Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

No Walk in the Park: The Viability and Fairness of Social Media Analysis for Parks and Recreation Policy Making

Afra Mashhadi[1], **Samantha G. Winder**[2], **Emilia H. Lia**[2], **Spencer A. Wood**[2]
[1] Computing and Software System, Bothell
[2] EarthLab, Seattle
University of Washington
Washington, USA
mashhadi@uw.edu, sgwinder@uw.edu, emmilia@uw.edu, spwood@uw.edu

Abstract

Recent years have seen an increase in the use of social media for various decision-making purposes in the context of urban computing and smart cities, including management of public parks. However, as such decision-making tasks are becoming more autonomous, a critical concern that arises is the extent to which such analysis are fair and inclusive. In this article, we examine the biases that exist in social media analysis pipelines that focus on researching recreational visits to urban parks. More precisely, we demonstrate the potential biases that exist in different data sources for estimating the number and demographics of visitors through a comparison of image content shared on Instagram and Flickr from 10 urban parks in Seattle, Washington. We draw a comparison against a traditional intercept survey of park visitors and a multi-modal city-wide survey of residents. We evaluate the viability of using more complex AI facial recognition algorithms and its capabilities for removing some of the presented biases. We evaluate the AI algorithm through the lens of algorithmic fairness and its impact on sensitive demographic groups. We show that despite the promising results, there are new sets of concerns regarding equity that arise when we use AI algorithms.

Introduction

Most cities globally oversee a large number of parks and green spaces, often covering hundreds or thousands of acres, with large annual budgets (typically $4M for parks and recreation agencies in US cities[1]). In return, well-managed parks provide innumerable benefits to the health and well-being of urban residents – particularly as a location for leisure activities, social activities, and relaxation – making them a critical municipal asset. City planners and managers who oversee urban parks and green spaces rely on information about the amount and character of park use for decisions regarding the maintenance of existing lands, and to plan for new parks or infrastructure that will serve urban residents. Since data are often lacking, practitioners are turning to user-generated content (UGC) as a source of data on urban parks and park visitors (Ilieva and McPhearson 2018). One early study into the potential for UGC in parks (Wood et al. 2013), found that the images posted on social media platforms and their geo-tagged locations are useful for estimating both the number and home locations of visitors, and thus their socio-economic backgrounds. More recent studies have employed similar techniques to map park visitor distributions, behaviors, and preferences (Donahue et al. 2018; Hamstead et al. 2018) based on the locations and content of images shared during their recreational visits (Richards and Friess 2015; Lee et al. 2019).

In the domain of urban computing (Zheng et al. 2014; Silva et al. 2019), location-based social media has been shown as a useful means of identifying and understanding semantic areas of cities (Cranshaw et al. 2012; Noulas et al. 2011; Zhang and Zhou 2018), providing valuable insights into people's opinions to support the well-being status of urban communities (De Choudhury, Sharma, and Kiciman 2016; Venerandi et al. 2015; Hecht and Stephens 2014). These results, demonstrating the utility of user generated content, have caught the attention of practitioners and policy-makers who aim to make cities more livable and equitable. However, there is an understudied risk that underlying biases in how these data are generated or analysed could lead decision-makers to unknowingly implement inequitable policies.

In order to successfully use this type of data for planning and management that promotes sustainability and equity of resources, we first need to know who produces these data and what portions of the population are underrepresented. Different social media platforms are known to attract different demographic users. Pinterest[2], for example, have a larger young female base than male. Instagram appeals more to urban, whereas Twitter accounts tend to belong to young, male and urban residents. The majority of Flickr users are male with a median age of 39 (Quercia, Aiello, and Schifanella 2018). Similarly, nearly all platforms impose a minimum user age, yet children are an important user-group for park managers. Together these platform-specific differences and biases towards specific demographics of users make it unlikely that social media are an accurate

[1] https://www.nrpa.org/publications-research/research-papers/agency-performance-review/

[2] As of August 2017, 58.9% of Pinterest's users in the United States were female.

portrayal of all park visitors. Although some statistical approaches exist to account for issues caused by platform popularity, we simply do not sufficiently understand the many biases that are necessary to perform such corrections. Using multiple data sources has been proposed as a way to help overcome these biases (Hausmann et al. 2018), but there is only preliminary evidence that it can work in practice (e.g., for estimating visitation (Wood et al. 2020)). In addition to data and population biases, AI algorithms could also reinforce the societal biases and create further discrimination and representation harms (Fuchs 2018; O'neil 2016; Noble 2018). Bolukbasi et al.(Bolukbasi et al. 2016), for instance, has shown that the popular word embedding space, Word2Vec, encodes societal gender biases. Other works have presented the societal bias of under-representation in identification of gender (Zhao et al. 2017) and race (Klare et al. 2012) in different settings.

In this paper, we study the viability of social media analysis as a means to map and measure demographics of visitors of urban parks in Seattle, Washington. We present a large-scale study of Social Media data, accompanied with intercept and multi-modal city-wide resident surveys. We leverage thousands of geo-tagged images that are publicly shared on two popular platforms – Flickr and Instagram – and estimate the number and demographic composition of the visitors by using a state-of-the-art face recognition algorithm (Face++[3]). In particular we examine platform and algorithmic biases through a comparative analysis of intercept and multi-modal surveys of Seattle park visitors. To our knowledge, there are no examples of studies assessing visitor demographics or visitation counts from image content.

Our results show that recreational metrics for estimating the visitation count is highly impacted by the popularity of the social media platform. We posit that the advances in AI provide opportunities to estimate the number of visitors by analysing the content of the photos publicly shared on social media platforms. However, our results of algorithmic fairness demonstrate that such techniques must be used with extreme caution as the AI consistently under-counts the number of visitors and heavily impacts a specific target group (children). By applying fairness criteria (Dwork et al. 2012), we demonstrate that the detection error impacts sensitive groups such as kids and non-white visitors differently than the non-sensitive groups (adult and white visitors).

The results of our study have implications for practitioners and the research community. For practitioners we offer guidelines regarding what performance in terms of utility and inclusion can be offered by the algorithm for different types of applications that require different sensitivity in regards with the visitation count. For the research community our study suggests an increasing need for techniques to collect self-identified demographic information from people. As a first step we call for a truly "in-the-wild" image dataset containing representative photographs of contributed and curated by people instead of the the web, so that it does not inherit the biases that are associated with the popularity of content on the web.

