media posts, rare studies have employed these advanced techniques in the park and tourism field. For example, Park (2020) utilized Latent Dirichlet Allocation (LDA), an unsupervised machine learning approach, to identify users' age and gender by profile descriptions.

In summary, previous research relies on traditional survey approaches to identify national park visitor demographics and origins. However, conducting visitor surveys in national parks is limited by financial resources, staff resources, and specific surveying time periods and areas. Additionally, during the COVID-19 pandemic, visitors are less likely to interact with surveyors because of safety concerns. Therefore, employing social media data as a new data collection approach to estimating visitor demographics will benefit park managers. Social media has been utilized in tourism and park research for understanding spatial and temporal distributions of visitors, park popularity, and visitor preferences (Hausmann et al., 2018; Tenkanen et al., 2017; Zhang et al., 2021). However, little is known about national park visitor demographics based on social media. Therefore, to fill the research gap, this study seeks to assess visitor demographics and origins through Twitter user profiles in a U.S. national park context and aims to uncover the representative issues of social media data.

3. Methodology

3.1. Data

This study selected Yellowstone National Park (YNP) as the study context due to its notable increasing visitation and data availability.

3.1.1. Yellowstone National Park Visitor Use Survey 2016

A visitor use study conducted in 2016 was utilized to establish the ground truth regarding visitor demographics and origins (National Park Service, 2019b). The survey was conducted by Resource Systems Group from August 4th through August 14th, 2016, at Yellowstone National Park (YNP, National Park Service, 2019b). The aims of this survey are to collect information about summer visitors,[1] trip motivations, park experiences, etc. in the park. The visitor use study was distributed at YNP as mail-back survey in two languages: English and Mandarin.

Five YNP entrances, including North entrance, Northeast entrance, East entrance, South entrance, and West entrance, were selected as the survey sample locations since the majority of visitors must pass through one of the entrances to access to the park. Visitor groups were intercepted by a timed-interval approach. The detailed sampling efforts can be found in the study report (National Park Service, 2019b).

Intercepted visitor groups were introduced to the study purposes and asked to participate. If a visitor group agreed, they were asked which adult member (older than 18 years old) within the group had the next birthday; the individual who had the next birthday was asked to fill out the questionnaire for the group.

The survey method had three phases: 1) distributing questionnaires on-site; 2) mailing reminder postcards; 3) mailing a replaced questionnaire for those participants who had not yet returned a completed questionnaire. During the sampling period, 2,265 visitor groups were intercepted to ask for participating in the study and 2,030 groups agreed by accepting a mail-back survey packet. Finally, 1,257 visitor groups completed and returned questionnaires. Therefore, the overall response rate was 55%.

3.1.2. Twitter data

Twitter was selected as the study platform due to its popularity and accessility to its user profiles. To match the study population (i.e., summer visitors) in the visitor use survey, we used the geotagged tweets that were continuously collected between June 1st to August 31st, 2016. The data collection utilized the Twitter Streaming API (https://developer.twitter.com) by setting up a geographical boundary of YNP and retrieving all the geolocated tweets that fell within. Since Twitter Streaming API randomly samples around 1% of public tweets in real-

[1] The summer season defined by YNP is from June to August.

time (Twitter Developer Platform, n.d.) and approximately 0.85% of tweets are geotagged (Sloan et al., 2013), suggesting that Twitter Streaming API will return all tweets inside YNP.

Twitter accounts for organizations or bloggers were filtered out through manual-checking. The filtered dataset included 3,847 unique tweets and was generated from 1,226 Twitter users. Although the number of Twitter profiles in the time period is low, the data collection period matches the period of survey data collection. Additionally, the count data of Yellowstone by National Park Service indicate that 2016 summer visitors account for over 60% of total visitors. Therefore, we assume that visitors captured by Twitter data can account for most visitors who used Twitter.

Beyond textual data, the dataset included user IDs, timestamps (when users posted their tweets), and specific geolocations in the form of latitude/longitude coordinates. Furthermore, self-reported names and profile photos of accessible Twitter users were collected.

3.2. Data analyses

3.2.1. Visitor demographics

This study estimated gender, age, and race from Twitter user profiles (Yin et al., 2018). The first name and the profile image of a user was utilized for determining the user's gender. First, the user's first name was matched with its occurrence in a first-name database, where each first name has a related probability of being a female or male (Longley & Adnan, 2016; Longley et al., 2015; Luo et al., 2016). The first name database is a collection of 23,363 first names generated from Facebook profile pages (Tang et al., 2011). A second database contains nicknames and related probabilities associated gender identification. If a Twitter user's first name appears in the first name database, the gender probability was calculated based on the fraction of occurrences labeled as male or female. If there was no match in the first name database, we continued the search on the nickname database. If the gender probability was less than 51%, no gender was assigned to the user. Specifically, if the gender probability is between 51% and 75%, gender was further determined by comparing it to the gender estimation from the user's image profile. Note that if no valid profile image exists, the gender identified by the first name was given priority. If the gender probability was more than 75%, gender was assigned to the Twitter user directly.

