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

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

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

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

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

Clean Version

Reddit: A novel data source for cultural ecosystem service studies

Abstract

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

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

1.0 Introduction

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

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

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

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

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

2.0 Methods

Here we present multiple methods for searching Reddit for data suitable for CES assessments via its Application Programming Interface (API), a computing interface that allows researchers to access a platform via code. First, we searched the site for all posts containing a specific keyword and compared these posts to those found using the same keyword search on Flickr. Second, we searched for posts on subreddits that are based around topics of interest to CES research. Third, we demonstrate a method for geocoding an estimated location for posts from Reddit as well as combining a place keyword search with another keyword, or within a subreddit to find posts linked to a particular location.

2.1 Data sources

2.1.1 Reddit

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

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

Service	**Subreddit**	**Extract of group description**	**Number of members (10th 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

Posts and comments from Reddit can be searched and returned through the Reddit API, with text and image posts, as well as their metadata including the, title, comments, date posted, how many upvotes (the number of people that like a post) a post has, and the ratio of upvotes to downvotes (the number of people that dislike a post). These data types are similar to data already being used in CES and social media studies derived from Flickr, Instagram and Twitter.

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

2.1.2 Flickr

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

2.2 Data collection and analysis

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

2.2.1 Keyword search

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

To compare the contents of the images posted on the two sites, we took a random sample of 1,000 images related to hiking from both sites (images listed as adult material were not included in the sample selection). The contents of the images were automatically tagged using the Google Cloud Vision API (Google Cloud Vision 2020). The Google Cloud Vision API is a machine learning algorithm that labels the content of images. The algorithm is based on a large pre-trained dataset and can label image contents into millions of predefined categories including objects and expressions. Here, we used the "imgrec" R package (Schwemmer 2019) to label each image with the 10 objects the algorithm first detects. To ensure that the image contents were accurate without manual validation we only kept labels that had a confidence score of > 0.6 (Gosal et al. 2019).

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

$${x^2}_i = \frac{\left(\mathrm{obs}_i - \mathrm{exp}_i\right)}{\mathrm{exp}_i} \quad (1),$$

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

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

2.2.2 Subreddit Search

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

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.”

2.2.3 Potential spatial uses for Reddit

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

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

3.0 Results

3.1 Full datasets

For each activity, the number of posts vary across each site. For hiking there were a similar number of posts uploaded to each site in 2020 with 145,036 hiking posts on Reddit and 148,535 on Flickr. There were also a similar number of posts relating to skiing across the two sites: 41,703 post on Reddit and 59,455 posts on Flickr. For camping, more posts were uploaded to Reddit (143,446) than Flickr (66,818); however for kayaking more posts were uploaded to Flickr (48,659) than Reddit (15,107). The number of uploads fluctuate across the year for both websites (Fig. 1). For hiking and skiing, even though there were a similar number of posts, Reddit had a greater quantity of unique users generating the posts. For hiking, Reddit had 88,075 unique users posting whilst Flickr had 9,392, while for skiing Reddit had 20,934 unique users whilst Flickr had 4,309.

