

Reddit

Fox, Nathan; Graham, Laura; Eigenbrod, Felix; Bullock, James M.; Parks, Katherine E.

DOI:

[10.1016/j.ecoser.2021.101331](https://doi.org/10.1016/j.ecoser.2021.101331)

License:

Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

Document Version

Peer reviewed version

Citation for published version (Harvard):

Fox, N, Graham, L, Eigenbrod, F, Bullock, JM & Parks, KE 2021, 'Reddit: a novel data source for cultural ecosystem service studies', *Ecosystem Services*, vol. 50, 101331. <https://doi.org/10.1016/j.ecoser.2021.101331>

[Link to publication on Research at Birmingham portal](#)

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

1 **Reddit: A novel data source for cultural ecosystem service studies**

2 **Abstract**

3 Social media sites have been gaining traction as a source of novel data for environmental research,
4 particularly for cultural ecosystem service (CES) assessments. However, Reddit, a discussion-based
5 site, has yet to establish itself as an important source of data for CES research, possibly due to
6 researchers not being aware of its potential applications or because Reddit posts lack georeferencing
7 information. Here, we demonstrate how researchers can search Reddit for CES datasets related to
8 recreation and how specific pages on Reddit may provide data for other CES such as aesthetics.
9 Using named-entity recognition, we developed an automated method of geocoding the approximate
10 location of where images in Reddit posts were taken. Furthermore, we compare posts from Reddit
11 and Flickr for a range of recreational activities and compare the content and textual metadata of
12 images relating to hiking. Though there is potential for Reddit data to be used in spatial analysis, we
13 highlight the limitations associated with georeferencing posts. We recommend that data from
14 Reddit is best suited to assessing general trends in CES, either for a given service or place. By
15 demonstrating the value of big data from Reddit we hope to encourage its inclusion in future CES
16 and environmental research.

17 **Key words:** Cultural ecosystem services, social media, Reddit, Flickr, aesthetic values, recreation

18 **1.0 Introduction**

19 Big data from social media sites has multiple benefits over conventional methods of data collection
20 for environmental studies, providing access to large spatio-temporal scale datasets, through
21 inexpensive and quick data collection methods (Barve 2014). The use of social media data is
22 therefore becoming more prominent in environmental research, ranging from the use of Twitter to
23 understand animal life-cycles (Hart et al. 2018) and prepare for natural hazards (Wang et al. 2016;
24 Mendoza et al. 2019), to Flickr being used to assess species niches (Peña-Aguilera et al. 2019) and
25 map invasive species (Allain 2019). One of the biggest applications for social media data in
26 environmental research has been the assessment of cultural ecosystem services (CES) (Ghermandi
27 and Sinclair 2019).

28 CES are the non-material benefits obtained through nature and are derived from the interaction of
29 biodiversity (biotic nature) and geodiversity (abiotic nature) (Gray 2011; Fox et al. 2020a). CES
30 include aesthetic value, recreational services and sense of place and can enhance physical and
31 mental well-being (Haines-Young and Potschin 2010). They can deliver multiple benefits for both
32 residents and tourists, supporting local and regional economies (Schirpke et al. 2016; King et al.
33 2017). However, the exploitation, destruction and consumption of natural landscapes by humans for
34 activities such as intensive agriculture, urban development and recreational activities can damage
35 ecosystems and reduce their capacity to provide CES (Figueroa-Alfaro and Tang 2017). Furthermore,
36 our understanding of CES is more limited than that of provisioning and regulating ecosystem services
37 (Milcu et al. 2013; Díaz et al. 2018), particularly because our interactions with CES are subjective and
38 vary between individuals, which makes obtaining practical measurements of their benefits and
39 values difficult (Daniel et al. 2012; Havinga et al. 2020). By developing a better understanding of the
40 natural and social drivers of CES we can help inform policy and management strategies to alleviate
41 the threats to their sustainable use (Clemente et al. 2019). Researchers, therefore, need to
42 understand better the supply and demand of these services over relevant temporal and spatial
43 scales (Langemyer et al. 2018). Here, social media datasets provide relatively quick and cost-
44 effective data collection for assessing CES, versus traditional methods, and provide novel approaches

45 to assessing how CES are generated as well as their perceived benefits and values over a range of
46 spatial and temporal scales (Wood et al. 2013; Ghermandi and Sinclair 2019; Fox et al. 2020b).

47 Due to the vast quantity of data available on social media websites can be viewed as a source of big
48 data and therefore benefit from the emergence of big data approaches to assessing human-nature
49 relationships (Retka et al. 2019). Social media sites, including Twitter and Weibo (microblogging
50 sites), Flickr, Instagram and Panaramio (image sharing sites), have already been widely used to
51 assess a range of CES. Aesthetic value has been assessed through textual metadata from Twitter
52 (Johnson et al. 2019), image and geographic distribution from Instagram (Guerrero et al. 2016; Chen
53 et al. 2020), and image content and geographic distribution from Flickr (Figueroa-Alfaro and Tang
54 2017; Tieskens et al. 2018). Recreational preferences have been studied using Flickr (van Zanten et
55 al. 2016; Graham and Eigenbrod 2019; Gosal et al. 2019) and Weibo (Zhang and Zhou 2019).
56 Furthermore, Flickr has also been used to assess changes in cultural values over time (Thiagarajah et
57 al. 2018) and identify trade-offs between multiple CES (Allan et al. 2015). However, some social
58 media sites have either ceased operating (e.g. Paramio) or introduced restrictions to accessing data
59 (e.g. Instagram) and therefore Flickr is becoming the main source of data for CES studies
60 (Langemeyer et al. 2018; Retka et al. 2019).

61 Metadata available from Reddit, the social news aggregation and discussion orientated social media
62 website, has been used in a vast array of scientific studies across a range of disciplines (Baumgartner
63 et al. 2020), including health and psychology (e.g. Jamnik and Lane 2017; Park et al. 2018),
64 technological development (e.g. Derczynski et al. 2017; Volske et al. 2017) and political studies (e.g.
65 Guimaraes et al. 2019). Reddit, which is broken up into different forums or "subreddits" themed
66 around different topics, allows for user to post a range of media such as images and text posts.
67 These posts, along with their associated metadata, draws parallels to the types of data from other
68 social media sites that are currently used in CES studies. However, there appears to have been little
69 to no application of big data from Reddit to assess any ecosystem service. A systematic review of the
70 applications of social media data in environmental research did not include any studies using Reddit
71 as a source of data (Ghermandi and Sinclair 2019). A search of the titles abstracts and keywords on
72 Web of Knowledge (<https://wok.mimas.ac.uk>) and Scopus (www.scopus.com) for "ecosystem
73 servic*" (the * denotes any end to the term e.g. service or services) AND "Reddit", carried out on
74 10th February, 2021, returned no results.

75 As there have been few studies comparing social media sources for CES, there is a need for a greater
76 understanding of the impacts of differences in data availability and biases among the various social
77 media sites used as data sources (Ostera-Roza et al. 2018). We therefore find it surprising that big
78 data from Reddit has yet to be explored in the context of CES, though we postulate that this may be
79 for two key reasons: researchers not being familiar with the website and its potential uses; and that
80 posts on the website are not georeferenced. In this paper, we aim to provide an overview of Reddit,
81 and to compare data from the site with that from another social media site, Flickr. We provide
82 examples of how data from Reddit can be used to assess recreational, aesthetic, spiritual and
83 cultural CES and address how Reddit can be a novel source of data for commonly used CES methods
84 such as assessing image contents and textual sentiment. We also provide an insight to the potential
85 uses and limitations of Reddit for spatial assessments.

86 **2.0 Methods**

87 Here we present multiple methods for searching Reddit for data suitable for CES assessments via its
88 Application Programming Interface (API), a computing interface that allows researchers to access a
89 platform via code. First, we searched the site for all posts containing a specific keyword and

90 compared these posts to those found using the same keyword search on Flickr. Second, we searched
 91 for posts on subreddits that are based around topics of interest to CES research. Third, we
 92 demonstrate a method for geocoding an estimated location for posts from Reddit as well as
 93 combining a place keyword search with another keyword, or within a subreddit to find posts linked
 94 to a particular location.

95 *2.1 Data sources*

96 *2.1.1 Reddit*

97 Reddit is a social media site consisting of over two million different communities called subreddits
 98 (Table 1). Subreddits are built around a topic, each with their own rules on posts and comments. The
 99 type of post is highly variable among subreddits. For example, the subreddit “r/EarthPorn” is limited
 100 to photographs of landscapes, accompanied by a text title and a comment section, whereas the
 101 subreddit “r/Culture” hosts primarily text-based posts with a title and a comment section.

102 Table 1. Selected examples of subreddits linked to cultural ecosystem services.

Service	Subreddit	Extract of group description	Number of members (10 th February 2021)	Primary metadata type
Aesthetic views	r/EarthPorn	“EarthPorn is your community of landscape photographers and those who appreciate the natural beauty of our home planet.”	20.9m	Images
	r/BotanicalPorn	“High quality images of plants (fungi are allowed!).”	167k	Images
Recreational activities	r/Outdoors	“Outdoors is for *all* outdoor experiences, not limited to any specific interest. Caving, mountain climbing, cycling, bushcraft, gardening, sailing, plants, birds, trees, going for a stroll -- it's all on topic here!”	2.7m	Images
	r/Hiking	“The hikers' subreddit.”	1.3m	Images
Tourism	r/Travel	“r/travel is a community about exploring the world. Your pictures, questions, stories, or any good content is welcome.”	5.7m	Images and text
Spirituality and sense of place	r/Spirituality	“Here, we discuss such things as personal transformation, the meaning of life, death, and moments of clarity.”	190k	Text
	r/Culture	“A subreddit dedicated to sharing and discussing the many aspects of culture”	6.3k	Text

103

104 Posts and comments from Reddit can be searched and returned through the Reddit API, with text
 105 and image posts, as well as their metadata including the, title, comments, date posted, how many
 106 upvotes (the number of people that like a post) a post has, and the ratio of upvotes to downvotes

107 (the number of people that dislike a post). These data types are similar to data already being used in
108 CES and social media studies derived from Flickr, Instagram and Twitter.

109 Accessing data from Reddit has multiple benefits for researchers. First, data from Reddit is freely
110 available. Second, the data is accessible across multiple software tools and programming languages.
111 For example, the Pushshift tool (Baumgartner et al. 2020) provides researchers with an accessible
112 method for querying and retrieving data. The tool also benefits researchers by providing built-in
113 functionality which overcomes Reddit's 100 object limit per search. For researchers familiar with
114 writing scripts, functionality for searching the Reddit API is available in multiple programming
115 languages: packages "RedditExtractor" (Rivera 2019) and "rreddit" (Kearney 2020) for the R
116 environment; "Python Reddit API Wrapper" (Boe 2020) within the Python environment; "jReddit"
117 (jReddit 2020) within the Java environment.

118 2.1.2 Flickr

119 Flickr is a popular social media site that hosts images and videos with up to 25 million uploads a day
120 (Ding and Fan 2019). Flickr has a broad user base, with a range of motivations for uploading
121 photographs (Oteros-Rozas et al. 2018), and therefore has potential as a source of data for a wide
122 range of CES. Posts on Flickr can have associated metadata that includes textual titles, description
123 and tags; spatial location in the form of latitude and longitude of where the image was taken; and
124 the time and date the image was taken. Flickr metadata is accessible through tools such as the
125 "photosearcher" package in the R environment (Fox et al. 2020b), and stand-alone software such as
126 the InVEST Recreational tool (Sharp et al. 2020).

127 2.2 Data collection and analysis

128 *A reproducible R file for the data collection methods has been included in the supplementary*
129 *material. To comply with API terms and privacy policies all data sets were anonymised, stored with*
130 *multiple layers of security and any unnecessary metadata was deleted.*

131 2.2.1 Keyword search

132 First, to find posts related to recreational activities, we searched the Pushshift tool (Baumgartner et
133 al. 2020) for any posts on Reddit containing a single keyword for four different activities; "hiking",
134 "camping", "skiing" and "kayaking", found in any textual metadata uploaded by the user e.g. the
135 posts title or description. We also constrained the search to any posts that were uploaded between
136 the 1st of January 2020 and the 1st of January 2021. We then repeated this query on Flickr, using the
137 photosearcher R package (Fox 2020b), ensuring that we made a comparable search using the same
138 keywords, again found in any of the posts textual metadata, and within the same uploaded date
139 range. We summarized the number of uploads per month as well as the mean character length of
140 the posts title and text, and the mean number of likes and comments on the images for each activity
141 across platforms. Furthermore, as posts on Reddit can be in a range of formats other than images
142 and text traditionally used in CES studies (e.g. links to other websites or videos), we calculate the
143 percentage of posts that were images or text.

