Passive crowdsourcing of social media in environmental research: A systematic map

Andrea Ghermandi⁎, Michael Sinclair

Department of Natural Resources & Environmental Management, Faculty of Management, University of Haifa, 199 Aba Khoushy Ave., Mount Carmel, Haifa, 3498838, Israel

ABSTRACT

The analysis of data from social media and social networking sites may be instrumental in achieving a better understanding of human-environment interactions and in shaping future conservation and environmental management. In this study, we systematically map the application of social media data in environmental research. The quantitative review of 169 studies reveals that most studies focus on the analysis of people’s behavior and perceptions of the environment, followed by environmental monitoring and applications in environmental planning and governance. The literature testifies to a very rapid growth in the field, with Twitter (52 studies) and Flickr (34 studies) being most frequently used as data sources. A growing number of studies combine data from multiple sites and jointly investigates multiple types of media. A broader, more qualitative review of the insights provided by the investigated studies suggests that while social media data offer unprecedented opportunities in terms of data volume, scale of analysis, and real-time monitoring, researchers are only starting to cope with the challenges of data’s heterogeneity and noise levels, potential biases, ethics of data acquisition and use, and uncertainty about future data availability. Critical areas for the development of the field include integration of different types of information in data mashups, development of quality assurance procedures and ethical codes, improved integration with existing methods, and assurance of long-term, free and easy-to-access provision of public social media data for future environmental researchers.

1. Introduction

Crowdsourced data have become an important component in various research fields within the environmental sciences. The term refers to data that is collected and made available to researchers by non-professionals and citizen organizations, as opposed to professional scientists and government agencies (Conrad and Hilchey, 2011). Much has been written, for instance, on Volunteered Geographic Information (VGI) that is actively collected and shared by non-expert “neogeographers” (Goodchild, 2007; Elwood et al., 2012) and on the prospects and challenges of citizen science projects (Tipaldo and Allamano, 2017; Velwaert and Caley, 2016).

Less attention has been given, however, to the emerging field of passive (or opportunistic) crowdsourcing, which involves data that is generated by non-professionals but is collected and shared independently of formal citizen science projects, such as through the upload of information to web-based social networking sites. In such applications, the passive users (or “produsers”) submit information as Web content but such information is used by researchers for purposes other than that which the users originally intended (Connors et al., 2012). Such approach allows for observational (rather than experimental) studies and relies on a semi-automated data collection process, in which users typically collect information through a sensor (e.g., smart phone camera) and subsequently choose to upload it (Muller et al., 2015). Similarly to automated digital sensors, surveillance and tracking devices, passive crowdsourcing offers a continuous and direct flow of data on human activities that are relevant to environmental research applications (Arts et al., 2015).

In spite of the novelty of the field and the relatively small number of papers published thus far, several reviews have been published focusing on specific applications of social media data in fields such as natural disasters preparedness and monitoring (Anson et al., 2017; Finch et al., 2016), conservation science (Di Minin et al., 2015), climate change (Auer et al., 2014) and urban sustainability (Ilieva and McPhearson, 2018). The fast pace of growth and development in social media-related research (Li et al., 2017), however, risks to render reviews rapidly outdated. The rationale of this study lies in the observed lack of a wide-ranging, systematic review of literature that uses social media as a...
source of environmental data.

The objective of this paper is to systematically map the state of the art in environmental research using passively crowdsourced data. The previously given definition of passive crowdsourcing relies on information that is voluntarily shared by users, albeit not for the purpose for which it is used by the researchers. Accordingly, we focus on data from social media, i.e., “websites and applications that enable users to create and share content or to participate in social networking”\(^1\) and exclude studies that rely on traces that are inadvertently left by Web users (e.g., through the use of search engines) (Ficetola GF., 2013) or are aimed at promoting environmental education or pro-environmental action through social media (Ballew et al., 2015; Pearson et al., 2016).

The remainder of this paper is organized as follows. Section 2 describes the methodology used for searching, filtering and synthesizing studies. Section 3 presents the main results of the quantitative review, focusing on the identification of temporal trends, types and sources of the investigated social media data, geographic regions, and fields of application of the studies. Section 4 builds upon the results of the quantitative review to provide a broader, more qualitative discussion of the strengths and weaknesses of social media data in environmental research, as identified through the analysis of the retrieved studies. Section 5 concludes by highlighting key areas for future research and development in the field of passive crowdsourcing.

2. Material and methods

2.1. Search strategy

We systematically reviewed the scientific literature following the PRISMA guidelines (Moher et al., 2009) using the Web of Science and Google Scholar as information sources. A resource-intensive search strategy was administered with high sensitivity (i.e., retrieval of all studies of relevance) and low specificity (i.e., proportion of studies of relevance among all retrieved studies) to reduce potential biases and increase repeatability (CEE – Collaboration for Environmental Evidence, 2013; Pullin and Stewart, 2006). The strategy relied on a standardized search string containing topic keywords and the names of popular social networking sites (Appendix S1).

The search in the Web of Science was restricted to 7 ISI Web of Science Categories (“Biodiversity conservation”, “Ecology”, “Environmental sciences”, “Environmental studies”, “Water resources”, “Geography”, and “Multidisciplinary sciences”) and to studies published in the year 2000 or later. All languages and citation indexes were included. The search was conducted on 18 October 2017 and resulted in the retrieval of 4137 studies. The results were analyzed to identify the most influential keywords and remove redundant ones.