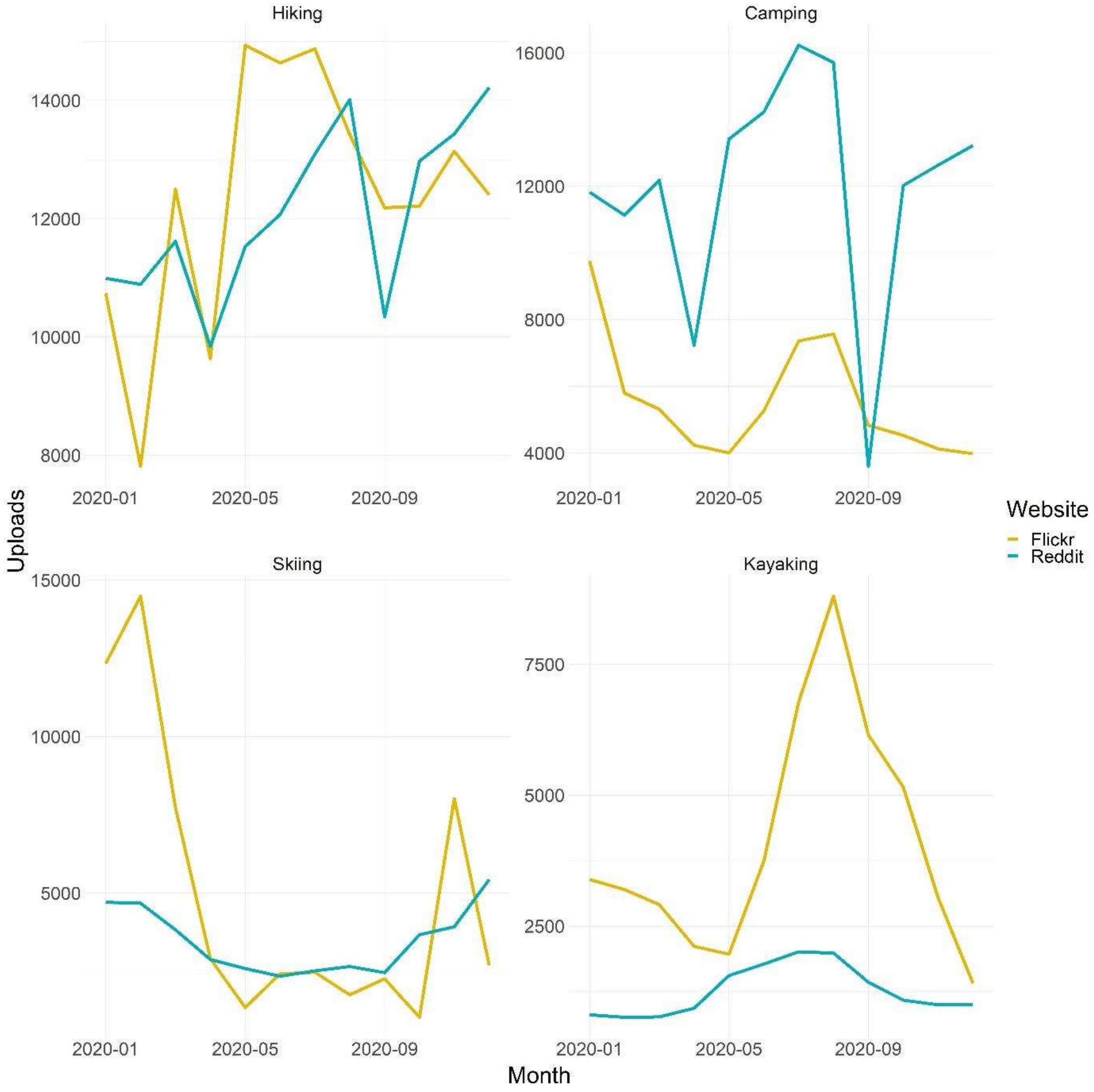

Fig.1: Uploads of posts including the words "hiking", "camping", "skiing" and "camping" to Reddit and Flickr between the 1st of January 2020 and the 1st of January 2021.

