SoftwareX 12 (2020) 100624

Contents lists available at ScienceDirect

SoftwareX

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

Original software publication

"photosearcher" package in R: An accessible and reproducible method for harvesting large datasets from Flickr

Nathan Fox [a,b,*], Tom August [b], Francesca Mancini [b], Katherine E. Parks [a], Felix Eigenbrod [a], James M. Bullock [a,b], Louis Sutter [c], Laura J. Graham [d,e]

[a] School of Geography and the Environment, University of Southampton, University Road, Southampton, SO17 1BJ, United Kingdom
[b] Centre for Ecology & Hydrology, Maclean Building, Crowmarsh Gifford, Wallingford OX10 8BB, United Kingdom
[c] Agroecology and Environment, Agroscope, Reckenholzstrasse 191, 8046 Zurich, Switzerland
[d] School of Geography, Earth and Environmental Sciences, University of Birmingham, United Kingdom
[e] Ecosystem Services and Management Program, International Institute for Applied Systems Analysis, Austria

ARTICLE INFO

Article history:
Received 25 March 2020
Received in revised form 5 August 2020
Accepted 6 November 2020

Keywords:
Biological datasets
Cultural ecosystem services
Flickr
R package
Social datasets
Social media

ABSTRACT

The social media website Flickr contains a wealth of spatial and temporal metadata, which can play an important role in environmental research including cultural ecosystem service and ecological assessments. However, the uptake of Flickr is potentially limited by issues with accessibility to the Flickr Application Planning Interface (API), which limits results and restricts searches. Here, we introduce *photosearcher*, an R package aimed at overcoming these challenges. We provide examples of how *photosearcher* can be used as an accessible and reproducible method of accessing large spatio-temporal datasets from the Flickr API.

Code metadata

Current code version	*v1.0*
Permanent link to code/repository used of this code version	https://github.com/ElsevierSoftwareX/SOFTX_2020_143
Code Ocean compute capsule	*NA*
Legal Code Licence	*GPL-3*
Code versioning system used	*git*
Software code languages, tools, and services used	*R*
Compilation requirements, operating environments & dependencies	*R (≥ 3.5.0) xml2, httr, dplyr, glue, clisymbols, crayon, sf. For compiling documentation: knitr, rmarkdown, LaTeX.*
If available Link to developer documentation/manual	https://docs.ropensci.org/photosearcher/
Support email for questions	nf2g13@soton.ac.uk

1. Motivation and significance

1.1. Scientific motivation

Biodiversity and social science datasets are key to many areas of environmental research, from understanding species distributions to guiding the management of cultural ecosystem services. Social media sites such as Flickr, Facebook, Twitter and Instagram and other online sites such as Wikipedia are becoming recognized as potential sources of data for, not only for cultural ecosystem service assessments but also increasingly for ecological questions [1–4]. First, due to high financial costs, time-intensive methods and logistical difficulties, biological datasets are often limited or incomplete across even small spatial scales [5,6]. By overcoming many of the limitations of extensive large-scale surveys, social media sites can provide large spatio-temporal datasets [7,8]. Furthermore, as natural ecosystems and protected

* Corresponding author at: School of Geography and the Environment, University of Southampton, University Road, Southampton, SO17 1BJ, United Kingdom.
E-mail address: nf2g13@soton.ac.uk (N. Fox).

https://doi.org/10.1016/j.softx.2020.100624

areas are at risk from overexploitation by people, understanding visitation rates can be useful for proactive conservation [9]. Here, social and demographic data provided by social media sites can represent actual visitation rates [2], which present opportunities to understand how humans interact with nature and how best to inform management choices relating to conservation and ecotourism.