In the situation that Twitter users' first names did not have matches in the first name database, facial recognition techniques were utilized on Twitter users' profile images to complement the estimation of gender, which were also used for age estimation. Instead of using the Microsoft Azure facial recognition service to assess the gender and age information (Yin et al., 2018), an open-source pre-trained model based on convolutional neural networks by Uchida (2019) was utilized. If the two approaches have an agreement, the user's gender information is retained. If they disagree, we used the results from the facial recognition instead of the gender estimated by first names (with a probability value less than 0.5). If two or more persons appeared in the image, we compared the gender information to that from the first name-based estimate; we retained the gender estimated from the first name-based approach as long as there is one person in the image with the same gender. It is worth noting that convolutional neural networks can provide reasonably accurate gender and age estimation (Dehghan et al., 2017), much more accurate than age estimation using other contemporary methods, such as using first names (Luo et al., 2016). However, due to the "black-box" nature of convolutional neural networks and the size of the training dataset, it is challenging to validate the accuracy of the gender and age estimation with absolute certainty.

Last names can be used as an indicator of users' race/ethnicity to some extent (Adjaye-Gbewonyo et al., 2014). U.S. Census Bureau provides a surname database, which is a collection of 162,255 surnames with self-reported race/ethnicity based on the 2010 census. Four race/ethnicity groups: non-Hispanic White, non-Hispanic Black, Asian, and American Indian and Alaskan Native, were assessed in this study. We matched the derived surnames in the name database and retrieved related probabilities for the four race/ethnicity groups, similar to the approach used for estimating a user's gender based on first names.

3.2.2. Visitor origins

Potential approaches to detecting the origins of Twitter user, in the existing literature, can be mainly categorized into two types, both replying on collecting the historical tweets from individual Twitter users over a longer period: (1) Twitter users' home locations can be estimated by performing spatial clustering methods on the collection of geolocations of the historical tweets, where the most frequently tweeted clusters are considered as Twitter users' home units, such as county, city, and state (Belcastro et al., 2021; Cao et al., 2020). (2) The home locations can also be inferred machine learning approaches based on Twitter users' network, the tweet content, and tweet context (Ajao et al., 2015; Flatow et al., 2015; Kotzias et al., 2016). In addition, a variety of text mining techniques were applied to improve the accuracy in the predicted Twitter home locations, such as identifying the spatial word usage in tweets (Chang et al., 2012).

In this paper, we have primarily relied on self-ported locations in Twitter users' profiles as their home locations. Each user's origin country was identified through a two-step process. First, from the users' Twitter profiles, we gathered their self-reported home locations. These locations lack consistency in formats as some may report specific cities and countries (e.g., Minnesota, USA), whereas some may write non-geographic locations (e.g., "*I'm everywhere man*"). Therefore, we used Google Maps API and a R package, "ggmap" (Kahle & Wickham, 2013), to extract longitudes and latitudes of geographically valid locations. Second, the longitudes and latitudes were reverse-geocoded to identify in which country they are located. For the reverse-geocoding, we used R packages "sp" (Pebesma & Bivand, 2005) and "rworldmap" (South, 2011). We used the same approach to assessed the U.S. states in which the visitors resided in.

3.2.3. Comparison of social media data with the survey data

Chi-square tests were conducted to compare the visitor demographics, including gender, age, and race, derived from Twitter user profiles and the visitor use survey. Phi (φ) was calculated for measuring the effect size of each Chi-square test (small if $\varphi = 0.10$, medium if $\varphi = 0.30$, and large if $\varphi = 0.50$).

4. Results

4.1. Visitor demographics

The genders of 897 Twitter users were identified through first names and profile photos (Table 1). About 47% of users were female and 53% were male. In the survey data, there were 3,893 visitors, of whom 50% were female and 50% were male. The Chi-square statistic is 1.82 and the p-value is 0.18, indicating that there is no statistically significant difference between the two data sources regarding vistor gender.

We identified age groups from 731 unique Twitter users (Table 1). Chi-square tests indicated statistically significant differences in all age groups between Twitter user profiles and the survey. Specifically, two age groups, 20–34 years old and 35–54 years old, estimated by Twitter user profiles are over-represented compared to those in the survey. In contrast, three age groups, under 20 years old, 55–64 years old, and older than 65 years old, are under-represented in Twitter user profiles. The value of Phi of 20–34 years old is 0.44, indicating a medium effect size of Chi-square test. The values of Phi of under 20 years old, 55–64 years old, and older than 65 years old are over 0.15, indicating relatively small effect sizes. The Phi value of the age group of 35–54 years old is only 0.06, meaning the result of Chi-square has a very small effect size.

About 650 unique users were examined by the probability of racial groups (Table 1). Significant differences exist regarding three racial groups, White, Black, and Asian, between Twitter user profiles and the

Table 1
Frequency and percentage of visitors by gender, age, and race.

