

# Digital Imaginations of National Parks in Different Social Media: A Data Exploration

Vuokko Heikinheimo<sup>1</sup>, Henrikki Tenkanen<sup>1</sup>, Tuomo Hiippala<sup>1,2</sup>, and  
Tuuli Toivonen<sup>1</sup>

<sup>1</sup>Digital Geography Lab, Department of Geosciences and Geography, University of Helsinki, Finland

<sup>2</sup>Department of Languages, University of Helsinki, Finland

Social media contains a wealth of information about human activities in different places. This information can complement data collection efforts in resource-scarce fields such as nature conservation. However, social media platforms differ in popularity, content, and access to data, and the choice of platform may greatly affect the resulting analysis. We explored Flickr, Instagram, and Twitter data from 39 Finnish national parks over a period of two years to assess the fitness-for-purpose of each platform for understanding place-based experiences of national park visitors. From Instagram, we extracted data using two different approaches: coordinate search and keyword search. Furthermore, we identified the languages used in Instagram data using the fastText library, and conducted preliminary content analysis of Flickr and Twitter data using Google Cloud Vision image annotation service. Instagram was the most popular platform in all national parks. Noteworthy, almost 50% of Twitter users had shared their geotagged national park post to Twitter via Instagram. Language identification from text content and content analysis of images provide basis for further exploration of the digital representations of national parks and place-related experiences of visitors.

**Keywords:** social media; content analysis; national parks; Flickr; Instagram; Twitter

## 1 Introduction

Information available in Web 2.0 provides new opportunities for geographic knowledge discovery (Menis and Guo, 2009; Stefanidis et al., 2013; Sui and Goodchild, 2011) especially in data and resource-scarce fields such as nature conservation (Arts et al., 2015; Di Minin et al., 2015). Data on where, when, and why people visit and appreciate different places has become easier to gather in the era of big data, as people share their experiences and observations online in social media. Geosocial media platforms such as Flickr, Instagram, and Twitter contain a wealth of text, image, and video content about people's opinions, observations, activities, and experiences. However, social media platforms differ in popularity and purpose of use, and the choice of platform can greatly influence the observed patterns. Furthermore, spatial context including the cultural and physical environment plays a role in defining what content is shared and by whom.

Nature-based tourism has become increasingly popular around the world, with an estimated total of 8 billion visits per year to terrestrial protected areas (Balmford et al., 2015). Thus far, social media usage has been found to correlate with official visitor statistics of nature destinations (Hausmann et al., 2017; Tenkanen et al., 2017; Wood et al., 2013). Furthermore, it has been shown that geotagged social media content corresponds to surveyed activities (Heikinheimo et al., 2017) and preferences in terms of biodiversity (Hausmann et al., 2017) in selected national parks. Geosocial media has great potential

---

V Heikinheimo, H Tenkanen, T Hiippala, and T Toivonen (2018): *Digital Imaginations of National Parks in Different Social Media: A Data Exploration*. In: R Westerholt, F-B Mocnik, and A Zipf (eds.), *Proceedings of the 1st Workshop on Platial Analysis (PLATIAL'18)*, pp. 45–52

<https://doi.org/10.5281/zenodo.1472745>



First Workshop on Platial Analysis (PLATIAL'18)  
Heidelberg, Germany; 20–21 September 2018

Copyright © by the authors. Licensed under Creative Commons Attribution 4.0 License.

to inform protected area visitor monitoring and management (Di Minin et al., 2015), but more research is needed regarding the inherent bias in social media data and the suitability of different platforms for extracting information about place-related experiences in nature destinations.

Meanings associated with specific places have been extracted and analysed from user-generated content especially in urban environments with Twitter as the main data source (Jenkins et al., 2016; Shelton et al., 2015; Steiger et al., 2015). In environmental sciences, social media – especially Flickr and Panoramio – have been used for understanding landscape values and ecosystem services (benefits people get from nature) (Oteros-Rozas et al., 2016; Richards and Friess, 2015; Van Berkel et al., 2018; van Zanten et al., 2016). Recent studies have compared the spatial and temporal patterns of different social media to official statistics (Levin et al., 2017; Tenkanen et al., 2017). However, most studies – especially those focusing on content analysis – have so far relied on single source of social media (mostly Flickr or Panoramio in environmental studies, and Twitter in urban studies).