144 To compare the contents of the images posted on the two sites, we took a random sample of 1,000
145 images related to hiking from both sites (images listed as adult material were not included in the
146 sample selection). The contents of the images were automatically tagged using the Google Cloud
147 Vision API (Google Cloud Vision 2020). The Google Cloud Vision API is a machine learning algorithm
148 that labels the content of images. The algorithm is based on a large pre-trained dataset and can label
149 image contents into millions of predefined categories including objects and expressions. Here, we

150 used the “imgrec” R package (Schwemmer 2019) to label each image with the 10 objects the
 151 algorithm first detects. To ensure that the image contents were accurate without manual validation
 152 we only kept labels that had a confidence score of > 0.6 (Gosal et al. 2019).

153 To compare the hiking images from Reddit and Flickr we used a chi-square test to compare the two
 154 sources of data in terms of their image content (frequency of Google Cloud Vision API labels). As the
 155 dataset is relatively large, some statistical tests may indicate statistical significance ($p < 0.05$)
 156 irrespective of real-world significance in the data. Furthermore, statistical significance does not
 157 provide information on the size of the effect (Kim 2017). Here, we primarily focus on the individual
 158 contribution of features (x^2_i eq. 1) to the total effect size $x^2 = \sum x^2_i$, enabling us to understand
 159 better the difference between the two datasets (Oakes and Farrow 2006).

160
$$x^2_i = \frac{(\text{obs}_i - \text{exp}_i)^2}{\text{exp}_i} \quad (1),$$

161 where obs_i and exp_i are the observed and expected values of feature i , respectively.

162 As textual metadata can be useful for understanding characteristics of CES or eliciting emotion of
 163 CES beneficiaries (Brindley et al. 2019; Hale et al. 2019), we also returned textual metadata for
 164 analysis for the random sample of hiking images. As images uploaded to Reddit can only contain a
 165 title, with no description text, the most comparable source of textual data for images from Flickr and
 166 Reddit are the comment sections. We summarised the number of comments for these images as
 167 well as the number of unique users interacting with the posts. The sentiment expressed in each
 168 comment was calculated using the Afinn dictionary (Nielsen 2011), which has previously been used
 169 to assess the sentiment value expressed in social media text posts (Koto and Adriani 2015). This
 170 dictionary ranks words on a -5 (negative sentiment) to +5 (positive sentiment) scale. The sum
 171 sentiment of each post was calculated, and the mean sentiment score of the posts on each site
 172 calculated. We also filtered out automated messages, weblinks and commonly used words such as
 173 “the” and “is” and calculated the most frequently used words in comments on the two sites.

174 *2.2.2 Subreddit Search*

175 A unique aspect to the Reddit API is the ability to search individual subreddits. Here, we searched
 176 four subreddits that are themed around the aesthetic value of nature; “r/EarthPorn”,
 177 “r/BotanicalPorn”, “r/WaterPorn” and “r/DesertPorn”, as well as posts from two subreddit about
 178 two recreational activities (“r/Birding” and “r/Scuba”) and two subreddits that discuss spirituality
 179 and culture (“r/Spirituality” and “r/Culture”). The results were limited to posts uploaded in the year
 180 2020. The aesthetic views subreddits have a set of rules that mean all posts on the subreddit are of
 181 photographs pertaining to nature. Table 2 summarises the submission rules for the “r/EarthPorn”
 182 subreddit, these rules are similar across the other aesthetic subreddits assessed, though the subject
 183 of the photograph differs. The rules for the recreational and spiritual subreddits allow for both
 184 images and discussion-based posts. To compare the contents of the images posted different
 185 subreddits, we took a random sample of 1,000 images posted on “r/EarthPorn” and
 186 “r/BotanicalPorn”. These images were then automatically tagged using the Google Vision Cloud API
 187 and the contents of the two sets of images were compared using a chi-square test.

188 Table 2. Selected rules for submissions to “r/EarthPorn” (as of the 10th February 2021).

Rule	Description
A photograph	“No Paintings, illustrations, gifs, videos, or interactive images.”

An image featuring a natural landscape	"Images must have visible land. Images with humans, machines, boats, roads, airplanes, farms, animals, buildings, or other man made objects in them will be removed."
A photograph you took (OC)	"Or one which you can provide and post the original source for. Do not rehost non OC images to reddit or imgur."
An unsilhouetted image	"Images where details in the landscape are not visible due to silhouetting will be removed."
The location of the area in the photo	"When it comes to location, the more specific the better. If you wish to not disclose the location you should at the very least name the state/country. Rule of thumb for naming only the location (e.g. a lake, mountain): if one can find the place immediately by searching it in google it's fine. For possibly ambiguous locations add state/country for safety."

189

190 2.2.3 Potential spatial uses for Reddit

191 As Reddit posts are not geolocated, it is not possible to directly map the distribution of the CES
192 expressed in the posts. Instead, we developed an automated method for estimating the
193 approximate location of images posted to Reddit, following a similar method to Harrington (2018).
194 The subreddit "r/EarthPorn" requires that posts must contain the image location in the title. To
195 extract the location name, we used named-entity recognition, a technique that classifies words in a
196 text into predefined categories, one of which is a named location, on the 1,000 images randomly
197 sampled from "r/EarthPorn" (Alfred et al. 2014). Named-entity recognition was carried out using the
198 "entity" R package (Rinker 2015). A subset of 10% of the name-entities were manually validated by
199 comparing the returned name-entity with the post title. The extracted location names were then
200 geolocated using the Google Maps API through the "ggmap" R package (Kahle and Wickham 2013).
201 Based on the place name, the Google Maps API provides an estimated latitude and longitude. The
202 global distribution of both sets of data was mapped and the percentage of uploads from each
203 continent was calculated.

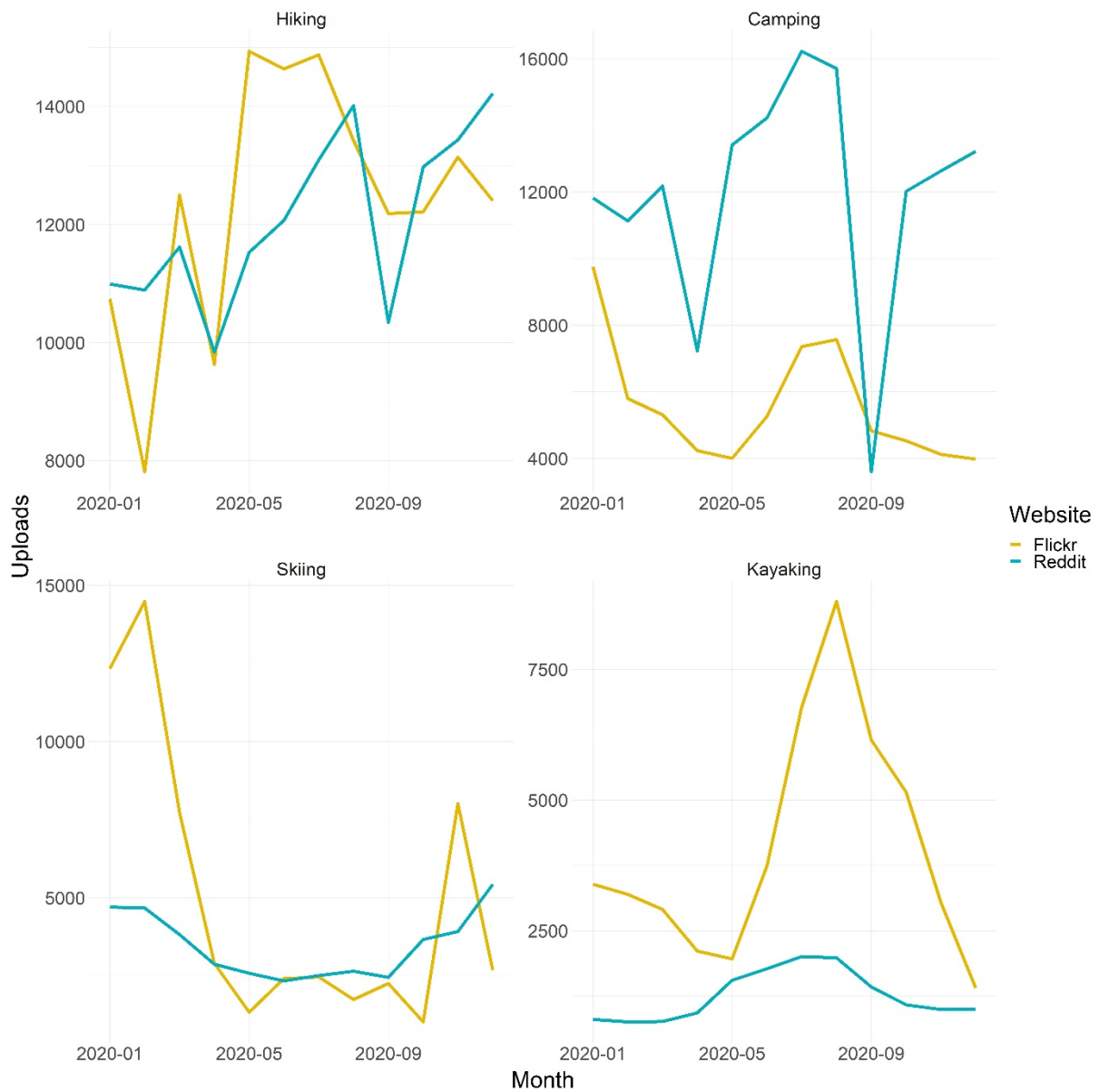
204 To assess whether Reddit posts can be used to assess general CES trends for a given location, we
205 also searched Reddit for posts containing given place names. We carried out two types of search;
206 first, we searched for posts containing a given place name as well as the term "hiking" and second,
207 we searched for posts containing a given place name within the subreddit "r/EarthPorn". The place
208 names were chosen to represent a range of scales; national ("USA" and "UK"), regional ("Wyoming"
209 and "Scotland") and National Park ("Yellowstone" and "Cairngorms"). The searches were carried out
210 for any post uploaded between the 1st of January 2010 and the 1st of January 2021. Total number of
211 posts was calculated.