A keyword-based search was also performed in Google Scholar on 19 October 2017 to reduce the risk of introducing a publication bias in the results. Given that Google Scholar conducts a keyword search on the whole text, rather than title or abstract, a stricter set of keywords were used to ensure that only studies in environmental fields were retrieved (Appendix S1). This resulted in the retrieval of bibliographic information for an additional 420 studies.

We tested the performance of the search against a training database containing 31 studies that had previously been identified by the authors through a preliminary, non-systematic, literature review. The search in WoS and Google Scholar retrieved, respectively, 77% and 26% of the studies in the training set. Of the 23% of studies (7 in total) that were not retrieved by either search, 3 are unpublished grey literature, 2 are papers recently published in relevant WoS categories but not included in the WoS database at the time of the search, 1 was published in a journal that is not included in the WoS, and 1 is published in a journal from a WoS category out with the 7 ISI Web of Science Categories included in our search. Duplicate studies were removed and the test studies that were not retrieved by the systematic search were included in the final database, which contained 4469 studies and was exported to the reference manager software Zotero 5.0.23 for further analysis.

2.2. Article screening and eligibility criteria

Screening of the studies was performed at two levels. First, given the large number of records retrieved, title-level screening was performed, which is consistent with the low specificity of the search. At this stage, records were retained whose title makes explicit reference to either (a) social media or social networking services, or (b) an application in the environmental science. Title-level screening was performed by one reviewer. Second, the title, abstract and keywords were examined. Studies were retained only when making explicit reference to both (a) and (b). Abstract-level screening was performed by two reviewers. An intercoder agreement assessment was performed, consisting in both reviewers analyzing in parallel a random sub-sample of 240 abstracts (i.e., 10% of all abstracts remaining after title-level screening). Agreement was tested for inclusion/exclusion decisions with Cohen’s kappa coefficient (Cohen, 1960). The overall agreement at this stage was 90% with a kappa coefficient equal to 0.57 indicating “moderate” agreement (Edwards et al., 2002), which was considered acceptable. At both levels of screening, records were conservatively retained when there was reasonable doubt regarding their relevance (Pullin and Stewart, 2006). The flow diagram for study selection (Moher et al., 2009) is given in Appendix S2.

In order to determine study eligibility, all social media identified based on the previously given definition were considered, including blogging sites, recommendation sites (e.g., TripAdvisor) and other user content sharing sites (e.g., online forums). Only studies using passively crowdsourced data, i.e., information uploaded by end users independently of the specific purpose of the study in which they are analyzed, were retained. This led to excluding various citizen science and VGI projects. Studies had to involve retrieval of data from one or more social networking site, either directly or relying upon data previously retrieved by others. Any analysis that generated information about how human activities impact the environment or how to affect this impact in the future was retained. Accordingly, applications focusing on natural and environmental disasters for real-time tracking or humanitarian relief were excluded unless explicitly aimed at better predicting future environmental damage and society’s resilience to it. Studies exclusively focusing on assessment of tourism flows and/or preferences were also excluded unless they involved a nature-based component.

2.3. Data extraction and synthesis

The 490 studies relevant for the third stage of analysis were examined based on their complete content. This stage was investigated independently by two reviewers. After full-text assessment, 169 studies were found to meet the eligibility criteria. We first investigated these studies for general bibliographic information and subsequently reviewed their full text following a “5WH” (“Who, When, Where, What, Why and How”) conceptual framework (Di Minin et al., 2015). Finally, we placed an emphasis on the identification of (1) strengths and prospects, and (2) weaknesses and challenges of social media data in environmental research, as well as (3) future research needs.

3. Results

The literature testifies to a rapid growth in the field with the number of papers growing from 1 in 2011 (no prior study met the eligibility criteria) to 61 studies published between 1 January and 17 October 2017, with 83% of the studies in the database published in 2015 or later.

---

\(^1\) Source: https://en.oxforddictionaries.com/definition/social_medias (accessed: 10 August 2018)
Most studies (83%) use data from a single social media site, with Twitter (52 studies) and Flickr (34 studies) being the most frequently used, while an increasing number of studies from China focus on Weibo. Popular social media such as Facebook and Instagram had very limited application, most likely due to restrictions on data access. The number of studies drawing from multiple social media sources has been growing consistently, making up 15% of the studies published in 2017.

Most studies (93%) directly retrieve the data from the social media, either using the sites’ Application Programming Interfaces (APIs) or by manual download. Reliance on the APIs’ structured interfaces appears to be the de facto standard approach for social media data retrieval in environmental research, as opposed to techniques based for instance on web data scraping. The remainder of the studies rely on secondary sources such as the InVEST recreation module’s database or the Yahoo Flickr Creative Commons 100 M dataset. The investigated media in most studies is either text (e.g., text of tweets) or metadata (e.g., geotags). Studies jointly analyzing different types of media (multimedia) has grown to make up 34% of all studies (Fig. 1). The latter category mainly comprises combined analyses of text and metadata such as geotagged location (n = 29) or photograph content and metadata such as geotagged location, titles and tags (n = 23).