Social media content shared from natural and semi-natural areas has specific qualities in comparison with data from urban areas. Firstly, photos shared from protected areas are likely to be related to nature or cultural heritage. Secondly, most people who share content from national parks are likely to be visitors (domestic or international) enjoying their leisure time in the park (i. e., not living or working in the area). Also, some people might “log off” from social media during their visit or reduce their social media activities in order to save battery of their mobile device. People might use social media differently while being in a national park compared with built-up areas, and this should be taken into account when analysing these data.

Social media content can be attached to a place – in our case the national park – in different ways and at different scales. Coordinates and georeferenced place-tags (or points of interest) are the most technical way for the user to place their content on a map, and the precision and accuracy of geotagged data varies between platforms (Hochmair et al., 2018). It is also important to acknowledge that only a small percentage of all social media content is geotagged (it is estimated that 1% of all tweets are geotagged). Much of the place-related information is shared using place names and hashtags within text in a social media post, as is common in regular human discourse (Goodchild, 2011). Furthermore, images shared on social media also contain a wealth of place-related information.

In this short paper, we investigate how places of nature recreation – national parks in specific – are represented in different social media and discuss the possibilities and limitations of characterizing national parks (or people’s imaginations of national parks) from digital content. We compare and contrast data collected from Flickr, Instagram, and Twitter in terms of data volume, user base, and content. Furthermore, we analyse in more detail the text content of Instagram data in order to find out what language people use when sharing their experiences on social media. We also conducted preliminary analysis of image content from Flickr and Twitter. This data exploration provides a basis for more in-depth analysis of place-based experiences in recreational areas, and material for discussing the following questions: Which platforms are most suitable sources of place-related information from recreation areas? What is the best way of acquiring such data? Who have generated the digital data about a place?

## 2 Material and Methods

### 2.1 Study Area

This study covers all 39 Finnish national parks that existed in 2015. According to the international definition, national parks are large protected areas designed to protect ecosystems and to provide recreational opportunities to visitors (IUCN). In Finland, national parks are free of charge to everyone. Facilities such as campfire sites, nature trails, wilderness cabins, and latrines are maintained by the state-owned national park organization. In 2015, the 39 national parks attracted 2,634,600 visitors with a 15% increase from 2014 (<http://www.metsa.fi/kansallispuistotyhteensa>).

### 2.2 Data Collection

**Spatial Search.** We collected geotagged social media data from three different geosocial media platforms: Flickr, Instagram, and Twitter. Data was retrieved via the Application Programming

**Table 1:** Social media data from the 39 National Parks in 2014 and 2015.

Dataset	Posts	Users		
		total	min	max
<b>Flickr</b>	<b>2 283</b>	<b>118</b>	0	57
<b>Instagram</b>	<b>7 627</b>	<b>4 137</b>	4	1 308
<b>Instagram, keyword search</b>	<b>110 176</b>	<b>42 931</b>		
<i>geotagged</i>	17 060	8 402		
<i>geotagged in Finland</i>	13 040	6 113		
<i>geotagged in national parks</i>	3 119	1 908		
<b>Twitter</b>	<b>5 567</b>	<b>729</b>		
<i>excluding bots</i>	3 979	728	1	130
<i>tweets with any link</i>	5 567	729		
<i>source: www.instagram.com</i>	653	356		
<i>source: www.twitter.com</i>	553	185		
<i>thereof images</i>	435	41		
<i>source: www.swarmapp.com</i>	100	64		
<i>source: www.youtube.com</i>	166	12		

Interface (API) of each platform using a spatial query based on bounding boxes (Flickr and Twitter) or buffer zones around point locations (Instagram). For more details on the data collection, see methods presented by Tenkanen et al. (2017).

**Keyword Search.** From Instagram, another dataset was collected using a keyword-based media search with a manually collected list of place-names. Altogether, 504 place names at different spatial scales (ranging from trail names to park names in different languages) were included in the keyword list. All place names mentioned in visitor surveys and websites (<http://luontoon.fi>) of each national park were included in the list. The list was complemented with a manual search of place-names used in Instagram.