[3]www.faceplusplus.com

Background

Recreational Studies

A number of recent studies have proposed that crowdsourced and ubiquitous data from social media can complement existing knowledge of park visitor distributions, behaviors, and preferences (Ilieva and McPhearson 2018; Ghermandi and Sinclair 2019). Studies spanning an impressive diversity of urban parks and protected areas have concluded that the popularity of parks is generally mirrored in the popularity of the same destinations on Flickr (Wood et al. 2013; Sessions et al. 2016; Levin, Lechner, and Brown 2017). The majority of studies quantify popularity by calculating total numbers of unique visitors per day who post geo-located Flickr photographs from a particular park over a given time period – termed "photo-user-days" (PUD) by Wood et al. (2013). A limited number of recent studies have expanded the methods to content from other platforms such as Twitter and Instagram (Tenkanen et al. 2017; Donahue et al. 2018; Hamstead et al. 2018). The consensus emerging from these studies is that social media data have the potential to inform estimates of absolute visitation at specific destinations and for multiple time periods.

The content (words and images) and other metadata associated with social media (such as the user's profile) may provide further opportunities for understanding park visitor's experiences, activities, and satisfaction during their recreational visits (Richards and Friess 2015; Lee et al. 2019), as well as the characteristics of the visitors themselves. In particular, home locations of visitors can be inferred by looking at users' public profiles (Wood et al. 2013) and the locations of all other content that the user has shared publicly on social media platforms (Martinez-Harms et al. 2018). In this way, the locations of photographs that are shared online can be a reliable source of data on the home locations of visitors across thousands of destinations (Keeler et al. 2015). These home location data are the most common existing method for inferring park visitor demographics in order to inform questions of equity. Works in other domains have also explored large sensor networks to estimate space density (Chen et al. 2018).

Urban Computing Studies

In the domain of urban computing, several studies have explored location-based social networks (LSBN) data. For example the LBSN data can be explored to help us better understand our perceived physical limits in urban environments, as well as to better understand city dynamics. Cranshaw et al.2012 presented a model to identify different regions of a city that reflect current patterns of collective activities. By doing so, they introduce new boundaries for neighborhoods. The main idea is to uncover the nature of local urban areas, which tend to be dynamic, considering the social proximity (obtained from the distribution of users who check-in) and the spatial proximity (obtained from geographical coordinates) of locations. Noulas et al. 2011 introduced a method to classify users and areas of a city exploring the types (categories) of places used by Foursquare. The method could be explored to discover communities of users

visiting similar type of places. This is useful for comparing urban areas within and between cities or in recommendation systems. Long et al. 2012 explored a dataset collected from Foursquare to introduce an approach based on a topic model to study the intrinsic relations among the different venues in an urban area. Considering a sequence of users' check-ins, they assume that the venues that appear together in several sequences will likely represent geographic topics, for example, indicating coffee shops people typically visit before going to a mall. Similarly, Frias-Martinez et al. 2012 explored a Twitter dataset and presented a technique that, by studying tweeting patterns, identifies the types of activities that are most common in a city. Their results suggest that geolocated tweets could be an essential data source to describe dynamic urban areas, which tend to be costly using other conventional approaches. For a full survey of literature on LSBN in urban computing we refer the readers to (Silva et al. 2019). This vast body of literature is a testimony to the importance of user-generated content in advancing our understanding of cities.

Demographic Research in Social Media

A growing body of literature deals with detecting demographic characteristics from social media data (as reviewed in (Cesare, Grant, and Nsoesie 2017)). Most of the current techniques rely on supervised learning approaches to detect race or ethnicity, and can be grouped into two categories: those that rely on features of a user's profile, and those that use the content of user's posts (Chen et al. 2015; Pennacchiotti and Popescu 2011; Ardehaly and Culotta 2014; Mislove et al. 2011; Alowibdi, Buy, and Yu 2013).

Techniques that detect demographics from users' profiles make use of profile images or text that users upload to describe themselves or where they live. Profile images appear to be a particularly good indicator of race or ethnicity. Pennacchiotti et al. (2011) for instance, obtained a higher precision (0.878) using profile photos to evaluate race, compared to a gradient boosted decision tree classifier that incorporated a combination of lexical features from users posts, and user activity measures (0.629). User profile descriptions also improve methods for predicting race and ethnicity. For example, Chen et al. (2015) observed that adding user descriptions into classifiers consistently improved accuracy, precision and recall for n-gram and name based models. Other research has found users' home locations to improve predictability of surnames or as a feature for calibrating supervised learning models (Mislove et al. 2011; Ardehaly and Culotta 2014). However, these methods are known to be confused by nicknames or arbitrary usernames, and the language used in profile descriptions is likely to follow a formal template, making it hard to detect age and the language skill of the writer. Furthermore such techniques limit results to the demographic information of those who are members of the social media platform. This has the potential to systematically omit certain groups, such as children, the elderly, or particular demographic groups who do not use a specific platform.

Methods

The goal of this research is to explore the novel question of whether geo-located images shared publicly on social media can offer an accurate portrait of urban park visitors and their demographics. To answer this question we investigate the prevalence of two types of biases that potentially arise in social media analyses: i) those that originate from how the data are generated and the underlying social media platform and ii) those that are caused by the algorithms that are used to analyze these data. In order to understand these two types of biases, we employ several data sources in a comparative study that focuses on two metrics that are commonly used by recreation planners as the basis for management decisions policies: visitation rates and visitor demographics.

Site Selection

This study examines ten city parks in Seattle, WA. The focal parks were selected using a stratified random sampling scheme, and represent a broad range of park types, neighborhoods, and user-groups across the city. We stratified using the Social Vulnerability Index (SVI) developed by the Centers for Disease Control and Prevention. This index combines 15 US census variables grouped into four themes (socioeconomic status, household composition, race/ethnicity/language, and housing/transportation) in order to rank census tracts by their relative vulnerability to hazardous events. We assigned a score ranging between 1 (low SVI) to 5 (high SVI) to each park according to the SVI of the surrounding census tract. We then dropped any parks with fewer than three average annual social media posts. This filtering excludes many small neighborhood areas that are common in the city. We then randomly selected two parks from each SVI category for inclusion. The selected parks are located across the city, and range from regional to pocket parks.

Data Sources

In this study we use data from two social media platforms (Flickr and Instagram) in order to estimate visitation rates and visitor demographics to the ten Seattle parks. We selected Flickr as it is traditionally and extensively used in recreational studies and Instagram as it is currently a very popular image sharing platform. We then compare our demographic results to two distinct traditional surveys, a visitor-intercept survey we conducted at these ten parks, and a larger scale multi-modal survey of Seattle residents conducted across the city.

Flickr: We queried the Flickr API for all geo-located photographs that were taken within the bounds of the ten study parks from January 2016–January 2019 (Table 1). These photographs contain metadata including a unique user identifier, date the photo was taken, and the latitude/longitude location. The location typically comes from a GPS receiver in the camera, but may also be manually assigned by the user by zooming and clicking on a web-map as the image is uploaded to Flickr.