Here, we develop an approach to accessing data from Flickr (flickr.com), an image and video hosting site with a large database of photographs accompanied by accessible metadata. Flickr has advantages as a source of data as it has an active user base with up to 25 million new uploads a day [10] and generally a wider demographic of users than other social media sites [11]. Furthermore, the photograph's metadata can be obtained by making calls to the Application Planning Interface (API), an interface for accessing the Flickr server. This metadata usually contains a georeferenced location as well as the time and date the image was taken and has the potential to be used as a primary source of data for answering ecological questions. Flickr has already been successfully used in cultural services studies, such as wildlife watching [12], recreational activities [13], landscape aesthetic qualities [14], and visitation rates in both protected areas [8] and national parks [15]. Additionally, Flickr has vast potential as a source of biodiversity data [16]. For example, it has been demonstrated as a successful tool for cross-validating Global Biodiversity Information Facility records [17] and assessing ecological niches [18]. It has been suggested that Flickr could be utilized to explore not just cultural ecosystem services, but wider ecological questions at a large scale [19]. However, due to some limitations, the potential of Flickr as a source of data for a wider range of studies has yet to be fully explored.

1.2. Current limitations

Flickr has specific limitations that need to be addressed when using it as a data source. For example, searching for photographs for a given spatial location is restricted to searching via either a bounding box or a Flickr specific location identifier. This has meant researchers have added additional steps to data manipulation in order to download image metadata for specific search boundaries [20]. Furthermore, searches for photographs through the Flickr API will only return 4,000 unique results per search criterion, limiting the ability to access data easily for spatially or temporally large searches. For searches that have more than 4,000 results, the API will appear to get metadata for all of them. However, the Flickr API only returns data for the first 4,000 images, after this the following pages of data are duplicates of the first 4,000. This means users can obtain what appears to be more than 4,000 results but end up having only the metadata for the first 4,000 unique images repeated multiple times. Some authors have limited their number of returns per query to fewer than 4,000 to get around this [21]. This workaround potentially omits the full range of data available and introduces biases, such as excluding early or new users of Flickr, or missing temporal patterns. Furthermore, the use of the API currently has limited accessibility and reproducibility. First, the API can only be accessed through a range of programming languages including Python, R and Java. To access datasets authors must be well versed in a programming language. Within R [22] there is a set of generic packages that allow harvesting data through APIs. However, researchers who want to use these packages need to have an extensive understanding of the Flickr API as well as the numerous R packages needed to call to it. Second, authors rarely provide complete methodologies or their code, limiting the ability to replicate studies. To increase the uptake of Flickr as a source of data, there is a need for an application which makes API calls more reproducible and more accessible to all.

1.3. Related work

The use of an R package for making calls to the Flickr API improves the reproducibility of studies using this data as well as giving users control over what they search. The existing R package "FlickrAPI" (cran.r-project.org/web/packages/FlickrAPI/index.html) provides some limited functionality of the Flickr API within the R environment. Other tools such as the Natural Capital Projects INVEST Recreational Tool (https://naturalcapitalproject.stanford.edu/) have also been developed to query the Flickr API. However, the FlickrAPI package only provides functions for obtaining information for a single known image and the IVEST tool only returns all images for an area. These tools do not provide functionality for searching based on criteria such as keywords or location. This, therefore, limits the functionality of these tools for ecological studies, which often require spatially explicit searches based on keywords, such as a target species. Furthermore, neither the FlickrAPI package nor the INVEST tool provides users with the functionality to download the raw images or return demographic data about Flickr users.

2. Software description

2.1. Software architecture

To overcome the challenges of using the Flickr API, we have developed the photosearcher R package (github.com/ropensci/photosearcher), aimed at facilitating reproducible requests to the Flickr API. The functions in this package make calls to the Flickr API and return both the raw photographs and their additional metadata in accessible formats, whilst overcoming the current limitations of larger spatial and temporal requests to the API.

2.2. Software functionalities

The photosearcher package provides a reproducible way of accessing geotagged photographs through search queries as well as several other functions that provide data sets useful for a range of ecological analysis. The *photo_search* function allows users to define a set of search criteria, which are then queried against the Flickr database. A data frame containing the metadata for the photographs matching the search criteria is then returned. To enable the use of Flickr across different disciplines, the *photo_search* argument *text* allows for searches to be defined by keywords. Searches for images will then only return photos that contain the keywords in their title, description or tags. Users can also limit the searches to find keywords in the photographs' tags only. As well as keywords, other search variables include minimum and maximum date the photograph was taken and a search location, provided as a bounding box, spatial layer or a Flickr specific location (where on earth identifier — woeid see: flickr.com/places/info/24865675). The ability to refine search parameters allows for a more focused approach to using Flickr's geotagged photographs by only returning those relevant to the study. The package also provides additional functionality for downloading images, getting user information and assessing related tags.

