Environment and Development Economics (2025), 1–19
doi:10.1017/S1355770X2510020X

RESEARCH ARTICLE

User-generated data to predict visitors in environmental areas

David Hervés-Pardavila, Ana Castro-Atanes and Maria L. Loureiro

ECOBAS Interuniversity Research Center-Facultade de Ciencias Económicas e Empresariais, Universidade de Santiado de Compostela, Santiago de Compostela, Spain
Corresponding author: David Hervés-Pardavila; Email: david.herves@rai.usc.es

(Received 30 April 2024; revised 1 June 2025; accepted 23 June 2025)

Abstract

The economic valuation of recreational ecosystem services is challenging due to difficulties in obtaining geo-tagged information of users. The objective of this study is to validate crowdsourced and user-generated content in order to predict visitation patterns to 16 national parks in Spain. The results may serve to encourage its utilization in the study of recreational demand in other countries, particularly developing countries, where on-site visitor information may be limited or expensive to gather. The present article employs a negative binomial regression model to evaluate the validity of two sources of data: Flickr and mobile phones. The accuracy of predictions exhibited variation across the 16 parks, indicating that site-specific characteristics, such as the seasonality of visitation patterns, may be of significance. The utilization of mobile phone data for modelling visitors yielded enhanced predictive capacity, as shown by the goodness of fit of the estimated models.

Keywords: Flickr; recreation; recreational ecosystem services; tourism; user-generated data

JEL classification: C2; C5; C8; Q57

1. Introduction

Revealed preference methods (Whitehead *et al.*, 2012; Cheng *et al.*, 2019) for valuing natural recreational areas frequently depend on visitation, occupancy, or user rates (Mwebaze and MacLeod, 2013; Parsons, 2017). This information is typically gathered on-site from various sources, such as visitor records, infrastructure-based counts, and data provided by the managing authority of the natural areas (Sessions *et al.*, 2016; Mancini *et al.*, 2018; Walden-Schreiner *et al.*, 2018; Fisher *et al.*, 2019; Owuor *et al.*, 2023). However, the process of collecting and analysing visitor data on-site can be challenging, time-consuming and expensive, particularly in the context of large-scale, spatially diverse and non-protected areas. These issues are more prevalent in low- and middle-income countries (LMICs) (Kim et al., 2019). Visitor monitoring is essential for

the valuation of recreational areas (Neuvonen *et al.*, 2010; Edwards *et al.*, 2011), distribution of resources and marketing (Buckley, 2009) and protection and sustainability (Cessford and Muhar, 2003; Buckley, 2011).

Recently, the use of social media, user-generated and geo-tagged data has emerged in the valuation of recreational ecosystem services (Da Mota and Pickering, 2020; Zhang *et al.*, 2022; Pickering *et al.*, 2023), holding the potential to address the aforementioned issues when utilized as a proxy for visitors of natural areas. However, such data should be validated before the implementation of valuation methods through a process of comparison to traditional data sources. On one hand, social media and user-generated data may not always provide a reliable representation of visitors to natural areas, often over-representing young, urban, well-educated and technology-friendly individuals (Keeler *et al.*, 2015; Tenerelli *et al.*, 2016). On the other hand, most studies of this nature use Flickr to approximate visitor data (Da Mota and Pickering, 2020; Wilkins *et al.*, 2021). Flickr is a website for sharing geolocated photographs which is less popular than other social media platforms such as X (formerly known as Twitter) and Instagram.[1] Furthermore, there is a dependence on the willingness of social media users to share their experiences, geo-locate the media, and maintain a public profile so that other users of the platform can access their content. Otherwise, researchers cannot identify them as visitors to natural areas.

Nevertheless, the previous statements do not imply that the aforementioned data sources are useless. Many studies have demonstrated that social media data can serve as a valuable proxy for tourism in natural areas (Wood *et al.*, 2013; Chun *et al.*, 2020; Sinclair *et al.*, 2022). For instance, a significant correlation has been found between Flickr users and visitors to mountain protected areas (Walden-Schreiner *et al.*, 2018). Owuor *et al.* (2023) conducted ordinary least squares, Spearman's correlation and time-series analyses on official visitor counts and various types of user-generated content such as X, Flickr, Google Maps, Wikipedia and TripAdvisor.

Similarly, Wood *et al.* (2020) used data from Flickr, X and Instagram alongside precipitation data in linear models with fixed and random effects. They concluded that "social media can be applied with moderate success to estimate visitation at sites that are unmonitored or otherwise lack on-site counts". Additionally, a wavelet analysis (Mancini *et al.*, 2018) identified Flickr as a reliable source for describing the temporal patterns of nature-based recreation.

Flickr data was further validated (Sessions *et al.*, 2016) by conducting a negative binomial regression analysis on visitor data from U.S. national parks. Some U.S. national parks were also used to test whether TripAdvisor reviews and climate data could forecast tourism demand using various techniques, including machine learning models (Khatibi *et al.*, 2018). In terms of correlation coefficients, Wilkins *et al.* (2021) conducted a comprehensive review of the use of social media to predict recreational activities in environmental areas, comparing the performance of various social media platforms reported in numerous articles. Interestingly, Flickr reported both the lowest and highest correlation coefficients, depending on the study. Fewer studies

[1]The website for Flickr is available at https://www.flickr.com/. The site for X (formerly known as Twitter) is https://twitter.com/home, and Instagram can be accessed at https://www.instagram.com/.

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

utilized Twitter and Instagram compared to Flickr, making it challenging to draw definitive conclusions.

Other researchers have explored the use of alternative data sources, such as mobile phone data, in travel cost modelling (Kubo *et al.*, 2020). Mobile phones have been widely used for several years, providing a data source that avoids the biases inherent in sub-samples of the population. Furthermore, the popularity of social media sites fluctuates over time, posing challenges for studies with long-term time series data. In contrast, the proportion of people who own and use mobile phones is expected to remain stable. Furthermore, mobile phone data can be considered a form of "passively" generated content, eliminating the need to rely on visitors' willingness to publicly share their experiences.

In a study conducted in South Korea, mobile phone data, along with geotagged tweets from X and photographs from Flickr, were compared with on-site surveys (Fisher *et al.*, 2019). The main results show that the correlations revealed by the mobile data were not as strong as those revealed by the other sources, although it is important to note that the sample size for the mobile data was much smaller. In a different study, Jaung and Carrasco (2021) used mobile phone data to examine the effect of weather on recreational ecosystem services.

Some studies highlighted a lack of research in LMICs compared to Europe and North America (Calcagni *et al.*, 2019; Ghermandi and Sinclair, 2019; Kim *et al.*, 2019), despite a tendency for protected areas to increase in LMICs (Balmford *et al.*, 2009; Kim *et al.*, 2019). Flickr has been used most frequently to study spatial visitation patterns in LMICs. In south-eastern Asia, a validation was carried out through on-site visits and surveys (Kim *et al.*, 2019), while in the Argentinian Andes the focus was on cultural perceptions (Rossi *et al.*, 2020). Another study in Argentina with similar objectives used Paronamio photographs (Martínez Pastur *et al.*, 2016). Other examples of Flickr use include Nepal (Bhatt and Pickering, 2022), Mexico (Ghermandi *et al.*, 2020) and India (Sinclair *et al.*, 2018). Other authors have used InVEST software (Natural Capital Project, 2025), which also utilizes Flickr photographs. This provides an understanding of the spatial patterns, finding tourism hotspots as well as predicting them based on natural and social infrastructure, in India (Bhalla *et al.*, 2022), Indonesia (Tussadiah *et al.*, 2021) and Morocco (Mouttaki *et al.*, 2021). In 2013, Wood *et al.* (2013) tested for differences between developed and developing countries when predicting visitation patterns with Flickr photographs. They concluded that income and type of attraction were significant parameters. Tenkanen *et al.* (2017) explored the potential of Instagram, Twitter and Flickr for monitoring visitors to natural areas in Finland and South Africa. They found that some country-specific differences overlapped with a better match in the most visited national parks.