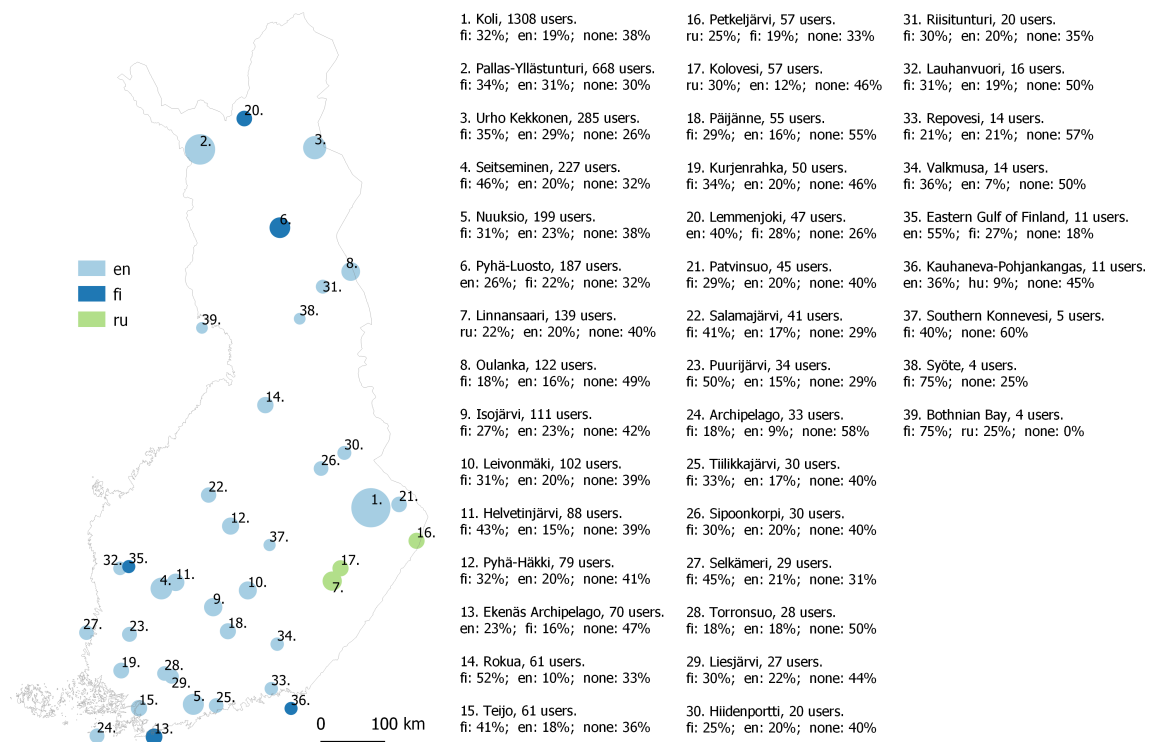
**Pre-processing.** All spatial datasets (Flickr, Instagram, and Twitter) were originally collected for larger areas, after which points intersecting the national park areas were selected. All datasets were subset to years 2014 and 2015, which was the most recent fully overlapping time period in all the collected datasets. Most active users (based on the number of posts per user) in each data set were manually checked to see if the data is generated by a machine, i. e., a bot. Posts generated by bots were removed.

## 2.3 Content Analysis

**Source of Twitter Data.** Twitter content was separately inspected in terms of data source. Any link shared on Twitter is automatically shortened by Twitter (<http://t.co>\*) in order to reduce the number of characters in a tweet. This happens if users share a link in their tweet or have shared their post first in Instagram or other social networking sites (link to the original post will be available as a shortened link). The source of each shortened link was detected using a custom script (Python 3.5) and labelled according to their original source.

**Languages.** For Instagram data, we applied automatic language identification to each post using a pre-trained model via the fastText library (Bojanowski et al., 2017), which supports a total of 176 languages out-of-the-box. To pre-process the data, we followed the procedure set out by Hiippala et al. (2018) to remove typical multilingual elements such as hashtags before retrieving predictions for each sentence in the caption. We then excluded very short sentences (< 7 characters) and predictions with a low confidence (< 0.5) from the results.

**Image Content.** Flickr and Twitter photo contents were labeled with up to 10 keywords using Google Cloud Vision (<https://cloud.google.com/vision/>) and the image annotation algorithm following the example of Richards and Tunçer (2017). Label detection was implemented in Python programming



**Figure 1:** Map of the most popular languages among Instagram users in Finnish national parks. The proportion of the first and second most popular languages among users, as well as the proportion of users for whom language could not be detected (“none”) are presented separately for each national park. The abbreviations used for different languages are “fi” for Finnish, “en” for English, and “ru” for Russian. Symbol size represents the number of Instagram users within each park.

language using the Google Cloud Vision library. Photographs uploaded on Instagram were not automatically analysed due to restrictions introduced in 2018. Further analysis will be done to summarize these results at a later stage, e.g., hierarchical clustering following Richards and Tunçer (2017) and Oteros-Rozas et al. (2016).