3. Illustrative examples

3.1. Spatial distribution and drivers of recreational cultural ecosystem services

The *photo_search* function returns a wealth of spatial, temporal and textural metadata. Here, we demonstrate the applications of this data by assessing recreational cultural ecosystem services, by searching for photographs of hiking in the contiguous USA.

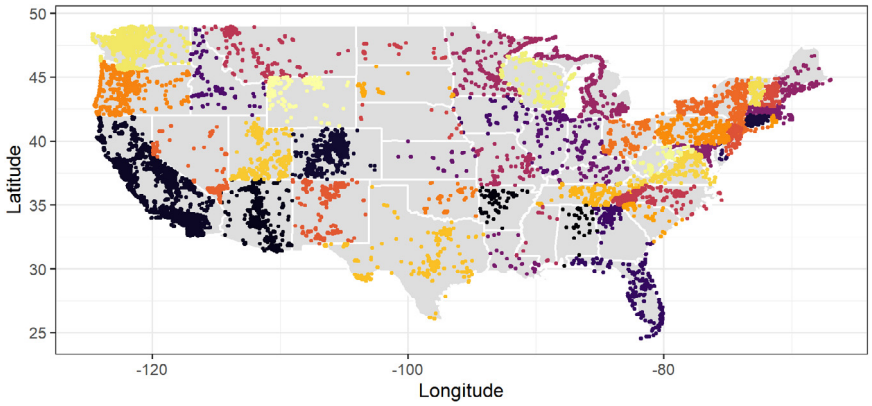

Fig. 1. Flickr photographs containing the word hiking in its title description or tag, 2015–2020 (points are coloured by state). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We then used the results of this search with *user_info* function to obtain social information on each of Flickr users. The general code is as follows (for a reproducible version see SI. 1):

```
area_photos <- photo_search(mindate_take = "2015-01-01", maxdate_taken = "2020-01-01", maxdate_uploaded = "2020-01-01", sf_layer = USAboundaries::us_states())
social_data <- user_info(user_id = area_photos$owner)
```

The *photo_search* function returned 160,923 photographs for hiking in the USA between 2015 and 2020 in 61 minutes (Fig. 1). To return metadata for this large number of photographs the bare minimum number of necessary calls to the Flickr API would be 644 (250 photographs per search). The *photo_search* function therefore makes a minimum of 10.55 calls a minute to the API returning metadata for approximately 2,637 photographs (NB in order to minimize errors the *photo_search* makes more than the minimum number of calls).

Like the *photo_search* function, *user_info* typically returns large social datasets in short periods of time. Here, the *user_info* function took just under 24 minutes to return information on 6,514 individuals, about 271 users per minute. Normally, to get a users' information you have to make a new call to the API for each individual, however, *user_info* function allows searches for multiple users at once, returning all available social data including hometown and occupation. The *user_info* function, therefore, provides an efficient method for obtaining large social datasets. Potential uses for the city datasets include network analysis to track travel route as well as to understand the social-economic drivers of supply and demand for cultural ecosystem services. By being able to assess rapidly where visitors travel from, protected area managers can inform visitor management plans. The social datasets could also be combined with ecological datasets for studies such as understanding human–wildlife interactions or ecotourism management. The hometown information can be plotted by geocoding their location with functions such as *geocode_OSM* function in the *tmap* R package (cran.r-project.org/web/packages/tmap/) (Fig. 2).