Research teams pursuing this field of research, as determined by the affiliation of the first author, are still limited to high-income (83%) and upper-middle-income economies (17%), with USA (27%), UK (15%) and China (12%) being the most represented (Li et al., 2017). The distribution of study sites shows a broader differentiation, with 57% and 17% of applications in, respectively, high-income and upper-middle-income economies and a substantial number of studies having a multi-national / continental (7%) or global (17%) scope (Fig. 2).

We classify each study based on its application into the three categories (1) Data on people, (2) Data on nature, and (3) Planning and governance (Fig. 3). The references of the 169 studies, their classification and the type of social media used are reported in the Supplementary Material.

3.1. Data on people

When it comes to analyzing people’s interactions with nature, social media can provide information regarding sensory impressions, emotional affinity, as well as reflexive and behavioral responses to specific stimuli (Cong et al., 2014). The largest group of studies in the database explores social media for assessing cultural ecosystem services, i.e., physical/experiential, intellectual/representative or spiritual/symbolic interactions with the natural environment (Haines-Young and Potschin, 2018). This category primarily includes studies focusing on the characterization of non-extractive recreational activities (e.g., hiking, walking, birdwatching, boating) (n = 34), including temporal and spatial patterns, and the aesthetic value of landscapes (n = 15). Specific applications aim at the identification of scenic routes (Baker et al., 2018).

---

**Fig. 1.** Distribution of the studies included in the synthesis over time based on type of social media data used and investigated media type.
2017; Sun et al., 2015), evaluation of factors contributing to eco-tourist satisfaction (Tenkanen et al., 2017; Hausmann et al., 2017a), extraction of points of interest (POI) or hot spots of cultural value (Figueroa-Alfar and Tang, 2017; Ghermandi, 2016; Lee et al., 2014; Levin et al., 2015; Peng and Huang, 2017), and keyword-based text analysis to identify types of recreational activities (Heikinheimo et al., 2017; Oteros-Rozas et al., 2018; Spalding et al., 2017) or differential patterns based on user type (e.g., residents versus tourists) (Garcia-Palomares et al., 2015; Hausmann et al., 2017b; Tenerelli et al., 2017). Geotagged photo counts show good correlation with observed visitation including both spatial and temporal patterns (Tenkanen et al., 2017), also in application to developing countries (Wood et al., 2013; Sinclair et al., 2018), and especially when data from different social media is combined (Ghermandi, 2016; Tenkanen et al., 2017). Social media analysis may be used to identify both overcrowded (e.g., exceeding tourist carrying capacity) and scarcely visited areas (Garcia-Palomares et al., 2015; Levin et al., 2015). The identification of hot and cold spots of visitation may support the development of strategies for the valorization of unexploited sites or the redirection of visitors in order to relieve pressure on wildlife populations in crowded areas (Ghermandi, 2016; Sinclair et al., 2018).

Less frequently social media data is applied to the analysis of extractive activities (e.g., recreational hunting) (Ebeling-Schuld and Darimont, 2017), passive values (Martinez Pastur et al., 2016; Oteros-Rozas et al., 2018), or trade-offs between cultural ecosystem services (Oteros-Rozas et al., 2018). Kothencz et al. (2017) and Palomino et al. (2016) explore the health implications of exposure to natural environments or lack thereof.

Social media data was used for the exploration of communication and (conflict) dynamics related to a range of environmental topics such as clean and renewable energy (Autry and Kelly, 2012; Hendriks et al., 2016), water governance (Mancilla-Garcia, 2015; Quinn et al. 2016), transmission of conservation research (Papworth et al., 2015), large infrastructure projects with significant environmental impacts (Hodges and Stocking, 2016; Jiang et al., 2016), UN Sustainable Development Goals (Wheeler and Quinn, 2017), and sustainable consumption (Cooper et al., 2012; Yeo, 2014). Moreover, a growing number of studies (n = 10) focuses on communication on climate change, including non-formal learning opportunities offered by social media (Andersson and Ohman, 2017; Robelia et al., 2011).

People’s perceptions of nature as well as the public’s understanding of events and topics of environmental interests can be gauged through the analysis of online conversations, as indicative of the public meaning-making of, among others, environmental accidents (Cha and
3.2. Data on nature

Studies in this category perform different types of environmental monitoring or characterization of land use/land cover (LULC) through the analysis of social media. This includes monitoring of physical and chemical environmental properties such as: monitoring of air quality (e.g., landfill odors) through text analysis of Weibo posts (Wang et al., 2017a; Cai et al., 2015), tweets (Riga et al., 2015) and blogs (Wang et al., 2017b; Ni et al., 2017); and mapping of flooding levels through analysis of tweets (Smith et al., 2017; Jongman et al., 2015; Fohringer et al., 2015), Flickr photos (Tkachenko et al., 2017; Sun et al., 2016; Fohringer et al., 2015) or YouTube videos (Michelsen et al., 2016). Other applications include monitoring of water flow velocity through YouTube videos (Le Boursicaud et al., 2016), mapping of heat waves (Jung and Uejio, 2017) and precipitation (Zhou and Xu, 2017).