Instagram: We used Instagram's *graphql* API to collect every image that was shared publicly and assigned to an In-

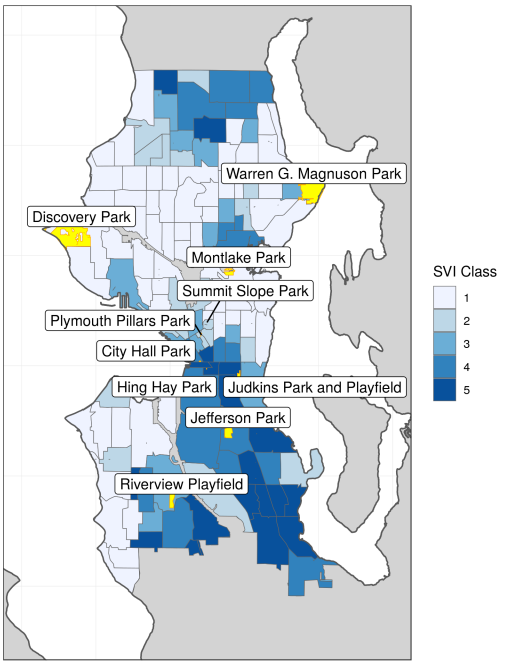


Figure 1: Social Vulnerability Index classes by census tract in Seattle (WA). The yellow regions depict the selected parks.

stagram location within one of the selected Seattle parks (Table 1). Since Instagram no longer provides an API endpoint for querying locations, we first manually searched for locations in the Instagram web interface, using the park name and major features as search terms. The number of Instagram locations that were available for any given park were correlated with park area, presumably because bigger parks have more locations for users to tag. We collected images that were uploaded from January 2016 to January 2019. Metadata available for each image include the Instagram users' identifier (referred to as short-code) and the date that the image was shared. Additionally, we collected images which were uploaded to Instagram between April-June 2019 from the selected sites in order to draw comparisons with our intercept survey data (below). We did not compare these data with Flickr because we did not have corresponding data from Flickr for these dates.

Intercept Survey: Between April and June, 2019, we conducted exit interviews at the study sites (Mashhadi 2019). Every park was surveyed on two weekdays and one weekend day, once in the morning, afternoon, and evening for approximately 4 hours each. During this time, the surveyor intercepted visitors at five randomly selected exits, and asked them to complete a written survey in English. Responses were voluntary and no compensation was provided. The survey contained questions about visitors' activities, demographics, experiences, and feelings about parks in their neighborhoods. We collected 165 surveys in total, and the number of respondents ranged from 7 to 39 surveys per park. Across all surveys, the overall response rate was 16% of exiting visitors. Approximately 41% of park users who were approached agreed to take the survey. The low response rate highlights the problem of gathering visitation data through traditional survey methodology and supports the quest of policy-makers for a more autonomous way of data collection.

Park Name	Instagram	Flickr
City Hall	168	133
Discovery	41994	7154
Hing Hay	2030	501
Jefferson	2737	559
Judkins	2453	164
W. Magnuson	44235	15771
Montlake	1022	152
Plymouth Pillars	1007	117
Riverview	341	119
Summit Slopes	364	250

Table 1: The number of collected Instagram and Flickr images per study site from January 2016–January 2019.

Multi-modal Survey: We draw comparisons with results of a multi-modal web and phone survey of Seattle residents that was conducted using address-based sampling during Spring 2019 (May-June 2019). The survey was commissioned by Seattle Parks and Recreation and was administered in English and Spanish capturing a total of 830 participants. The survey was then weighted such that it is representative of the demographic population of the residents in each given census tracts. The survey contained questions about residents' satisfaction and experiences in local parks and recreational facilities, as well as the frequency of visitation to local parks and the purpose of visit. It also included information regarding the demographics of the respondents. In order to use these data for our demographic comparison, we first used participants' zip-codes to select only the respondents that live in the same neighborhood as our selected parks. We then used a specific question that asked the participants to "indicate how often you or your family visit the local neighborhood park", and selected those who responded "10 or more times/year". Our filtered dataset is composed of 72 participants.

Privacy and Ethical Considerations We acknowledge that there are privacy and ethical concerns associated with the use of social media as data for research, particularly for the development of tools for inferring user demographics. Accordingly, we took several steps to avoid risks to human subjects since participants no longer opt into studies in a traditional sense (Ang et al. 2013). First, we purposely did not consider any profile photos or declared profile information such as name email, username and home location of the users. Second, we did not download the photographs shared on the social media platforms. Instead we kept our data in form of URLs (belonging to Facebook or Flickr) along with a numerical short-code for users. The APIs that we used for data collection do not allow for reverse short-code lookup

which means that our collected data cannot be used to link back to a specific user. Finally, this research was reviewed and approved by the University of Washington Institutional Review Board.

Metrics

In order to study the platform biases across Instagram and Flickr, we employ several metrics of visitor numbers and demographics based on social media images. Visitation is estimated by computing PUD per park. This is the most common UGC-based metric in the recreation literature, and it quantifies the unique number of social media users that post photographs from a specific location per day. To serve as ground-truth, we generate human-labelled data on the number and demographics of visitors by asking crowd-workers on Amazon's Mechanical Turk to count the total number of people *facing* the camera in each photograph.[4] Unlike visitor counts, assigning demographic labels to people is a challenging task that we cannot assign to crowd-workers. This is because racial and ethnic identity is complex and evaluations by others may not match an individuals' self-identification. As we have no means of collecting self-identified demographic information from Flickr or Instagram users by directly contacting them, we follow the methodology described in (Pennacchiotti and Popescu 2011) and bound our definition of demographics to binary values of *white (vs non-white)* and *children (vs adults)*. Our labelled dataset is composed of 500 images uniformly selected across the ten parks from both Instagram and Flickr.

Algorithm

In order to estimate park visitation and demographic distribution of the visitors, we use a facial recognition algorithm to study the *content* of users' posts and photographs. This allows us to broaden our visitation information beyond the person who posted on social media to people whose faces are captured in the posts. To this end we are interested in using an off-the-shelf algorithm that is accessible to policy makers. Currently there are many commercial facial recognition algorithms available, with some specifically designed for providing analytic information in the context of smart cities, such as DeepVisionAI. [5] We use one of such algorithm, Face++, that has been evaluated in the past by the research community (Jung et al. 2018) and has been shown to have a high accuracy in some contexts. In particular, we use the $Detection$ API of Face++ which detects faces that appear in the photos. This API does not *match* the faces of individuals with their identities — that is if a face has appeared in two different photographs it does not detect that it is the same person. However, a unique feature of this API is that it offers ethnicity detection in addition to gender and age, a feature that is currently not offered by any other off-the-shelf commercial algorithms. At the time of this study, the races in Face++ Detection API were defined as African, White, Indian or Asian.

[4]In hiring the crowd workers we followed minimum wage regulation of our state ($14 USD per hour) at the time the study.

