Understanding human activities in green areas with social media data

Vuokko Heikinheimo
Digital Geography lab,
Department of Geosciences and
Geography, University of Helsinki,
PO Box 64, 00014 University of
Helsinki, Finland
Vuokko.heikinheimo@helsinki.fi

Abstract

Up-to-date information about human-nature interactions are urgently needed to inform sustainable land use planning and nature conservation. Large amounts of content-rich geographic data are produced continuously by users of different social media platforms across the globe containing information about the whereabouts and activities of people. Such data, combined with other sources of data, have potential to provide new and useful information about human presence, activities, observations and movements at different spatial and temporal scales. Despite many examples in other fields, location-based social media data have not been widely used in nature conservation. This work aims to understand the potential and biases in geographic social media data in order to inform conservation-related decision making across scales. Main objectives of the work are to 1) extract meaningful patterns related to human-nature interactions in green areas from location-based social media, while 2) understanding the biases and limitations of the data. Firstly, the aim is to position location-based social media data among other sources of user-generated geographic information and to identify the useful elements and limiting factors of using such data in conservation science. Secondly, the aim is to understand who the data represents in order to derive further information about green area users. Lastly, user-generated data is combined and contrasted with other data sources to understand the spatial and temporal patterns of human actions, and potential threats in areas of high conservation value at regional and global scales.

Keywords: Social media data, human-nature interactions, green areas, bias

1 Introduction

Understanding patterns of human-nature interactions is crucial for sustainable land use planning and nature conservation (Venter *et al.*, 2016). However, spatially and temporally accurate data on threats affecting biodiversity persistence are lacking (Joppa *et al.*, 2016), and datasets needed to inform conservation decision-making are limited and often biased (Di Minin & Toivonen, 2015).

Spatial data generated by non-experts have recently become a valuable resource both in academia and the society in addition to more traditional data produced by scientists and other authorities (See *et al.*, 2016; Goodchild, 2007). Geographic information generated by crowds, such as geotagged photos and other location-based social media data provide diverse information about human activities across the globe. These user-generated data, as opposed to official data sources (such as census data, visitor counts and surveys), may provide complementary information about the values, observations and activities of different groups of people especially in regions where official data is collected rarely.

Social media, in general, refers to computer-based applications used for networking and sharing digital content. Here, social media data refers to the spatial attributes (location), temporal attributes (time), and relevant content (text, photos, and video) generated by users of different social media platforms (such as Flickr, Instagram and Twitter). Publicly shared content in different social media services can be often accessed in large quantities through Application Programming Interfaces (APIs).

Social media data is used in various fields of science, also increasingly in environmental sciences, to explore spatial and temporal patterns of human activities. However, issues of data quality and data ownership might limit the use and trustworthiness of these data sources in systematic analyses and decision-making support (Sui & Goodchild, 2011) and not many studies, especially in the context of environmental studies, have aimed to validate observed patterns or to systematically asses evident gaps in the data, for example related to spatial coverage of the available data in different platforms.

This work aims to underpin the potential and limitations of using user-generated geographic information in environmental studies. Furthermore, this work aims to understand how user-generated geographic information can complement traditional data sources in the study of human-nature interactions. Focusing on natural and semi-natural environments such as national parks and urban green areas at different spatial scales, the aim is to map and analyse nature recreation and human pressure using publicly shared social media posts in combination with other available data. The general goals of this wor are to 1) extract meaningful patterns related to human-nature interactions in green areas in order to inform

conservation-related decision making at different spatial scales, while 2) understanding the biases in location-based social media data. The work includes the development of automated workflows for data processing and analysis, comparisons between user-generated data and official datasets, and accounting for gaps, inaccuracies and bias in the user-generated data at different spatial scales. The main objectives and related questions are the following:

- Analyzing the potential and limitations of different social media data for studying human activities in green areas: What kind of information do we get from social media, what biases are included in the data and how is it useful for studying human activities in green areas? How does social media data compare with other information sources from focus areas?
- Understanding the park visitors: Whose views and observations are presented in user-generated content from national parks? Where do the visitors come from and how do they move within and between green areas?
- Mapping conservation opportunities and threats: Can we characterize national parks and national park visitors based on social media data? Can we identify human pressure on the environment from social media data? What tradeoffs between nature recreation and nature conservation can we discover on a regional/global scale?

2 Related work

Data and tools related to the information age (Castells, 2000) and the big data revolution (Mayer-Schönberger & Cukier, 2013; Kitchin, 2014) have opened up new possibilities for geographic knowledge discovery (Mennis & Guo, 2009; Crampton *et al.*, 2013). Before, recreational use patterns and preferences related to green spaces have been studied using surveys (Tyrväinen, Mäkinen & Schipperijn, 2007), activity diaries (Mytton *et al.*, 2012), GPS tracking (Korpilo, Virtanen & Lehvävirta, 2017) and public participatory GIS (PPGIS) (Brown, Schebella & Weber, 2014; Laatikainen *et al.*, 2015). However, these methods are often costly to implement (Kwan, 2013) and often limited to a specific case study area (Ives *et al.*, 2017). Recently, large amounts of geographic 'big data', such as location-based social media data, have become available for capturing information about people's movement and activities in unprecedented volumes. This "location-based story telling" (Sui & Goodchild, 2011) in various online platforms such as Facebook, Twitter and Instagram, has fundamentally transformed the notion of geographic information in recent years.