The aim of this study is to assess the suitability of two types of user-generated data for capturing the number visitors to Spanish national parks between 2015 and 2023.[2] We aim to compare the most widely adopted source, Flickr, with the alternative considered

[2]Outdoor activities in Spain were heavily affected by COVID-19 restrictions from March 2020 to mid-2021, including a strict national lockdown, curfews and travel limitations (Pozo *et al.*, 2022, Tapia-Serrano *et al.*, 2022). Although most outdoor restrictions ended by mid-2021, entry requirements such as vaccination or testing remained until early 2023.

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

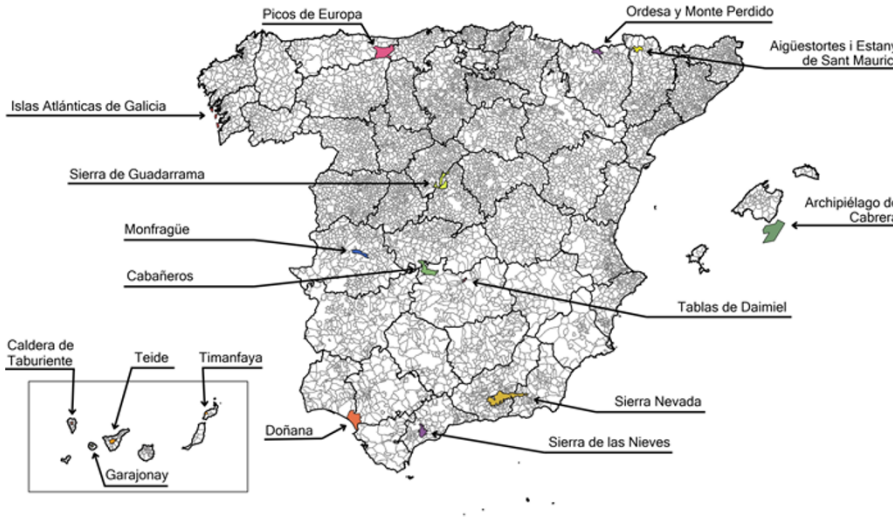

Figure 1. Map showing the locations of Spanish national parks, with provincial boundaries in black and municipal boundaries in grey.

to be the most promising: mobile phone data. This approach helps to mitigate some of the biases introduced by Flickr, such as the over-representation of certain population groups and variability in its popularity.

2. Data

The analysis of user- or device-generated data, obtained from social network websites or mobile phones, raises some ethical implications that need to be addressed. This work aims to respect both the terms of service (TOS) of the platforms from which the data were obtained, as well as the dignity and privacy of their users. A statement on this aspect of the article is provided in online appendix B. Next, we list the different sources of data and provide a brief explanation of their characteristics and how they were obtained.

2.1. Visitors to Spanish national parks

Spain is home to 16 national parks (figure 1), each of which employs its own unique methods of counting visitors (Red de Parques Nacionales, 2017). *Picos de Europa*, for instance, is one of the largest parks and includes permanent residents within its boundaries, as well as roads that traverse the park but are not intended for visitor use. While all primary access points are equipped with vehicle and pedestrian counters, secondary access points lack such devices.

Similarly, the *Ordesa y Monte Perdido* park relies on a combination of information from bus companies and vehicle and hiker counters to estimate visitor numbers. *Aigüestortes i Estany de Sant Maurici* uses periodic surveys and tracks nights spent in the park to supplement its counting methods. By contrast, *Doñana* National Park can

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

only be accessed via organized tours with local guides, which simplifies the counting process. In *Cabañeros* National Park, for example, 16 signposted routes are monitored, and visitor numbers are recorded at the entrances and tourist information offices. Manual surveys are considered more reliable than automated detection of vehicles and hikers.

National parks located on islands, comprising six of the total, often have more confidence in their measurements (Red de Parques Nacionales, 2017). They combine data from visitor offices, pressure counters for pedestrians, vehicle counters and shipping companies.

Despite these efforts, automated counts frequently underestimate the number of people in the parks. Park authorities assert that their data provide a reasonable estimate of the actual visitors but they acknowledge that it remains an estimate.

Visitor data for Spanish national parks is managed by the Autonomous Agency for National Parks (*Organismo Autónomo de Parques Nacionales*) (Red de Parques Nacionales, n.d) and made available upon request. These datasets are spatially aggregated by each park and, since 2015, data points have been temporally aggregated by month. For our study, we requested monthly data from 2015 up to the most recent year available, 2023. It is important to note that there are missing data points, particularly in 2018 and 2019, when eight parks had no observations in one of the two years. Besides this, *Sierra de las Nieves* only provided data for 2022 and 2023, while *Sierra Nevada* had data available for 2015-2017, 2022 and 2023.

2.2. Flickr photographs

Flickr has been widely used in the literature, enabling us to compare our findings with those of other researchers. Flickr also offers the advantage of relatively easily extracting photographs taken within specific coordinates. We used the Flickr API (Application Programming Interface) to retrieve public photographs that matched specific criteria. We searched for content matching the approximate coordinates of each national park between 2015 and 2023. To focus the analysis on photographs captured within the boundaries of each park (as shown in figure 1), we clipped the search results to match the exact polygon shape of each park.

To address the potential for bias introduced by highly active users uploading multiple photographs on the same day, we employed the concept of Flickr-user-days (FUD). One FUD represents a unique owner posting at least one photograph in a national park on a given day. For instance, an individual sharing 10 pictures on the same day would be counted as 1 FUD. Someone sharing two pictures on two consecutive days would be counted as 2 FUD. This standardized method is commonly used in the previously referenced literature.

To align the Flickr data with visitor data, we grouped FUD by month and park name, aggregating all FUD for the same park and month.

2.3. Mobile phone data

Mobile phone network data introduce a novel element to our study, as it has not been explored as extensively as the other two sources. It also offers significant advantages,

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

particularly in avoiding some of the biases associated with social media data, such as the over-representation of young, urban populations. However, this source only became available in July 2019, resulting in fewer data points per park compared to the Flickr data.

The Spanish National Institute of Statistics (INE) publishes tourism measurement data derived from mobile phone usage in collaboration with the country's three main telephone companies (Instituto Nacional de Estadística, 2024). This collaboration began with the first release in May 2022, which included data from July 2019 onwards, and has continued with monthly publications. The measurement process relies on the antennas to which each mobile phone is connected and records of previous connections.