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

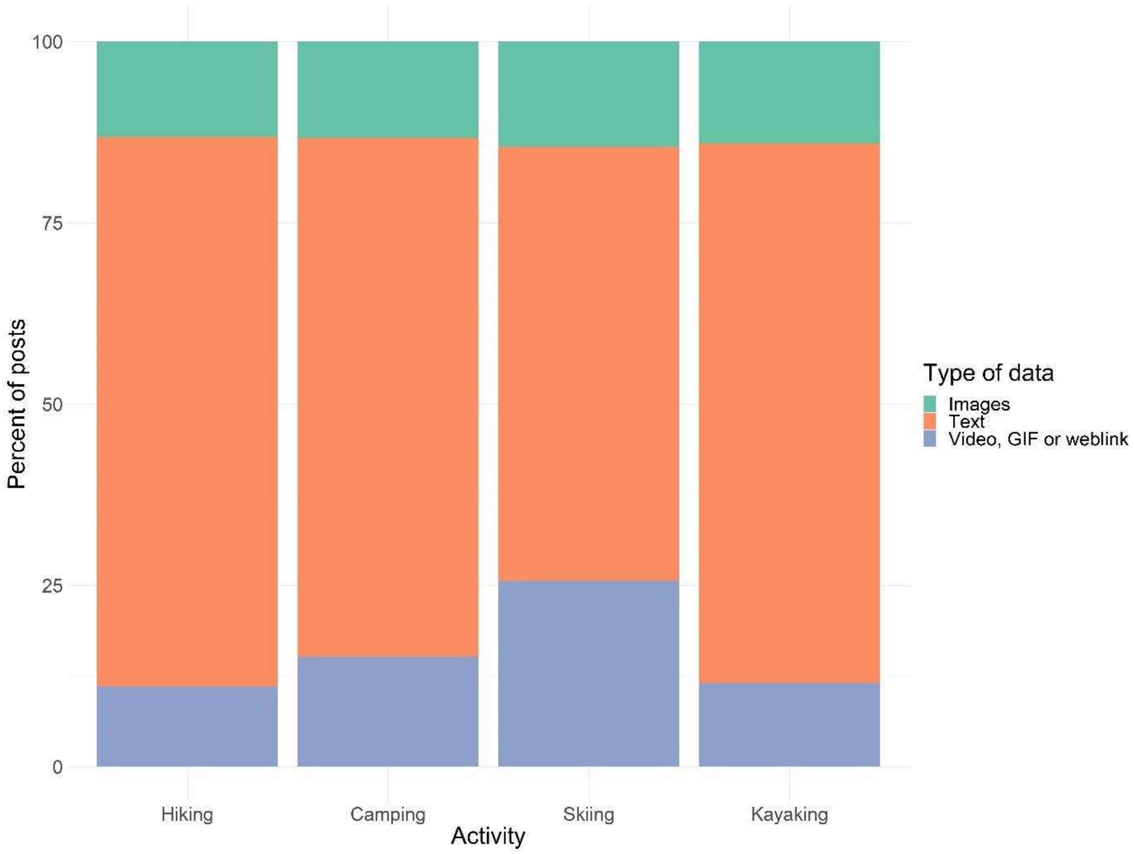

Fig. 2: Types of posts uploaded to Reddit.

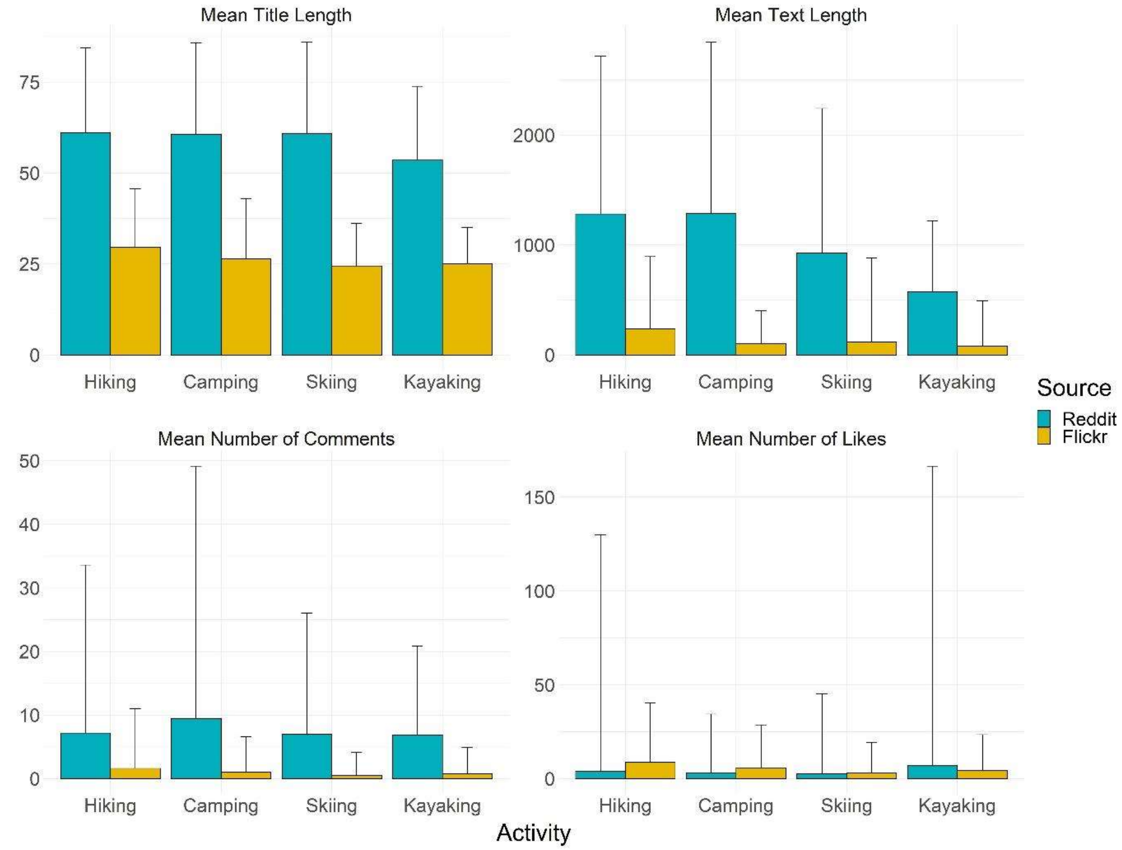

Fig. 3. Summary of posts made on Reddit and Flickr (mean + 0.5 standard deviations).