## 3 Results

### 3.1 Data Volume

**Coordinate Search.** Instagram and Twitter contained information from each of the 39 national parks, Instagram being the most popular platform (Table 1). Flickr had the highest ratio of posts per user. Based on the manual check of most active users, Twitter was the only data source where automatically generated data, i.e., a bot, could be identified. The number of Instagram users for each national park are presented in Figure 1.

**Keyword Search.** The keyword search of Instagram data resulted in a relatively large dataset out of which 16% contained coordinate information. Out of these geotagged place-name search results, the majority (76%) were located within Finland, and almost one fifth (18%) had their coordinates within one of the national parks. The keyword search dataset and coordinate search dataset from Instagram had 1867 records in common (25% of coordinate-search data within national parks were found also via the keyword search).

**Origins of Twitter Data.** After excluding the most evident bot (a Twitter account posting automatically Finnish numbers with a random geotag), the Twitter dataset contained 3979 tweets by 728 users. Over 80% of the remaining tweets contained one or several URL-links. There were 210 different source websites, most common being instagram.com (content from Instagram), twitter.com (mostly photos

shared via Twitter), swarmapp.com (Foursquare check-ins) and youtube.com (videos). Over half of the Twitter users in this dataset had shared content originally generated in other social media platforms.

### 3.2 Content

**Languages.** Language could be identified for 56% of all captions (the rest were excluded due to length or low confidence associated with automatic language identification) and for 65% of all users in the coordinate-search-based Instagram data. Most captions were monolingual, i. e., they were written in a single language. Only 1% of the posts were bilingual. Across the entire dataset, Finnish and English were the dominant languages, followed by Russian and Swedish. Altogether, 31% of the captions were primarily or entirely written in Finnish, 21% in English, 6% in Russian, and 1% in Swedish. Popular languages across all national parks are visualized in Figure 1.

## 4 Discussion and Conclusion

In this paper we explored the volume and properties of social media data shared from Finnish national parks within a two year time period. Previous research has examined the relationship between the temporal patterns of social media activity and official visitor statistics (Tenkanen et al., 2017) across several parks, and compared the results of manual content classification and traditional surveys about national park visitor preferences and activities (Hausmann et al., 2017; Heikinheimo et al., 2017). This data exploration sets grounds for upcoming work where we will utilise automated content analysis methods to understand place-related experiences in national parks, and compare the results with official visitor information.

In terms of data volume, all of the platforms contained information from the observed national parks, but Instagram was clearly the most prominent source of social media data during the observed time period. Furthermore, Instagram usage correlates well with temporal visitation rates in the parks (Tenkanen et al., 2017). However, access to Instagram data through the Instagram API has been hindered since 2016, which affects the use of these data for scientific use. In addition to data availability, spatial context plays a role in determining which social media platform is most fit for capturing place-based experiences of people. While Flickr has been popular data source in environmental studies (Levin et al., 2017; Richards and Friess, 2015; Richards and Tunçer, 2017), it has clearly the smallest user-base among the studied platforms.

Interestingly, many of the geotagged tweets in our data set had originally been generated in other location-based social media platforms (Instagram, Foursquare, or Youtube). Furthermore, 10% of all tweets contained an image (shared originally in Twitter). Twitter is most often used as a source of text-based analysis (Jenkins et al., 2016; Steiger et al., 2015), but also the image content shared in Twitter (or in Instagram via Twitter) contains a wealth of useful information. This also means that some of the geotagged Instagram data can still be accessed programmatically, despite the changes in the Instagram API.

We used two different approaches for collecting Instagram data from the study areas: a coordinate-based spatial query and a keyword-based place-name query. Unsurprisingly, the keyword search captured a lot of data that was not related to our areas of interest due to keywords with multiple purposes. For example, the keyword “Koli” is the name of a popular national park in Eastern Finland but also a language dialect in India, among other meanings. Trough the combination of the place-name query and a spatial query (selecting keyword search results within Finland or within National Park borders), we probably managed to exclude a lot of irrelevant content located outside Finnish national parks, but also potentially lost relevant non-geotagged information and relevant content posted outside the borders of the national parks. As Goodchild (2011) has argued, an intelligent strategy would be needed in order to develop a search which captures meanings of different places and place names correctly from human-generated content.