According to the technical report provided by INE (Instituto Nacional de Estadística, 2022), the coverage area of each antenna can vary significantly. In densely populated areas, it can be a few hundred metres, whereas in rural areas it can be several kilometres. This results in approximate localization. While the specific names of the telephone companies involved are not disclosed, it is noted that they collectively represent 75 per cent of the Spanish market.

Data acquisition and the definition of tourism depend on whether they pertain to outbound tourism (foreigners entering the country) or domestic tourism. In both cases, the information provided to the companies by INE is completely anonymous, and each telephone is monitored for as long as it remains switched on and connected.

The technical report on outbound tourism provided by INE (Instituto Nacional de Estadística, 2022) defines tourists as mobile phones belonging to international operators that are detected in Spain between 10 pm and 6:00 am, and remain so thereafter. The trip is considered to end when the mobile phone is no longer detected in Spain during these hours. Mobile phones detected solely outside of these hours are still counted as foreign tourists under two conditions: they have been detected for at least three hours, and they have not been detected more than eight times in the previous eight weeks in the same municipality. Data are aggregated by month and municipality. These administrative divisions are shown in figure 1.

According to this definition, if a national park spans multiple municipalities and assuming the visitor does not stay overnight, a mobile phone visiting the park would only be reported if the visitor spent at least three hours in any of these municipalities. Additionally, the dataset includes information on the tourists' origin country. To ensure anonymity, countries with fewer than 30 travellers per month are disregarded.

The technical report on domestic tourism defines tourists as mobile phones devices departing from their usual area (which does not have to be contiguous) that are detected in a municipality located in a different province between 10 pm and 6 am, and are subsequently detected in that same municipality. Excluding intra-province travel aims to simplify measurement by eliminating routine displacements within densely populated provinces. To ensure anonymity, municipalities with fewer than 30 travellers per month are disregarded.

If a mobile phone is only detected in another municipality during the day, it is still counted as a tourist under the following conditions: the mobile phone was detected in the usual area the previous night, the detection lasted for at least three hours, and the phone was not detected there more than eight times in the previous eight weeks.

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

To maintain consistency in notation, we created a new variable named mobile-phone-user-days (MPUD). This variable was computed to represent the number of days each mobile phone was detected in the vicinity of the national parks.

The original data provided by INE was grouped by month and municipality. In order to align with our analysis, we aggregated the data by month and park, summing all MPUD values for each month and park.

3. Methods

In this section, we describe the statistical calculations employed to address our main scientific question. Our hypothesis is that recreational demand and its temporal variability can be predicted using information generated by users or devices. To test our hypothesis, we performed regressions of visitor estimators, such as FUD and MPUD (sections 2.2 and 2.3), against on-site counts (section 2.1). We evaluate the most suitable regression approach, assess the model's goodness-of-fit and examine the significance of the predictors.

All of our calculations were conducted locally on a desktop computer using Python 3 (Python Software Foundation, 2024) programming language. Some Python libraries that are specialized in statistical calculations were employed in different steps of the work, such as Scipy (Virtanen *et al.*, 2020), Statsmodels (Seabold and Perktold, 2010) and Matplotlib (Hunter, 2007) for plotting.

Our analysis begins with an exploratory data analysis, aiming to visualize global trends and assess the popularity of the two social media sources. We will also investigate seasonal patterns in visitor numbers. Additionally, we will compute Spearman's correlation coefficient, since this metric does not assume a linear relationship between variables.

Next, we begin studying the predictability of recreational demand by our data sources. When modelling count data, the Poisson distribution is often considered. However, visitor counts in recreational areas typically show high levels of overdispersion, making a negative binomial distribution a more suitable choice (Cameron and Trivedi, 2013). This distribution captures the observed variance w in terms of the mean μ: $w = \mu + \alpha\mu^p$ where $p = 2$ leads to the so-called NB2 model, the most common specification (Cameron and Trivedi, 2013) for high overdispersion. We have implemented a significance test to calculate the value of α for the NB2 model, the details of which can be found in online appendix A. The value of alpha is then used to fit the dependent variable (i.e., the number of on-site visitors, as provided by the parks' authorities) against a set of independent variables $\vec{x'}$ multiplied by $\vec{\beta}$ coefficients,

$$\hat{\mu} = NB2\left(\exp\left(\vec{x'}\vec{\beta}\right), \alpha\right).$$

We test three models. Each one yields a different expression for $\vec{x'}\vec{\beta}$ (table 1). The subscript i represents different national parks, and the subscript t controls for time. The variables Park_i and Season_t are sets of dummy variables used to incorporate fixed effects for each park and season, respectively. Given the climate of Spain, we defined winter as December-January-February, spring as March-April-May, and summer as June-July-August-September. To avoid perfect collinearity, the omitted category is fall for October-November, and the coefficients of these seasonal dummy variables are to

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

Table 1. Variables involved in each of the three negative binomial regression models

Model	Expression	Data sources
1	$\text{Visitors}_{it} \sim \ln \text{FUD}_{it} + \text{Park}_i + \text{Season}_t + \text{COVID}_t$	Flickr
2	$\text{Visitors}_{it} \sim \text{MPUD}_{it} + \text{Park}_i + \text{Season}_t + \text{COVID}_t$	Mobile phone
3	$\text{Visitors}_{it} \sim \ln \text{FUD}_{it} + \text{MPUD}_{it} + \text{Park}_i + \text{Season}_t + \text{COVID}_t$	Both

be interpreted with respect to this baseline. COVID_t is another dummy variable, taking the value 1 for the period from July 2020 to June 2021, to account for strict social distancing and travel restriction during the pandemic, and 0 otherwise. At this juncture, we opted to exclude observations from the dataset for the months of March to June 2020, inclusive. This period corresponds to the lockdown, during which outdoor activities were either prohibited or highly restricted. These models were specified based on Spearman's correlations between the independent variables (section 4.1). We also employed the Akaike Information Criterion (AIC) and log-likelihood estimations to ensure that applying the natural logarithm improved the fit for FUD but not for MPUD (not shown here).

The three models can be fitted using the complete dataset as a pooled panel data. However, park-specific regressions may be more appropriate due to their different characteristics and the seasonalities of each of them. In this last case, the subscript *i* in table 1 disappears, and the models become time series. A log-likelihood test can be used to determine whether the pooled or park-specific approach is better. The null hypothesis (H_0) posits that the pooled model is preferred. If the sum of the 16 different log-likelihood functions ($\sum_{i=1}^{16} \mathcal{L}_i\left(\vec{\beta}\right)$) is greater than the log-likelihood function of the pooled model, denoted as $\mathcal{L}_{pooled}(\vec{\beta})$, then the park-specific approach provides a better representation of the data. Conversely, if $\mathcal{L}_{pooled}(\vec{\beta})$ is larger, then the pooled model is deemed superior. We test the significance of the differences using a log-likelihood ratio test,

$$LR = -2 \cdot [H_0 - H_a] = -2 \cdot \left[\mathcal{L}_{\text{pooled}}\left(\vec{\beta}\right) - \sum_{i=1}^{16} \mathcal{L}_i\left(\vec{\beta}\right)\right]. \tag{1}$$

LR is χ^2 distributed with $\sum_{i=1}^{16} G_i - G_{pooled}$ degrees of freedom. G_{pooled} is the number of coefficients in the pooled model, and G_i is the number of coefficients in each single park regression.