A separate group of studies is aimed at biological and biodiversity-related environmental monitoring such as the discovery, organization and mapping of occurrence records of species in their spatial distribution (Daume and Galaz, 2016; Dylewski et al., 2017; ElQadi et al., 2017), including invasive species (Daume, 2016; Daume et al., 2014; Jovanovic and Vukelic, 2015). For instance, online photographs or videos may be investigated to identify geographic and temporal breeding patterns and phenology of species (Atsumi and Koizumi, 2017) or other aspects related to their behavior and ecology (Dylewski et al., 2017).

Social media data may also be applied for the determination of land cover from analysis of place names and other information uploaded to micro-blogs (Jendryke et al., 2017; Zeng et al., 2017), photo libraries (Antoniou et al., 2016; Mackaness and Chaudhry, 2013; Sitthi et al., 2016) and “folksonomies” from blogs (Derungs and Purves, 2016), or different types of land use (e.g., utilization for recreational, residential or work purposes) (Brown and McCarty, 2017; Liu et al., 2017; Soliman et al., 2017; Zhou and Zhang, 2016). Geotagged photos can also be processed to monitor the extent of snow cover (Giuliani et al., 2016) and geomorphometry for producing Digital Elevation Models (Gschwend and Purves, 2012). The density and content of tweets has also been explored to investigate population distributions (Lin and Cromley, 2015; Patel et al., 2017).

3.3. Planning and governance

Studies in this category are aimed at improving the current practices in the fields of hazard preparedness and management, urban planning, corporate social responsibility (CSR) and monitoring of illegal activities. Several studies (n = 18) explore changes in moods of tweets to manage or plan response to hazards such as heat waves (Jung and Uejio, 2017), wildfires (Kent and Capello, 2013; Truelove et al., 2015; Wang et al., 2016a), landslides (Pennington et al., 2015) and droughts (Tang et al., 2015). Social media data can be integrated in decision support systems (Shook and Turner, 2016) and are generally investigated for their potential to enhance communities’ resilience (Deng et al., 2016; Lopez-Cuevas et al., 2017; Wang et al., 2016b) or assist in post-disaster recovery (Kryvasheiyeu et al., 2016; Yan et al., 2017).

Applications of social media in urban planning include uncovering urban functions (Arribas-Bel et al., 2015; Chen et al., 2017; Tu et al., 2017) and areas of interest for more sustainable urban development (Garcia-Palomares et al., 2015; Hu et al., 2015), including patterns of use of urban green spaces (Roberts, 2017). This may involve monitoring mobility within specific areas (e.g., cyclists or pedestrians in a city) (Luo et al., 2016; Sun et al., 2017; Zhang and Feick, 2016) and their revealed preferences toward the promotion of sustainable transportation, including reducing exposure to air pollution (Sun and Mobasher, 2017).

Finally, several studies explore the potential for social media to promote an improved environmental governance through improved CSR practices (Tseng, 2017), e.g., improved sustainability communication (Cavalcante and Dante, 2016; Lee, 2017) and sustainability-related supply chain risk assessment (Wu et al., 2017), or identification of illegal activities such as trade of protected species (Hinsley et al., 2015).
4. Discussion

The results of the quantitative review of the literature presented in Section 3 testify to a growing interest in the application of social media data in the environmental sciences. In this section, we build upon the insights provided by the investigated studies to provide a broader, more qualitative discussion of the identified strengths and weaknesses of using social media data in environmental research. The discussion is organized around six key dimensions, starting from the three Vs (volume, velocity, variety) of geographically referenced big data (Figueroa-Alfaro and Tang, 2017) and subsequently focusing on additional aspects that are peculiar to social media data.

4.1. Volume: “big” data and large scales

The most frequently reported strength of social media data analysis lies in the size of the available data samples, which is often several orders of magnitude larger than that obtainable by traditional data collection, such as surveys (Martin and Schuurman, 2017). This is often associated with a less labor-intensive, less costly and less time-consuming procedure, especially if automated retrieval is performed (Antoniou et al., 2016; Soliman et al., 2017; Yan et al., 2017). Such data richness is however more apparent in areas with large populations or in natural areas with major tourist attractions, than remote locations or areas with minor tourist attractions (Jongman et al., 2015; Richards and Friess, 2015; Tenkanen et al., 2017; Wood et al., 2013; Zhou and Xu, 2017).

Social media data has an additional advantage in the fact that it is relatively easily applicable at large scales, such as entire populations, ecosystems or biomes (Ghermandi, 2018; Levin et al., 2015; Spalding et al., 2017; van Zanten et al., 2016). In application to nature-based recreation, it allows to analyze patterns of activity also in areas that lie outside of the limited network of sites that can be effectively monitored using conventional techniques (Levin et al., 2015; Mancini et al., 2018; Orsi and Genelleti, 2013). In LULC applications, social media data may reflect a “social sensing” of land use, in contrast to “remote sensing” which provides data about land cover but not its utilization and the functions it performs (Soliman et al., 2017; Zhou and Zhang, 2016). Similarly, social media can provide information about the revival of human activities during post-disaster recovery, not just physical reconstruction (Yan et al., 2017). As such, it may usefully complement conventional data sources.