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

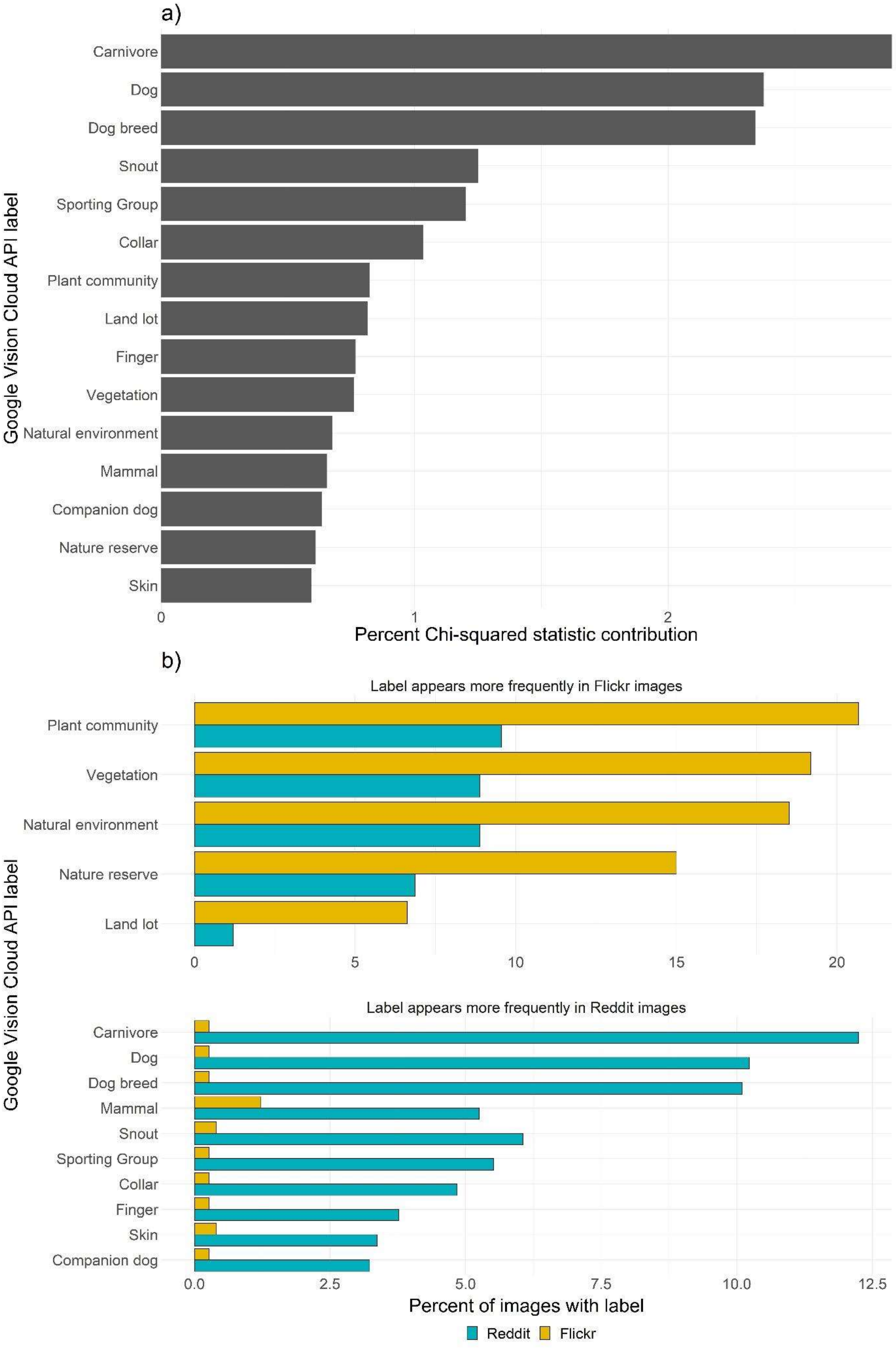
a)
Google Vision Cloud API label
Carnivore
Dog
Dog breed
Snout
Sporting Group
Collar
Plant community
Land lot
Finger
Vegetation
Natural environment
Mammal
Companion dog
Nature reserve
Skin
0
1
2
Percent Chi-squared statistic contribution
b)
Label appears more frequently in Flickr images
Plant community
Vegetation
Natural environment
Nature reserve
Land lot
0
5
10
15
20
Label appears more frequently in Reddit images
Carnivore
Dog
Dog breed
Mammal
Snout
Sporting Group
Collar
Finger
Skin
Companion dog
0.0
2.5
5.0
7.5
10.0
12.5
Percent of images with label
Reddit
Flickr