212 3.0 Results

213 3.1 Full datasets

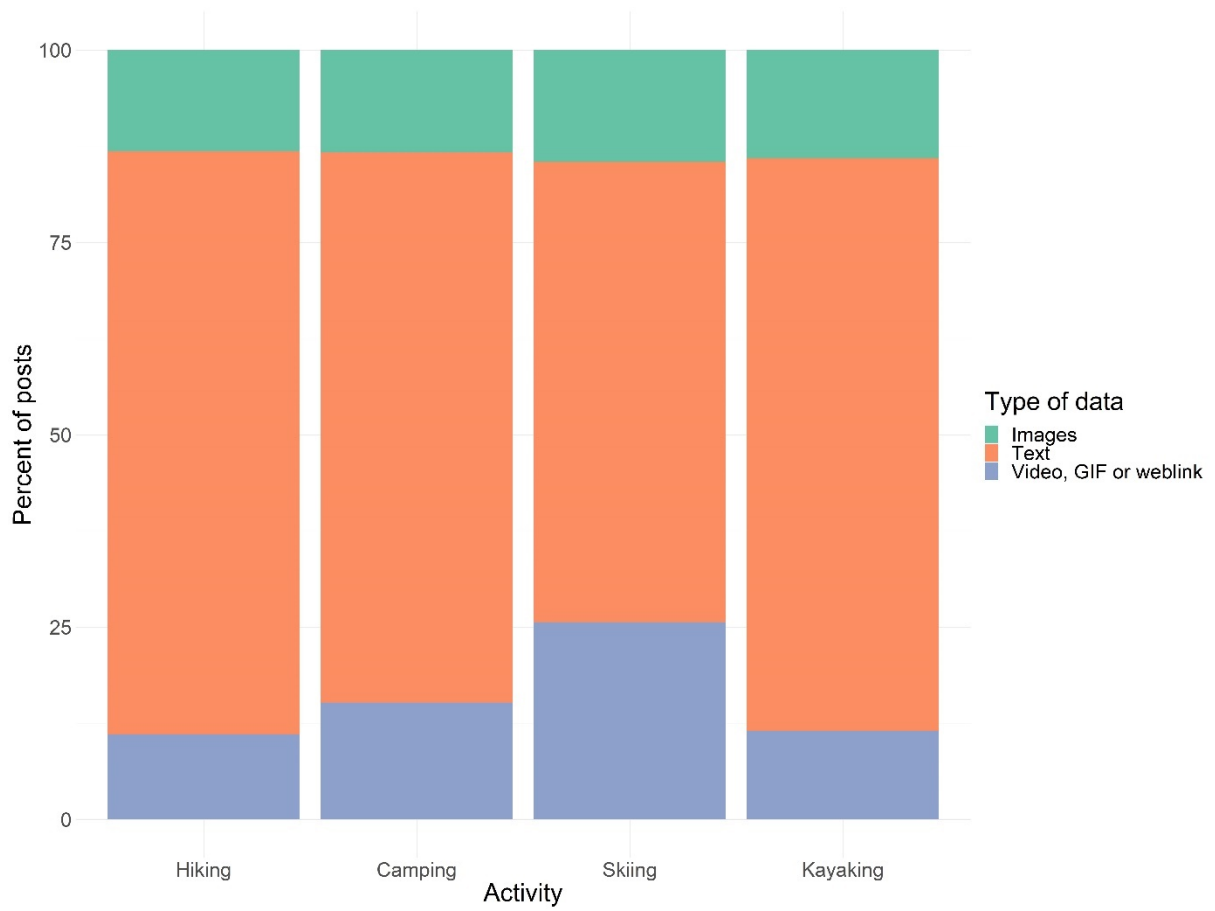
214 For each activity, the number of posts vary across each site. For hiking there were a similar number
215 of posts uploaded to each site in 2020 with 145,036 hiking posts on Reddit and 148,535 on Flickr.
216 There were also a similar number of posts relating to skiing across the two sites: 41,703 post on
217 Reddit and 59,455 posts on Flickr. For camping, more posts were uploaded to Reddit (143,446) than
218 Flickr (66,818); however for kayaking more posts were uploaded to Flickr (48,659) than Reddit
219 (15,107). The number of uploads fluctuate across the year for both websites (Fig. 1). For hiking and
220 skiing, even though there were a similar number of posts, Reddit had a greater quantity of unique

221 users generating the posts. For hiking, Reddit had 88,075 unique users posting whilst Flickr had
 222 9,392, while for skiing Reddit had 20,934 unique users whilst Flickr had 4,309.
 223



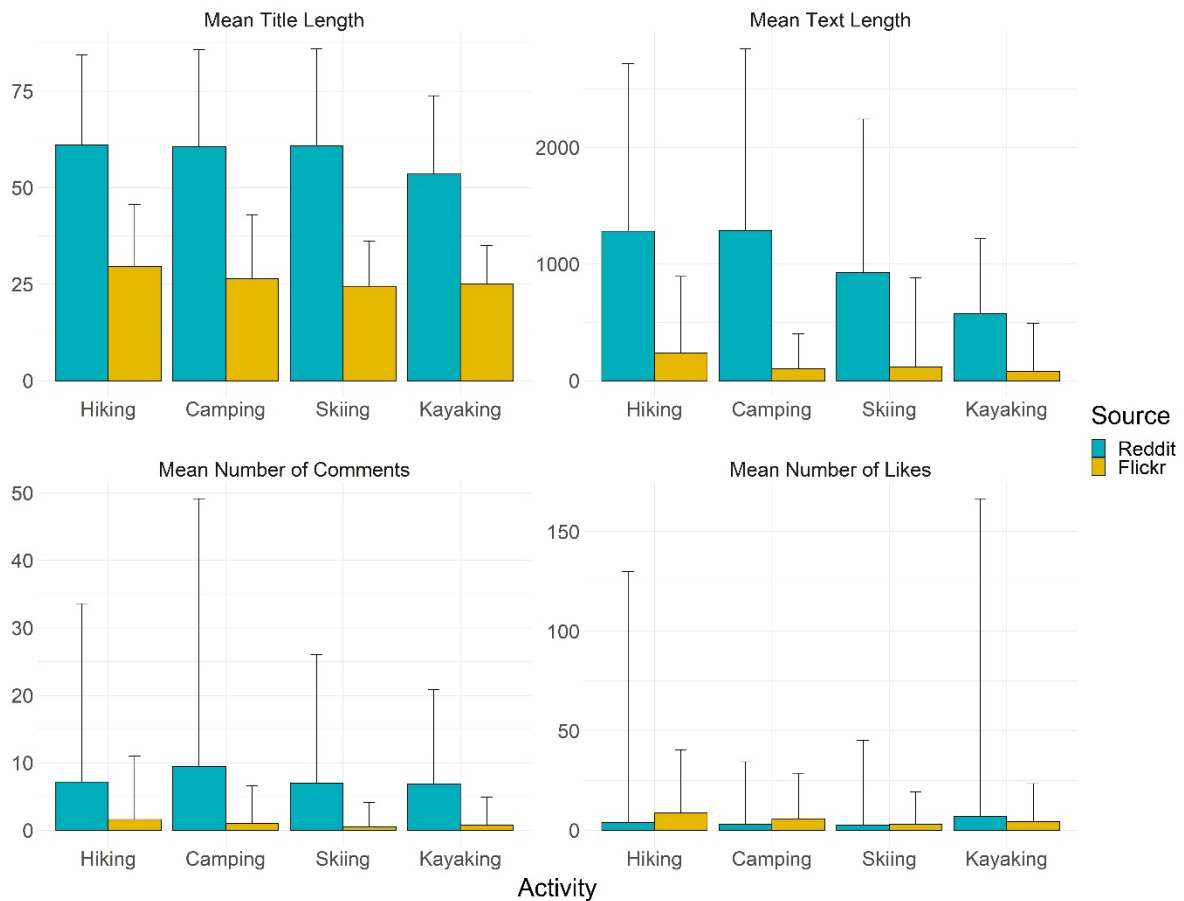
224
 225 Fig.1: Uploads of posts including the words “hiking”, “camping”, “skiing” and “camping” to Reddit
 226 and Flickr between the 1st of January 2020 and the 1st of January 2021.

227 For each activity that we searched, many of the posts uploaded to Reddit were text based (Fig. 2).
 228 Only around 15% of the posts returned via a keyword search from Reddit were images. Compared to
 229 posts uploaded to Flickr, posts on Reddit, in general, have longer titles and text descriptions as well
 230 as a higher number of comments (Fig. 3). Posts relating to hiking, camping and skiing on Flickr have,
 231 on average, more likes than posts on Reddit, though Kayaking posts on Reddit have a higher mean
 232 number of likes than those on Flickr.



233

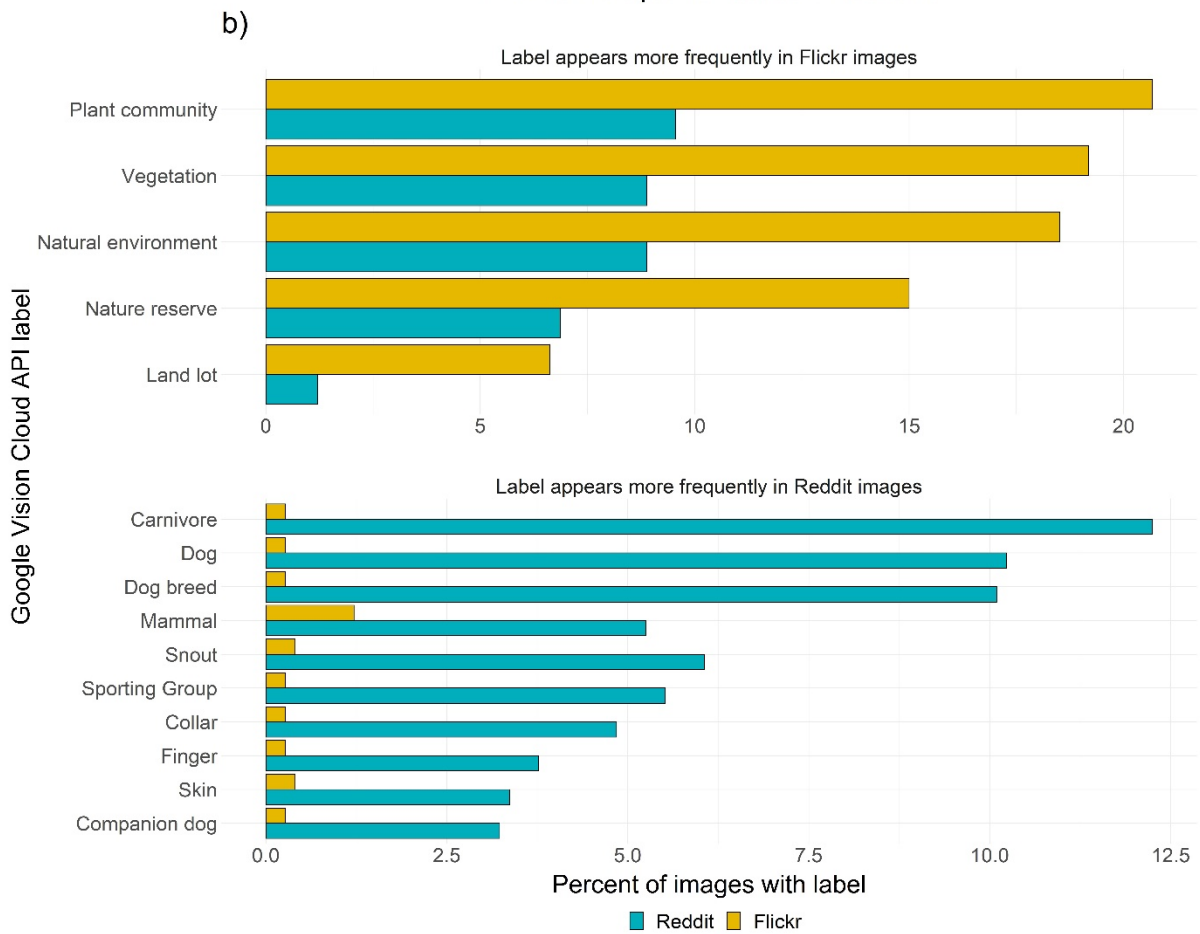
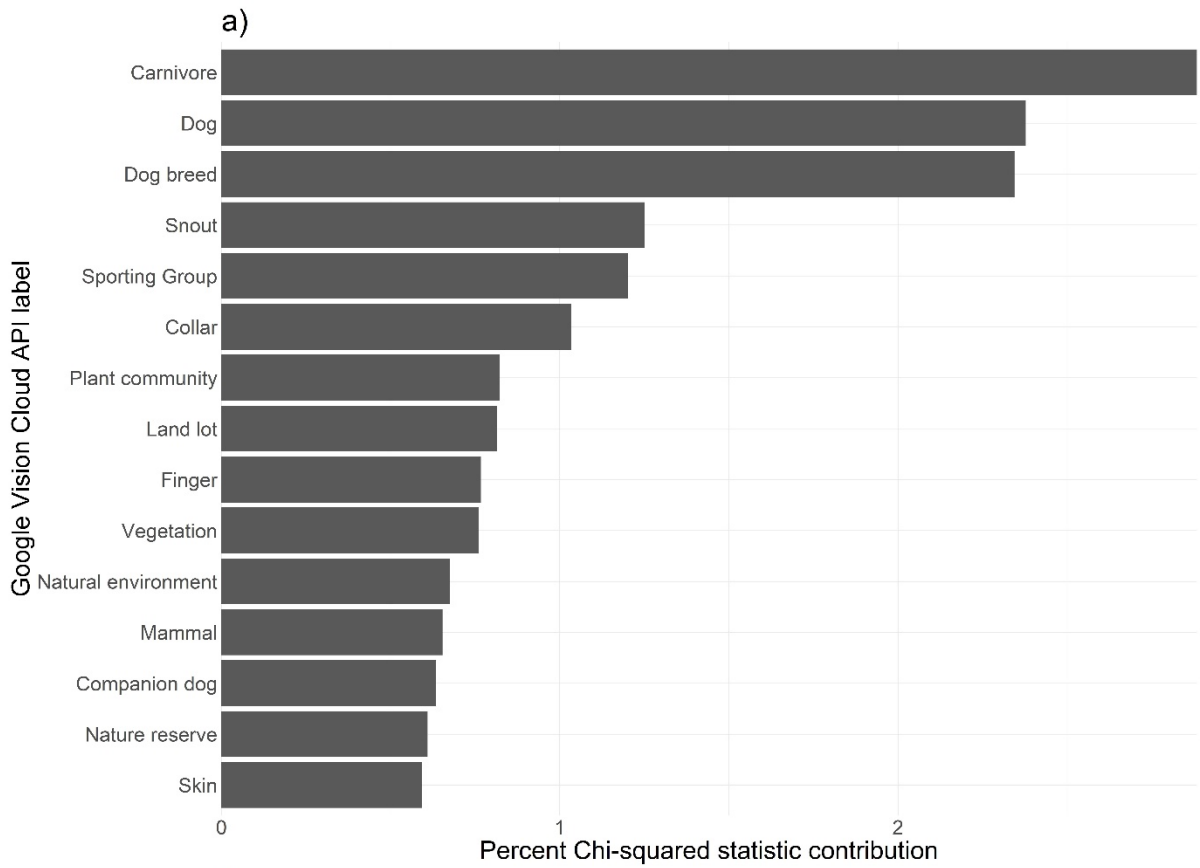
234 Fig. 2: Types of posts uploaded to Reddit.