4.2. Velocity: near real-time but short time span

Social media data allows for a timely and (near) real-time monitoring and analysis of land use changes (Arribas-Bel et al., 2015; Siththi et al., 2016; Zhou and Zhang, 2016), pollution (Riga et al., 2015), species distribution (Daume et al., 2014; ElQadi et al., 2017), environmental perceptions (Jiang et al., 2016; Williams et al., 2015), online communication for early warning to natural hazards or to assess communities’ preparedness to them (Lopez-Cuevas et al., 2017; Ripberger et al., 2014), and cultural ecosystem services such as aesthetic value (Figueroa-Alfaro and Tang, 2017) and recreation (Becken et al., 2017; Sessions et al., 2016). Social media effectively introduces a new type of data source that relies on first-hand observations (or “eyewitnesses” accounts) of ongoing events (Arribas-Bel et al., 2015; Fohringer et al., 2015), although the identification of honest eyewitness accounts is not trivial (Truelove et al., 2015). For such real-time applications, media from microblogging sites has advantages over those relying on video uploads or reviews, which are more likely to be uploaded at a later date (Michelsen et al., 2016).

The relatively short time span for which social media data is available may be problematic for applications that require long-term monitoring, such as analysis of LULC changes (Derungs and Purves, 2016) or occurrence of invasive species (Daume, 2016). Spatial and temporal patterns should be considered jointly since they may influence one another (Wang et al., 2016a). In the analysis of tweets, for instance, care should be used in dealing with retweets, which may not be representative of the space-time location of the reported event (Jongman et al., 2015). Geographic information embedded in the data may be used to infer users’ mobility over time (Chua et al., 2016; Luo et al., 2016; Sun et al., 2017). Such application however presents issues insofar as the data generally represents incidentally recorded locations, i.e., event-based movement data rather than actual paths (Prager and Wiegood, 2014). The suitability for deriving trajectories depends on the frequency on sampling and how sampled locations are identified (Vu et al., 2015).

The analysis of temporal patterns also needs to account for changes in the number of users over time, whereby a general increase in social media users in recent years and changes in the popularity of individual media both play a role (Cody et al., 2015). The use of social media over time is very dynamic and not only subjected to social trends, but also to market decisions and corporate strategies (Oteros-Rozas et al., 2018). Some media have more limited temporal cover than others (van Zanten et al., 2016). Studies relying on real-time streaming of data are often limited in their temporal scope, which exposes them to biases from the existence of seasonal patterns (Arribas-Bel et al., 2015) or confounding events occurring during data retrieval (Autry and Kelly, 2012; Kirilenko and Stepchenkova, 2014).

4.3. Variety: heterogeneity and noise

Several studies explicitly acknowledge the heterogeneous and noisy nature of social media data and in some cases raise concerns about deliberate attempts to spread misinformation and the biases these may generate if not properly controlled for (Boulos et al., 2011; Starbird et al., 2015). Accordingly, most studies are limited to exploratory research and correlation and regression analysis (Jiang et al., 2016).

Insofar as online photographs are concerned, limiting the analysis to metadata, such as geotags and time stamps, overlooks the fact that the subject of some photographs may not be relevant for the analysis (Figueroa-Alfaro and Tang, 2017; Su et al., 2016). There may be differences in the levels of such “noise” in the data depending on how individual social media platforms are used. For instance, for LULC classification, Panoramio was found to have a higher percentage of relevant photographs than Flickr (Antoniou et al., 2016), the former hosting proportionally more images of outdoor scenes (Giozzi et al., 2016; Zielstra and Hochmair, 2013). After irrelevant photos are removed, however, Flickr is found to contain more relevant images and to better capture cultural ecosystem services than Panoramio (Oteros-Rozas et al., 2018). Biodiversity photos are more frequent in Flickr than in Instagram, while pictures including people are more frequent in Instagram than in Flickr (Tenkanen et al., 2017). One should notice, however, that such results are likely very site- or region-specific. Human and/or machine-aided quality control, filtering and validation procedures are essential for data cleaning (Muller et al., 2015) and may involve automatic image processing (de Albuquerque et al., 2015) or analysis of the direction of the photos (Siththi et al., 2016).

Lack of quality assurance is often mentioned as a limitation, given that data is not acquired following best-practice standards (Daume, 2016; Kent and Capello, 2013; Muller et al., 2015). Accuracy of geotags is frequently mentioned as a limiting factor, but relatively few studies attempt to quantitatively characterize it (Kirilenko et al., 2015; Leibovici et al., 2017; Senaratne et al., 2017; Tenerelli et al., 2016). Geotagging error tends to be lower in Panoramio than Flickr, possibly due to the higher quality prerequisite standards demanded (Zielstra and Hochmair, 2013). Higher quality standards requirements may, on the other hand, bias the sample of photos toward users with high-quality
equipment or a self-selected community of spatially aware users (Oteros-Rozas et al., 2018). Spatial accuracy tends to be lower in remote regions (Heikinheimo et al., 2017) and may in general be influenced by mobile coverage (Chua et al., 2016), recording device and weather conditions (Richards and Friess, 2015), GPS visibility (Zielstra and Hochmair, 2013), and manual geotagging procedures. Users may manually geotag photos based on the position of the subject, which may be at a substantial distance from where the photo was taken (Senaratne et al., 2017), especially in mountainous areas (Oteros-Rozas et al., 2018). Zoom settings may complicate ex post manual geotagging by either users or researchers (Zielstra and Hochmair, 2013). Retrieval of information from titles and tags may help geotagging photos but is only suitable for a sub-set of photos (Figueroa-Alvaro 2017), provides information at granular rather than fine scale, and may still bias the analysis if the tags are indicative of the object in the field of view rather than the location at which the photo was taken (Mackness and Chaudhry, 2013). Geotagging of tweets based on spatial references in the text is limited by the general lack of reliable spatial information in most tweets (Senaratne et al., 2017).