After deciding whether to use pooled or park-specific regression, the pseudo R^2 measurements are computed to assess the goodness of fit. The deviance R^2 is a generalization of the sum of squares for count data:

$$R^2_{DEV} = 1 - \frac{D\left(y, \hat{\mu}\right)}{D\left(y, \overline{y}\right)}. \tag{2}$$

Here, y represents the observed counts and $\hat{\mu}$ the expected counts, as calculated by the NB2 regression. $D\left(y, \hat{\mu}\right)$ is the deviance of the complete model incorporating all regressors in table 1. $D\left(y, \overline{y}\right)$ is the deviance of an intercept only model. R^2_{DEV} measures the proportional reduction in the deviance between the observed counts and the

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

expected counts due to the inclusion of new regressors, with respect to an intercept-only model. R^2_{DEV} is defined between 0 and 1: for example, $R^2_{DEV} = 0.65$ means that, when the regressors are included, the deviance between the predicted and observed counts has decreased by 65 per cent compared to the deviance of an intercept-only model.

Finally, we perform a robustness check. We fitted the NB2 models using data from years 2021 and 2022 to predict visitation in 2023. The accuracy of the prediction can be evaluated using the mean absolute percentage error (MAPE),

$$\text{MAPE}_i = \frac{100}{T}\sum_{t}^{T}\left|\frac{y_{it} - \widehat{\mu_{it}}}{y_{it}}\right|, \tag{3}$$

where i stands for each one of the 16 parks, t is the month, y_{it} is the observed visitation to park i in month t in 2023, and $\widehat{\mu_{it}}$ is the predicted number of visits to park i in 2023.

4. Results

4.1. Exploratory data analysis

Approximately 15 per cent of the on-site visitor observations had missing data. This corresponds to the winter months, when access to some parks is prohibited, and to parks for which data became available only after 2015. After removing these observations, the dataset consisted of 1,443 observations. The dataset's summary statistics are presented in table A2 in appendix A. Table 2 shows the total number of on-site visitors, FUD and MPUD each year. The number of visitors remained close to 15 million during the period 2015–2017, with a moderate decline to 12–13 million observed in the years 2018 and 2019. The impact of the 2020 pandemic is evident, with visitor numbers falling below 7 million. After 2020, visitor numbers gradually recovered to pre-pandemic levels. Flickr data show a notable decrease between 2015 (1,252 FUD) and 2019 (509 FUD). The post-pandemic recovery of FUD is much weaker. In 2023, the number of FUD (299) was similar to that in 2020, when outdoor recreation and international travel were highly restricted. This behaviour indicates a gradual decline in the popularity of the website, rather than real trends in visitation. MPUD data collection started in July 2019. Therefore, this year's MPUD is only 16.9 million. There was a strong post-pandemic increase in MPUD, especially from 2021 (21.45 million) to 2022 (40.77 million). Note the significant differences in the order of magnitude between the two predictors. MPUD clearly overestimates on-site visitors, whereas FUD underestimates them. This was expected due to the limitations and characteristics of the data sources, as described in section 2.

Figure A1 in the appendix shows the seasonality of each national park. Parks located in the northernmost regions tend to have the majority of visits in the summer: *Picos de Europa*, *Islas Atlánticas de Galicia* and *Ordesa y Monte Perdido* are good examples of this behaviour. Parks with a hot-summer Mediterranean climate experience their highest number of visitors in spring and autumn. See, for example *Monfragüe* or *Cabañeros*. In the Canary Islands, all four parks present a homogeneous distribution.

Spearman correlations (ρ_s) between the three variables in table 2 were computed for the entire dataset. On-site visitors were found to be highly correlated with FUD

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

Table 2. The total number of visitors provided by the park authorities, Flickr-user-days (FUD) and mobile-phone-user-days (MPUD) each year

Year	Visitors ($\times 10^6$)	Flickr (FUD)	Telephone (MPUD) ($\times 10^6$)
2015	14.4	1252	–
2016	15.01	1241	–
2017	15.32	916	–
2018	12.9	741	–
2019	12.1	509	16.9
2020	6.4	224	11.97
2021	11.5	405	21.45
2022	13.9	420	40.77
2023	15.01	299	44.41

($\rho_s = 0.68$) and moderately correlated with MPUD ($\rho_s = 0.33$). Of the two predictors, FUD shows a small correlation with MPUD ($\rho_s = 0.18$).

4.2. Negative binomial regression

First, we evaluate the log-likelihood ratio test in (equation 1) and determine which approximation, pooled or park-specific, exhibits stronger predictive capabilities. For all the models in table 1, the LR follows a χ^2 distribution with $\sum_{i=1}^{16} G_i - G = 75$ degrees of freedom. The critical value for testing whether to reject the null hypothesis is $\chi^2_{c,0.05}(75) = 96.2$. Notably, the null hypothesis is rejected in all cases. For models 1 (Flickr data only), 2 (mobile phone data only) and 3 (both), the LR values are $LR_1 = 3297$, $LR_2 = 1309$ and $LR_3 = 1429.6$, respectively. All of these values are positive, indicating that the park-specific approach provides a better representation of the data and is well above the critical value required to reject the null hypothesis. Consequently, we conclude that park-specific regression provides a more suitable approximation for predicting visitor numbers.

Figure 2 plots the results of park-specific regressions and shows the pseudo R^2_{DEV} defined in (equation 2). An $R^2_{DEV} = 0.6$ indicates that the deviance between the observations and the predictions has decreased by 60 per cent compared to the deviance of an intercept-only model. In each of the 16 plots, estimated normalized visitors are plotted against on-site normalized visitors. Normalization is performed by dividing by the maximum number of on-site visitors in the park. Some parks, such as *Ordesa y Monte Perdido* (third row, first column), have estimations using any of the user-generated sources that are close to the 1-1 line. This indicates a good match between predictions and observations and, therefore, $R^2_{DEV} > 0.8$. Other parks, such as *Monfragüe* (second row, fourth column in figure 2), show dispersion on both sides of the 1-1 line and $R^2_{DEV} < 0.35$. If the points in figure 2 are located below the 1-1 line, the predictions underestimate the on-site visitors and vice versa. When R^2_{DEV} is high for a model (Flickr data, mobile phones, or both), it tends to be similarly high for the others. However, the majority of the parks exhibit higher R^2_{DEV} (goodness-of-fit) when phone data are used for prediction purposes. This is true whether used on its own or with Flickr data. For