235

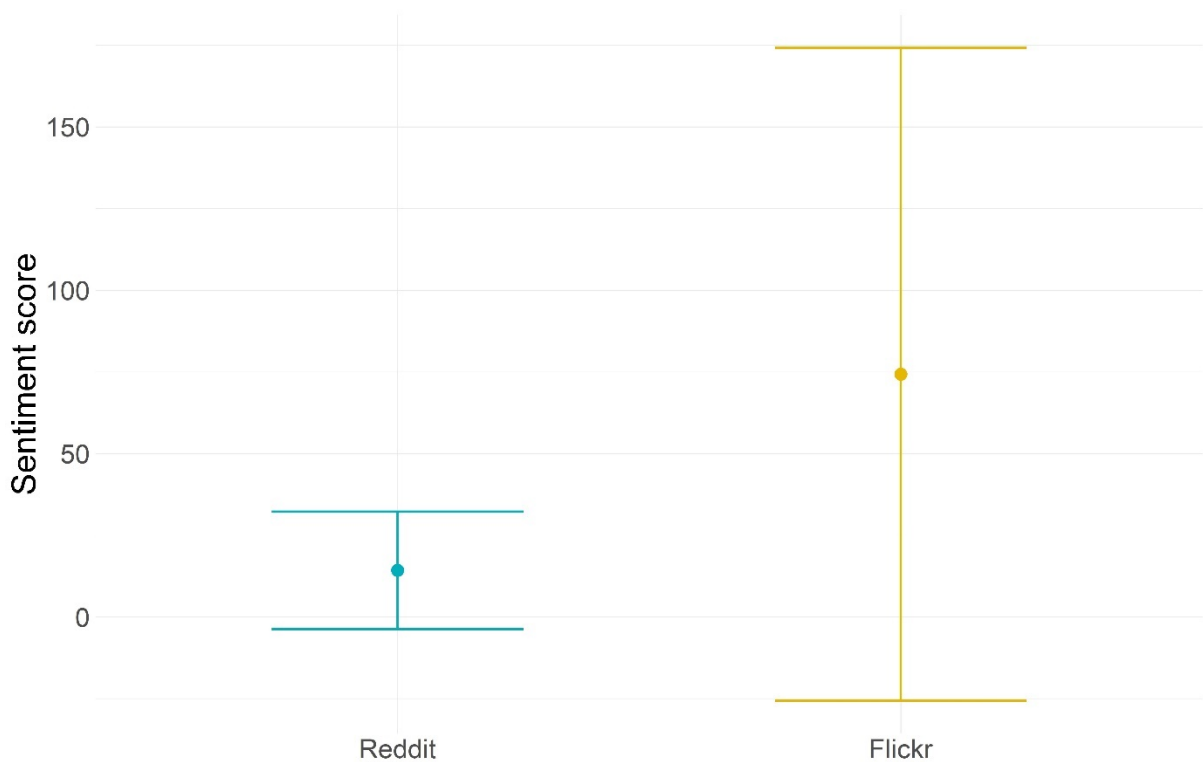
236 Fig. 3. Summary of posts made on Reddit and Flickr (mean + 0.5 standard deviations).

237 While the majority of the most labelled objects were common between the two sets of images (e.g.
 238 tree and mountain), there was an overall significant difference in the contents of the two sets of
 239 photographs labelled by the Google Cloud Vision API, ($\chi^2 = 3,127.5, df = 1230, N = 13,582, p <$
 240 0.001) The 15 Google Cloud Vision API labels (1.22% of the total number of unique labels) that had
 241 the highest contribution to the total χ^2 effect size contributed 17.42% of the total χ^2 value (Figure
 242 4a). Of these 15 labels, five (“plant community”, “vegetation”, “natural environment”, “nature
 243 reserve” and “land lot”) appeared more frequently in the images from Flickr (Fig. 4b). Though more
 244 frequent in Flickr images, the Google Cloud Vision API labels such as “plant community” and “natural
 245 environment” were present in 71 and 66 Reddit images, respectively. The other ten highest
 246 contributing labels, relating to dog walking, sports and people were more frequently photographed
 247 in Reddit images, with the labels such as “dog” and “dog breed” only being tagged in two of the
 248 Flickr images.



250 Figure 4: a) The 15 Google Cloud Vision API labels which had the greatest contribution to the overall
251 Chi-squared statistic (larger values indicate a larger difference between Reddit and Flickr); b) The
252 percentage of Reddit and Flickr images that the 15 labels appeared in.

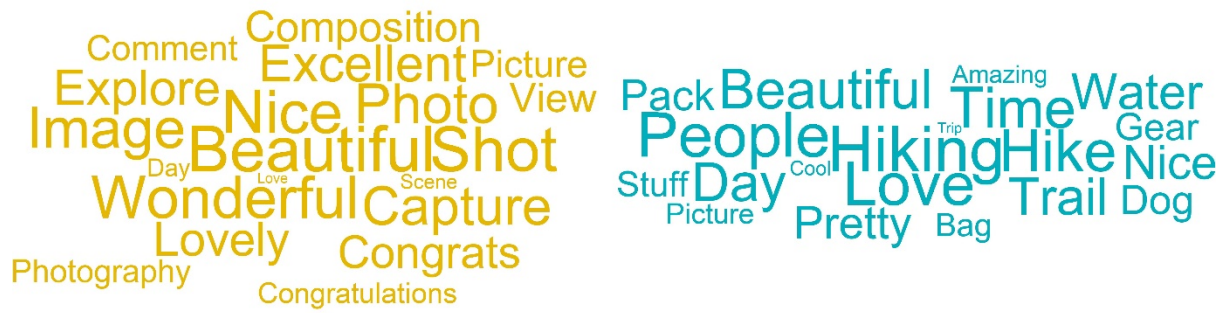
253 For the 1,000 hiking images from Reddit, 702 posts had comments, while for the 1,000 Flickr images
254 only 116 posts had comments. The 6,602 comments on the Reddit post were made by 4,142 unique
255 users, while the 1,702 Flickr comments were made by 1,119 unique users. A sentiment score could
256 be calculated for 642 Reddit comments and 108 Flickr comments, those where a score could not be
257 calculated did not contain any words in the AFINN dictionary. In general, the sentiment expressed in
258 Flickr comments was far higher than those on Reddit (Fig. 5). Only 1.90% of Flickr images expressed a
259 negative or neutral sentiment, whilst 11.66% of Reddit comments expressed a negative or neutral
260 sentiment. Many of the non-unique Flickr comments are "awards", a small sticker accompanied by a
261 text phrase, while on Reddit they were automatically generated messages from moderators of the
262 subreddit. After filtering, the most used words in Flickr and Reddit comments suggest that Flickr
263 users more frequently comment general positive comments regarding the picture, such as
264 "wonderful" and "excellent", while Reddit users more frequently comment regarding features of the
265 photograph, such as "trail", "water" and "dog" (Fig. 6).



266 Fig. 5: Mean +/- 0.5 standard deviations for the AFINN sentiment score expressed in the comments of
267 hiking images on Reddit and Flickr.
268

Flickr

Reddit



269

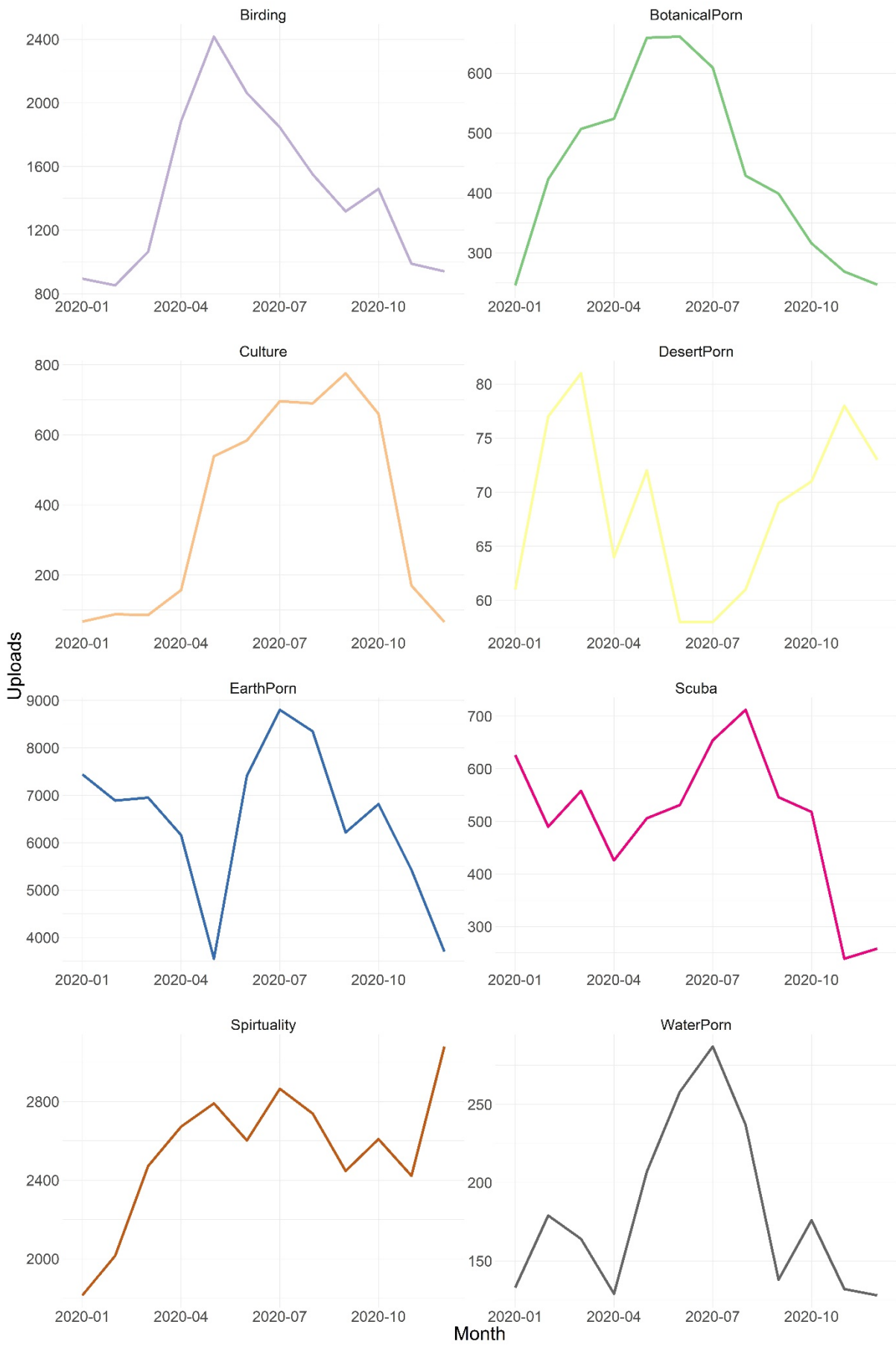
270 Fig. 6: The 20 most frequently used word in Flickr and Reddit comments after filtering.

271 *3.2 Subreddit search*

272 Of the subreddits relating to aesthetic values, “r/EarthPorn” was the most popular of the four we
273 searched, with 77,717 photographs uploaded in 2020. The subreddit “r/BotanicalPorn” had 5,289
274 uploads, “r/WaterPorn” 2,168 and “r/DesertPorn” 823. The number of uploads to each subreddit
275 varies by month (Fig. 7). The subreddits “r/Spirituality” and “r/Culture” also had a relatively large
276 number of uploads during the year 2020 with 30,528 and 4,579 uploads, respectively. Furthermore
277 the recreational based subreddits “r/Birding” had 17,280 post and “r/Scuba” had “6,064” posts.