Demographic	Twitter		Survey		Chi Square Statistics	p-value	Phi (φ)
	Frequency	Percentage	Frequency	Percentage			
Gender							
Female	426	47%	1946	50%	1.82	0.18	
Male	471	53%	1947	50%			
Total	897	100%	3893	100%			
Age Group							
Under 20	6	<1%	780	26%	160.69	<0.0001	0.19
20 to 34	435	60%	468	12%	885.15	<0.0001	0.44
35 to 54	285	39.%	1209	31%	17.97	<0.0001	0.06
55 to 64	4	<1%	624	16%	125.41	<0.0001	0.16
65+	1	<1%	585	15%	123.05	<0.0001	0.16
Total	731	100%	3900	100%			
Race							
White		71%		82%	43.35	<0.0001	0.09
Black		11%		<1%	231.70	<0.0001	0.23
Asian		8%		15%	22.64	<0.0001	0.07
American Indian or Alaska		0.7%		2%	4.75	0.029	0.03
Native Hawaiian or Pacific		NA		1%	NA	NA	
Total	650	100%	3888	100%			

survey. However, the proportions of White (71%) and Asian (8%) by Twitter user profiles are significantly lower than the racial groups from the survey (82% and 15%). In comparison, the proportion of Black visitors on Twitter (13%) is higher than the proportions revealed by the survey (<1%). About 0.7% of Twitter users are identified as American Indian or Alaska compared to 2% of the total visitors by the survey and there is no significant difference in the percentage. No Twitter users were identified as Native Hawaiian/Pacific.

The value of Phi for Black group is 0.23, indicating a small effect size of Chi-square test, while the values of Phi of White and Asian groups are less than 0.1, suggesting very small effect sizes of Chi-square tests.

4.2. Visitor origins

This study identified the origins of country/states of 833 unique Twitter users. U.S. visitors came from 47 states (no identified visitors were from Alaska, Hawaii, and Vermont) and the District of Columbia and comprised 79% of total visitation, while the U.S. visitors in the survey came from 50 states and the District of Columbia and account for 83% of total visitor population in the survey.

Table 2 presents the top 10 states of residence of domestic visitors. Both data sources identified California as the state with most domestic visitors. In addition, both data sources included Texas, Colorado, Utah, New York, and Minnesota in the top 10 list. However, in the top 10 states, only Twitter data included Kansas, Florida, Illinois, and Ohio, while only the survey included Washington, Connecticut, and Wyoming.

Table 2
Top 10 states of residency of domestic visitors.

Twitter			Survey		
Country	Percent of U.S. visitors (N = 662)	Percent of all visitors (N = 833)	Country	Percent of U.S. visitors (N = 2,891)	Percent of all visitors (N = 3,483)
California	14%	11%	California	8%	7%
Texas	7%	6%	Utah	6%	5%
Colorado	7%	6%	Texas	5%	4%
Kansas	5%	4%	Washington	5%	4%
Utah	4%	3%	Minnesota	5%	4%
New York	4%	3%	Colorado	5%	4%
Florida	4%	3%	Connecticut	4%	3%
Illinois	3%	3%	Montana	4%	3%
Minnesota	3%	2%	New York	4%	3%
Ohio	3%	2%	Wyoming	4%	3%

Table 3 reports the top 10 countries of residency of international visitors in a descending order. Both data sources included U.K., Canada, France, Italy, The Netherlands, and Germany in the top 10 countries of residence, while only Brazil, Mexico, Thailand, and Belgium were in the top10 countries identified by Twitter data and only the visitor survey includes China, Spain, Australia, and Switzerland in the top 10 list.

5. Discussions

This study utilized machine learning-based techniques to identify visitor demographics (gender, age, and race) and origins by Twitter user profiles and compared them with a traditional visitor survey. The results suggested that, not surprisingly, there were statistically significant differences between the two data sources regarding age groups and racial groups.

5.1. Similarities and differences between twitter data and survey data

5.1.1. Gender

Twitter profile data indicated that 53% of the YNP visitors during the study period were male and showed no statistically significant difference from the result reported in the survey. However, it is worth noting that Mislove et al. (2011) and Alowibdi et al. (2013) indicated that the population of male Twitter users had a higher proportion by detecting gender using first names and facial recognition techniques.

5.1.2. Age

Statistically significant differences existed in all age groups between Twitter user profiles and the survey. The results from Twitter revealed that the group of young (20–34 years old) and middle-aged adults (35–54 years old) had a statistically significant higher proportion than the proportion of the same age groups in the survey, which is not surprising and is consistent with statistics from Pew Research Center (2009) in that 55% of Twitter users are 18–49 years-old and young adults are more likely to share their experiences online. The results of age groups in this study are consistent with Yin et al. (2018) that Twitter users at age 20–34 years were over-represented. Another reason to explain the low percentage of older users is that predicting older adults is more challenging (Morgan-Lopez et al., 2017).

The survey has higher proportions in two age groups, 55–64 years old and older than 65 years old, than the proportions of the same age

Table 3
Top 10 countries of residency of international visitors.

Twitter			Survey		
Country	Percent of international visitors (N = 126)	Percent of all visitors (N = 833)	Country	Percent of international visitors (N = 594)	Percent of all visitors (N = 3,483)
Canada	15%	2%	China	34%	6%
U.K.	15%	2%	Italy	11%	2%
France	7%	1%	Canada	10%	2%
Italy	6%	1%	France	8%	1%
Mexico	5%	1%	The Netherlands	7%	1%
Brazil	4%	1%	Germany	7%	1%
Germany	4%	1%	U.K.	5%	<1%
Thailand	4%	1%	Spain	4%	<1%
Belgium	3%	<1%	Australia	3%	<1%
The Nertherlands	3%	<1%	Switzerland	2%	<1%

groups revealed by Twitter. This finding confirmed that older age groups less likely present online sharing behaviors and the result is also consistent with Pew Research Center (2019).