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

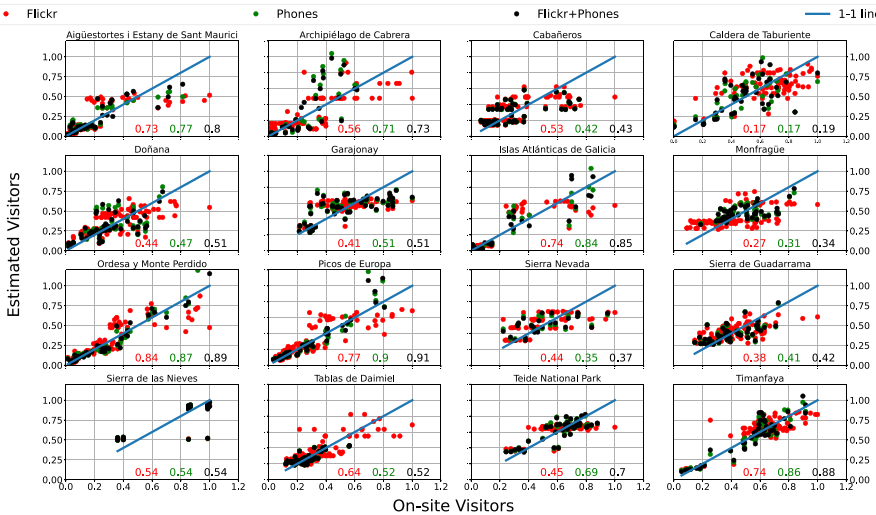

Figure 2. Park-specific NB2 regressions using three different user-generated data sources.
Notes: Flickr photographs (red, Model 1 in table 1), mobile phone data (green, Model 2 in table 1) and both of them (black, Model 3 in table 1). Visitors are normalized using the maximum number of on-site visitors at each park. R^2_{DEV} is depicted in the bottom right corner of each subplot.

example in *Teide National Park* (fourth row and third column), $R^2_{DEV} = 0.69$ when mobile phone data are used to predict visitor numbers, $R^2_{DEV} = 0.45$ when Flickr data are used and $R^2_{DEV} = 0.7$ when both sources are used. There are only three exceptions where Flickr data produces better predictions: *Cabañeros, Sierra Nevada and Tablas de Daimiel*. Figure A1 in the appendix shows that these three parks have a more even distribution of visitors throughout the year, with no visible maximum in summer.

The second and third columns of table 3 show the β values of the coefficients that multiply the Flickr ($\ln FUD$) and mobile phone (MPUD) data regressors in the negative binomial park-specific regressions and their significance levels according to the p-value. Note that the value of β_{Flickr} is several orders of magnitude larger than β_{Phones} because typical Flickr-user-days are much smaller than mobile-phone-user-days. The interpretability of the coefficients can be explained in terms of the average marginal effect (AME) (Cameron and Trivedi, 2013), which is proportional to the β coefficients. For example, in the case of *Timanfaya National Park*, the AME of the Flickr data regressor is $AME = 0.131 \cdot \overline{E[y|x]} = 16394$. This means that if the $\ln FUD$ increases by 1, the conditional mean, i.e., the mean number of predicted visitors, increases by 16,394 units. For mobile phone data, the AME equals $4.4 \cdot 10^{-6} \cdot \overline{E[y|x]} = 0.473$, indicating that an increase of 1,000 in MPUD is expected to result in 473 more visitors to the area.

When Flickr data are used in the negative binomial regression, β_{Flickr} is not significant in the following five parks: *Cabañeros, Monfragüe, Sierra Nevada, Sierra de las Nieves* and *Teide National Park*. However, when phone data are used in the regressions, only three parks present non-significant coefficients, as can be observed in the third column of the table: *Cabañeros, Sierra Nevada* and *Sierra de las Nieves*. β_{Flickr} is significant but the sign is negative for *Garajonay* park. A negative coefficient indicates

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

Table 3. Results of the 16 park-specific regressions

National Park	β_{Flickr}	β_{Phones}	$R^2_{DEV,Flickr}$	$R^2_{DEV,Phones}$
Aigüestortes i Estany de Sant Maurici	0.13*	3.92×10^{-6}***	0.73	0.77
Archipiélago deCabrera	0.47***	1.75×10^{-6}***	0.56	0.71
Cabañeros	0.06	3.61×10^{-6}	0.53	0.42
Caldera de Taburiente	0.18***	18.1×10^{-6}***	0.17	0.17
Doñana	0.16**	2.86×10^{-6}***	0.44	0.47
Garajonay	−0.078**	15.3×10^{-6}***	0.41	0.51
Islas Atlánticas de Galicia	0.12*	3.1×10^{-6}***	0.74	0.84
Monfragüe	−0.056	22×10^{-6}***	0.27	0.31
Ordesa y Monte Perdido	0.34***	19×10^{-6}***	0.84	0.87
Picos de Europa	0.203***	8.3×10^{-6}***	0.77	0.9
Sierra Nevada	0.019	0.24×10^{-6}	0.44	0.35
Sierra de Guadarrama	0.121***	2.4×10^{-6}	0.38	0.41
Sierra de las Nieves	-8.8×10^{-16}	1.6×10^{-6}	0.54	0.54
Tablas de Daimiel	0.21***	19.9×10^{-6}**	0.64	0.52
Teide National Park	0.034	0.57×10^{-6}***	0.45	0.69
Timanfaya	0.131***	4.4×10^{-6}***	0.74	0.86

Notes: The second column shows the value of the coefficient associated with the natural logarithm of Flickr data ($\ln FUD$) in Model 1. The third column is analogous to the second column, except that β_{Phones} is the value of the coefficient related to *MPUD* in the Model 2 regression (mobile phone data). Significance levels according to p-values are reported as ($p < 0.01$ ***1%, $p < 0.05$ **5%, $p < 0.1$ *10%). Fourth and fifth columns collect R^2_{DEV} values from figure 2.

that a smaller number of visitors is expected when the number of Flickr users in the area increases. This implies that using Flickr data to predict visitor numbers in this area is not suitable. Table 3 shows that negative binomial regressions are statistically more consistent when mobile phone data are used than when Flickr data are used.

In the fourth and fifth columns of table 3, one can note that some of the parks with non-significant coefficients exhibit moderately high goodness-of-fit in figure 2. For example, *Cabañeros* park had $R^2_{DEV} = 0.53$ for Flickr data. However, the β_{Flickr} is non-significant. In this case, the 53 per cent reduction in the deviance compared to an intercept-only model is due to the dummy variables related to seasons and the COVID-19 pandemic. Nevertheless, in any national park, for R^2_{DEV} to exceed 0.55, the coefficients related to Flickr and phone regressors must be significant. Hence, the effect of the dummy variables is not strong enough to deliver R^2_{DEV} values above this threshold.

In data analysis, a common procedure is to fit a model using a training set and then test it on a test set. In this study, a training set of data from 2021 and 2022 was employed to fit the regression and predict visitation in 2023. A plot of the measured versus estimated values for the test data are depicted in figure 3. MAPE, as defined in (equation 3), is used to evaluate the accuracy of the predictions. Now, the points in figure 3 are not as close to the 1-1 line as before. The increase in prediction errors was expected, because only a fraction of the dataset was engaged in the regression. Interestingly, some