4.4. Potential biases

One major concern with social media data regards the representativeness of the sampled users with respect to the population of interest. Sampled users may be biased in gender, age, socioeconomic status, education and motivations, although previous studies are not consistent with respect to the direction of the bias (Chua et al., 2016; Heikinheimo et al., 2017; Keeler et al., 2015). Previous findings suggest that social media users in national parks tend to be younger than the general population of visitors (Hausmann et al., 2017a, 2017b; Heikinheimo et al., 2017). Moreover, different social media attract users with different characteristics. For instance, Heikinheimo et al. (2017) found in a survey of national park visitors that Instagram users tend to be younger than users of other social media, while Agyzakov et al. (2017) note that Foursquare users are primarily young professionals. Social media data may thus be better suited to the analysis of eco-tourism, which is more attractive to young and active travelers, than general tourism patterns (Lu and Stepchenkova, 2012). Methods used by ecologists to correct for sampling effects when assessing species richness may be useful to mitigate biases in social media data (Levin et al., 2015). The analysis of socio-economic and demographic biases in social media data is hampered by the limited information available to researchers (Antoniou et al., 2016; Oteros-Rozas et al., 2018), especially compared to what can be collected through surveys (Roberge, 2014). Analysis of user profiles with deep learning algorithms (Hausmann et al., 2017a) or distribution probabilities for race, age and gender based on name databases (Lansley and Longley, 2016; Luo et al., 2016) may help to compensate such limitations. Such tools however require a general understanding of the demographic structure of the study area, which is not always available for regions suffering from data scarcity.

In large-scale studies, data may be biased due to the existence of a digital divide insofar as it affects the access to digital devices, technologies and supporting infrastructure (Patel et al., 2017; Richards and Friess, 2015), either physically or due to a lack of technical knowhow (Arts et al., 2015). Photographs or videos of animals often involve specialized equipment, which may not be widely available in developing regions. Dylewski et al., (2017), for instance, note a lack of correlation between YouTube videos of shrikes (a group of songbirds) and proportion of smart phone users by country, suggesting that it can be attributed to the use of more specialized devices by birdwatchers. Comparison between countries is complicated by differences in internet penetration and uptake of social media in general (Ghermandi, 2016) or specific social media (Levin et al., 2015). This may lead to over-representation of international visitors in developing countries (Sessions et al., 2016), especially concerning that travelers are more likely to share information than those staying at home (Becken et al., 2017). Techniques relying on mobility entropy (Yan et al., 2017), text-mining approaches (de Albuquerque et al., 2015) or analysis of the entire corpus of data uploaded by single individuals (Bojic et al., 2015; Ghermandi, 2018; Li et al., 2013; Sinclair et al., 2018) have been proposed to differentiate between local and international visitors or identify the home location of users.

Users may not necessarily be treated equally in the analysis. Unless implementing weighing systems (Yan et al., 2017) or controls for multiple uploads during one single visit to a specific location (Wood et al., 2013), the analysis may be biased toward very active users. When analyzing large areas, the application of such techniques may be problematic since they rely on defining individual “locations”, whose spatial delimitation may be ambiguous (Lee, 2017). In this matter, the identification of hotspots or points of interest through spatial analysis may help to define the natural boundaries of individual attractions (Orsi and Geneletti, 2013; Peng and Huang, 2017). In the analysis of the text of (micro-)blogs, posts or reviews, users who write more extensively (Cong et al., 2014; Yeo, 2014), have better argumentation skills (Andersson and Ohman, 2017), or are perceived as opinion leaders (Dalrymple et al., 2013) may be overly influential. The fixed structure of microblogs that are limited in the number of characters allows for a more straightforward analysis than other media (Roberts, 2017) but may limit the depth of insight of the individual posts (Andersson and Ohman, 2017). Some users may have a dominant position in the social communication network (e.g., traditional media accounts, institutional users) (Liu and Zhao, 2017). The number of views or “likes” of a message can affect its perceived authority and thus one’s perceptions of environmental issues such as climate change (Spartz et al., 2017). Finally, one should take into account that some users may have multiple accounts (Zoomers et al., 2016), some accounts may have multiple users (e.g., governmental agencies) (Kay et al., 2015; Liu and Zhao, 2017) or represent automated services (Palomino et al., 2016), and some users may deliberately attempt to spread misinformation (Cong et al., 2014).