5.1.3. Racial groups

Twitter user profiles revealed a significant higher proportion of Black visitors compared to the results of the survey. The possible explanation is that Twitter is functioned as an important communication platform for the Black population (Pinckney et al., 2018). Yin et al. (2018) and Pew Research Center (2019) indicated that over 20% of Twitter users are Black. Furthermore, the proportion of Asian visitors in Twitter user profiles is lower than the proportion of Asian visitors in the survey. Chinses accounts for a large portion of Asian visitors. However, Chinese visitors may not have a Twitter account since Twitter platform is blocked in China. In addition, the survey languages with English and Mandrian can be a plausible reason to yield high response rate from Chinese visitor groups.

5.1.4. Visitor origins

In terms of visitor origins, although the two data sources included similar countries and states of the top 10 residencies, the orders were slightly different. Additionally, the survey showed that Chinese visitors accounted for more than 30% of international visitors (National Park Service, 2019b), while Twitter user profiles indicated only 0.7% of total international visitors as Chinese, because China has a restriction on Twitter usage and most Chinese visitors do not have Twitter accounts. Therefore, this study confirmed the issue of geographical bias in social media user population.

According to the results, the demographics of social media users showed significant differences from the ones collected by the traditional survey, suggesting that related studies have to consider such biases.

5.2. Contributions

Methodologically, identifying visitor demographics from Twitter profiles is a significant contribution of this study. This research also filled the gap that few studies examined social media users' demographics for national park research. Additionally, the approaches employed in this study could be adapted to other social media platforms.

Practically, this study provides an opportunity to understand the representativeness of Twitter users for studying visitor demographics compared to a traditional survey approach. Many previous studies explored visitor behaviors and attitudes by textual and visual materials from social media platforms without considering the representative issues of social media users. The findings from this study suggested that we should carefully draw findings and conclusions when using social media data as they are not generalized and may be biased towards certain user groups.

5.3. Ethical issues

All Twitter user information, including their profile information, twitter handles, and tweets are public data per Twitter user agreement. Twitter users can set their accounts to be private, which means that their tweets are only visible to their followers and cannot be downloaded by the Twitter Streaming API. Twitter does impose restrictions on the usage the download Twitter data. Specifically, the owner of the downloaded Twitter data is not allowed to post those tweets to Twitter, modifying the profiles, or adding content to the tweets, without getting express and informed consent from the users. Given the sensitive nature of the geographic location information in tweets, the owner of the data cannot share detailed location history of individuals, such as the trajectory of the tweeted locations because it could infringe Twitter users' privacy. Any violation to the Twitter Developer Agreement are not putting Twitter users' privacy at risk, it can also bear legal consequence from both the Twitter company and individual Twitter users (https://developer.twitter.com/en/developer-terms/policy).

However, prediction of user demographics raises ethical issues since Twitter user profiles, including reported first/last names, profile photos, and geolocations, were tightly connected to users' private information. When social media content is public, it blurs the boundaries between the public and private domains. The issues related to the definition of privacy in social media have been debated in the literature (Roberts, 2015). A general agreement is that if social media data are publicly accessible, it is potentially ethical for research use, while it is potentially unethical if social media accounts are restricted by users for privacy protections and without user consent (Woodfield & Iphofen, 2017, pp. 1–12).

In this study, Twitter Streaming API only allows the authors to collect Twitter user profiles that are publicly accessible, while some researchers indicated that private versus public data should be discussed from the perspective of the type of data (Williams et al., 2017). Images and videos generated by users on social media platforms are generally considered to be more sensitive and more private than textual data, even though these images and videos are publicly available, since these user-generated content might contain identifiable information about individuals. In this study, self-reported real names and profile photos have been employed to identify Twitter users' gender, age groups, and racial groups. However, to avoid privacy concerns, the data were only used at an aggregated level.

From a technical perspective, the algorithms to identify Twitter users' demographics could also raise ethical issues. For example, Leslie (2020) indicated that algorithms trained by datasets with demographic biases have resulted in algorithmic discrimination and do not work well for some segments of the population. Although there are no documented demographic biases about the algorithms employed in this study, the concerns remain valid.

5.4. Limitations

The social media-based analyses have several limitations. First, we only employed one social media platform to infer visitor demographics and origins. Collecting and comparing such information from multiple social media platforms may help reduce platform bias (Tenkanen et al., 2017).

Secondly, the accuracy of the machine learning approaches can affect the demographics captured by Twitter user profiles. Based on a review by Cesare et al. (2018), the average accuracies of predicting gender, age, and race are all over 70%. Although we did not examine the accuracy of the predicted gender, age, and race, the approaches employed in this study could lead to the differences in visitor demographic compositions from the traditional survey.