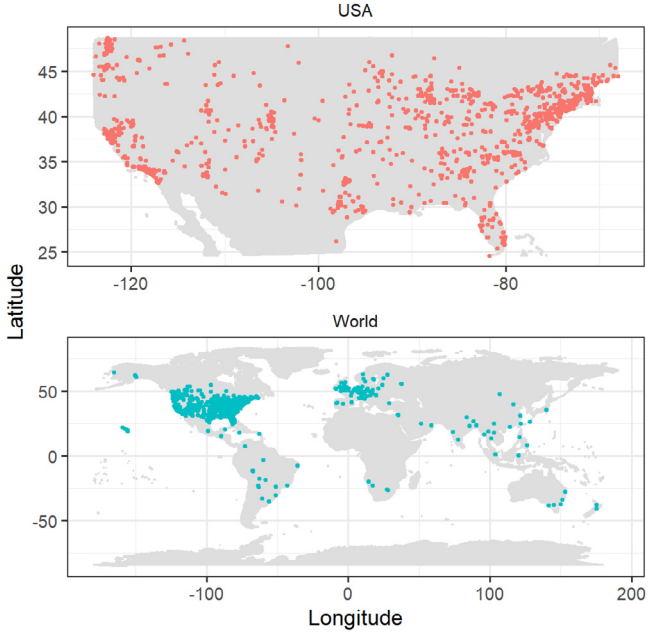

Fig. 2. Geocoded hometowns of people undertaking hiking in the USA, 2015–2020.

Table 1
Search terms used, the number of results, time taken to search for the results, and number of API calls needed.

Species	Text search	Results returned	Time
Barn owl — *Tyto alba*	Common name	17,436	10.14 minutes
	Latin name	3,529	1.12 minutes
Red fox — *Vulpes Vulpes*	Common name	25,225	14.14 minutes
	Latin name	7,793	3.14 minutes
Brown bear — *Ursus arctos*	Common name	21,555	10.40 minutes
	Latin name	5,170	1.53 minutes

3.2. Spatial and temporal distribution of species

To demonstrate the ease of using the photosearcher package for obtaining large ecological datasets, we utilize the *photo_search* function to find images metadata containing either the common or Latin names of a number of species (Table 1). The Flickr metadata can contain the complete date and time data, allowing for investigation of temporal distributions such as migratory patterns, diurnal cycles and floral phenology. Flickr may be best suited to large charismatic species that are easily identifiable by the public, such as some birds [16]. The following piece of code outlines the basic search used (for a reproducible document see SI. 1).

```
species_name <- photo_search(mindate_take = "2000-01-01", maxdate_taken = "2020-01-01", maxdate_uploaded = "2020-01-01", text = <species common or Latin name>, has_geo = TRUE)
```

The *photo_search* function was able to return large datasets in short periods of time — i.e. returning 25,225 unique geotagged data points globally for the red fox in just over 14 minutes. These

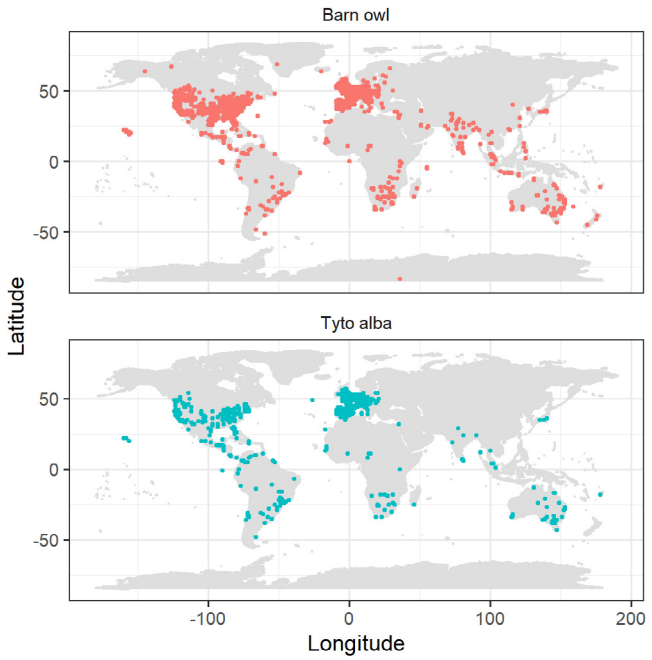

Fig. 3. Spatial distribution of Flickr photographs of barn owls.