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

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

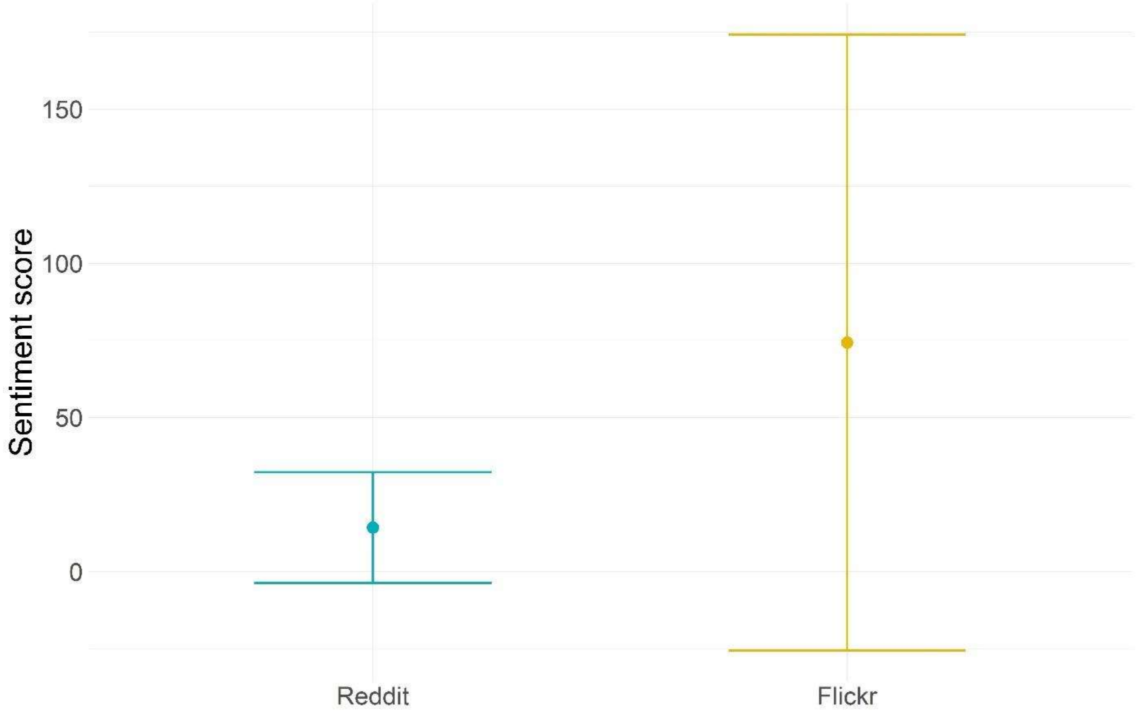

Fig. 5: Mean +- 0.5 standard deviations for the AFINN sentiment score expressed in the comments of hiking images on Reddit and Flickr.

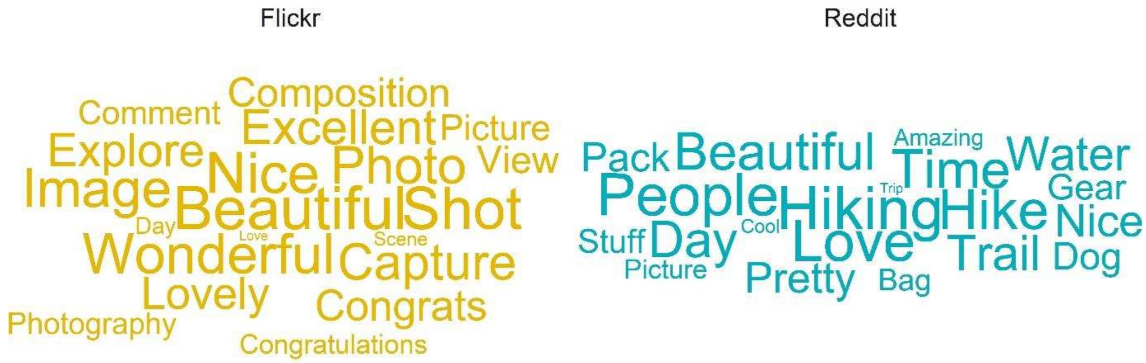

Fig. 6: The 20 most frequently used word in Flickr and Reddit comments after filtering.