Thirdly, the populations captured by the survey and the Twitter data could be different. The target population of the visitor survey focused on summer visitors, while the population captured by Twitter data was visitors who posted their experiences on Twitter during the summer. Another issue is that the data collection period of the survey was from August 4th through August 14th, 2016, and did not consider visitors in June and July, although the population of the survey study is summer visitors in YNP. Therefore, the differences in visitor demographic compositions could be caused by the target populations and the data collection periods of the two data sources.

Fourthly, the survey languages include English and Mandrian. The survey specifically targeted Asian population with mandarin as survey lanauge, which may yield higher response rate from Chinese visitor groups.

5.5. Future research

This section highlights several directions of future research. First, to fully understand the representativeness issue of Twitter, questions about social media platforms usage during traveling (e.g., *which social media platform do you use for sharing your national park experience? How often do you use Twitter (or other social media platforms) during national park traveling?*) could be asked in visitor use studies. This approach will help researchers assess the differences between the demographics captured by surveys and the demographics of visitors who really use social media during national park visitation. Additionally, future research can estimate visitor demographics by multiple social media platforms and fully understand the representativeness of various social media platforms.

Secondly, collaborations between park researchers and computer science experts could be established to develop more accurate algorithms to predict visitor demographics. Although there were significant differences between Twitter user profiles and the traditional survey regarding visitor demographics, the approaches in this study provide a promising opportunity for newer and smaller national parks (limited by financial resources and staff) and an efficient way during the COVID-19 pandemic to estimate visitor demographics. Therefore, more accurate techniques will benefit park managers in visitor management.

Thirdly, social media platforms allow researchers to understand visitor demographics at longer periods. In this study, the authors utilized Yellowstone National Park Visitor Use Study 2016, which only collected demographic information from summer visitors. Rare studies conducted a visitor use survey at an entire year or at different seasons. Therefore, social media user profiles help researchers and park managers understand visitor characteristics that is not limited by traditional data collection periods. The consistent monitoring of visitor demographics can help park managers establish strategies to attract more diverse visitor groups and foster diversity, equity, and inclusion in national parks (Byrne et al., 2009; Park et al., 2021).

Fourthly, the approaches to identify user demographics can be combined with other social media analytics (e.g., sentiment analysis, topic modeling) to investigate visitor behaviors and attitudes towards their experiences by visitor groups. The combinations of approaches to identify user demographics and textual and visual analyses can help park managers meet the preferences and needs of various visitor groups, which can contribute to achieve the mission of NPS and alleviate the unbalanced visitor population.

Lastly, researchers should pay much attention to ethical issues related to social media data and machine learning algorithms in the park and tourism field.

6. Conclusions

This study sought to assess the similarities and differences in visitor demographic and origins between Twitter user profiles and a traditional visitor survey. The results indicated that there were significant differences in age groups and racial groups between the two data sources. Currently, social media data are unable to replace the traditional survey approach and can only be a complementary source to understand visitor demographics and origins. This study fills the methodological gap by employing advanced social media analytics to predict visitor demographics and helps park researchers understand the representative issue of social media platforms. Finally, further attention should be paid to ethical issues raised by social media data.

Grant

The Center for Social Data Analytics (C-SoDA) and the Institute for Computational and Data Sciences (ICDS) in Penn State University provided seed funding for this project. This work is also supported in part by the National Science Foundation (Award # SES-1823633), the USDA National Institute of Food and Agriculture and Multistate Research Project #PEN04623 (Accession #1013257), and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (Award #P2C HD041025).

CRediT authorship contribution statement

Yun Liang: Conceptualization, Formal analysis, Writing – original draft. **Junjun Yin:** Conceptualization, Methodology, Formal analysis, Writing – review & editing. **Soyoung Park:** Methodology, Formal analysis. **Bing Pan:** Conceptualization, Methodology, Investigation, Writing – review & editing. **Guangqing Chi:** Conceptualization, Writing – review & editing. **Zachary Miller:** Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The views expressed in this paper are the responsibility of the authors and do not necessarily represent the official opinions or policy of the National Park Service.

References

Adjaye-Gbewonyo, D., Bednarczyk, R. A., Davis, R. L., & Omer, S. B. (2014). Using the bayesian improved surname geocoding method (BISG) to create a working classification of race and ethnicity in a diverse managed care population: A validation study. *Health Services Research, 49*(1), 268–283. https://doi.org/10.1111/1475-6773.12089

Ajao, O., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science, 41*(6), 855–864.

Alowibdi, J. S., Buy, U. A., & Yu, P. (2013). Empirical evaluation of profile characteristics for gender classification on twitter. In *2013 12th international conference on machine learning and applications* (Vol. 1, pp. 365–369). https://doi.org/10.1109/ICMLA.2013.74

An, J., & Weber, I. (2016). #greysanatomy vs. #yankees: Demographics and hashtag use on twitter. In *Proceedings of the international AAAI conference on web and social media* (Vol. 10, pp. 523–526), 1.

Belcastro, L., Marozzo, F., & Perrella, E. (2021). Automatic detection of user trajectories from social media posts. *Expert Systems with Applications, 186*, Article 115733.

Bergsma, S., Dredze, M., Van Durme, B., Wilson, T., & Yarowsky, D. (2013). Broadly improving user classification via communication-based name and location clustering on twitter. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 1010–1019). https://aclanthology.org/N13-1121.