results reiterate that generally this method does not result in exceptionally long search times. Furthermore, the results demonstrate that large spatial and temporal searches would require a large number of API calls, for example, a global study for barn owls would require 70 calls to the API and searches for brown bears would require 87. As *has_geo* = *TRUE*, the returned metadata contains a latitude and longitude information, here we map the distributions of the photographs tagged with species names (Fig. 3). Users should be aware that species distributions based on Flickr photographs may have erroneous points. First, Flickr users may misidentify species. To overcome the issue of mistagged images users should properly define their search criteria i.e. using the Latin name, or with a shapefile of its known distribution, or users can use classification techniques to confirm which photographs have positive sightings. Second, some distributions can be influenced by visitor attractions such as zoos and museums. These erroneous points can be removed using *CoordinateCleaner* R package (cran.r-project.org/web/packages/CoordinateCleaner/index.html). Furthermore, the temporal metadata can be used to assess change in species over time (Fig. 4). Here we demonstrate that sightings of brown bears vary monthly, with fewer sightings occurring during periods of known hibernation. This temporal metadata could be combined with the spatial data to assess migratory patterns, or with photograph contents (accessible via the *download_images* function) to assess animal behaviour or plant phenology.

4. Impact

photosearcher provides a more accessible and reproducible method of accessing the Flickr API, as well as overcoming limitations that prevent researchers from obtaining datasets. By creating *photosearcher* within the R environment it is freely available to all researchers. Furthermore, by consolidating the code into user-friendly functions the *photosearcher* package expands the accessibility of the Flickr dataset to non-data scientists. The simple functions also allow researchers to share their methods in a transparent and reproducible manner. However, we note that as people can add new uploads, edit metadata or delete their images, a search for the same criteria on two different occasions may return a different number of results. By providing arguments for limiting searches by the date they were uploaded the *photo_search* function helps to minimize any changes between repeated searches. This – combined with the ability to share the arguments used in the function calls or a full reproducible document (SI. 1) – makes *photosearcher* well suited to producing replicable results when working with Flickr data.

The *photosearcher* package allows researchers to obtain the full range of data available. To overcome the API limit of 4,000 results per query, *photos_search* requires the user to provide a minimum and maximum search date for when the photographs were taken. If the number of photographs matching the users defined criteria is less than 4,000, the metadata is returned. However, if the number of photographs is greater than 4,000 the metadata for the first 4,000 photographs are returned chronologically. The function then extracts the maximum date on which these images were taken and carries out a new search using this as the *min-date_taken* argument. The function does not assume that the new search contains fewer than 4,000 images and therefore checks whether the new search contains more than 4,000 results. In this way, the package will continue to dynamically split the initial search into new searches until it returns all available unique images from the initial search. The only time where all data may not be returned is if there were more than 4,000 images for a given second. As this process is automated it means users do not have to make additional calls manually to test which range of dates will return fewer than 4,000 results. Through using an automated method of splitting the searches, the *photo_search* function provides users with time and cost-efficient method of data collection. Furthermore, unlike the other software such as the INVEST tool, the *photo_search* function returns the full available metadata available for each photograph. This metadata can be useful for novel research by helping filter results to overcome some of the limitations of social media data. For example, by returning a Flickr-derived measure of spatial accuracy, users of the photosearcher package can quickly filter the returned results based on the accuracy of the spatial reference. Moreover, the anonymous user ID allows users to calculate visitation metrics such as photo–user–days [2], to overcome bias introduced by very active users. We have also provided an option to allow to supply a shapefile to search for a specific area. The *photo_search* function automatically transforms the provided shapefile to a bounding box which is then sent to the Flickr API to search for photographs. The function then extracts and returns only the responses from the original shapefile.

The other functions available in photosearcher are also designed to be useful in novel ecological assessments. For example, by returning the ID of the user uploading the images, additional analyses can be carried out using their publicly available data, returned by function *user_info*. Furthermore, the *download_images* function allows users to download the images themselves, which could be used for additional analysis or validation. The returned images could be classified by hand or through machine learning techniques to answer a range of ecological questions including the distribution of ecosystem services [19] and identifying plant species [23]. The plant species data set [23] was derived from the outputs of the *photo_search* function.