Automatic language identification helps us to understand who is sharing content in social media and for which audience. As such, geotagged language information may reveal the linguistic landscape of an area and give hints about the origins of visitors (Hiippala et al., 2018). In our case, the simple language detection revealed that national parks as largely visited by national visitors. Language identification is also a crucial pre-processing step for further text analysis methods, such as topic

modelling and sentiment analysis. In conjunction with other types of user-related information, language helps us to better understand who has generated the data and for what purpose, and whose digital imaginations of a place are represented in social media.

Social media data analysis from well-monitored national parks has the potential to provide new information about gaps and bias in different social media data (Hausmann et al., 2017; Tenkanen et al., 2017; Wood et al., 2013). National park visitor surveys and visitor counting serve as ground-truth information for patterns observed in online social media. Our further work will focus on comparing online social media content to experiences measured with more traditional methods such as visitor surveys across multiple parks.

Using location-based social media data in research requires constant reflection about data quality and research ethics (boyd and Crawford, 2012). Here (and in many other studies) social media has been mined from online sources without specific consent from users who have generated the data. However, just the fact that these data are online does not necessarily justify their use in respect to a new purpose (boyd and Crawford, 2012). In research, one should constantly consider the potential benefit and harm to anyone involved, and to ensure the protection of personal information (Monkman et al., 2017).


In sum, understanding place-based experiences from social media can benefit from applying existing machine learning methods as well as from the development of new, more efficient ways of automatically extracting and analysing place-related information. Well-monitored national parks, such as those in Finland, provide a convenient test environment, and an interesting application case for observing collective experiences and emerging phenomena from geosocial media. In future work we aim to deepen general understanding about place-related experiences of people in national parks based on social media data, and to provide further insights about which platforms are most suitable for extracting these information, what is the best way of acquiring such data, and who have generated the digital data about a place.


## Funding


We thank the Kone Foundation for support.

## ORCID

Vuokko Heikinheimo  <https://orcid.org/0000-0001-5119-0957>

Henrikki Tenkanen  <https://orcid.org/0000-0002-0918-4710>

Tuomo Hiippala  <https://orcid.org/0000-0002-8504-9422>

Tuuli Toivonen  <https://orcid.org/0000-0002-6625-4922>

## References

- Arts, Koen; van der Wal, René; and Adams, William M: *Digital technology and the conservation of nature*. *Ambio*, 44, 2015, 661–673. doi: 10.1007/s13280-015-0705-1
- Balmford, Andrew; Green, Jonathan MH; Anderson, Michael; et al.: *Walk on the wild side: estimating the global magnitude of visits to protected areas*. *PLoS Biology*, 13(2), 2015, e1002074. doi: 10.1371/journal.pbio.1002074
- Bojanowski, Piotr; Grave, Edouard; Joulin, Armand; and Mikolov, Tomas: *Enriching word vectors with subword information*. *Transactions of the Association for Computational Linguistics*, 5, 2017, 135–146
- boyd, danah and Crawford, Kate: *Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon*. *Information, Communication & Society*, 15(5), 2012, 662–679. doi: 10.1080/1369118X.2012.678878
- Di Minin, Enrico; Tenkanen, Henrikki; and Toivonen, Tuuli: *Prospects and challenges for social media data in conservation science*. *Frontiers in Environmental Science*, 3, 2015. doi: 10.3389/fenvs.2015.00063