[5]https://www.deepvisionai.com/smart-city

In order to measure the viability of the algorithm as a source of inclusive insights for decision makers, we first define the outcome of the facial recognition algorithm as a binary classification problem. In this study, focusing on park and recreation management, we define the favorable outcome (positive class) as being detected by the facial recognition algorithm, and unfavourable outcome (negative class) as not being detected by the algorithm. Park managers and planners often allocate resources such as funding and staffing to parks according to the frequency and type of visitor use. Thus, it is important to ensure that algorithms that are potentially being used to understand visitor use are correctly detecting number of people that are represented in the data source.

Outputs of the binary classifier are organized in a contingency table. We define the True Positive (TP) photographs as ones where the algorithm accurately counted the same number of people appearing in the photograph, based on the human-labeled count. We define the False Positive (FP) group as those where the algorithm over-counted the number of people that appear in the image. The False Positive group demonstrates cases where false information about the number of visitors could lead to misdirected management decisions and the potential to poorly allocate resources. In the domain of face recognition this could happen when an algorithm mistakes similar patterns such as paintings and flyers for human faces. Similarly, we define the False Negative (FN) group as photographs wherein faces should have been detected by the algorithm but were not (i.e., the algorithm under-counted the number of people compared to the crowd-labels). The FN group is the most important classification output in our context as it potentially reduces the equity of management decisions related to resource allocation as well as recognition harm (Whittaker et al. 2018). Finally, we define the True Negative (TN) group as those photographs which did not have any people appearing in them and the algorithm correctly did not count any visitors. Figure 2 presents the favourable and unfavourable outcome along with some example images. Additionally, we quantify the performance of the Face++ algorithm based on $precision$ and $recall$. We compute precision as Positive Predictive Value (PPV): that is the fraction of positive cases correctly predicted (TP) to be in the positive class out of all predicted positive cases (favourable outcome). We compute recall as True Positive Rate (TPR), or the fraction of positive cases correctly predicted to be in the positive class (TP), out of all actual positive cases ($TP + FN$).

Analysis

In this section we present the results of our analyses of visitor counts and demographics. We present comparisons of estimates based on image content with the more widely used PUD visitation metric and demographic composition of surveyed park visitors.

Visitation count Comparing the two sources of social media shared from Seattle Parks, we find that PUD – a metric commonly used to study recreational visits – is substantially higher according to Instagram compared to Flickr. Fig-

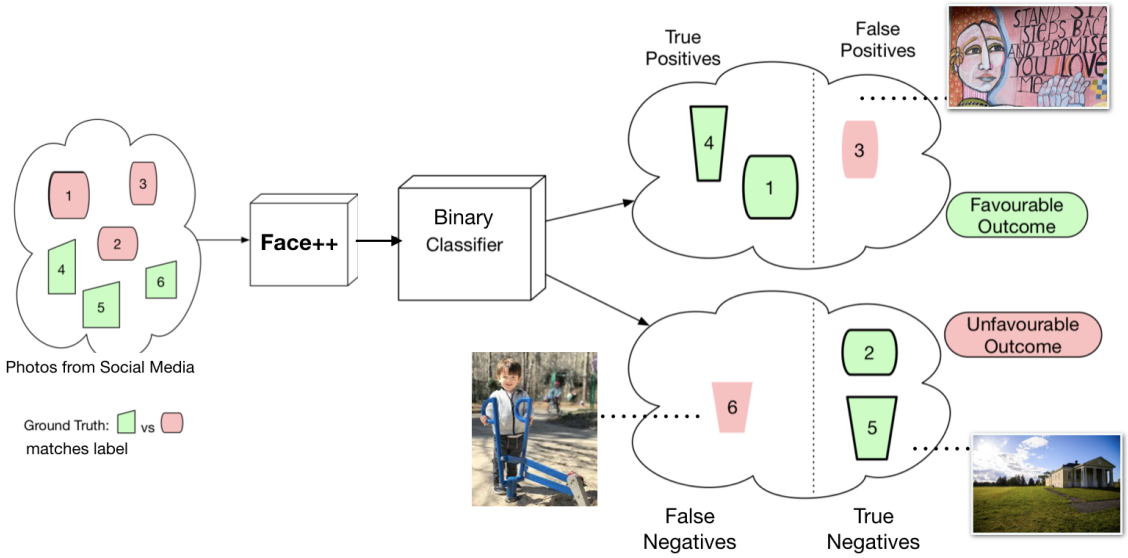


Figure 2: Illustration of the output classes of binary classification where the matching is done based on the comparison of crowd-workers label and Face++ visitor count.

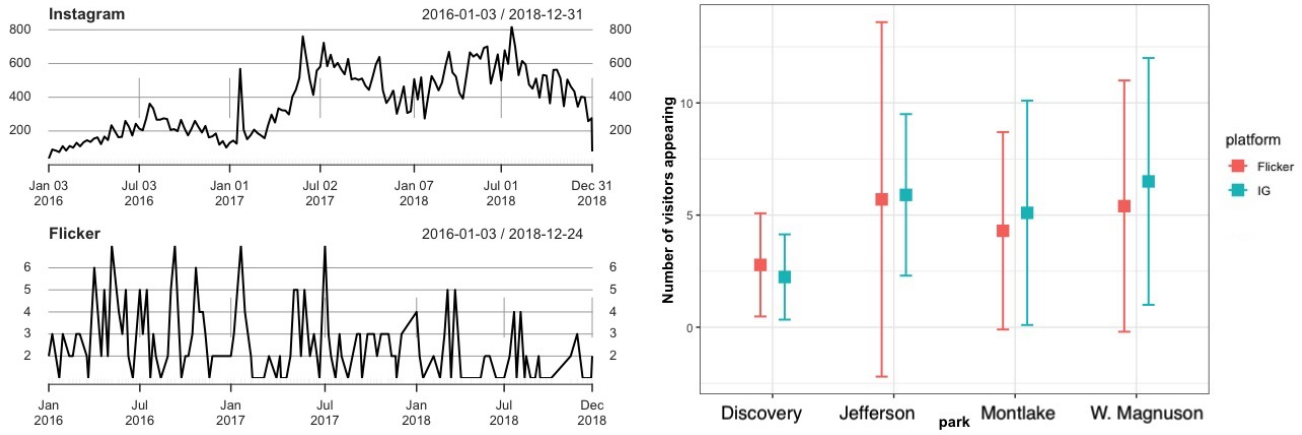


Figure 3: (a) Weekly aggregate of PUD based on photos posted to Instagram (top) and Flickr (bottom) from the ten studied parks from January 2016–January 2019. (b) Average number of people who appear in 50 random photographs selected from each of the four most popular city parks, based on the labelled data. Bars show standard errors.

ure 3(a) demonstrates this result in terms of total weekly PUD across the ten study locations. Furthermore, for both platforms, the long-term trend in PUD appears to reflect the popularity of each platform, with Flickr showing a downward trend and Instagram becoming more popular overtime. Responses to our on-site intercept survey support this observation: 59% of survey respondents stated that they use social media, and out of those, 62% said that they share content on Instagram, as opposed to only 2% who said they share content on Flickr.