Regarding the content of uploads, observer effects may be at play. Analysts should consider that interesting (EI-Qadi et al., 2017) or unusual (Becken et al., 2017; Daume et al., 2014) events or subjects are more likely to be shared. The perceived value of a trip (Wood et al., 2013) or a landscape (Dunkel, 2015) might thus influence whether an individual takes or shares photographs, resulting in a bias against local or regular visitors. The sample of shared photos is also not necessarily representative of the population of photos taken: pre-selection by the users may create biases but may also be useful to reveal their preferences (Donaire et al., 2014). Some types of recreational activities, such as diving (Spalding et al., 2017), surfing (Wood et al., 2013) or rock climbing (Tenerelli et al., 2016), are less suitable for sharing in social media and, in general, one might expect a bias in favor of land-based as opposed to marine or water ecosystems (Howarth, 2014). More charismatic, popular or observable species attract more social media activity, which may introduce a bias in studies aiming at eliciting users’ preferences for biodiversity from geotagged photographs (Mancini et al., 2018; Tenkanen et al., 2017). Moreover, photographs may tend to be concentrated around areas of high human populations or human activity such as roads (Stafford et al., 2010).

Environmental communication in text-based media may also suffer from biases ensuing from social dynamics such as the amplification of tendencies such as homophily and segregation (Williams et al., 2015) or dominance by specific groups or countries (Kirilenko and Stepchenkova, 2014). Previous research found that Facebook users are more inclined to like positive than negative news, while Twitter users are more likely to tweet about negative news (Papworth et al., 2015). Polarized views are better represented than neutral ones (Jang and Hart, 2015) and discussions rapidly become consensus-oriented or turn into generalizations (Cooper et al., 2012), although there are examples of elaborated discussions taking place (Andersson and Ohman, 2017). Natural language processing and text mining methods (e.g., Latent
Dirichlet Allocation, text reduction and factor analysis) may assist in data preprocessing (Ferrari et al., 2011; Kirilenko et al., 2012; van Zanten et al., 2016) but they are not well developed for languages other than English (Daume and Galaz, 2016).

Other biases may arise during the analysis of the data. Keyword-based text analyses are strongly influenced by the dictionaries used (Leibovici et al., 2017), but these may not be available for testing, as is often the case for commercial software (Palomino et al., 2016). For instance, specific terms may be reflective of subjective rather than objective experiences (Gschwend and Purves, 2012) or be used with humor or sarcasm (Jang and Hart, 2015). Previous work shows little agreement in sentiment analyses conducted with different software applications (Palomino et al., 2016). Unless straightforward objective criteria are used (Richards and Friess, 2015), a researcher bias may be introduced when the analyst builds an ad hoc dictionary of keywords (Cody et al., 2015; Roberge, 2014), performs a manual photograph content analysis (Martinez Pastur et al., 2016; Oteros-Rozas et al., 2018), or selects specific blogs for analysis of discussions in the blogosphere (Sharman, 2014).

4.5. Ethics in data acquisition and use

Several studies acknowledge that the way social media data and personal information is used requires consideration of privacy issues and ethical use. Although the analysis typically relies on public data, users do not “volunteer” the information and are unaware of the purpose of use, which raises the question of whether the implicit consent to publish the data is sufficient or rather an expressed consent is needed (Arts et al., 2015; Connors et al., 2012). Studies show a lack of uniformity in the level of privately identifiable information released and it has been argued that in some cases sufficient information may be available for ill-intentioned individuals to pursue criminal offenses, or “cybercasing” (Friedland and Choi, 2011). Even in the absence of individual privacy violations, use of customers’ data without consent may constitute a violation of the rights of research subjects (Buchanan, 2017) and may lead to consequences for website owners, especially in light of the recently heightened concerns with online privacy (Krotov and Silva, 2018).

Generally speaking, however, passive crowdsourcing of social media is less intrusive than other forms of crowdsourcing (Heikinheimo et al., 2017; Jiang et al., 2016) and surveying (Kirilenko and Stepchenkova, 2014; Roberts, 2017; Yeo, 2014). As opposed to survey respondents, or participants to citizen science projects, the fact that social media users are unaware of being analyzed may result in the collection of data that is unfiltered through unconscious experiences (Cooper et al., 2012), as arising as a naturally occurring discussion (Dunkel, 2015). As an expression of revealed rather than stated preferences, social media data is not exposed to respondent biases such as interviewer biases (Martinez Pastur et al., 2016). In retrospective studies, social media analyses, unlike surveys, do not suffer from re-collection bias (Dunkel, 2015; Shook and Turner, 2016; Vu et al., 2015), unless they focus on blogs and reviews, which are usually not real-time accounts (Cong et al., 2014).