- Goodchild, Michael F: *Formalizing place in geographic information systems*. In: Burton, Linda M; Matthews, Stephen A; Leung, ManChui; Kemp, Susan P; and Takeuchi, David T (eds.), *Communities, neighborhoods, and health: expanding the boundaries of place*, New York, NY: Springer, 2011. 21–34. doi: 10.1007/978-1-4419-7482-2\_2
- Hausmann, Anna; Toivonen, Tuuli; Slotow, Rob; et al.: *Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas*. *Conservation Letters*, 11(1), 2017. doi: 10.1111/conl.12343
- Heikinheimo, Vuokko; Minin, Enrico Di; Tenkanen, Henriikki; et al.: *User-generated geographic information for visitor monitoring in a national park: a comparison of social media data and visitor survey*. *ISPRS International Journal of Geo-Information*, 6(3), 2017, 85. doi: 10.3390/ijgi6030085
- Hiippala, Tuomo; Hausmann, Anna; Tenkanen, Henriikki; and Toivonen, Tuuli: *Exploring the linguistic landscape of geotagged social media content in urban environments*. *Digital Scholarship in the Humanities*, 2018. doi: 10.1093/llc/fqy049
- Hochmair, Hartwig H; Juhász, Levente; and Cvetojevic, Sreten: *Data quality of points of interest in selected mapping and social media platforms*. *Proceedings of the 14th International Conference on Location Based Services*, 2018, 293–313. doi: 10.1007/978-3-319-71470-7\_15
- Jenkins, Andrew; Croitoru, Arie; Crooks, Andrew T; and Stefanidis, Anthony: *Crowdsourcing a collective sense of place*. *PLoS ONE*, 11(4), 2016, e0152932. doi: 10.1371/journal.pone.0152932
- Levin, Noam; Lechner, Alex Mark; and Brown, Greg: *An evaluation of crowdsourced information for assessing the visitation and perceived importance of protected areas*. *Applied Geography*, 79, 2017, 115–126. doi: 10.1016/j.apgeog.2016.12.009
- Mennis, Jeremy and Guo, Diansheng: *Spatial data mining and geographic knowledge discovery – an introduction*. *Computers, Environment and Urban Systems*, 33(6), 2009, 403–408. doi: 10.1016/j.compenvurbsys.2009.11.001
- Monkman, Graham George; Kaiser, Michel; and Hyder, Kieran: *The ethics of using social media in fisheries research*. *Reviews in Fisheries Science & Aquaculture*, 26(2), 2017, 235–242. doi: 10.1080/23308249.2017.1389854
- Oteros-Rozas, Elisa; Martín-López, Berta; Fagerholm, Nora; Bieling, Claudia; and Plieninger, Tobias: *Using social media photos to explore the relation between cultural ecosystem services and landscape features across five European sites*. *Ecological Indicators*, 2016. doi: 10.1016/j.ecolind.2017.02.009
- Richards, Daniel R and Friess, Daniel A: *A rapid indicator of cultural ecosystem service usage at a fine spatial scale: content analysis of social media photographs*. *Ecological Indicators*, 53, 2015, 187–195. doi: 10.1016/j.ecolind.2015.01.034
- Richards, Daniel R and Tunçer, Bige: *Using image recognition to automate assessment of cultural ecosystem services from social media photographs*. *Ecosystem Services*, 31(C), 2017, 318–325. doi: 10.1016/j.ecoser.2017.09.004
- Shelton, Taylor; Poorthuis, Ate; and Zook, Matthew: *Social media and the city: rethinking urban socio-spatial inequality using user-generated geographic information*. *Landscape and Urban Planning*, 142, 2015, 198–211. doi: 10.1016/j.landurbplan.2015.02.020
- Stefanidis, Anthony; Crooks, Andrew; and Radzikowski, Jacek: *Harvesting ambient geospatial information from social media feeds*. *GeoJournal*, 78(2), 2013, 319–338. doi: 10.1007/s10708-011-9438-2
- Steiger, Enrico; Westerholt, René; Resch, Bernd; and Zipf, Alexander: *Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data*. *Computers, Environment and Urban Systems*, 54, 2015, 255–265. doi: 10.1016/j.compenvurbsys.2015.09.007
- Sui, Daniel and Goodchild, Michael F: *The convergence of GIS and social media: challenges for GI-Science*. *International Journal of Geographical Information Science*, 25(11), 2011, 1737–1748. doi: 10.1080/13658816.2011.604636

Tenkanen, Henrikki; Di Minin, Enrico; Heikinheimo, Vuokko; et al.: *Instagram, Flickr, or Twitter: assessing the usability of social media data for visitor monitoring in protected areas*. Scientific Reports, 7(17615), 2017. doi: 10.1038/s41598-017-18007-4

Van Berkel, Derek B; Tabrizian, Payam; Dorning, Monica A; et al.: *Quantifying the visual-sensory landscape qualities that contribute to cultural ecosystem services using social media and LiDAR*. Ecosystem Services, 31, 2018, 326–335. doi: 10.1016/j.ecoser.2018.03.022

Wood, Spencer A; Guerry, Anne D; Silver, Jessica M; and Lacayo, Martin: *Using social media to quantify nature-based tourism and recreation*. Scientific Reports, 3, 2013, 2976. doi: 10.1038/srep02976

van Zanten, Boris T; Van Berkel, Derek B; Meentemeyer, Ross K; et al.: *Continental-scale quantification of landscape values using social media data*. Proceedings of the National Academy of Sciences, 113(46), 2016, 12974–12979. doi: 10.1073/pnas.1614158113