278

279

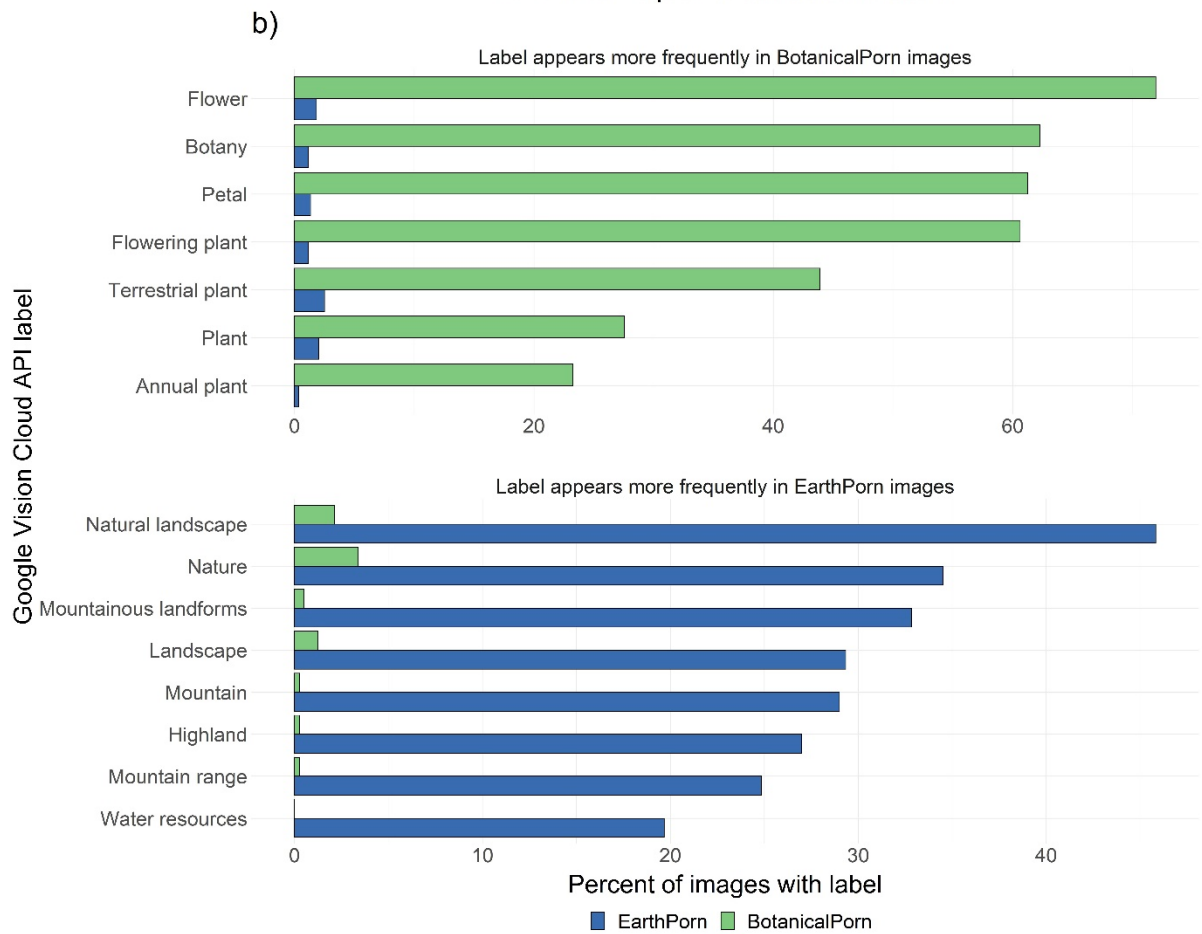
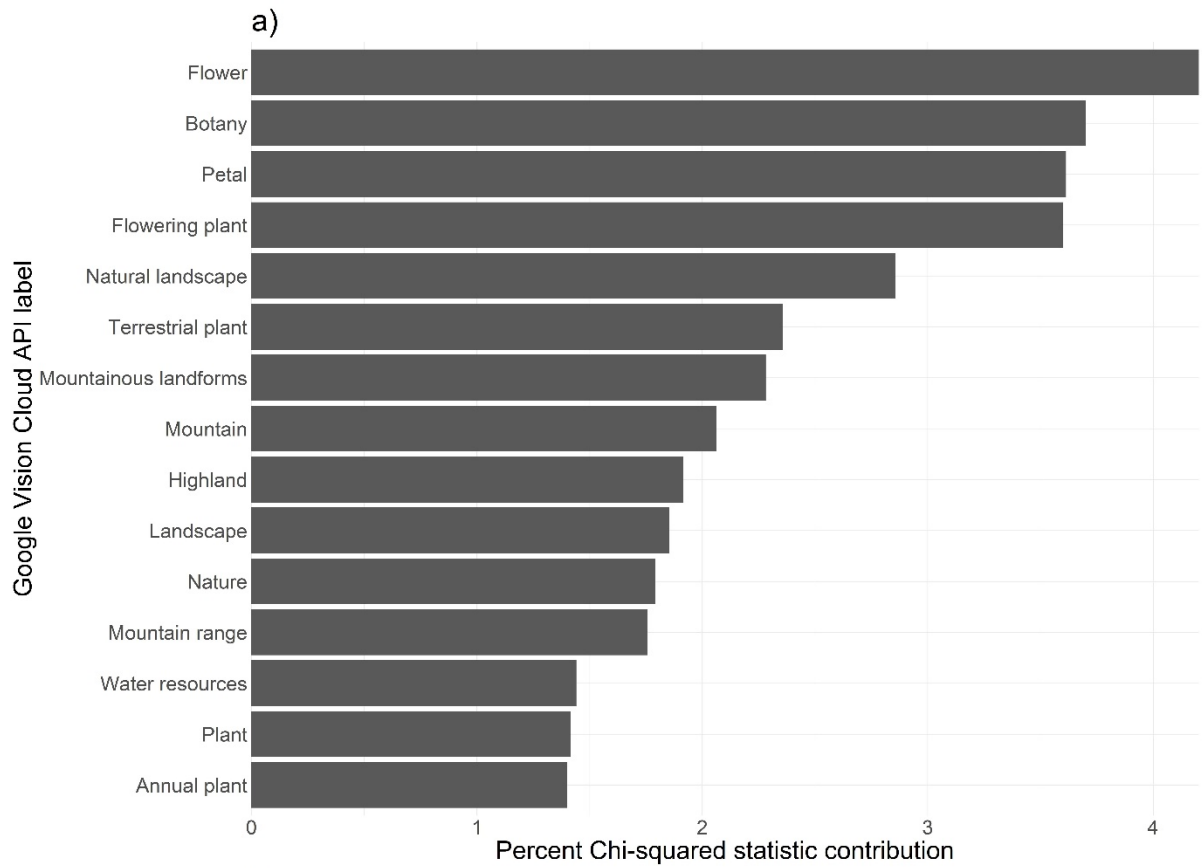


280
281

Figure 7: Uploads of posts to the subreddits “r/Birding”, “r/BotanicalPorn”, “r/Culture”,

282 “r/DesertPorn”, “r/EarthPorn”, “r/Scuba”, “r/Spirituality” and “r/WaterPorn” between the 1st of
283 January 2020 and the 1st of January 2021.

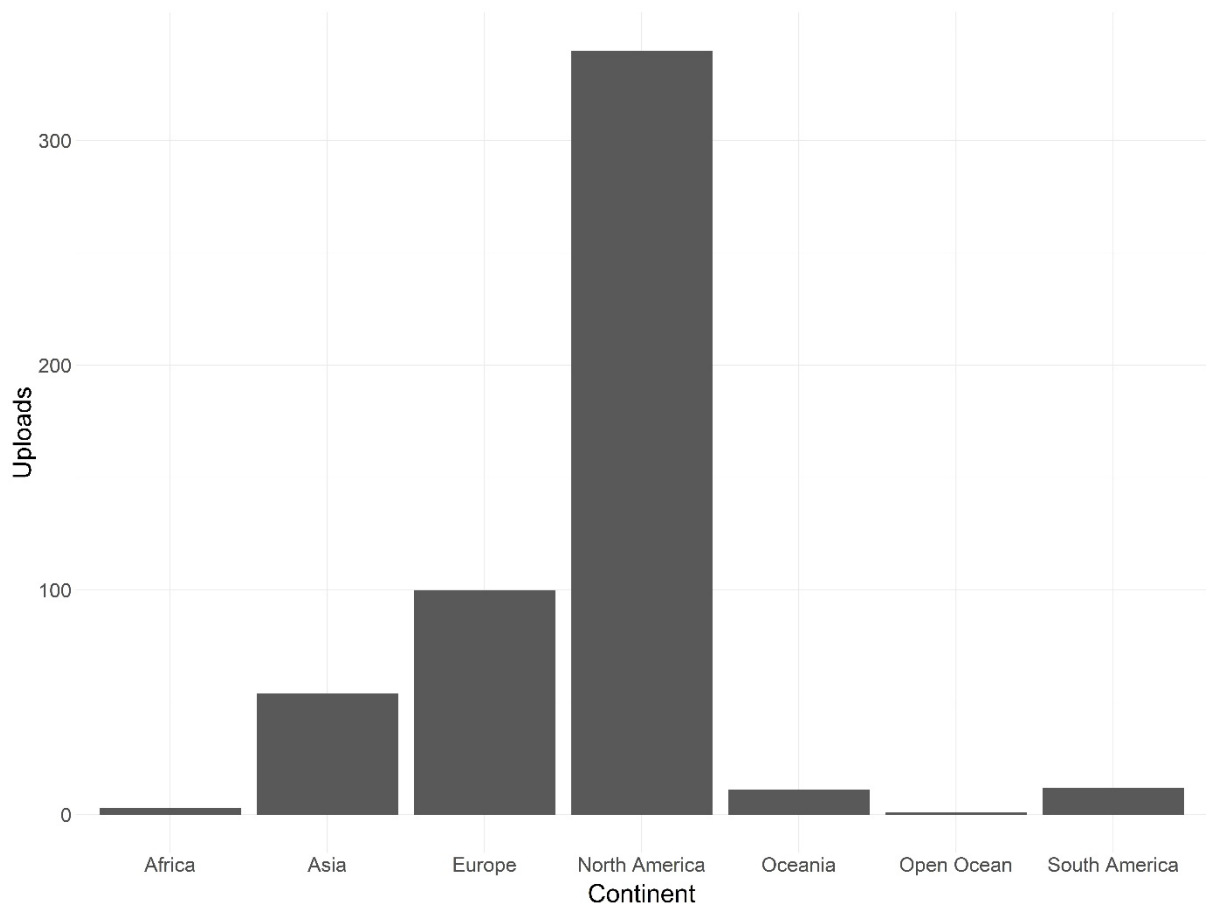
284 There was a large contrast between the labelled objects in images from the “r/EarthPorn” and
285 “r/BotanicalPorn” subreddits, with an overall significant difference in the contents of the two sets of
286 photographs labelled by the Google Cloud Vision API, ($x^2 = 10,205.5$, $df = 765$, $N = 13,196$, $p <$
287 0.001). The 15 Google Cloud Vision API labels (1.95% of the total number of unique labels) that had
288 the highest contribution to the total x^2 effect size contributed 36.26% of the total x^2 value (Figure
289 8a). Of these 15 labels, seven, all relating to plants and flowers, appeared more frequently in the
290 images from “r/BotanicalPorn” (Fig. 8b). The other highest contributing labels, relating to
291 landscapes, were more frequently photographed in “r/EarthPorn” images.



293 Figure 8: a) The 15 Google Cloud Vision API labels which had the greatest contribution to the overall
 294 Chi-squared statistic (larger values indicate a larger difference between Reddit and Flickr); b) The
 295 percentage of “r/EarthPorn” and “r/BotanicalPorn” subreddit images that the 15 labels appeared in.

296 *3.3 Potential spatial uses for Reddit*

297 Our automated method for estimating image location returned a latitude and longitude for 574
 298 “r/EarthPorn” subreddit images (57.4%) (Fig. 9). The vast majority of images (65.26%) were
 299 distributed across North America. Overall, there were fewer images taken in the other continents,
 300 with Europe and Asia having relatively higher numbers of images than Oceania, South America and
 301 Africa.