Location-based social media data is often discussed in the context of Volunteered Geographic Information (VGI). The concept VGI, coined by Goodchild in 2007, is widely used to describe geographic datasets generated by non-experts. Vast amounts of spatial data are continuously created in collaborative projects such as the OpenStreetMap, social networks such as Twitter and other location-aware platforms on the web which host user-generated content. However, the term VGI does not fully capture the nature of more spontaneously generated data (See *et al.*, 2016) such as tweets and Flickr photos which have originally been shared for other purposes than mapping and research. These passively shared data evidently require special consideration related to the ethical use of data, and representativeness of the results. Thus, there is a need to further position social media among other sources of user-geographic information and authorative data sources.

Social media data has been used in many application fields of geography to study spatial phenomena, especially in the urban context. The study of population dynamics in cities (Longley & Adnan, 2016; Steiger *et al.*, 2015), spatial diffusion (Crampton *et al.*, 2013) and humanitarian response applications (Crooks *et al.*, 2013) are only a few examples of existing research from the fields of geography and geographic information science.

However, examples in environmental studies, especially in conservation science are still limited (Di Minin, Tenkanen & Toivonen, 2015). Studies focusing more on human-nature interactions include the quantification of visitation rates (Wood *et al.*, 2013; Levin, N., Kark, S. and Crandall, 2015), assessment of cultural ecosystem services and people's interests (Richards & Friess, 2015; Roberge, 2014), and the extraction of species data (Barve, 2014; Stafford *et al.*, 2010) from social media. Also, only a few studies have used social media data to understand human-nature interactions in urban environments. These include methodological development for studying cultural ecosystem services based on social media content analysis with a case study from urban mangroves in Singapore (Richards & Friess, 2015; Thiagarajah *et al.*, 2015) and tourism crowding (including parks) based in check-in data from Shanghai (Shi, Zhao & Chen, 2017).

Existing studies are often limited to only a single social media platform, and lack comparisons and validation against other data sources. Most studies using social media data for environmental studies rely solely on one platform (mostly Flickr). Flickr might be the most suitable platform when looking at biodiversity features, while Instagram or other available data sources might reflect better the activities present in the area of interest (Hausmann *et al.*, 2017b). Studies with more in-depth and advanced content analysis have thus far been limited to smaller study sites (for example Richards & Tunçer, 2017) and there is great potential to scale up such analysis to continental and even global scales.

A critical approach is needed when using social media as a source of information (Boyd & Crawford, 2012). Firstly, ethical issues related to using people's personal data need to be considered even when using publicly available content. Secondly, it can be difficult to assess how representative the captured social media users are of the population in question (Longley & Adnan, 2016). Furthermore, the data is often biased both spatially and temporally in relation to infrastructure, mobile phone coverage, and popular events, and potentially towards certain socioeconomic classes. This work aims to provide a deeper understanding of these issues in the context of conservation-related questions.

3 Methods

The study sites of this research consists of green area networks at different scales; urban green areas from the city of Helsinki, Finland, individual protected areas from Finland and South Africa and the global protected area network and key

biodiversity areas worldwide. National parks with regular visitor monitoring schemes provide a test environment for comparing social media user counts and content to official statistics.

Main material for the study consist of openly shared location-based social media posts from different social media platforms including (but not restricted to) Instagram, Twitter, and Flickr. Data is retrieved from platform APIs using existing packages and custom-made scripts in the Python Programming Language.

Social media data are used in conjunction with, and contrasted to official visitor statistics, surveys and other available data from the focus areas. Official visitor statistics and visitor survey data from national park authorities are used together with social media content from different platforms. Global analysis from projected areas and key biodiversity areas is done using additional data from the International Union for the Conservation of Nature (IUCN), and Birdlife International.

The first part of the study is focused on the potential and limitations of social media data for environmental studies through analyzing the different components of the data in terms of precision, accuracy and fit-for-purpose. Accuracy and precision of spatial and temporal information are assessed trough methods of data exploration and comparisons to ancillary data sets. Manual and automated content analysis of texts and images is used to explore if the content shared from green areas is thematically meaningful for green area management and conservation, and to produce further analyses of the spatial and temporal patterns of observed content categories (for example related to a specific activity or species within a national park). Results are compared with existing information about park visitation from case study sites with existing reference data. For example, activities revealed from social media photo content are compared with surveyed activities in a case study from a Finnish National Park.