Despite the differing popularity of the two platforms, we find that when we consider the average number of people that appear in photographs (counted by the crowd-workers), there is a consistency across parks and between social media platforms. Figure 3(b) presents the mean and standard deviation of visitors count for samples of 50 photographs for the top four parks. A t-test comparison indicates that there are no significant differences in the average number of people that appear in images. This result is particularly important to researchers who use social media for estimating park visitation as it suggests that there is potential to create metrics that are more robust and agnostic to the online platform. However, we find that many people in images are undetected by Face++, which leads to an underestimate of park visitation rates. Figure 4 shows this observation by comparing the number of faces in each image as measure by Face++ versus crowd-labels. We expand on this observation in the next Algorithm Performance section.

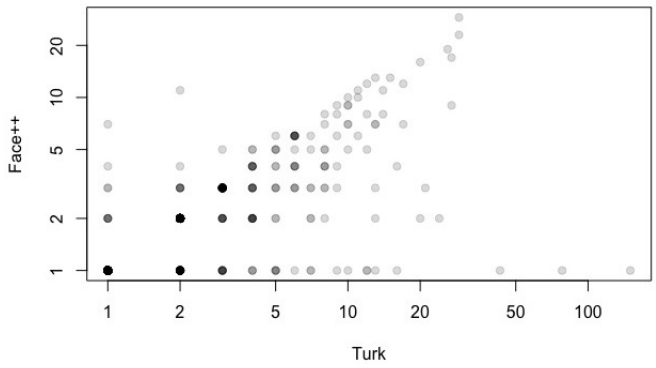


Figure 4: Number of park visitors in the photograph according to crowd-workers (x-axis) vs Face++ (y-axis). Every point presents a photograph in our sample dataset. The shades of each point corresponds to the number of overlapping points, the darker the shade the greater number of data points with the specific crowd-workers and Face++ count.

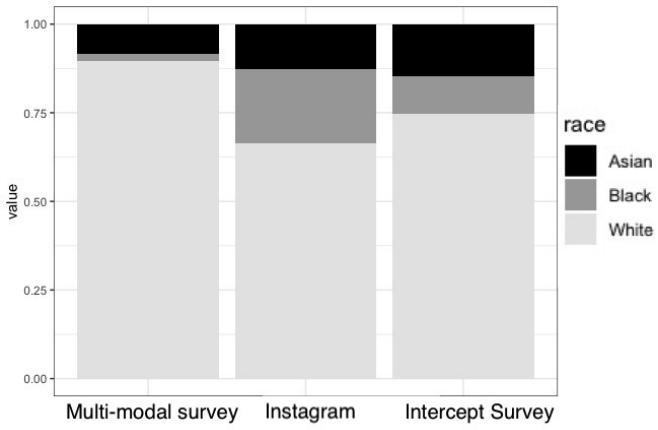


Figure 5: Race of the park visitors as detected by Face++ in Instagram images in 2019, compared with the intercept survey, and the multi-modal survey respondents.

Visitors Demographics Looking at the demographic composition of the park users according to content from different social media platforms, we observe a slight difference in the age distribution of people in Flickr images, compared to Instagram images. According to the demographic detection feature of Face++, Flickr images contain a greater number of children (Figure 6). Our manual inspection of some of the images indicates that Flickr is more often used for photographing sports events that included children. The racial composition of people in the studied images also differs across the two platforms ($\tilde{\chi}^2 = 14.25$ (p= 0.002)).

In a comparison of visitor demographics based on social media content versus visitor surveys we find that a greater portion of survey respondents report their race as White compared to those identified by Face++ (Figure 5). A chi-square test examining the relationship between race and detection method (survey responses vs social media content) found that the results varied by detection method ($\tilde{\chi}^2 = 8.9$, $p = .01$). This difference is largest between people in social media images compared to the multi-model survey respondents, over the same time period of April–June, 2019.

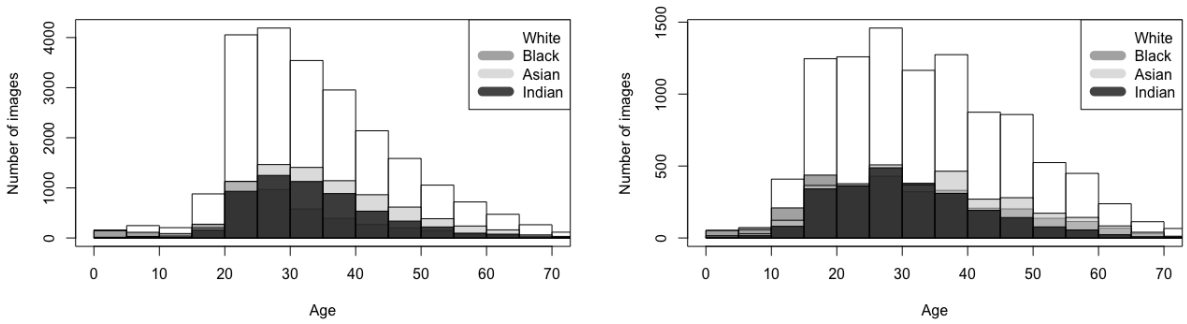
Our crowd-labelled sample data indicates that 39% of the people appearing in social media from parks were children ($mean = 0.28$ per photo, $sd = 0.40$). This result differs from our on-site survey respondents who reported the number of adults and children in their party. The 165 respondents reported being members of parties with 250 total adults and 46 total children: a lower ratio than what we observe from our images. In the multi-modal survey, 28% of the respondents answered selected "Children's Playground or Recreational activities for Children" as their main park usage. While the comparison reported here demonstrates some of the limitations of the traditional survey in being inclusive (as we discuss in the Limitations Section), it also demonstrates the potential to use AI to understand a more demographically diverse set of visitors. It is worth noting that the multi-modal survey respondents capture the racial composition of the local residents where as our intercept survey and Instagram images captures the respondents regardless of whether they live in immediate vicinity of the park.

Algorithm Evaluation

This section follows-up on visitation and demographics results according to the Face++ algorithm with an evaluation of the viability of such an approach in terms of accuracy and fairness. We are particularly interested in answering two research questions: *1*. for what type of decisions can this method be appropriately used? and *2*. What demographic groups does the selected algorithm exclude?