302
 303 Figure 9: Estimated locations of a subset of photographs from the “r/EarthPorn” subreddit.

304 When searching the Reddit API for posts relating to a place name as a keyword the number of posts
 305 vary depending on the spatial scale and location (Table 3). For both searches containing a separate
 306 keyword (“hiking”) and those from a specific subreddit (“r/EarthPorn”) a large number of posts were
 307 returned.

308 Table 3: Number of posts, when searching Reddit with a location name as a criterion.

Country	Scale	Search Criteria	Number of Posts
USA	National	Text = “USA” AND “hiking”	13,148
		Subreddit = any	12,336
	Regional	Text = “USA”	12,336
		Text = “Wyoming” AND “hiking”	1,209

		Subreddit = any	
		Text = "Wyoming"	3,399
		Subreddit = "r/EarthPorn"	
	National park	Text = "Yellowstone" AND "hiking"	2,794
		Subreddit = any	
		Text = "Yellowstone"	4,334
		Subreddit = "r/EarthPorn"	
UK	National	Text = "UK" AND "hiking"	8,196
		Subreddit = any	
		Text = "UK"	5,539
		Subreddit = "r/EarthPorn"	
	Regional	Text = "Scotland" AND "hiking"	2,528
		Subreddit = any	
		Text = "Scotland"	5,539
		Subreddit = "r/EarthPorn"	
	National park	Text = "Cairngorms" AND "hiking"	87
		Subreddit = any	
		Text = "Cairngorms"	131
		Subreddit = "r/EarthPorn"	

309

310 4.0 Discussion

311 The main aim of this paper was to understand the potential applications for Reddit as a
312 complementary or alternative source of CES data from social media sites. Here, we explored two
313 methods of searching the Reddit API: a keyword search and searching specific subreddits. In general
314 we were able to return a relatively large number of posts relating to a range of CES (recreation,
315 aesthetic, spirituality and culture). Searches made via the keywords search showed that Reddit has a
316 comparable number of available posts on recreational CES to Flickr. However, the posts returned via
317 a keyword search on Reddit are primarily text based, which is unsurprising given that Reddit is
318 marketed as a discussion-based social media site. The two sites had similar numbers of posts for
319 hiking and skiing, though Reddit had more posts about camping and Flickr had more posts about
320 kayaking. This suggests that the choice of site may depend on the activity of interest and thus the
321 suitability for CES research is context dependent. Furthermore, even when the posts had a similar
322 number of uploads between sites, the posts on Reddit were contributed by a far greater quantity of
323 unique users. This gives rise to the potential for posts to be generated by a more diverse user base
324 than Flickr. There are however socio-demographic biases associated with social media sites (Duggan
325 and Smith 2018; Rekta et al. 2019), and these need to be explored fully before making
326 generalisations about the wider population.

327 The biggest limitation of Reddit is that the posts do not have geotagged locations. Our automatic
328 method for estimating the approximate location of a photograph calculated latitude and longitude
329 for 57.4% of the Reddit posts. From our analysis of landscape photographs, the distribution of
330 images uploaded to the "r/EarthPorn" subreddit are primarily concentrated in North America,
331 though many images were also from Europe and Asia. Harrington (2018) estimated the distribution
332 of the Reddit users base through geolocating statements in their comments and found that the
333 demographic was primarily people living in North America, followed by Europe and Asia. Harrington
334 (2018) also provides a potential method of establishing user origins, a key feature in understanding
335 CES interaction from Flickr (Wood et al. 2013; Sinclair et al. 2020). The demographic of users and
336 distribution of posts may have implications for studies that wish to assess CES across different

337 continents, with previous studies assessing CES in North America potentially missing out on the
338 wider range of photographs available from Reddit.

339 A potential issue with Reddit, as well as other social media sites such as Flickr, is the potential biases
340 introduced by the demographics of their users. For example, though Reddit has a large user base
341 with high socio-demographic diversity, with an estimate that around 6% of internet users were
342 active on Reddit, there is bias towards male users (8% of male internet users compared to 4%
343 female) and a bias to younger users, with a higher percentage users aged 18-49 than those over 50
344 (Duggan and Smith 2018). Furthermore, the users of both Reddit and Flickr are concentrated in
345 western, developed countries. Where studies are at a global or super-continental scale, data from
346 Reddit and Flickr should therefore be used in combination with each other and with other sources of
347 data that are popular in other areas of the world. For example, in China where Flickr is banned and
348 Reddit is not a popular social media site, alternative social media sites such as Weibo (Zhang and
349 Zhou 2019), or travel comment portals websites such as Tuniu Travel (Dai et al. 2019), should be
350 used to bridge the gap in CES data. At local and regional scales other sources of data may also help
351 to complement social media data such as on-site survey data (Sinclair et al. 2020), online surveys
352 (Moreno-Llorca et al. 2020) and national statistics (Graham and Eigenbrod 2019). Future work
353 should begin to assess the respective biases in these alternative sources to ensure they are
354 comparable. Furthermore, both Flickr and “r/EarthPorn” are related to images pertaining to high-
355 end photography, which may restrict the demographics to only those with access to such technology
356 (Chen et al. 2020). One possible source of data that we suggest needs exploring is other subreddits
357 focused on natural landscapes, such as “r/AmatureEarthPorn”, which do not restrict uploads to high-
358 quality images and therefore may have greater representation of landscapes from a wider
359 demographic.

360 There are however several caveats to geocoding Reddit post locations. First, the extracted location
361 name from the named-entity recognition may not be correct due to ambiguity in the text, spelling or
362 language differences, or capitalizations (Goyal et al. 2018). Given that posts on “r/EarthPorn” are
363 predominantly in the English language, this may not have been a significant issue in our analyses.
364 The issue with multi-part names being extracted to a single word place name means that the finer
365 spatial scale of the location is lost. The rules of the subreddit specifies that place names included in
366 the post title should be as specific as possible. However, the named-entity recognition method often
367 identified the location as the regional (i.e. state) or country part of the place name, losing the finer
368 detail of the image’s location. Though the named-entity recognition method can correctly recognise
369 and extract places names with multiple parts (e.g. "Ocean Beach", "San Francisco" was correctly
370 identified), for many multi-part place names the finer location detail can be lost. For example, “Mt.
371 St. Helens, Washington” was extracted as “Washington”. The automated extraction of the landscape
372 image place name presented here may be best suited for generalising large-scale distributions.
373 However, as the Reddit posts normally contain specific location details in their titles, studies that
374 wish to assess spatial distribution on a finer scale may find success in manually extracting the place
375 name.

376 Second, the high number of available geocoding algorithms, as well as the potential for ambiguity in
377 the named entity locations extracted from the Reddit comments, can introduce errors in the
378 geocoded results (McDonald et al. 2018). For example, there are multiple locations globally named
379 “Portland”; without more context the geocode algorithm may not correctly code the location. Third,
380 though the geocoding method can provide a latitude and longitude with a high spatial accuracy,
381 when geocoding is based on a general location name, the location will be plotted to a single point
382 within that region. For example, multiple photographs taken in completely different areas of the

383 Badlands National Park, US, all containing “Badlands National Park” in their title, will all be
384 aggregated to the same point location. Furthermore, though this method was successful on posts to
385 “r/EarthPorn”, other subreddits may not stipulate that a location must be present in the text. We
386 suggest that future studies using Reddit data for spatial analysis should consider methods for
387 reducing geocoding inaccuracies (McDonald et al. 2018). Another possible source of geocoding a
388 posts location is the Google Cloud Vision API which can estimate the location of an image; however
389 this process is currently only capable of locating popular sites.

390 Due to the limitations of geocoding Reddit posts, we do not recommend using posts from Reddit to
391 assess the spatial variation of CES in a similar manner to those from Flickr, Twitter or Instagram (e.g.
392 Graham and Eigenbrod 2019; Chen et al. 2020). Instead, one potential method for getting CES data
393 for a location without the need for geocoding posts is searching for a given name place alongside
394 other keywords or within a subreddit. This method has previously been used in CES studies from
395 Flickr, for example Thiagarajah et al. (2015) searched Flickr for photographs based on the place
396 names of four mangrove sites in Singapore, while Roberts (2017) queried Twitter posts for any
397 containing the names of urban green spaces in Birmingham, UK. Here, we showed that searches for
398 Reddit posts with a relevant study site as a key word provides a relatively large dataset across spatial
399 scales and locations. Though we have demonstrated that Reddit data has the potential for spatial
400 studies, we acknowledge these limitations do restrict the use of Reddit’s data to assess spatial
401 variations in CES and therefore suggest that Reddit posts are more suited to generalising CES based
402 on a given search criteria e.g. a place name or specific activity. However, these limitations do not
403 hinder the use of data for studies that assess CES through content analysis and textual analysis.

404 We have shown that photographs associated with hiking from both Reddit and Flickr can both be
405 used in the same image content analysis techniques, thus illustrating their potential for CES studies
406 which use content analysis of images, without additional spatial analysis (e.g. Thiagarajah et al.
407 2015). Oakes and Farrow (2006) demonstrated that words with the highest percentage contribution
408 of the total χ^2 value, relative to the other words in the set, best highlight the differences in two
409 groups of words. Here, the small number of labels contributing to a high percentage of the total
410 χ^2 value indicates that, in general, many images contain similar scenes, but the difference between
411 the two sites is driven by a small number of features identified with the Google Cloud Vision API. The
412 differences between the two sites may be reflected in the user’s motivations for undertaking hiking.
413 As the reasons to undertake hiking are multifaceted (Wilcer et al. 2019), the difference in
414 demographics between users of Reddit and Flickr suggests they may be undertaking hiking or
415 uploading images to each site for different reasons. For example, results from our subset of images
416 suggest that Reddit users are more likely to participate in hiking for physical activity and dog walking,
417 whilst Flickr users are more likely to undertake hiking to access aesthetic views.

418 We have demonstrated that, as the contents of images from Reddit and Flickr can provide
419 essentially the same information about CES, Reddit may be a valuable additional source of data for
420 assessing aesthetic landscape qualities (e.g. Oteros-Rozas et al. 2018) or recreational preferences
421 (e.g. Gosal et al. 2019; Lee et al. 2019). The difference in contents may also be down to the
422 motivations to upload to each platform. Kipp et al. (2017) found that Flickr users have multiple
423 motivations for uploading photographs including wanting to get an opinion on their photographs
424 and because they have an interest in a particular subject. However, as one of the main features of
425 Reddit is the ranking of posts through user votes (Duggan and Smith 2013), further work should be
426 undertaken to assess whether the relative motivations for uploading to Reddit are similar to other
427 social media sites. Furthermore, our searches were only carried out in the English language, and

428 therefore may introduce bias into the conclusion drawn about the motivations for undertaking
429 different recreational activities.

430 Comparison of image content from uploads to the subreddits “r/EarthPorn” and “r/BotanicalPorn”,
431 which focus on photographs of different aspects of nature, demonstrated distinctions between the
432 two - and therefore provide unique sources of data for assessing the role of different aspects of
433 nature to CES. Building on this, “r/WaterPorn” and “r/DesertPorn” may help to provide a robust
434 dataset for untangling the contributions of geodiversity to CES by providing unique insights into
435 peoples' opinions on abiotic features (Fox et al. 2020a). Furthermore, subreddits are not just useful
436 for assessing aesthetic CES, but can also provide a large source of data for spirituality and recreation.
437 There is a far larger range of subreddits available than accessed here, each with a unique theme that
438 can help to understand CES, for example “r/Travel” (a discussion board for travel) could be a useful
439 source of data for understanding the links between tourism and CES and “r/CityPorn” (images of
440 cityscapes and urban areas) may help to investigate urban ecosystem services, although this may
441 require some content filtering to remove purely architectural images. As our keyword searches
442 return significantly more text-based posts than images, researchers should familiarise themselves
443 with the different subreddit as potential sources of images, for example, titles of posts in
444 “r/EarthPorn” generally do not contain words like “landscape” or “view” and would therefore not be
445 returned through a keyword search looking for images relating to aesthetics. The results presented
446 here demonstrate that Reddit has the potential to be a significant source of image data and may be
447 beneficial to CES studies that incorporate content analysis.