Brooker, P., Barnett, J., Cribbin, T., & Sharma, S. (2016). Have we even solved the first 'big data challenge?' Practical issues concerning data collection and visual representation for social media analytics. In H. Snee, C. Hine, Y. Morey, S. Roberts, & H. Watson (Eds.), *Digital methods for social science*. London: Palgrave Macmillan. https://doi.org/10.1057/9781137453662_3.

Byrne, J., Wolch, J., & Zhang, J. (2009). Planning for environmental justice in an urban national park. *Journal of Environmental Planning and Management, 52*(3), 365–392.

Cao, Y., Stewart, K., Factor, J., Billing, A., Massey, E., Artigiani, E., Wagner, M., Dezman, Z., & Wish, E. (2020). Using socially-sensed data to infer ZIP level characteristics for the spatiotemporal analysis of drug-related health problems in Maryland. *Health & Place, 63*, Article 102345.

Cardoso, A. S., Renna, F., Moreno-Llorca, R., Alcaraz-Segura, D., Tabik, S., Ladle, R. J., & Vaz, A. S. (2022). Classifying the content of social media images to support cultural ecosystem service assessments using deep learning models. *Ecosystem Services, 54*, Article 101410. https://doi.org/10.1016/j.ecoser.2022.101410

Cesare, N., Grant, C., Nguyen, Q., Lee, H., & Nsoesie, E. O. (2018). *How well can machine learning predict demographics of social media users? ArXiv:1702.01807 [Cs]*. http://arxiv.org/abs/1702.01807.

Cessford, G., & Muhar, A. (2003). Monitoring options for visitor numbers in national parks and natural areas. *Journal of Natural Conservation, 11*, 240–250.

Chang, H. W., Lee, D., Eltaher, M., & Lee, J. (2012). @ phillies tweeting from philly? Predicting twitter user locations with spatial word usage. In *2012 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 111–118). IEEE.

Chua, A., Servillo, L., Marcheggiani, E., & Moere, A. V. (2016). Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tourism Management, 57*, 295–310. https://doi.org/10.1016/j.tourman.2016.06.013

Chun, J., Kim, C.-K., Kim, G. S., Jeong, J., & Lee, W.-K. (2020). Social big data informs spatially explicit management options for national parks with high tourism pressures. *Tourism Management, 81*, Article 104136. https://doi.org/10.1016/j.tourman.2020.104136

Dehghan, A., Ortiz, E. G., Shu, G., & Masood, S. Z. (2017). *Dager: Deep age, gender and emotion recognition using convolutional neural network*. https://arxiv.org/abs/1702.04280v2.

Di Minin, E., Tenkanen, H., & Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science, 3*. https://doi.org/10.3389/fenvs.2015.00063

Ednie, A., Gale, T., Beeftink, K., & Adiego, A. (2020). Connecting protected area visitor experiences, wellness motivations, and soundscape perceptions in Chilean Patagonia. *Journal of Leisure Research*, 1–27. https://doi.org/10.1080/00222216.2020.1814177

Fisher, D. M., Wood, S. A., Roh, Y.-H., & Kim, C.-K. (2019). The geographic spread and preferences of tourists revealed by user-generated information on Jeju Island, South Korea. *Land, 8*(5), 73. https://doi.org/10.3390/land8050073

Flatow, D., Naaman, M., Xie, K. E., Volkovich, Y., & Kanza, Y. (2015). On the accuracy of hyper-local geotagging of social media content. In *Proceedings of the eighth ACM international conference on web search and data mining* (pp. 127–136).

Floyd, D. (1999). Race, ethnicity and use of the national park system. *Social Sciences Research Review, 1*(2), 1–24.

Gstaettner, A. M., Lee, D., & Weiler, B. (2020). Responsibility and preparedness for risk in national parks: Results of a visitor survey. *Tourism Recreation Research, 45*(4), 485–499. https://doi.org/10.1080/02508281.2020.1745474

Hadwen, W. L., Hill, W., & Pickering, C. M. (2007). Icons under threat: Why monitoring visitors and their ecological impacts in protected areas matters. *Ecological Management and Restoration, 8*(3), 177–181. https://doi.org/10.1111/j.1442-8903.2007.00364.x

Hamstead, Z. A., Fisher, D., Ilieva, R. T., Wood, S. A., McPhearson, T., & Kremer, P. (2018). Geolocated social media as a rapid indicator of park visitation and equitable park access. *Computers, Environment and Urban Systems, 72*, 38–50. https://doi.org/10.1016/j.compenvurbsys.2018.01.007

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., & Minin, E. D. (2018). Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conservation Letters, 11*(1), Article e12343. https://doi.org/10.1111/conl.12343

Joseph, K., Landwehr, P. M., & Carley, K. M. (2014). Two 1%s don't make a whole: Comparing simultaneous samples from twitter's streaming API. In W. G. Kennedy, N. Agarwal, & S. J. Yang (Eds.), *Social computing, behavioral-cultural modeling and prediction* (pp. 75–83). Springer International Publishing. https://doi.org/10.1007/978-3-319-05579-4_10.