Algorithm Performance

In order to answer *Question 1*, we begin by considering the scenario presented in the Algorithm section that describes how information about park visitation could be used to allocate limited staffing or funding resources. Depending on the decision context, planners may require more or less precise estimates of the number of visitors. To measure the viability of the selected algorithm across a spectrum of scenarios that vary in the acceptable level of error, we introduce a variable β that corresponds to the sensitivity threshold that a manager or policy-makers requires. When $\beta = 0$ the algorithm is strictly required to match $exactly$ the number of people who truly appear in the photographs. A greater value of β corresponds with less correspondence and the sensitivity is relaxed. For example, imagine a scenario where there are 10 people present in a photograph, but the algorithm has only counted 9. Under the condition where $\beta = 0$ the image is considered a False Negative. Alternatively, when β equals any value greater than 1, the photograph would be classified as a positive match (favorable outcome). Figure 7(a) illustrates the performance of the algorithm in terms of precision and recall for increasing sensitivity thresholds β and Figure 7(b) illustrates this trend in terms of the error exhibited in False Positive and False Negative groups. We find the



(a) Instagram (b) Flickr

Figure 6: Age and race distributions of individuals in photographs posted to Instagram (a) and Flickr (b) as detected by the Face++ algorithm.

$precision = 0.80$ (for $\beta = 0$) which resonates with the findings from previous studies (Jung et al. 2018). However, we find for the same threshold the $recall = 0.3$ indicating that many visitors were undetected and that Face++ underestimates park visitation rates.

Two additional key observations can be made from the results presented in Figure 7. First, the False Positive ratio (over-detection) is low in all situations, and for managers with a small degree of flexibility in β it disappears. In other words, the threshold of $\beta = 2$ is enough for all data points in the False Positive class to be regrouped as True Positive. The False Negative curve however suggests that there are also many photographs which are under-counted by the algorithm and the difference in the count is large. Only when β is set to greater than 10 do we see a larger drop in the False Negative ratio. Our manual examination of these photographs confirms that the algorithm fails to detect subjects in photographs that capture group activities. We expand on this observation in the next section.

Algorithmic Fairness

Turning our attention to *Q*uestion 2 about whether the algorithm successfully detects parks visitors from different demographics, we use a non-discrimination fairness criterion that can be applied after the processing stage (Verma and Rubin 2018). Post-processing has the advantage that it works for any black-box classifier regardless of its inner workings as it is unnecessary to access the training data. This approach is best suited to our work since we are exploring applications of off-the-shelf AI models for practitioners who would not have knowledge or control over the training process. Specifically, we measure fairness in terms of equal opportunity (Hardt, Price, and Srebro 2016). A classifier satisfies this definition if both protected and unprotected groups have equal False Negative rates — where the probability of a subject in a positive class ($Y = 1$) has a negative predictive value ($\hat{Y} = 0$). In our example, the protected group corresponds to the photographs containing children or non-white subjects, and unprotected group is simply those photographs where all the subjects are adult and white (based on the ground-truth labels).

$$EO_{race} = \frac{P\{\hat{Y}=0|Y=1,A=Non-White\}}{P\{\hat{Y}=0|Y=1,A=White\}}$$

$$EO_{age} = \frac{P\{\hat{Y}=0|Y=1,A=Kids\}}{P\{\hat{Y}=0|Y=1,A=Non-Kids\}}$$

where $\hat{Y}$ is the binary outcome of the classifier, Y is the actual label and A is the protected attribute. A classifier is considered fair if the metric of equal opportunity (EO) is between 0.8–1.2 (Dwork et al. 2012).

Figure 8 presents this fairness criteria for an increasing threshold of β for the two protected groups of race and age. We observe that for most values of β the algorithm achieves fairness in terms of $race$, in the sense that there are almost equal numbers of photographs that are under-detected by the algorithm (i.e., put in False Negative class) regardless of the race of the people who appear in the images. However, the algorithm does not satisfy fairness criteria for the of age people in images. Photographs containing children are twice as likely to be under-detected by the algorithm compared with images that lack any children. We also observe that for both sensitive attributes (age and race) the algorithm is fairest for smaller values of β. As the sensitivity of the classifier is relaxed by increasing β there is a corresponding increase in the disparity of false negative ratio of non-whites/whites and kids/adults.

These results are based on an analysis of photographs from all 10 study sites. The fairness of the algorithm could indeed vary across parks within the study and elsewhere across the city. To investigate this we measure the fairness and performance for different parks categorized according to their SVI into two bins of low (less than 0.2) and high (greater than 0.8) vulnerability. We observe no statistically significant differences in the fairness of the algorithm for parks that differ in SVI. This result indicates that the detection rate based on our sample of photographs used in this analysis is independent of neighborhood composition. However caution must be taken in generalizing these results and

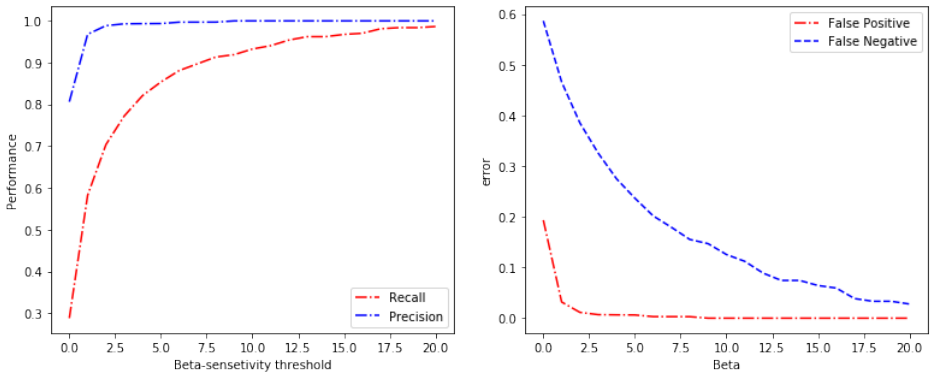


Figure 7: (a) Precision and Recall performance of the algorithm for increasing sensitivity threshold Beta. (b) Classification error in terms of false positive and false negative for the increasing value of β.

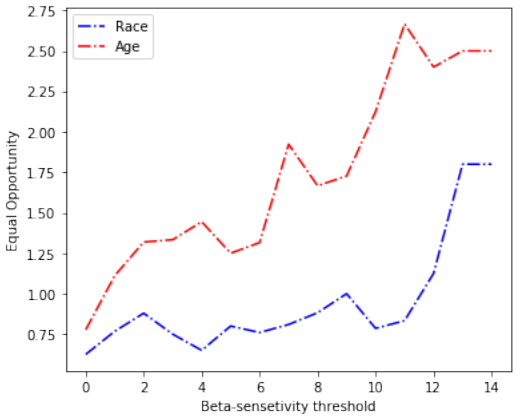


Figure 8: Equal opportunity for the sensitive attributes of race and age for increasing threshold of β.