3.2 Subreddit search

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

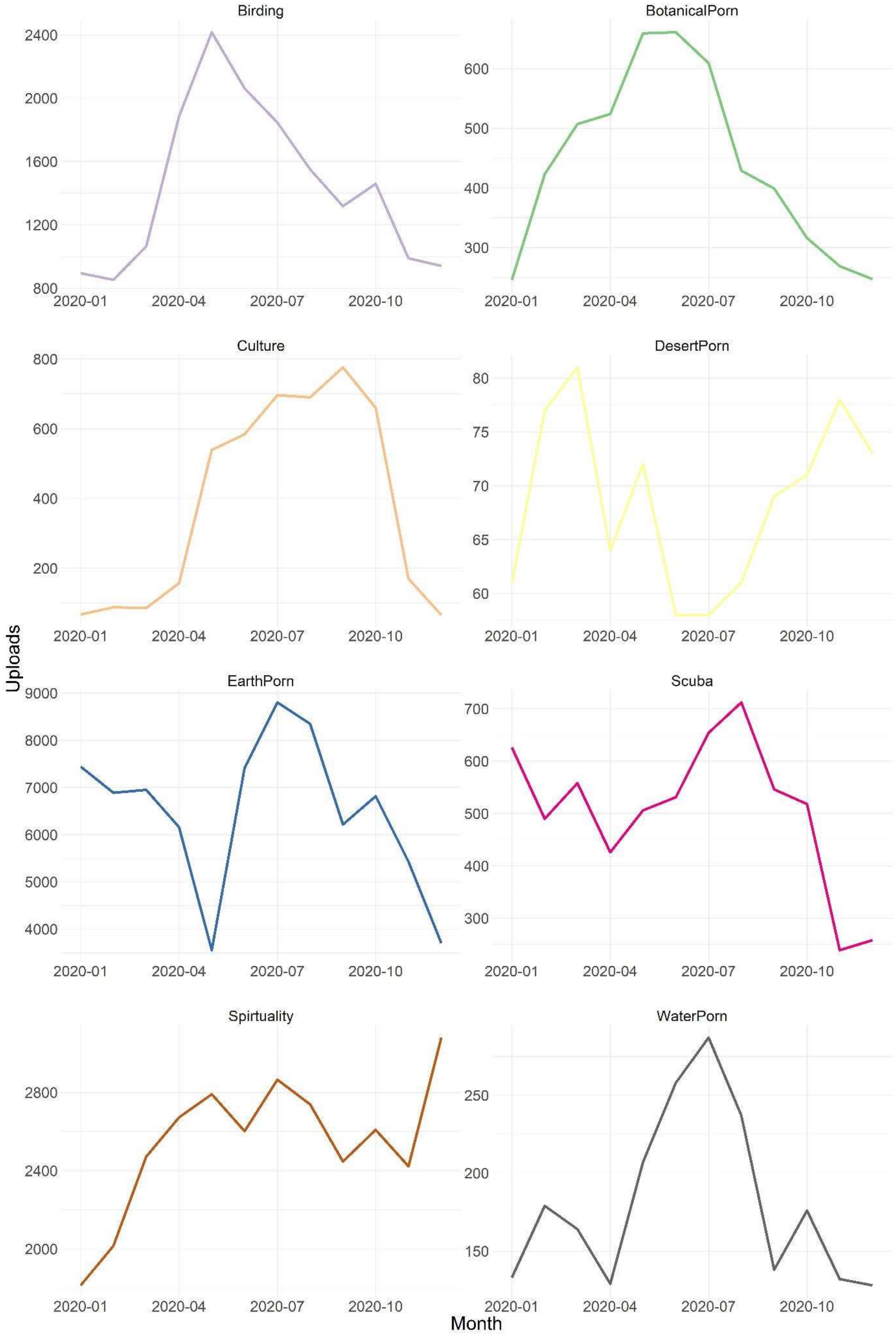

Figure 7: Uploads of posts to the subreddits “r/Birding”, “r/BotanicalPorn”, “r/Culture”,