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

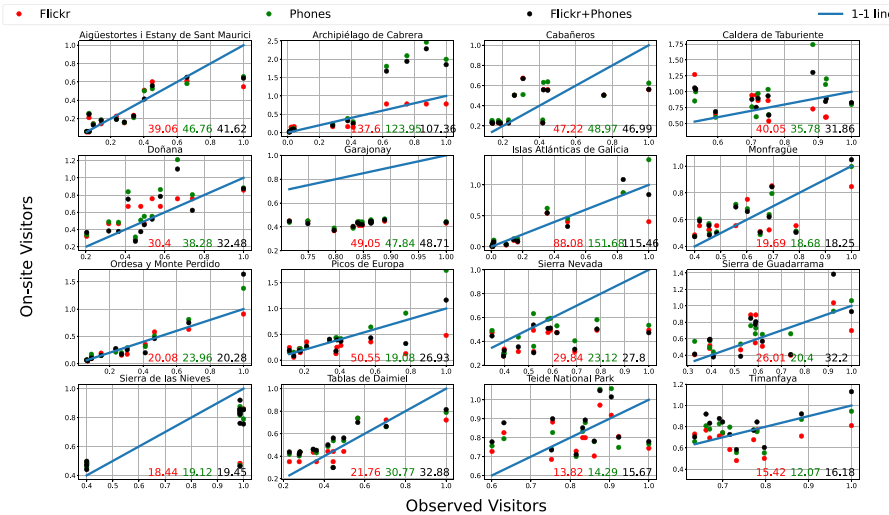

Figure 3. Park-specific NB2 regressions using three different user-generated data sources.
Notes: Flickr photographs (red, Model 1 in table 1), mobile phone data (green, Model 2 in table 1) and both of them (black, Model 3 in table 1). The NB2 model was trained using data from 2021–2022 to predict data from 2023. Visitors are normalized using the maximum number of on-site visitors at each park. MAPE is depicted in the bottom right corner of each subplot.

parks that did not exhibit high goodness-of-fit previously now have low MAPE. For example, *Sierra de Guadarrama* and *Monfragüe* presented R^2_{DEV} below 0.45 in figure 2 for any of the data sources employed. However, the MAPE values for *Monfragüe* are 19.7 per cent using Flickr data, 18.7 per cent using phone data and 18.25 per cent using both. For *Sierra de Guadarrama*, MAPE is 26.01 per cent for Flickr data, 20.4 per cent for phone data and 32.2 per cent for both sources as regressors. Conversely, *Islas Atlánticas de Galicia* had $R^2_{DEV} = 0.84$ with mobile phone data but now has a high error (MAPE=151 per cent). It is now more difficult to determine which data source is better for predicting natural recreation. Flickr data yields predictions with smaller errors than phone data in seven parks. However, in only four of these seven parks (*Ordesa y Monte Perdido, Sierra de las Nieves, Tablas de Daimiel* and *Teide National Park*) is MAPE below 25 per cent. In the other three parks, MAPE is higher than 25 per cent and the predictions are not as accurate. Mobile phone data make more accurate predictions than Flickr data in five parks. MAPE is below 25 per cent in four of them (*Picos de Europa, Guadarrama, Sierra Nevada* and *Timanfaya*). In the remaining four parks, the most effective method of predicting visitor numbers is to include both data sources. However, of these four parks, only *Monfragüe* has MAPE below 25 per cent.

When the entire dataset was used for the regression analysis, mobile phone data proved to be a better predictor than Flickr data, despite having fewer observations. When the dataset was split into training and test sets, both data sources had equal number of observations. However, in this last case, it is unclear which source is the more suitable predictor. Rather than being a limitation of the methodology, this behaviour

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

could be the consequence of data scarcity. The training set could have been larger by adding data from 2020 and earlier. However, there is likely to be a structural break between the pre- and post-pandemic years.

Overall, the user-generated data performed best in predicting visitation to parks located in the mountainous regions in the north of the country: *Aigüestortes i Estany de Sant Maurici* , *Ordesa y Monte Perdido* and *Picos de Europa*. These three parks exhibit high R^2_{DEV} and low MAPE. *Timanfaya*, located in the Canary Islands, also exhibits this behaviour. The climate in northern Spain differs from that in other regions of the country, characterized by greener and more lush landscapes. The beauty of the surrounding sea in islands may also increase visitors' willingness to share photographs on social media. However, landscape attractiveness is influenced by various subjective factors, and other reasons could underlie this finding. Furthermore, table A1 shows that the number of FUD in these four parks in northern Spain and *Teide National Park* is unexpectedly high relative to the number of visitors. In contrast, *Sierra de Guadarrama* and *Garajonay* are under-represented on Flickr. This reinforces the idea that some parks may be more visually appealing than others, leading to a higher "willingness to share" among visitors on social media and a better fit for our models. The reason why our modelling performs better in some parks than in others is difficult to determine. Tenkanen *et al.* (2017) pointed out that social media data tends to perform robustly in more visited parks. As a result, Spearman's correlations were computed between high R^2_{DEV} or low MAPE and the mean visitors of each park. Additionally, one new variable was defined:

$$\textit{Seasonality Index}_i = \frac{\text{Summer visitors}_i}{\text{Total visitors}_i}, \tag{4}$$

where i stands for each one of the 16 parks. Therefore, two statements are tested:

(1) User-generated data are more suitable for predicting visitation to the most popular parks, as previously found by Tenkanen *et al.* (2017).
(2) User-generated data are more suitable for predicting visitation in parks where the majority of visitors are concentrated in the summer.

The correlation coefficients between mean number of visitors of each park, Seasonality Index and R^2_{DEV} and MAPE values are depicted in table 4. The Spearman's correlation coefficient between the mean number of on-site visitors of each park and the MAPE values in figure 3 using mobile phone data is −0.44. The mean number of visitors is less correlated with MAPE when using Flickr data or with R^2_{DEV}. This indicates that statement 1 above is very weak. The Spearman's correlation coefficient between the Seasonality Index of each park and the R^2_{DEV} values of figure 2 is 0.46 (Flickr data) and 0.5 (phone data). The Spearman's correlation coefficient between the Seasonality Index of each park and the MAPE values in figure 3 is 0.46 for Flickr data. This last result suggests that the second statement announced above is stronger than the first. However, the limited number of observations in the dataset means that robust conclusions cannot be drawn.

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

Table 4. Correlation matrix showing Spearman's correlation coefficients between R^2_{DEV}, MAPE, total number of visitors and seasonality index (equation 4) of each park

	$R^2_{DEV,\text{Flickr}}$	$R^2_{DEV,\text{phones}}$	MAPE_{Flickr}	MAPE_{phones}	Seasonality Index	Total visitors
$R^2_{DEV,\text{Flickr}}$	1.000	0.934			0.458	0.1
$R^2_{DEV,\text{phones}}$	0.934	1.000			0.522	0.29
MAPE_{Flickr}			1.000	0.796	0.464	−0.19
MAPE_{phones}			0.796	1.000	0.304	−0.44
Seasonality Index	0.458	0.522	0.464	0.304	1.000	
Total visitors	0.1	0.29	−0.19	−0.44		1.000

5. Discussion

This study is not unique in finding that social media and user-generated data can complement on-site data. Sessions *et al.* (2016) used Flickr photographs to model visitor numbers with negative binomial regression. They also introduced ln FUD as a regressor, as has been done in the present work. They obtained a β coefficient of 0.649 for all parks pooled. This is in the same order of magnitude as most of the coefficients in table 3. Wood *et al.* (2020) found $R^2 = 0.79$ when predicting visitation to unmonitored sites using social media and calendar data in a pooled model with fixed effects. This measurement of goodness of fit aligns with our highest R^2 values in the park-specific regressions. We also reported a Spearman's correlation (ρ_s) of 0.68 between Flickr-user-days and visitor counts. Owuor *et al.* (2023) reported $\rho_s = 0.77$ in Cartinthia, Austria, also using Flickr. In a similar manner, Walden-Schreiner *et al.* (2018) found $\rho_s = 0.25$ in Argentina and Australia and Levin *et al.* (2017) measured $\rho_s = 0.5$. In a 2021 review, Wilkins *et al.* (2021) compared Spearman's correlations of 35 studies using Flickr and eight using Instagram. The highest correlation reported, for both platforms, was around 0.8. For mobile phone data, this study obtained $\rho_s = 0.33$ with on-site visitor counts, while Fisher *et al.* (2019) obtained a higher value of 0.56. Other authors (Khatibi *et al.*, 2018) tested less-explored data sources such as TripAdvisor and different techniques in 70 national parks in the US. Their MAPE values were below 25 per cent for the majority of the parks. Further research efforts could follow this approach, testing machine learning models such as Support Vector Regression (Cortes and Vapnik, 1995) or wavelet analysis (Mancini *et al.*, 2018).