5. Conclusions

The R package *photosearcher* provides an easily accessible and reproducible method for accessing large datasets from Flickr. The simple skill set needed to use the photosearcher package will increase opportunities for use of Flickr data by non-data scientists. By addressing the challenges and limitations associated with

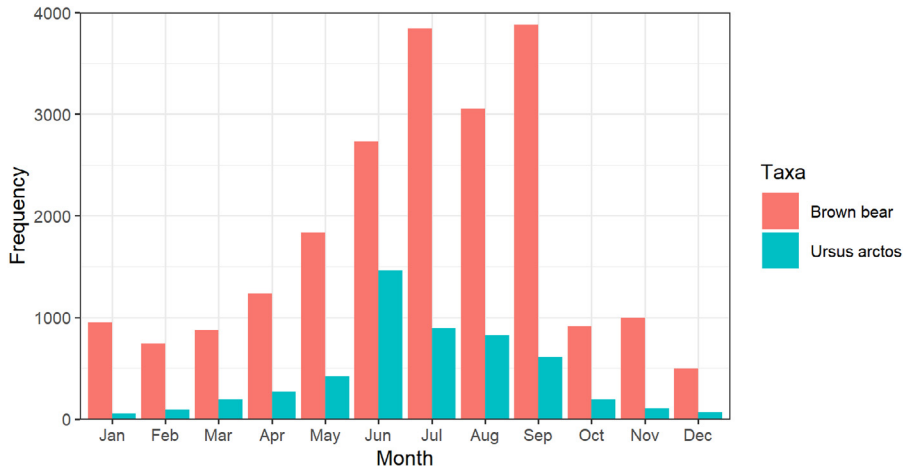

Fig. 4. Temporal distribution of Flickr photographs of brown bears. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

API access photosearcher provides the basis for a standardized method for API calls. The *photosearcher* package provides both a quick and inexpensive method of gathering large quantities of data, with the methods presented here demonstrating how the package can help provide extensive biological and social data. We hope that the package allows future studies to build upon the current use of Flickr in cultural ecosystem service research, whilst facilitating users to answer a wider array of ecosystem service and ecological questions

CRediT authorship contribution statement

Nathan Fox: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft. **Tom August:** Conceptualization, Methodology, Software, Writing - review & editing. **Francesca Mancini:** Conceptualization, Methodology, Software, Writing - review & editing. **Katherine E. Parks:** Conceptualization, Methodology, Software, Writing - review & editing, Supervision, Funding acquisition. **Felix Eigenbrod:** Conceptualization, Methodology, Software, Writing - review & editing, Supervision. **James M. Bullock:** Conceptualization, Methodology, Software, Writing - review & editing, Supervision, Funding acquisition. **Louis Sutter:** Conceptualization, Methodology, Software, Writing - review & editing. **Laura J. Graham:** Conceptualization, Methodology, Software, Writing - review & editing, Supervision.

Declaration of competing interest

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

Acknowledgements

This work was supported by the Natural Environmental Research Council [grant number NE/L002531/1] and by the Centre for Ecology and Hydrology, United Kingdom [grant number NEC06895]

References

[1] Li L, Goodchild MF, Xu B. Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. Cartogr Geogr Inf Sci 2013;40(2):61–77. http://dx.doi.org/10.1080/15230406.2013.777139.

[2] Wood SA, Guerry AD, Silver JM, Lacayo M. Using social media to quantify nature-based tourism and recreation. Sci Rep 2013;3. http://dx.doi.org/10.1038/srep02976.

[3] Byczek C, Longaretti PY, Renaud J, Lavorel S. Benefits of crowd-sourced GPS information for modelling the recreation ecosystem service. PLoS One 2018;13(10):1–23. http://dx.doi.org/10.1371/journal.pone.0202645.

[4] Mittermeier JC, Roll U, Matthews TJ, Grenyer R. A season for all things : Phenological imprints in wikipedia usage and their relevance to conservation. PLOS Biol 2019;17(3):e3000146. http://dx.doi.org/10.1371/journal.pbio.3000146.

[5] Hjort J, Heikkinen RK, Luoto M. Inclusion of explicit measures of geodiversity improve biodiversity models in a boreal landscape. Biodivers Conserv 2012;21(13):3487–506. http://dx.doi.org/10.1007/s10531-012-0376-1.

[6] Wetzel FT, Bingham HC, Groom Q, et al. Unlocking biodiversity data : Prioritization and filling the gaps in biodiversity observation data in Europe. Biol Conserv 2018;221(2017):78–85. http://dx.doi.org/10.1016/j.biocon.2017.12.024.