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

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

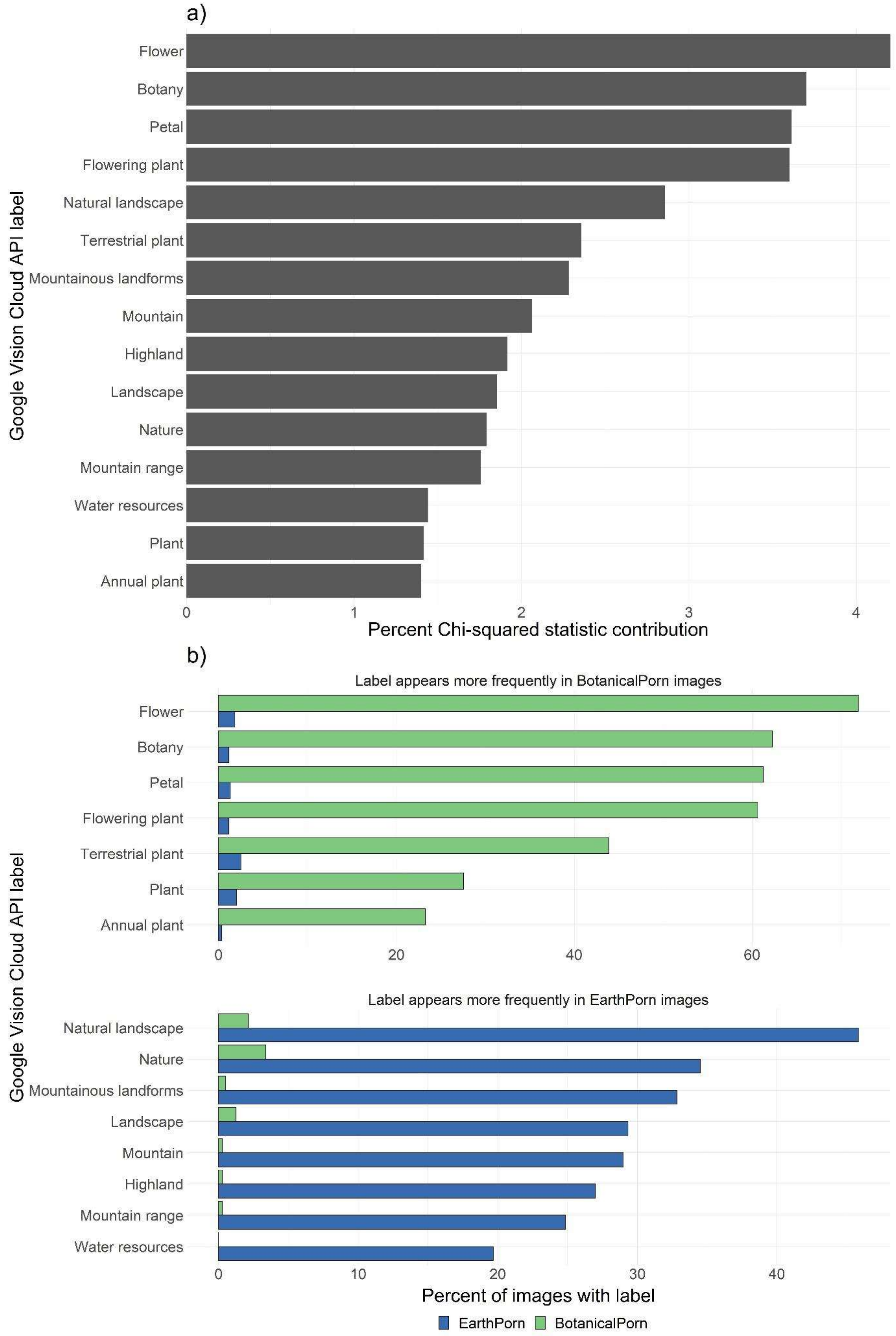
a)
Google Vision Cloud API label
Flower
Botany
Petal
Flowering plant
Natural landscape
Terrestrial plant
Mountainous landforms
Mountain
Highland
Landscape
Nature
Mountain range
Water resources
Plant
Annual plant
0
1
2
3
4
Percent Chi-squared statistic contribution
b)
Label appears more frequently in BotanicalPorn images
Google Vision Cloud API label
Flower
Botany
Petal
Flowering plant
Terrestrial plant
Plant
Annual plant
0
20
40
60
Label appears more frequently in EarthPorn images
Natural landscape
Nature
Mountainous landforms
Landscape
Mountain
Highland
Mountain range
Water resources
0
10
20
30
40
Percent of images with label
EarthPorn
BotanicalPorn