448 Studies can also use textual metadata to assess CES (e.g. Roberts 2017; Hale et al. 2019; Johnson et
449 al. 2019). Flickr images tend to have description metadata that the uploader provides, which has
450 been demonstrated to be useful in textual analysis such as sentiment analysis (Brindley et al. 2019)
451 or eliciting information on CES from the text (Hale et al. 2019). A disadvantage of photographs
452 uploaded to Reddit is that images do not have an equivalent description by the uploader, therefore
453 we only compare the comment sections of the two websites. As many posts on Reddit have
454 comments and because Reddit is a discussion-based platform, this large online database may help to
455 understand the opinions of thousands of individuals. As the perception of the CES can only be drawn
456 from those that comment (Dai et al. 2019), having a larger number of unique individuals interacting
457 with CES related posts may enable the results to be generalised to the wider population and
458 therefore better help to inform policy, planning and management (Dunkel 2015). Here, the text
459 comments from the two sites vary regarding the sentiment expressed, with Flickr images having a
460 more positive associated sentiment score, but also a large variability within the score. The subset
461 analysed here also showed very few negative comments on Flickr, whilst on Reddit a negative
462 sentiment was more frequently expressed. Moreover, the actual text contained within the
463 comments differs between the two sources, with comments on Flickr tending to be more general
464 appraisals of the photograph, while Reddit comments are more often a discussion around the image
465 themselves, thus potentially providing richer information on the users' perspective of CES. Having
466 access to a wider range of opinions, both positive and negative, may help to better generalise
467 attitudes to CES.

468 As Reddit is designed to be a discussion-based forum it may contribute to richer information on the
469 users' perspective of CES. For example, the “r/Spirituality” subreddit encourages users to contribute
470 to the discussion of any aspects of spirituality regardless of religion or ideology, thus providing the
471 potential to assess the opinions of people from a wide range of backgrounds. Furthermore, Reddit
472 comments can be longer than most other social media sites (e.g. Twitter has a 280-character limit
473 and Instagram has a 300-character comment limit) and therefore a user can discuss their opinions in

474 greater detail (Gkotsis et al. 2017). The discursive nature of Reddit provides researchers a unique
475 opportunity to assess which aspects of a certain image or video people appreciate. There is also
476 scope for this interactive and discussion-based platform to be used in experimental studies in which
477 researchers post content and monitor feedback. Though as with all social media-based studies, we
478 recommended that the ethics of these studies be discussed in further detail. We suggest that Reddit
479 data is particularly useful for studies that wish to analyse users' comments in conjunction with the
480 metadata available for each image for a more robust assessment of CES.

481 For studies carrying out image content or textual analysis we suggest that combining Reddit data
482 alongside other sources of data, would be useful in CES because (1) images and text from Reddit can
483 provide comparable data used to assess aspects of CES; (2) Reddit potentially contains additional
484 data previously overlooked; (3) they have different geographical biases (e.g. Reddit to North
485 America, Flickr to Europe and Weibo to Asia). We therefore suggest that a more holistic approach of
486 assessing CES would be to include cross-platform analysis including multiple sources (Retka et al.
487 2019). However, we note that Reddit may not be suitable for integrating into studies assessing
488 spatial variations in CES. Data integration, the bringing together of data from multiple sources, could
489 be implemented to allow data from social media sites to be analysed as a complete unit. Data
490 integration methods, which control for differing biases and sizes of datasets, have been successfully
491 used in other scientific fields such as species distribution modelling (Issac et al. 2019) and those
492 using satellite imagery (Aires 2014). As accessing data from Reddit requires a similar skill level as
493 accessing datasets from other social media websites, data integration of these multiple sources is
494 feasible. The tools and software used in this manuscript make these datasets more accessible and
495 reproducible for non-data scientists and enable us to start to bridge the gap in integrating multiple
496 sources. We therefore recommend that CES and wider environmental science studies make use of
497 these tools to include the vast amount of data from Reddit alongside other social media data sources
498 in their future studies.

499 **5.0 Conclusion**

500 We have demonstrated that posts from Reddit can be used in commonly applied CES assessment
501 methods, such as image content analysis and textual analysis, which leverage the power of big data
502 from social media sites. The results from this study show that Reddit can provide a large source of
503 data similar to Flickr. However, the posts available on Reddit are not geolocated and the geocoding
504 of a post's location has several limitations meaning that Reddit is not as suited to assessing the
505 spatial variation of CES as other social media sites. The large quantity of data available on Reddit is
506 most appropriate for assessing general trends in CES through image content analysis and textual
507 analysis. The discursive nature of Reddit provides a unique opportunity to assess a wide range of CES
508 including recreational activities, aesthetic views, spirituality and culture. We argue that Reddit
509 should be more widely considered as a useful source of data for CES studies and we hope that this
510 paper sets a precedent for including big datasets from Reddit in future studies.

511 **6.0 References**

512 Aires, F., 2014. Combining datasets of satellite-retrieved products. Part I: Methodology and water
513 budget closure. *Journal of Hydrometeorology*, 15(4), pp.1677-1691. [https://doi.org/10.1175/JHM-D-](https://doi.org/10.1175/JHM-D-13-0148.1)
514 [13-0148.1](https://doi.org/10.1175/JHM-D-13-0148.1)

515 Alfred, R., Leong, L.C., On, C.K. and Anthony, P., 2014. Malay named entity recognition based on
516 rule-based approach. *International Journal of Machine Learning and Computing*, 4(3), 300-306.
517 <https://doi.org/10.7763/IJMLC.2014.V4.428>

518 Allan, J.D., Smith, S.D., McIntyre, P.B., Joseph, C.A., Dickinson, C.E., Marino, A.L., Biel, R.G., Olson,
519 J.C., Doran, P.J., Rutherford, E.S. and Adkins, J.E., 2015. Using cultural ecosystem services to inform
520 restoration priorities in the Laurentian Great Lakes. *Frontiers in Ecology and the Environment*, 13(8),
521 pp.418-424. <https://doi.org/10.1890/140328>

522 Allain, S., 2019. Mining Flickr: a method for expanding the known distribution of invasive species.
523 *Herpetological Bulletin*, 148, pp.11-14. <https://doi.org/10.33256/hb148.1114>

524 Barve, V., 2014. Discovering and developing primary biodiversity data from social networking sites: A
525 novel approach. *Ecological Informatics*, 24, pp.194-199. <https://doi.org/10.1016/j.ecoinf.2014.08.008>

526 Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. and Blackburn, J., 2020, May. The pushshift
527 reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol.
528 14, pp. 830-839).

529 Boe, B. PRAW: The Python Reddit API Wrapper. 2020-, <https://github.com/praw-dev/praw/> [Online;
530 accessed 2020-01-01]

531 Brindley, P., Cameron, R.W., Ersoy, E., Jorgensen, A. and Maheswaran, R., 2019. Is more always
532 better? Exploring field survey and social media indicators of quality of urban greenspace, in relation to
533 health. *Urban Forestry & Urban Greening*, 39, pp.45-54. <https://doi.org/10.1016/j.ufug.2019.01.015>

534 Chen, Y., Caesemaeker, C., Rahman, H.T. and Sherren, K., 2020. Comparing cultural ecosystem
535 service delivery in dykelands and marshes using Instagram: A case of the Cornwallis (Jijuktu'kwejk)
536 River, Nova Scotia, Canada. *Ocean & Coastal Management*, 193, p.105254.
537 <https://doi.org/10.1016/j.ocecoaman.2020.105254>

538 Clemente, P., Calvache, M., Antunes, P., Santos, R., Cerdeira, J.O. and Martins, M.J., 2019.
539 Combining social media photographs and species distribution models to map cultural ecosystem
540 services: The case of a Natural Park in Portugal. *Ecological Indicators*, 96, pp.59-68.
541 <https://doi.org/10.1016/j.ecolind.2018.08.043>

542 Dai, P., Zhang, S., Chen, Z., Gong, Y. and Hou, H., 2019. Perceptions of Cultural Ecosystem
543 Services in Urban Parks Based on Social Network Data. *Sustainability*, 11(19), p.5386.
544 <https://doi.org/10.3390/su11195386>

545 Daniel, T.C., Muhar, A., Arnberger, A., Aznar, O., Boyd, J.W., Chan, K.M.A., Costanza, R., Elmqvist,
546 T., Flint, C.G., Gobster, P.H., Gret-Regamey, A., Lave, R., Muhar, S., Penker, M., Ribe, R.G.,
547 Schauppenlehner, T., Sikor, T., Soloviy, I., Spiereburg, M., Taczanowska, K., Tam, J., von der Dunk,
548 A., 2012. Contributions of cultural services to the ecosystem services agenda. *Proc. Natl. Acad. Sci.*
549 109, 8812-8819. <https://doi.org/10.1073/pnas.1114773109>

550 Derczynski, L., Nichols, E., van Erp, M. and Limsopatham, N., 2017, September. Results of the
551 WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd*
552 *Workshop on Noisy User-generated Text* (pp. 140-147).

553 Díaz, S., Pascual, U., Stenseke, M., Martín-López, B., Watson, R.T., Molnár, Z., Hill, R., Chan,
554 K.M.A., Baste, I.A., Brauman, K.A., Polasky, S., Church, A., Lonsdale, M., Larigauderie, A., Leadley,
555 P.W., van Oudenhoven, A.P.E., van der Plaats, F., Schröter, M., Lavorel, S., Aumeeruddy-Thomas, Y.,
556 Bukvareva, E., Davies, K., Demissew, S., Erpul, G., Failer, P., Guerra, C.A., Hewitt, C.L., Keune, H.,
557 Lindley, S., Shirayama, Y., 2018. Assessing nature's contributions to people. *Science* (80-.). 359,
558 270-272.

559 Ding, X. and Fan, H., 2019. Exploring the Distribution Patterns of Flickr Photos. *ISPRS International*
560 *Journal of Geo-Information*, 8(9), p.418. <https://doi.org/10.3390/ijgi8090418>

561 Duggan, M. and Smith, A., 2013. 6% of online adults are reddit users. *Pew Internet & American Life*
562 *Project*, 3, pp.1-10.

563 Dunkel, A., 2015. Visualizing the perceived environment using crowdsourced photo geodata.
564 *Landscape and urban planning*, 142, pp.173-186. <https://doi.org/10.1016/j.landurbplan.2015.02.022>

565 Figueroa-Alfaro, R.W. and Tang, Z., 2017. Evaluating the aesthetic value of cultural ecosystem
566 services by mapping geo-tagged photographs from social media data on Panoramio and Flickr.
567 *Journal of Environmental Planning and Management*, 60(2), pp.266-281.
568 <https://doi.org/10.1080/09640568.2016.1151772>

569 Fox, N., Graham, L.J., Eigenbrod, F., Bullock, J.M. and Parks, K.E., 2020a. Incorporating geodiversity
570 in ecosystem service decisions. *Ecosystems and People*, 16(1), pp.151-159.
571 <https://doi.org/10.1080/26395916.2020.1758214>

572 Fox, N., August, T., Mancini, F., Parks, K.E., Eigenbrod, F., Bullock, J.M., Sutter, L. and Graham, L.J.,
573 2020b. "photosearcher" package in R: An accessible and reproducible method for harvesting large
574 datasets from Flickr. *SoftwareX*, 12, p.100624. <https://doi.org/10.1016/j.softx.2020.100624>

575 Ghermandi, A. and Sinclair, M., 2019. Passive crowdsourcing of social media in environmental
576 research: A systematic map. *Global environmental change*, 55, pp.36-47.
577 <https://doi.org/10.1016/j.gloenvcha.2019.02.003>

578 Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T.J., Dobson, R.J. and Dutta, R., 2017.
579 Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific*
580 *reports*, 7, p.45141. <https://doi.org/10.1038/srep45141>

581 Google Cloud Vision, 2020. Documentation for the Google Cloud Vision API [WWW Document]. URL
582 www.cloud.google.com/vision/.