The results indicate that site-specific characteristics are important when considering multiple environmental areas or national parks with different landscape, climate and socioeconomic characteristics, as visitor estimations in one area may not be applicable to others. User-generated data sources can better capture seasonal visitation patterns in countries where visits are mostly concentrated in the summer (see figure A1 in appendix A). Furthermore, estimates derived from these sources using a negative binomial regression model may be more accurate for parks exhibiting such seasonal behaviour and located in the mountainous regions in the north of the country.

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

Another topic beyond the scope of this article is the quality of on-site visitor data Red de Parques Nacionales (2017). The authorities at some parks, such as *Ordesa y Monte Perdido, Teide National Park, Aigüestortes i Estany de Sant Maurici* and *Timanfaya*, are confident in their visitor counting methodology. Visitor estimates for these parks are among the most accurate of the 16 parks studied. Conversely, the authorities of *Caldera de Taburiente* explain that most of their visitors arrive by cruise ship, resulting in a flood of tourists arriving all at once and crowding the information desks. Therefore, the priority is to reduce their waiting time, rather than precise counting. The automatic counters located at different points also do not seem to provide reliable data. The park authorities of *Sierra de Guadarrama* and *Garajonay* also report inaccurate data from their automatic counters. The inability of user-generated data to reproduce the seasonal trends observed in these three parks, as well as the inability to produce accurate estimators using negative binomial regression, indicates that data quality is a limiting factor. This finding reinforces our hypothesis that user-generated data can be a groundbreaking tool for complementing or improving traditional recreational data. This is especially true in non-protected areas in LMICs, unlike the 16 areas studied in this article. For these areas, data quality would be a much greater limiting factor.

6. Conclusions

The use of user-generated data is becoming increasingly prevalent in research and policy analysis. These datasets are particularly important when official data are scarce, as is the case with rural tourism and other recreational activities, which rely on visitor statistics to determine their economic value.

This paper provides a comparison of two popular user-generated data sources (Flickr and Mobile Phone data) in terms of statistical accuracy for the modelling of visitation to national parks in Spain. The best statistical fit of the estimated models corresponds with using both sources. If a single source needs to be used, mobile data produces the finest predictions. Despite having fewer data points, it was as effective or more effective than Flickr in many of the studied parks. However, when the sample was divided into a training set and a test set, individual differences across parks do not allow for the generalization of a single data source. User-generated data might produce more accurate predictions of visitor numbers in natural areas where most visitors are concentrated in the summer.

Our findings show that social media and user-generated data can complement on-site statistics effectively. This helps to allocate resources more effectively, inform economic valuation and protect hot spots where increased human activity poses a threat to biodiversity and conservation. Although the study was conducted in Spain, our results suggest that it would be worthwhile trying it in other locations. This is particularly relevant in low- and middle-income countries where protected areas have experienced rapid growth and there is a lack of infrastructure for counting or monitoring recreational demand.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S1355770X2510020X.

Appendix A contains all the additional figures and tables referenced in the text.

Appendix B contains a statement on the legal and ethical implications of using user-generated content for research purposes.

Acknowledgements. The authors would like to thank *Organismo Autónomo de Parques Nacionales* for providing data on the number of monthly visitors to Spain's national parks between 2015 and 2023. The authors would also like to thank the journal's editors and anonymous reviewers for their valuable comments and recommendations. The authors would like to acknowledge research support from Pablo Coello Pulido with the Python script for downloading Flickr photographs.

Funding statement. The authors would like to thank *Programa de Ciencias Mariñas de Galicia* for funding part of this research, and the project PID2022-142642OB-I00 from the State Reseach Agency .

Competing interest. None declared.

References

Balmford A, Beresford J, Green J, Naidoo R, Walpole M and Manica A (2009) A global perspective on trends in nature-based tourism. *PLoS Biology* **7**, e1000144.

Bhalla P, Bhattacharya P, Areendran G and Raj K (2022) Ecotourism spatio-temporal models to identify visitation patterns across the Indian Himalayan region. *GeoJournal* **87**, 1777–1792.

Bhatt P and Pickering CM (2022) Spatial and temporal patterns of visitation in the Annapurna Conservation Area, Nepal: Insights from geolocated social media images. *Mountain Research and Development* **42**, R16–R24.

Buckley R (2009) Parks and tourism. *PLoS Biology* **7**, e1000143.

Buckley R (2011) Tourism and environment. *Annual Review of Environment and Resources* **36**, 397–416.

Calcagni F, Amorim Maia AT, Connolly JJT and Langemeyer J (2019) Digital co-construction of relational values: Understanding the role of social media for sustainability. *Sustainability Science* **14**, 1309–1321.

Cameron AC and Trivedi PK (2013) *Regression Analysis of Count Data*. 2nd ed. Cambridge University Press.

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

Cheng X, Van Damme S, Li L and Uyttenhove P (2019) Evaluation of cultural ecosystem services: A review of methods. *Ecosystem Services* **37**, 100925.

Chun J, Kim CK, Kim GS, Jeong J and Lee WK (2020) Social big data informs spatially explicit management options for national parks with high tourism pressures. *Tourism Management* **81**, 104136.

Cortes C and Vapnik V (1995) Support-vector networks. *Machine Learning* **20**, 273–297.

Da Mota VT and Pickering C (2020) Using social media to assess nature-based tourism: Current research and future trends. *Journal of Outdoor Recreation and Tourism* **30**, 100295.

Edwards D, Jensen FS, Marzano M, Mason B, Pizzirani S and Schelhaas MJ (2011) A theoretical framework to assess the impacts of forest management on the recreational value of European forests. *Ecological Indicators* **11**, 81–89.

Fisher DM, Wood SA, Roh YH and Kim CK (2019) The geographic spread and preferences of tourists revealed by user-generated information on Jeju Island, South Korea. *Land* **8**, 73.

Ghermandi A, Camacho-Valdez V and Trejo-Espinosa H (2020) Social media-based analysis of cultural ecosystem services and heritage tourism in a coastal region of Mexico. *Tourism Management* **77**, 104002.

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

Hunter JD (2007) Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9**, 90–95.

Instituto Nacional de Estadística (2022) Measurement of outbound tourism using mobile phone positioning. Available at https://www.ine.es/en/experimental/turismo_moviles/exp_moviles_turismo_emisor_en.pdf.

Instituto Nacional de Estadística (2024) Instituto nacional de estadística. Available at https://www.ine.es (in Spanish).

https://doi.org/10.1017/S1355770X2510020X Published online by Cambridge University Press

Jaung W and Carrasco LR (2021) Using mobile phone data to examine weather impacts on recreational ecosystem services in an urban protected area. *Scientific Reports* **11**, 5544.