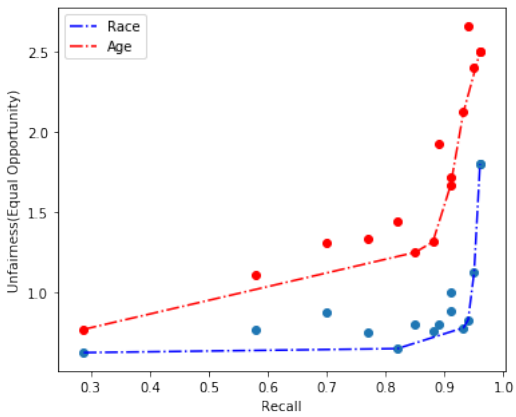


Figure 9: Pareto frontier presenting the trade off between unfairness and utility.

further research is needed to ensure that other confounding effects such as type of camera do not play a part in excluding low income areas of the city.

Discussion

Management Implications

Information on the amount and character of park use is critical for successful park management. Managers rely on visitation estimates from specific locations to inform decisions about where to allocate resources and how to improve the recreational opportunities that are available to communities. As social media becomes an increasingly popular source of information on park use, there is growing recognition that changes in the underlying popularity of social media platforms will likely bias PUD estimates of visitation over time (Wood et al. 2020). In this study, we observe that the average number of people in photographs shared from a set of parks in Seattle, WA is the same for images posted to both Flickr and Instagram. This novel result suggests that there is potential to develop new content-based visitation metrics that are less impacted by the choice of social media platform. While this result is a promising one, we note that the method's performance and fairness depend on the choice of β, or the sensitivity required by the decision maker. The particular AI algorithm that we selected for this study often under-counts the number of visitors in photographs with children.

The methodological biases that we observe in this study could have important consequences for urban planners and park managers who are considering using social media and AI to understand who is currently being served by local green-spaces. This study shows how representations such as a Pareto frontier that allow practitioners to visualize trade-offs between utility and inclusiveness are helpful for guiding decision-making processes. For example in the context of our study Figure 9 presents the trade-off between $recall$ and $(un)fairness$ (as measured by equal opportunity). In the figure we can see that there is potential for managers to apply the AI-based methods used in this study to estimate visitation ± 2, with algorithm recall of 70%, while staying fair and inclusive in terms of age and race.

This study indicates that the content of social media im-

ages may provide useful information about the characteristics of park visitors. This is of great interest to planners and managers who grapple with how to meet growing and changing demands for urban parks and how to improve the equity of the benefits that are provided by urban green spaces. This study demonstrates that AI may improve existing approaches for understanding who uses urban parks in terms of race and age of visitors, if they are used appropriately. There is intriguing evidence that the content of photographs shared from parks in Seattle, WA captures a greater proportion of non-white visitors than traditional methods for surveying park users. Given the known limitations and biases of structured surveys, it may be informative to pair in-person interviews with social media image analyses, with careful attention to the many biases in who uses and shares content on social media, and biases in the algorithms that are used to classify the content.

Theoretical Implications

From the theoretical perspective, first and foremost our work highlights a gap in the literature for understanding the impact of AI algorithms when applied to real-world situations. We believe it is important for interdisciplinary researchers to investigate the result of applying new computational approaches such as facial recognition on user-generated data and the potential impact on people from social science perspective. Furthermore, a venue is needed to promote the ethical usage of such techniques and posit methodologies for handling data and for the analysis of it. Some research question that arise from our study include: what methods should be used in order to ensure that a person is not counted more than once in two different images without employing privacy invasive facial matching? What are the impacts overcounting the same people have on the demographic distribution of the data? What type of behaviours in terms of social group dynamic would these data capture? These are all questions that we believe researchers in the $ICWSM$ community could investigate and derive these types of interdisciplinary work forward.

Finally, if AI models are to be used in future decision makings, more attention is needed to ensure that interpretability is an integral part of these models where user-centric, human-friendly explanations could be provided to justify the decisions that were made (Miller 2019). We foresee the multi-disciplinary field of interpret-able machine learning to include recreational studies in addition to the social sciences (Du, Liu, and Hu 2019).

Finally our work highlights the need for a truly in-the-wild dataset for demographic detection which could be accompanied with demographic labels. Current databases for training AI include the MS-Celeb-1 dataset of 1 million celebrity head-shots matched with demographic information (available from their portfolio or Google FreeBase data-dump (Guo et al. 2016)). VGGFace2 (Cao et al. 2018) contains 3.31 million images of 9,131 subjects where images are downloaded from Google Image Search with variations in pose, age, ethnicity, and profession (e.g., actors, athletes, politicians). However these datasets are often curated based on images gathered on the Internet and thus inherit the biases that are associated with search, ranking and popularity of web content. Further research is also needed in defining the context and target domains in which these algorithms can be applied (Mashhadi 2020).

Limitations

We acknowledge that this work has several limitations. First, the Face++ AI algorithm that we used to classify image content is proprietary and we do not know on which data it was trained. Furthermore, the AI algorithm did not classify races such as Hispanic and Native American, which in some areas of the city we studied are majority minority ethnic groups. Similar mismatches are arising in other works in computational social science where an existing challenge is the definition and boundary of ethnicity. For example in (Chang et al. 2010; Chen et al. 2015) both authors used the term ethnicity to refer to a classification system that includes both racial and ethnic identities (black, white, Asian, Hispanic) whereas others (Ardehaly and Culotta 2014) used the same classification system but referred to it as race. We also acknowledge that our intercept survey was conducted in English only and required in-person interactions. This means that our survey results — like the results of most intercept surveys — could be biased towards English speaking people and those who are most content to interacting with the interviewer. Such population might also be less active or present on social media.

Conclusion

In this paper we present the results of a study using the content of Instagram and Flickr photographs as a source of data for estimating the number and characteristics of people who visit urban parks in Seattle, WA. We show that AI techniques for counting the number of people in images may overcome limitations faced by previous studies that rely solely on numbers of posts. We also present evidence that advances in AI could be leveraged to enable park and recreation managers to improve existing approaches for surveying visitors by pairing structured interviews with information about visitors who share about themselves online. However, as researchers and practitioners are investing efforts into social media analyses there needs to be careful thought given to how to responsibly handle both the data and potential sources of bias. In all respects, Instagram and Flickr are not unique cases. These data and techniques must be used carefully and cautiously to avoid the potential for underlying biases to produce misleading results.

Acknowledgments

The funding for this work was provided by a Thought Leadership and Innovation grant from the Bullitt Foundation promoting responsible human activities and sustainable communities in the Pacific Northwest.

References

Alowibdi, J. S.; Buy, U. A.; and Yu, P. 2013. Language independent gender classification on twitter. In *Proceedings of*

the 2013 IEEE/ACM international conference on advances in social networks analysis and mining, 739–743. ACM.

Ang, C. S.; Bobrowicz, A.; Schiano, D. J.; and Nardi, B. 2013. Data in the wild: Some reflections. *interactions* 20(2):39–43.

Ardehaly, E. M., and Culotta, A. 2014. Using county demographics to infer attributes of twitter users. In *Proceedings of the joint workshop on social dynamics and personal attributes in social media*, 7–16.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 67–74. IEEE.