583 Gosal, A.S., Geijzendorffer, I.R., Václavík, T., Poulin, B. and Ziv, G., 2019. Using social media,
584 machine learning and natural language processing to map multiple recreational
585 beneficiaries. *Ecosystem Services*, 38, p.100958. <https://doi.org/10.1016/j.ecoser.2019.100958>

586 Goyal, A., Gupta, V. and Kumar, M., 2018. Recent named entity recognition and classification
587 techniques: a systematic review. *Computer Science Review*, 29, pp.21-43.
588 <https://doi.org/10.1016/j.cosrev.2018.06.001>

589 Graham, L.J. and Eigenbrod, F., 2019. Scale dependency in drivers of outdoor recreation in
590 England. *People and Nature*, 1(3), pp.406-416. <https://doi.org/10.1002/pan3.10042>

591 Gray, M., 2011. Other nature: geodiversity and geosystem services. *Environmental Conservation*,
592 38(3), pp.271-274. <https://doi.org/10.1017/S0376892911000117>

593 Guerrero, P., Møller, M.S., Olafsson, A.S. and Snizek, B., 2016. Revealing cultural ecosystem
594 services through Instagram images: The potential of social media volunteered geographic information

595 for urban green infrastructure planning and governance. *Urban Planning*, 1(2), pp.1-17.
596 <https://doi.org/10.17645/up.v1i2.609>

597 Guimaraes, A., Balalau, O., Terolli, E. and Weikum, G., 2019. Analyzing the traits and anomalies of
598 political discussions on reddit. In *Proceedings of the International AAAI Conference on Web and*
599 *Social Media* (Vol. 13, pp. 205-213).

600 Haines-Young, R. and Potschin, M., 2010. The links between biodiversity, ecosystem services and
601 human well-being. *Ecosystem Ecology: a new synthesis*, 1, pp.110-139.

602 Hale, R.L., Cook, E.M. and Beltrán, B.J., 2019. Cultural ecosystem services provided by rivers across
603 diverse social-ecological landscapes: A social media analysis. *Ecological Indicators*, 107, p.105580.
604 <https://doi.org/10.1016/j.ecolind.2019.105580>

605 Harrigian, K., 2018. Geocoding without geotags: A text-based approach for reddit. arXiv preprint
606 arXiv:1810.03067.

607 Hart, A.G., Carpenter, W.S., Hlustik-Smith, E., Reed, M. and Goodenough, A.E., 2018. Testing the
608 potential of Twitter mining methods for data acquisition: Evaluating novel opportunities for ecological
609 research in multiple taxa. *Methods in Ecology and Evolution*, 9(11), pp.2194-2205.
610 <https://doi.org/10.1111/2041-210X.13063>

611 Havinga, I., Bogaart, P.W., Hein, L., Tuia, D., 2020. Defining and spatially modelling cultural
612 ecosystem services using crowdsourced data. *Ecosyst. Serv.* 43, 101091.
613 <https://doi.org/10.1016/j.ecoser.2020.101091>

614 Isaac, N.J., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., Freeman, S.N.,
615 Golding, N., Guillera-Aroita, G., Henrys, P.A. and Jarvis, S., 2020. Data integration for large-scale
616 models of species distributions. *Trends in ecology & evolution*, 35(1), pp.56-67.
617 <https://doi.org/10.1016/j.tree.2019.08.006>

618 Jamnik, M.R. and Lane, D.J., 2017. The use of Reddit as an inexpensive source for high-quality data.
619 *Practical Assessment, Research, and Evaluation*, 22(1), p.5. <https://doi.org/10.7275/swgt-rj52>

620 Johnson, M.L., Campbell, L.K., Svendsen, E.S. and McMillen, H.L., 2019. Mapping Urban Park
621 Cultural Ecosystem Services: A Comparison of Twitter and Semi-Structured Interview Methods.
622 Sustainability, 11(21), p.6137. <https://doi.org/10.3390/su11216137>

623 jReddit. 2020. jReddit. <https://github.com/jReddit/jReddit> [Online; accessed 2020-01-01]

624 Kahle, D., Wickham, H., 2013. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-
625 161. URL <http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf>

626 Kearney, M.W., 2020. rreddit: Collecting reddit data. R package version 0.0.1.
627 <https://github.com/mkearney/rreddit>

628 Kim, H.Y., 2017. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test.
629 Restorative dentistry & endodontics, 42(2), pp.152-155. <https://doi.org/10.5395/rde.2017.42.2.152>

630 King, H.P., Morris, J., Graves, A., Bradbury, R.B., McGinlay, J. and Bullock, J.M., 2017. Biodiversity
631 and cultural ecosystem benefits in lowland landscapes in southern England. *Journal of Environmental*
632 *Psychology*, 53, pp.185-197. <https://doi.org/10.1016/j.jenvp.2017.08.002>

633 Kipp, M.E., Beak, J. and Choi, I., 2017. Motivations and intentions of flickr users in enriching flickr
634 records for library of congress photos. *Journal of the Association for Information Science and*
635 *Technology*, 68(10), pp.2364-2379. <https://doi.org/10.1002/asi.23869>

636 Langemeyer, J., Calcagni, F. and Baro, F., 2018. Mapping the intangible: Using geolocated social
637 media data to examine landscape aesthetics. *Land use policy*, 77, pp.542-552.
638 <https://doi.org/10.1016/j.landusepol.2018.05.049>

639 Lee, H., Seo, B., Koellner, T. and Lautenbach, S., 2019. Mapping cultural ecosystem services 2.0–
640 Potential and shortcomings from unlabeled crowd sourced images. *Ecological Indicators*, 96, pp.505-
641 515. <https://doi.org/10.1016/j.ecolind.2018.08.035>

642 McDonald, Y.J., Schwind, M., Goldberg, D.W., Lampley, A. and Wheeler, C.M., 2017. An analysis of
643 the process and results of manual geocode correction. *Geospatial health*, 12(1), p.526.
644 <https://doi.org/10.4081/gh.2017.526>

645 Mendoza, M., Poblete, B. and Valderrama, I., 2019. Nowcasting earthquake damages with Twitter.
646 EPJ Data Science, 8(1), p.3. <https://doi.org/10.1140/epids/s13688-019-0181-0>

647 Milcu, A.I., Hanspach, J., Abson, D., Fischer, J., 2013. Cultural ecosystem services: A literature
648 review and prospects for future research. *Ecol. Soc.* 18, 44-88. [https://doi.org/10.5751/ES-05790-](https://doi.org/10.5751/ES-05790-180344)
649 180344

650 Sinclair, M., Mayer, M., Woltering, M., Ghermandi, A., 2020. Valuing nature-based recreation using a
651 crowdsourced travel cost method: A comparison to onsite survey data and value transfer. *Ecosyst.*
652 *Serv.* 45, 101165. <https://doi.org/10.1016/j.ecoser.2020.101165>

653 Moreno-Llorca, R., Méndez, P.F., Ros-Candeira, A., Alcaraz-Segura, D., Santamaría, L., Ramos-
654 Ridaio, Á.F., Revilla, E., García, F.J.B. and Vaz, A.S., 2020. Evaluating tourist profiles and nature-
655 based experiences in Biosphere Reserves using Flickr: Matches and mismatches between online
656 social surveys and photo content analysis. *Science of The Total Environment*, p.140067.
657 <https://doi.org/10.1016/j.scitotenv.2020.140067>

658 Oakes, M.P. and Farrow, M., 2006. Use of the chi-squared test to examine vocabulary differences in
659 English language corpora representing seven different countries. *Literary and linguistic*
660 *computing*, 22(1), pp.85-99. <https://doi.org/10.1093/lc/fql044>

661 Oteros-Rozas, E., Martín-López, B., Fagerholm, N., Bieling, C. and Plieninger, T., 2018. Using social
662 media photos to explore the relation between cultural ecosystem services and landscape features
663 across five European sites. *Ecological Indicators*, 94, pp.74-86.
664 <https://doi.org/10.1016/j.ecolind.2017.02.009>

665 Park, A., Conway, M. and Chen, A.T., 2018. Examining thematic similarity, difference, and
666 membership in three online mental health communities from Reddit: a text mining and visualization
667 approach. *Computers in human behavior*, 78, pp.98-112. <https://doi.org/10.1016/j.chb.2017.09.001>

668 Peña-Aguilera, P., Burguillo-Madrid, L., Barve, V., Aragón, P. and Jiménez-Valverde, A., 2019. Niche
669 segregation in Iberian Argiope species. *The Journal of Arachnology*, 47(1), pp.37-44.
670 <https://doi.org/10.1636/0161-8202-47.1.37>

671 Retka, J., Jepson, P., Ladle, R.J., Malhado, A.C., Vieira, F.A., Normande, I.C., Souza, C.N.,
672 Bragagnolo, C. and Correia, R.A., 2019. Assessing cultural ecosystem services of a large marine
673 protected area through social media photographs. *Ocean & Coastal Management*, 176, pp.40-48.
674 <https://doi.org/10.1016/j.ocecoaman.2019.04.018>

675 Rinker, T.W., 2015. entity: Named entity recognition version 0.1.0. University at Buffalo. Buffalo, New
676 York. <http://github.com/trinker/entity>

677 Rivera, I., 2019. RedditExtractoR: Reddit Data Extraction Toolkit. R package version 2.1.5.
678 <https://CRAN.R-project.org/package=RedditExtractoR>

679 Roberts, H.V., 2017. Using Twitter data in urban green space research. Appl. Geogr, 81, pp.13-20.
680 DOI:10.1016/j.apgeog.2017.02.008

681 Schirpke, U., Timmermann, F., Tappeiner, U. and Tasser, E., 2016. Cultural ecosystem services of
682 mountain regions: Modelling the aesthetic value. Ecological indicators, 69, pp.78-90.
683 <https://doi.org/10.1016/j.ecolind.2016.04.001>

684 Schwemmer, C., 2019. imgrec: An Interface for Image Recognition. R package version 0.1.1.
685 <https://github.com/cschwem2er/imgrec>

686 Sharp, R., Douglass, J., Wolny, S., Arkema, K., Bernhardt, J., Bierbower, W., Chaumont, N., Denu,
687 D., Fisher, D., Glowinski, K., Griffin, R., Guannel, G., Guerry, A., Johnson, J., Hamel, P., Kennedy, C.,
688 Kim, C.K., Lacayo, M., Lonsdorf, E., Mandle, L., Rogers, L., Silver, J., Toft, J., Verutes, G., Vogl, A. L.,
689 Wood, S, and Wyatt, K. 2020, InVEST 3.8.7.post9+ug.ga50c7f5 User's Guide. The Natural Capital
690 Project, Stanford University, University of Minnesota, The Nature Conservancy, and World Wildlife
691 Fund. [https://storage.googleapis.com/releases.naturalcapitalproject.org/invest-
692 userguide/latest/index.html](https://storage.googleapis.com/releases.naturalcapitalproject.org/invest-userguide/latest/index.html)

693 Sinclair, M., Mayer, M., Woltering, M. and Ghermandi, A., 2020. Valuing nature-based recreation
694 using a crowdsourced travel cost method: A comparison to onsite survey data and value transfer.
695 Ecosystem Services, 45, p.101165. <https://doi.org/10.1016/j.ecoser.2020.101165>

696 Thiagarajah, J., Wong, S.K., Richards, D.R. and Friess, D.A., 2015. Historical and contemporary
697 cultural ecosystem service values in the rapidly urbanizing city state of Singapore. Ambio, 44(7),
698 pp.666-677. <https://doi.org/10.1007/s13280-015-0647-7>

699 Tieskens, K.F., Van Zanten, B.T., Schulp, C.J. and Verburg, P.H., 2018. Aesthetic appreciation of the
700 cultural landscape through social media: An analysis of revealed preference in the Dutch river
701 landscape. Landscape and Urban Planning, 177, pp.128-137.
702 <https://doi.org/10.1016/j.landurbplan.2018.05.002>

703 van Zanten, B.T., Van Berkel, D.B., Meentemeyer, R.K., Smith, J.W., Tieskens, K.F. and Verburg,
704 P.H., 2016. Continental-scale quantification of landscape values using social media
705 data. *Proceedings of the National Academy of Sciences*, 113(46), pp.12974-12979.
706 <https://doi.org/10.1073/pnas.1614158113>

707 Völske, M., Potthast, M., Syed, S. and Stein, B., 2017, September. Tl; dr: Mining reddit to learn
708 automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization* (pp.
709 59-63).

710 Wang, Z., Ye, X. and Tsou, M.H., 2016. Spatial, temporal, and content analysis of Twitter for wildfire
711 hazards. *Natural Hazards*, 83(1), pp.523-540. <https://doi.org/10.1007/s11069-016-2329-6>

712 Wood, S.A., Guerry, A.D., Silver, J.M., Lacayo, M., 2013. Using social media to quantify nature-based
713 tourism and recreation. *Sci. Rep.* 3, 1-7. <https://doi.org/10.1038/srep02976>

714 Zhang, S. and Zhou, W., 2018. Recreational visits to urban parks and factors affecting park visits:
715 Evidence from geotagged social media data. *Landscape and urban planning*, 180, pp.27-35.
716 <https://doi.org/10.1016/j.landurbplan.2018.08.004>

717