Keeler BL, Wood SA, Polasky S, Kling C, Filstrup CT and Downing JA (2015) Recreational demand for clean water: Evidence from geotagged photographs by visitors to lakes. *Frontiers in Ecology and the Environment* **13**, 76–81.

Khatibi A, Belem F, Silva A, Shasha D, Almeida J and Goncalves M (2018) Improving tourism prediction models using climate and social media data: A fine-grained approach. *Proceedings of the International AAAI Conference on Web and Social Media* **12**.

Kim Y, Kim CK, Lee DK, Lee HW and Andrada RIT (2019) Quantifying nature-based tourism in protected areas in developing countries by using social big data. *Tourism Management* **72**, 249–256.

Kubo T, Uryu S, Yamano H, Tsuge T, Yamakita T and Shirayama Y (2020) Mobile phone network data reveal nationwide economic value of coastal tourism under climate change. *Tourism Management* **77**, 104010.

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

Mancini F, Coghill GM and Lusseau D (2018) Using social media to quantify spatial and temporal dynamics of nature-based recreational activities. *PloS One* **13**, e0200565.

Martínez Pastur G, Peri PL, Lencinas MV, García-Llorente M and Martín-López B (2016) Spatial patterns of cultural ecosystem services provision in southern Patagonia. *Landscape Ecology* **31**, 383–399.

Mouttaki I, Khomalli Y, Maanan M, Bagdanavičiūtė I, Rhinane H, Kuriqi A, Pham QB and Maanan M (2021) A new approach to mapping cultural ecosystem services. *Environments* **8**, 56.

Mwebaze P and MacLeod A (2013) Valuing marine parks in a small island developing state: A travel cost analysis in Seychelles. *Environment and Development Economics* **18**, 405–426.

Natural Capital Project (2025) Invest 3.15.0 [Stanford University, University of Minnesota, Chinese Academy of Sciences, The Nature Conservancy, World Wildlife Fund, Stockholm Resilience Centre and the Royal Swedish Academy of Sciences]. Available at https://naturalcapitalproject.stanford.edu/software/invest.

Neuvonen M, Pouta E, Puustinen J and Sievänen T (2010) Visits to national parks: Effects of park characteristics and spatial demand. *Journal for Nature Conservation* **18**, 224–229.

Owuor I, Hochmair HH and Paulus G (2023) Use of social media data, online reviews and Wikipedia page views to measure visitation patterns of outdoor attractions. *Journal of Outdoor Recreation and Tourism* **44**, 100681.

Parsons GR (2017) Travel cost models. In Champ PA, Brown TC Boyle KJ (Eds.) *A Primer on Nonmarket Valuation*, Springer, Dordrecht, pp.187–233.

Pickering C, Olafsson AS and Hansen AS (2023) Editorial for Special Issue of the Journal of Outdoor Tourism and Recreation on social media and other user created content for outdoor recreation and nature-based tourism research. *Journal of Outdoor Recreation and Tourism* **44**, 100727.

Pozo R, Wilby M, Diaz J and Rodriguez Gonzalez AB (2022) Data-driven analysis of the impact of COVID-19 on Madrid's public transport during each phase of the pandemic. *Cities* **127**, 103723.

Python Software Foundation (2024) Python Language Reference, version 3.x. Python Software Foundation. Available at https://www.python.org/.

Red de Parques Nacionales (2017) Diagnóstico inicial del conteo de visitantes en la red de parques nacionales. Available at https://portal-miteco-stage.adobecqms.net/content/dam/miteco/es/ceneam/grupos-de-trabajo-y-seminarios/red-parques-nacionales/anexo3bis-diagnostico-conteo-rpn_tcm30-502162.pdf (in Spanish).

Red de Parques Nacionales (n.d.) Organismo autónomo de parques nacionales. Available at https://www.miteco.gob.es/en/parques-nacionales-oapn.html (in Spanish).

Rossi SD, Barros A, Walden-Schreiner C and Pickering C (2020) Using social media images to assess ecosystem services in a remote protected area in the Argentinean Andes. *Ambio* **49**, 1146–1160.

Seabold S and Perktold J (2010) Statsmodels: Econometric and statistical modeling with Python Proceedings of the 9th Python in Science Conference SciPy Proceedings. Available at https://www.statsmodels.org/v0.10.2/.

Sessions C, Wood SA, Rabotyagov S and Fisher DM (2016) Measuring recreational visitation at US national parks with crowd-sourced photographs. *Journal of Environmental Management* **183**, 703–711.

Sinclair M, Ghermandi A and Sheela AM (2018) A crowdsourced valuation of recreational ecosystem services using social media data: An application to a tropical wetland in India. *Science of the Total Environment* **642**, 356–365.

Sinclair M, Ghermandi A, Signorello G, Giuffrida L and De Salvo M (2022) Valuing recreation in Italy's protected areas using spatial big data. *Ecological Economics* **200**, 107526.

Tapia-Serrano MA, Sánchez-Oliva D, Sevil-Serrano J, Marques A and Sánchez-Miguel PA (2022) 24-h movement behaviours in Spanish youth before and after 1-year into the COVID-19 pandemic and its relationship to academic performance. *Scientific Reports* **12**, 16660.

Tenerelli P, Demšar U and Luque S (2016) Crowdsourcing indicators for cultural ecosystem services: A geographically weighted approach for mountain landscapes. *Ecological Indicators* **64**, 237–248.

Tenkanen H, Di Minin E, 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**, 17615.

Tussadiah A, Sujiwo AS, Andesta I and Daeli W (2021) Assessment of coastal ecosystem services and its condition for policy management plan in East Nusa Tenggara, Indonesia. *Regional Studies in Marine Science* **47**, 101941.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat I, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F and van Mulbregt P (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272.

Walden-Schreiner C, Rossi SD, Barros A, Pickering C and Leung YF (2018) Using crowdsourced photos to assess seasonal patterns of visitor use in mountain-protected areas. *Ambio* **47**, 781–793.

Whitehead J, Haab T and Huang J-C (eds.) (2012) *Preference Data for Environmental Valuation: Combining Revealed and Stated Approaches*, Abingdon: Routledge.

Wilkins EJ, Wood SA and Smith JW (2021) Uses and limitations of social media to inform visitor use management in parks and protected areas: A systematic review. *Environmental Management* **67**, 120–132.

Wood SA, Guerry AD, Silver JM and Lacayo M (2013) Using social media to quantify nature-based tourism and recreation. *Scientific Reports* **3**, 2976.

Wood SA, Winder SG, Lia EH, White EM, Crowley CS and Milnor AA (2020) Next-generation visitation models using social media to estimate recreation on public lands. *Scientific Reports* **10**, 15419.

Zhang H, Huang R, Zhang Y and Buhalis D (2022) Cultural ecosystem services evaluation using geolocated social media data: A review. *Tourism Geographies* **24**, 646–668.

Cite this article: Hervés-Pardavila D, Castro-Atanes A and Loureiro ML (2025) User-generated data to predict visitors in environmental areas. *Environment and Development Economics*, 1–19. https://doi.org/10.1017/S1355770X2510020X