[7] van Zanten B, Van Berkel DB, Meentemeyer RKT, Smith JW, Tieskens KF, Verburg PH. Continental-scale quantification of landscape values using social media data. Proc Natl Acad Sci 2016;113(46):12974–9. http://dx.doi.org/10.1073/pnas.1614158113.

[8] Kim Y, Kim C, Kun D, Lee H, Ii R, Andrada T. Quantifying nature-based tourism in protected areas in developing countries by using social big data. Tour Manag 2019;72(2018):249–56. http://dx.doi.org/10.1016/j.tourman.2018.12.005.

[9] Hadwen BWL, Hill W, Pickering CM. Icons under threat : Why monitoring visitors and their ecological impacts in protected areas matters. Ecol Manag Restor 2007;8(3):177–81. http://dx.doi.org/10.1111/j.1442-8903.2007.00364.x.

[10] Ding X, Fan H. Exploring the distribution patterns of flickr photos. Int J Geo-Inf 2019;8(9):418. http://dx.doi.org/10.3390/ijgi8090418.

[11] Oteros-rozas E, Martín-lópez B, Fagerholm N, Bieling C, Plieninger T. Using social media photos to explore the relation between cultural ecosystem services and landscape features across five European sites. Ecol Indic 2018;94:74–86. http://dx.doi.org/10.1016/j.ecolind.2017.02.009.

[12] Mancini F, Coghill GM, Lusseau D. Quantifying wildlife watchers ' preferences to investigate the overlap between recreational and conservation value of natural areas. J Appl Ecol 2019;56:387–97. http://dx.doi.org/10.1111/1365-2664.13274.

[13] Graham LJ, Eigenbrod F. Scale dependency in drivers of outdoor recreation in England. People Nat 2019;1(June):406–16. http://dx.doi.org/10.1002/pan3.10042.

[14] Figueroa-alfaro RW, Tang Z. Evaluating the aesthetic value of cultural ecosystem services by mapping geo-tagged photographs from social media data on Panoramio and Flickr. J Environ Plan Manag 2017;60(2):266–81. http://dx.doi.org/10.1080/09640568.2016.1151772.

[15] Tenkanen H, Di Minin E, Heikinheimo V, et al. Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. Sci Rep 2017;7(1):1–11. http://dx.doi.org/10.1038/s41598-017-18007-4.

[16] Barve V. Ecological informatics discovering and developing primary biodiversity data from social networking sites : A novel approach. Ecol Inform 2014;24:194–9. http://dx.doi.org/10.1016/j.ecoinf.2014.08.008.

[17] Wittich HC, Seel M, Wäldchen J, Rzanny M, Mäder P. Recommending plant taxa for supporting on-site species identification. BMC Bioinformatics 2018;19:190. http://dx.doi.org/10.1186/s12859-018-2201-7.

[18] Peña Aguilera P, Burguillo-Madrid L, Barve V, Aragon P, Jimenez-Valverde A. Niche segregation in Iberian Argiope species Niche segregation in Iberian Argiope species. J Arachnol 2019;47:37–44. http://dx.doi.org/10.1636/0161-8202-47.1.37.

[19] Richards DR, Tunçer B. Using image recognition to automate assessment of cultural ecosystem services from social media photographs. Ecosyst Serv 2018;31:318–25. http://dx.doi.org/10.1016/j.ecoser.2017.09.004.

[20] Lee H, Seo B, Koellner T, Lautenbach S. Mapping cultural ecosystem services 2. 0 - potential and shortcomings from unlabeled crowd sourced images. Ecol Indic 2019;95:505–15. http://dx.doi.org/10.1016/j.ecolind.2018.08.035.

[21] van Zanten BT Van, Berkel DB Van, Meentemeyer RK, Smith JW, Tieskens KF. Continental-scale quantification of landscape values using social media data. Proc Natl Acad Sci USA 2016;113(46):12974–9. http://dx.doi.org/10.1073/pnas.1614158113.

[22] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019, URL https://www.R-project.org/.

[23] August T, Affouard A, Bystriakova N, et al. AI validated plant observations from social media: Flickr images from central London 2011-2019 (version 1.1) [data set]. Zenodo 2019. http://dx.doi.org/10.5281/zenodo.3514685.