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

3.3 Potential spatial uses for Reddit

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

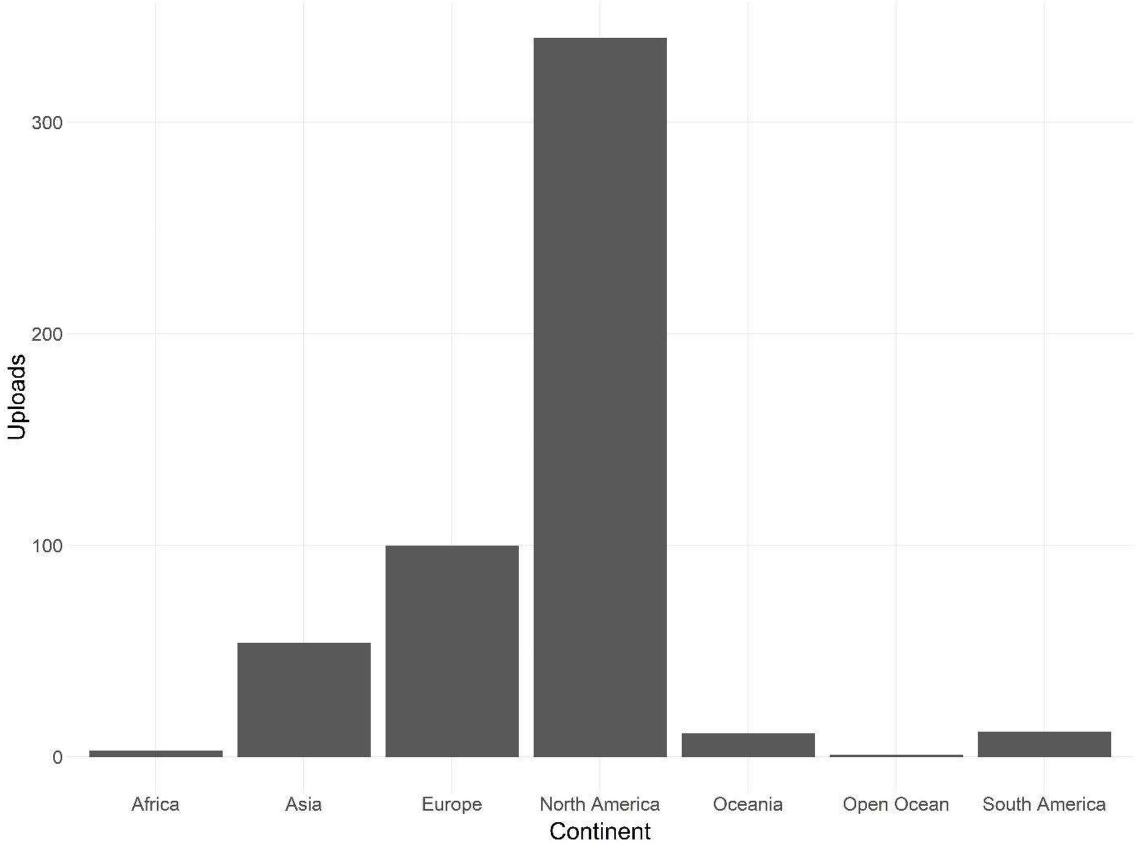

Figure 9: Estimated locations of a subset of photographs from the “r/EarthPorn” subreddit.

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

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”
		Subreddit = any
			13,148

		Text = “USA”
		Subreddit = “r/EarthPorn”
			12,336

	Regional	Text = “Wyoming” AND “hiking”	1,209

		Subreddit = any	

		Text = “Wyoming”
		Subreddit = “r/EarthPorn”
			3,399

	National park	Text = “Yellowstone” AND “hiking”
		Subreddit = any
			2,794

		Text = “Yellowstone”
		Subreddit = “r/EarthPorn”
			4,334

UK	National	Text = “UK” AND “hiking”
		Subreddit = any
			8,196

		Text = “UK”
		Subreddit = “r/EarthPorn”
			5,539

	Regional	Text = “Scotland” AND “hiking”
		Subreddit = any
			2,528

		Text = “Scotland”
		Subreddit = “r/EarthPorn”
			5,539

	National park	Text = “Cairngorms” AND “hiking”
		Subreddit = any
			87

		Text = “Cairngorms”
		Subreddit = “r/EarthPorn”
			131

4.0 Discussion

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

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

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

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

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

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

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

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

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

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

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

As Reddit is designed to be a discussion-based forum it may contribute to richer information on the users’ perspective of CES. For example, the “r/Spirituality” subreddit encourages users to contribute to the discussion of any aspects of spirituality regardless of religion or ideology, thus providing the potential to assess the opinions of people from a wide range of backgrounds. Furthermore, Reddit comments can be longer than most other social media sites (e.g. Twitter has a 280-character limit and Instagram has a 300-character comment limit) and therefore a user can discuss their opinions in greater detail (Gkotsis et al. 2017). The discursive nature of Reddit provides researchers a unique opportunity to assess which aspects of a certain image or video people appreciate. There is also scope for this interactive and discussion-based platform to be used in experimental studies in which researchers post content and monitor feedback. Though as with all social media-based studies, we recommended that the ethics of these studies be discussed in further detail. We suggest that Reddit data is particularly useful for studies that wish to analyse users' comments in conjunction with the metadata available for each image for a more robust assessment of CES.

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

5.0 Conclusion

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