Jung, S.-G., An, J., Kwak, H., Salminen, J., & Jansen, B. (2017). Inferring social media users' demographics from profile pictures: A face analysis on twitter users. In *ICEB 2017 proceedings (dubai, UAE)*. https://aisel.aisnet.org/iceb2017/22.

Kahle, D., & Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal, 5*(1), 144. https://doi.org/10.32614/RJ-2013-014

Kotzias, D., Lappas, T., & Gunopulos, D. (2016). Home is where your friends are: Utilizing the social graph to locate twitter users in a city. *Information Systems, 57*, 77–87.

Krymkowski, D. H., Manning, R., & Valliere, W. A. (2014). Race, ethnicity, and visitation to national parks in the United States: Tests of the marginality, discrimination, and subculture hypotheses with national-level survey data. *Journal of Outdoor Recreation and Tourism, 7*, 35–43.

Leslie, D. (2020). *Understanding bias in facial recognition technologies*. https://doi.org/10.5281/zenodo.4050457.

Liang, Y., Kirilenko, A. P., Stepchenkova, S. O., Ma, S., & David. (2020). Using social media to discover unwanted behaviours displayed by visitors to nature parks: Comparisons of nationally and privately owned parks in the Greater Kruger National Park, South Africa. *Tourism Recreation Research, 45*(2), 271–276. https://doi.org/10.1080/02508281.2019.1681720

Liu, W., & Ruths, D. (2013). What's in a name? Using first names as features for gender inference in twitter. In *2013 AAAI spring symposium series*.

Longley, P. A., & Adnan, M. (2016). Geo-temporal twitter demographics. *International Journal of Geographical Information Science, 30*(2), 369–389.

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of twitter usage. *Environment and Planning A: Economy and Space, 47*(2), 465–484.

Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via twitter: A case study of chicago. *Applied Geography, 70*, 11–25.

Lu, Y., & Zheng, Q. (2021). Twitter public sentiment dynamics on cruise tourism during the COVID-19 pandemic. *Current Issues in Tourism, 24*(7), 892–898. https://doi.org/10.1080/13683500.2020.1843607

Mangachena, J. R., & Pickering, C. M. (2021). Implications of social media discourse for managing national parks in South Africa. *Journal of Environmental Management, 285*, Article 112159. https://doi.org/10.1016/j.jenvman.2021.112159

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for demographic and social science research: Tools for data collection and processing. *Sociological Methods & Research, 46*(3), 390–421. https://doi.org/10.1177/0049124115605339

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., & Rosenquist, J. (2011). Understanding the demographics of twitter users. *Proceedings of the International AAAI Conference on Web and Social Media, 5*(1), 554–557.

Nadeau, J., Wardley, L. J., & Rajabi, E. (2021). Tourism destination image resiliency during a pandemic as portrayed through emotions on Twitter. *Tourism and Hospitality Research, 14673584211038316*. https://doi.org/10.1177/14673584211038317

National Park Service. (2016). *About the national park service*. https://www.nps.gov/aboutus/aboutus.htm#:~:text=The%20mission%20of%20the%20National,of%20this%20and%20future%20generations.

National Park Service. (2019a). *Yellowstone national park summer 2018 visitor use surveys*. https://www.nps.gov/yell/learn/management/upload/2018-Yellowstone-Visitor-Use-Surveys-FINAL-REPORT_WEB-RESOLUTION.pdf.

National Park Service. (2019b). *Yellowstone national park visitor use study summer*, 2016 https://www.nps.gov/yell/getinvolved/upload/R-YELL_VUS_FINAL-Report.pdf.

National Park Service Office of Communications. (2020). *National park visitation tops 327 million in 2019*. https://www.nps.gov/orgs/1207/2019-visitation-numbers.htm#:~:text=WASHINGTON%20%E2%80%93%20America's%20national%20parks%20continue,record%20keeping%20began%20in%201904.

Oktay, H., Firat, A., & Ertem, Z. (2014). Demographic breakdown of twitter users: An analysis based on names. In *Academy of science and engineering (ASE)*.

Park, S. (2020). Building A foundation for online public communication campaigns against orphanage tourism. *The Pennsylvania State University*.

Park, S. B., Kim, J., Lee, Y. K., & Ok, C. M. (2020). Visualizing theme park visitors' emotions using social media analytics and geospatial analytics. *Tourism Management, 80*, Article 104127. https://doi.org/10.1016/j.tourman.2020.104127

Park, S., Ok, C., & Chae, B. (2016). Using Twitter data for cruise tourism marketing and research. *Journal of Travel & Tourism Marketing, 33*(6), 885–898. https://doi.org/10.1080/10548408.2015.1071688

Park, K., Rigolon, A., Choi, D., Lyons, T., & Brewer, S. (2021). Transit to parks: An environmental justice study of transit access to large parks in the U.S. West. *Urban Forestry and Urban Greening, 60*, Article 127055. https://doi.org/10.1016/j.ufug.2021.127055

Pebesma, E., & Bivand, R. S. (2005). Classes and methods for spatial data: The sp Package. *R News, 5*(2), 9–13.

Pettebone, D., & Meldrum, B. (2018). The need for a comprehensive socioeconomic research program for the National Park Service. *George Wright Forum, 35*(1), 22–31.