Cesare, N.; Grant, C.; and Nsoesie, E. O. 2017. Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv preprint arXiv:1702.01807*.

Chang, J.; Rosenn, I.; Backstrom, L.; and Marlow, C. 2010. epluribus: Ethnicity on social networks. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Chen, X.; Wang, Y.; Agichtein, E.; and Wang, F. 2015. A comparative study of demographic attribute inference in twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Chen, Y.; Liu, X.; Gao, W.; Wang, R. Y.; Li, Y.; and Tu, W. 2018. Emerging social media data on measuring urban park use. *Urban forestry & urban greening* 31:130–141.

Cranshaw, J.; Schwartz, R.; Hong, J.; and Sadeh, N. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Sixth International AAAI Conference on Weblogs and Social Media*.

De Choudhury, M.; Sharma, S.; and Kiciman, E. 2016. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing*, 1157–1170. ACM.

Donahue, M. L.; Keeler, B. L.; Wood, S. A.; Fisher, D. M.; Hamstead, Z. A.; and McPhearson, T. 2018. Using social media to understand drivers of urban park visitation in the twin cities, mn. *Landscape and urban planning* 175:1–10.

Du, M.; Liu, N.; and Hu, X. 2019. Techniques for interpretable machine learning. *Communications of the ACM* 63(1):68–77.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Frias-Martinez, V.; Soto, V.; Hohwald, H.; and Frias-Martinez, E. 2012. Characterizing urban landscapes using geolocated tweets. In *2012 International conference on privacy, security, risk and trust and 2012 international confernece on social computing*, 239–248. IEEE.

Fuchs, D. J. 2018. The dangers of human-like bias in machine-learning algorithms. *Missouri S&T's Peer to Peer* 2(1):1.

Ghermandi, A., and Sinclair, M. 2019. Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change* 55:36–47.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 87–102. Springer.

Hamstead, Z. A.; Fisher, D.; Ilieva, R. T.; Wood, S. A.; McPhearson, T.; and 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.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

Hausmann, A.; Toivonen, T.; Slotow, R.; Tenkanen, H.; Moilanen, A.; Heikinheimo, V.; and Di Minin, E. 2018. Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conservation Letters* 11(1):e12343.

Hecht, B., and Stephens, M. 2014. A tale of cities: Urban biases in volunteered geographic information. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Ilieva, R. T., and McPhearson, T. 2018. Social-media data for urban sustainability. *Nature Sustainability* 1:553–565.

Jung, S.-G.; An, J.; Kwak, H.; Salminen, J.; and Jansen, B. J. 2018. Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race. In *Twelfth International AAAI Conference on Web and Social Media*.

Keeler, B. L.; Wood, S. A.; Polasky, S.; Kling, C.; Filstrup, C. T.; and Downing, J. A. 2015. Recreational demand for clean water: evidence from geotagged photographs by visitors to lakes. *Frontiers in Ecology and the Environment* 13(2):76–81.

Klare, B. F.; Burge, M. J.; Klontz, J. C.; Bruegge, R. W. V.; and Jain, A. K. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7(6):1789–1801.

Lee, H.; Seo, B.; Koellner, T.; and Lautenbach, S. 2019. Mapping cultural ecosystem services 2.0 – Potential and shortcomings from unlabeled crowd sourced images. *Ecological Indicators* 96.

Levin, N.; Lechner, A. M.; and Brown, G. 2017. An evaluation of crowdsourced information for assessing the visitation and perceived importance of protected areas. *Applied Geography* 79:115–126.

Long, X.; Jin, L.; and Joshi, J. 2012. Exploring trajectory-driven local geographic topics in foursquare. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, 927–934.

Martinez-Harms, M. J.; Bryan, B. A.; Wood, S. A.; Fisher, D. M.; Law, E.; Rhodes, J. R.; Dobbs, C.; Biggs, D.; and Wilson, K. A. 2018. Inequality in access to cultural ecosystem services from protected areas in the Chilean biodiversity hotspot. *Science of The Total Environment* 636:1128–1138.

Mashhadi, A. 2019. Seattle city park visitor survey.

Mashhadi, A. 2020. A privacy-preserving framework for collecting demographic information. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, 1–8. New York, NY, USA: Association for Computing Machinery.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media*.

Noble, S. U. 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press.

Noulas, A.; Scellato, S.; Mascolo, C.; and Pontil, M. 2011. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *Fifth International AAAI Conference on Weblogs and Social Media*.

O'neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pennacchiotti, M., and Popescu, A.-M. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 430–438. ACM.

Quercia, D.; Aiello, L. M.; and Schifanella, R. 2018. Diversity of indoor activities and economic development of neighborhoods. *PloS one* 13(6):e0198441.

Richards, D. R., and Friess, D. A. 2015. A rapid indicator of cultural ecosystem service usage at a fine spatial scale: Content analysis of social media photographs. *Ecological Indicators* 53:187–195.

Sessions, C.; Wood, S. A.; Rabotyagov, S.; and Fisher, D. M. 2016. Measuring recreational visitation at U.S. National Parks with crowd-sourced photographs. *Journal of Environmental Management* 183:703–711.

Silva, T. H.; Viana, A. C.; Benevenuto, F.; Villas, L.; Salles, J.; Loureiro, A.; and Quercia, D. 2019. Urban computing leveraging location-based social network data: a survey. *ACM Computing Surveys (CSUR)* 52(1):17.

Tenkanen, H.; Minin, E. D.; Heikinheimo, V.; Hausmann, A.; Herbst, M.; Kajala, L.; and Toivonen, T. 2017. Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports* 7.

Venerandi, A.; Quattrone, G.; Capra, L.; Quercia, D.; and Saez-Trumper, D. 2015. Measuring urban deprivation from user generated content. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 254–264. ACM.

Verma, S., and Rubin, J. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1–7. IEEE.

Whittaker, M.; Crawford, K.; Dobbe, R.; Fried, G.; Kaziunas, E.; Mathur, V.; Myers West, S.; Richardson, R.; Schultz, J.; and Schwartz, O. 2018. Ai now report 2018. ai now institute, new york university.

Wood, S. A.; Guerry, A. D.; Silver, J. M.; and Lacayo, M. 2013. Using social media to quantify nature-based tourism and recreation. *Scientific reports* 3:2976.

Wood, S. A.; Winder, S. G.; Lia, E. H.; White, E.; Crowley, C.; and Milnor, A. 2020. Next-generation visitation models using social media to estimate recreation on public lands. *In review for Scientific reports*.

Zhang, S., and Zhou, W. 2018. Recreational visits to urban parks and factors affecting park visits: Evidence from geotagged social media data. *Landscape and Urban Planning* 180:27–35.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zheng, Y.; Capra, L.; Wolfson, O.; and Yang, H. 2014. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5(3):38.