4.6. Data ownership and future availability

Social media data is generated bottom-up, which represents a shift from statistics generated using centralized, top-down approaches such as governmental population censuses (Derungs and Purves, 2016). Nevertheless, data is obtained from private entities, who manage the web services and could filter or censor the data in ways that are not verifiable (Agyrzakov et al., 2017; Liu and Zhao, 2017). This may create biases in cross-country comparisons, given that governments practice different levels of web censorship (Levin et al., 2015). Photo-sharing websites often pose restrictions on the acceptable formats and quality of uploads (Antoniou et al., 2016), which may limit applications that require high-quality photographs such as the identification of animal species (Daume and Galaz, 2016). Some API methods (e.g., Twitter’s Standard Search API) only allow for a sampling of the relevant social media data (Daume, 2016; Palomino et al., 2016). Social media companies may have different policies regarding the sharing of information by the users: Sina-Weibo, for instance, encourages users to share location information (Wang et al., 2016b). There is also no guarantee that the data will be available in the future, given that services may be terminated (e.g., Panoramio was discontinued in November 2016) and conditions of the APIs may change (e.g., Instagram API changed in 2016 and again in 2018). The rapid evolution of APIs, mobile device apps and social networking services requires regular updates to the data harvesting software and strategy (Kirilenko and Stepchenkova, 2014) and is overall a non-inclusive social development process (Arts et al., 2015).

5. Conclusions

The analysis of the rapidly growing body of studies using social media data in various environmental disciplines supports the notion that this new data source offers unprecedented opportunities to extend the scope, scale and depth of research, especially insofar as the interactions between humans and the environment are concerned, but, at the same time, presents environmental researchers with a range of issues involving potential biases, big data management and rapidly evolving frameworks with which they are generally not familiar.

We briefly highlight below five key areas, which might be critical in determining whether social media analysis will in the coming years spearhead a new age in environmental research or rather fade out after being a briefly “en vogue” technique (Muller et al., 2015):

1) Data mashups: research so far suggests that there are unrealized benefits to be derived from the integration of data from multiple social media as well as the combined analysis of the available data at multiple levels (e.g., analyzing photographs’ metadata jointly with their content and accompanying text). Benefits range from mitigating some of the self-selection biases of social media users, both in terms of socio-demographic characteristics (different social media appeal to different swathes of society) and differential between-country adoption levels, to revealing experiences, motivations and preferences that cannot be fathomed from a unidimensional analysis. This might well require the development of procedures for combined data integration and analysis, similar to those already done in other big data fields (Arts et al., 2015).

2) Quality assurance: while it is tempting to think that the sheer strength of relying on large sample sizes will compensate for the poor quality of individual data items, this may not always be the case (Muller et al., 2015). Computational advances in deep learning algorithms and natural language processing will undoubtedly play a growing role in social media data analysis, which in turn will call for more complex and increasingly multidisciplinary efforts on the part of environmental scientists. The development of standards for quality assessment and data preprocessing and “cleaning” would represent an important improvement compared to the state of the art.

3) Integration with existing methods: rather than being an alternative to them, reliance on crowdsourced social media data may provide a useful complement to “conventional” techniques (e.g., remote sensing data, hydrological and meteorological models, surveys), including citizen science projects. They may provide a platform for extending results calibrated for a test region to larger and currently unmonitored areas. Integration might also be aimed at mitigating the biases that may exist in social media data, for instance by

---

involving underrepresented actors such as elderly, rural and indigenous communities (Oteros-Rozas et al., 2018).

4) Ethical codes: as pointed out by Arts et al. (2015), the current lack of protocols and framework of good practice for dealing with potentially sensitive information may be a stimulus for the observed rapid growth of the field but in the long-term hamper its sustainability. While the ethical and privacy issues involved in dealing with social media data are by no means limited to their applications in the environmental sector, they are relatively new to environmental researchers who are not necessarily used to deal with such extensive amounts of individual citizens’ data (Mol, 2008). Multi-sectorial cooperation and self-regulation would therefore be beneficial (Arts et al., 2015).

5) Long-term availability: whether social media analysis will leave a long-lasting mark in environmental research will largely be determined by whether such data will be made available to researchers in a reliable and sustained way. Many researchers may not have the inclination to keep up with the extremely dynamic evolution of APIs and the changes in policy and strategy by social networking companies. In this sense the establishment of easy-to-access, international, and open repositories of publicly available data such as the Yahoo Flickr Creative Commons 100 M database represents an important step forward.

Social media and social networking services are forces that shape contemporary society and have become important research tools in many disciplines, including environmental research. While the analysis of social media data is not a panacea for the many environmental applications in which it has been explored, it offers the promise of new tools and formerly unavailable dimensions on which to test theories and search for empirical evidence. In particular, the ubiquitous nature of social media in developed and, increasingly, developing countries as well as their suitability for large-scale analyses, offer unique opportunities to rely on such passively crowdsourced data in analyzing and addressing global environmental problems. In spite of its already fairly conspicuous size, research on environmental applications of social media data is only in its infancy and may well play a crucial role in understanding how humans perceive and interact with the natural environment as well as shaping future nature conservation efforts and environmental management.

Acknowledgements

This research (Grant No. 2751/16) was supported by the Israel Science Foundation within the ISF-UGC joint research program framework.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doihttps://doi.org/10.1016/j.gloenvcha.2019.02.003.

References

Anon, S., Watson, H., Wadhwa, K., Metz, K., 2017. Analysing social media data for disaster preparedness: understanding the opportunities and barriers faced by humanitarians. Int. J. Disaster Risk Reduct. 21, 131–139.
Cooper, G., Green, N., Burningham, K., Evans, D., Jackson, T., 2012. Unravelling the threads: discourses of sustainability and consumption in an online forum. Environ. Commun. 6 (1), 101–118.
A. Ghermandi, M. Sinclair