Pew Research Center. (2019). *Sizing up twitter users*. Retrieved February 02, 2023, from https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

Pew Research Center. (2021). *The behaviors and attitudes of U.S. Adults on twitter*. Retrieved September 22, 2022, from https://www.pewresearch.org/internet/2021/11/15/1-the-views-and-experiences-of-u-s-adult-twitter-users/.

Pickering, C., Walden-Schreiner, C., Barros, A., & Rossi, S. D. (2020). Using social media images and text to examine how tourists view and value the highest mountain in Australia. *Journal of Outdoor Recreation and Tourism, 29*, Article 100252. https://doi.org/10.1016/j.jort.2019.100252

Pinckney, H. P., Mowatt, R. A., Outley, C., Brown, A., Floyd, M. F., & Black, K. L. (2018). Black spaces/white spaces: Black lives, leisure, and life politics. *Leisure Sciences, 40*(4), 267–287. https://doi.org/10.1080/01490400.2018.1454361

Rice, W. L., Taff, B. D., Miller, Z. D., Newman, P., Zipp, K. Y., Pan, B., Newton, J. N., & D'Antonio, A. (2020). Connecting motivations to outcomes: A study of park visitors' outcome attainment. *Journal of Outdoor Recreation and Tourism, 29*, Article 100272. https://doi.org/10.1016/j.jort.2019.100272

Roberts, L. (2015). Ethical issues in conducting qualitative research in online communities. *Qualitative Research in Psychology, 12*(3), 314–325.

Sinclair, M., Mayer, M., Woltering, M., & Ghermandi, A. (2020). Using social media to estimate visitor provenance and patterns of recreation in Germany's national parks. *Journal of Environmental Management, 263*, Article 110418. https://doi.org/10.1016/j.jenvman.2020.110418

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological Research Online, 18*(3), 74–84.

South, A. (2011). rworldmap: A new R package for mapping global data. *The R Journal, 3*(1), 35–43.

Tang, C., Ross, K., Saxena, N., & Chen, R. (2011). What's in a name: A study of names, gender inference, and gender behavior in Facebook. In J. Xu, G. Yu, S. Zhou, & R. Unland (Eds.), *Database Systems for adanced applications* (pp. 344–356). Springer. https://doi.org/10.1007/978-3-642-20244-5_33.

Tarrant, M. A., & Cordell, H. K. (1999). Environmental justice and the spatial distribution of outdoor recreation sites: An application of geographic information Systems. *Journal of Leisure Research, 31*(1), 18–34. https://doi.org/10.1080/00222216.1999.11949849

Teles da Mota, V., & Pickering, C. (2020). Using social media to assess nature-based tourism: Current research and future trends. *Journal of Outdoor Recreation and Tourism, 30*, Article 100295. https://doi.org/10.1016/j.jort.2020.100295

Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., & Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports, 7*(1), Article 17615. https://doi.org/10.1038/s41598-017-18007-4

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H., & Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation, 233*, 298–315. https://doi.org/10.1016/j.biocon.2019.01.023

Twitter Developer Platform. Volume streams (n.d.), Retrieved September 22, 2022, from https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/introduction.

Uchida, Y. (2019). *Keras implementation of a CNN network for age and gender estimation.* https://github.com/yu4u/age-gender-estimation.

Weber, J., & Sultana, S. (2013). Why do so few minority people visit national parks? Visitation and the accessibility of "America's best idea. *Annals of the Association of American Geographers, 103*(3), 437–464.

Wilkins, E. J., Wood, S. A., & Smith, J. W. (2021). Uses and limitations of social media to inform visitor use management in parks and protected areas: A systematic review. *Environmental Management, 67*(1), 120–132. https://doi.org/10.1007/s00267-020-01373-7

Williams, M. L., Burnap, P., Sloan, L., Jessop, C., & Lepps, H. (2017). Users' views of ethics in social media research: Informed consent, anonymity, and harm. In K. Woodfield (Ed.), *The ethics of online research* (Vol. 2, pp. 27–52). Emerald Publishing Limited. https://doi.org/10.1108/S2398-601820180000002002.

Woodfield, K., & Iphofen, R. (2017). *The ethics of online research*. Introduction to (Vol. 2). Bingley: Emerald Publishing Limited.

Xiao, X., Lee, K. J., & Larson, L. R. (2022). Who visits U.S. National parks (and who doesn't)? A national study of perceived constraints and vacation preferences across diverse populations. *Journal of Leisure Research*, 1–22. https://doi.org/10.1080/00222216.2021.1899776

Yin, J., Chi, G., & Van Hook, J. (2018). Evaluating the representativeness in the geographic distribution of twitter user population. In *Proceedings of the 12th workshop on geographic information retrieval* (pp. 1–2). https://doi.org/10.1145/3281354.3281360

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). Inferring international and internal migration patterns from Twitter data. In *Proceedings of the 23rd international conference on world wide web* (pp. 439–444). https://doi.org/10.1145/2567948.2576930

Zhang, H., van Berkel, D., Howe, P. D., Miller, Z. D., & Smith, J. W. (2021). Using social media to measure and map visitation to public lands in Utah. *Applied Geography, 128*, Article 102389. https://doi.org/10.1016/j.apgeog.2021.102389