Understanding bias in social media data includes the analysis of age- and language groups as well as the differentiation between data generated by locals and visitors. Questions related to visitors' social media usage are conducted in selected national parks in Finland in order to find out the proportion of national park visitors who are active in social media and to link this information to demographic background variables. Furthermore, different platforms are compared in terms of their information content. Based on earlier findings from Kruger National park in South Africa, Flickr contains more information related to biodiversity features, whereas Instagram posts portray more often human activities (Hausmann *et al.*, 2017a). In this work, such differences in shared content will be explored further in different spatial and temporal contexts.

The last sections of the work apply the earlier findings for answering questions related to conservation opportunities and threats in the global protected area network and key biodiversity areas using a combination of user-generated and official data sources.

The work has potential to bridge the gap between recent advances in social media data analytics and information needs in conservation science. The work will likely reveal new information about spatial and temporal patterns of human activities in green areas worldwide, especially from areas with no systematic visitor monitoring in place.

References

Barve, V. (2014) Discovering and developing primary biodiversity data from social networking sites: A novel approach. *Ecological Informatics*, 24, 194–199.

Boyd, D. & Crawford, K. (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662–679.

Brown, G., Schebella, M.F. & Weber, D. (2014) Using participatory GIS to measure physical activity and urban park benefits. *Landscape and Urban Planning*, 121, 34–44.

Castells, M. (2000) *The rise of the network society*. Blackwell Publishers.

Crampton, J.W., Graham, M., Poorthuis, A., Shelton, T., et al. (2013) Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, 40(2), 130–139.

Crooks, A., Croitoru, A., Stefanidis, A. & Radzikowski, J. (2013) # Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1), 124–147.

Goodchild, M.F. (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211–221.

Hausmann, A., Toivonen, T., Heikinheimo, V., Tenkanen, H., et al. (2017a) Social media reveal that charismatic species are not the main attractor of ecotourists to sub-Saharan protected areas. *Scientific Reports*., 7(1), 763.

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., et al. (2017b) Social Media Data Can Be Used to Understand Tourists' Preferences for Nature-Based Experiences in Protected Areas. *Conservation Letters*, 11(1).

Ives, C.D., Oke, C., Hehir, A., Gordon, A., et al. (2017) Capturing residents' values for urban green space: Mapping, analysis and guidance for practice. *Landscape and Urban Planning*, 161, 32–43.

Joppa, L.N., O'Connor, B., Visconti, P., Smith, C., et al. (2016) Filling in biodiversity threat gaps. *Science*, 352(6284), 416–418.

Kitchin, R. (2014) Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 1–12.

Korpilo, S., Virtanen, T. & Lehvävirta, S. (2017) Smartphone GPS tracking—Inexpensive and efficient data collection on recreational movement. *Landscape and Urban Planning*, 157, 608–617.

Kwan, M.-P. (2013) Beyond space (as we knew it): toward temporally integrated geographies of segregation, health, and accessibility: Space--time integration in geography and

GIScience. *Annals of the Association of American Geographers*, 103(5), 1078–1086.

Laatikainen, T., Tenkanen, H., Kyttä, M. & Toivonen, T. (2015) Comparing conventional and PPGIS approaches in measuring equality of access to urban aquatic environments. *Landscape and Urban Planning*, 144, 22–33.

Levin, N., Kark, S. and Crandall, D. (2015) Where have all the people gone? Enhancing global conservation using night lights and social media. *Ecological Applications*. 25 (8), 2153–2167. Longley, P.A. & Adnan, M. (2016) Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, 30(2), 369–389.

Mayer-Schönberger, V. & Cukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, Massachusetts, Houghton Mifflin Harcourt.

Mennis, J. & Guo, D. (2009) Spatial data mining and geographic knowledge discovery-An introduction. *Computers, Environment and Urban Systems*, 33(6), 403–408.

Di Minin, E., Tenkanen, H. & Toivonen, T. (2015) Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, 3.

Di Minin, E. & Toivonen, T. (2015) Global protected area expansion: creating more than paper parks. *BioScience*, 65(7), 637–638.

Mytton, O.T., Townsend, N., Rutter, H. & Foster, C. (2012) Green space and physical activity: an observational study using Health Survey for England data. *Health & place*. 18 (5), 1034–1041.Richards, D.R. & Friess, D.A. (2015) A rapid indicator of cultural ecosystem service usage at a fine spatial scale: Content analysis of social media photographs. *Ecological Indicators*, 53, 187–195.

Richards, D.R. & Tunçer, B. (2017) Using image recognition to automate assessment of cultural ecosystem services from social media photographs. Ecosystem Services. [Online] Available from: doi:10.1016/J.ECOSER.2017.09.004 [Accessed: 4 October 2017].

Roberge, J.M. (2014) Using data from online social networks in conservation science: Which species engage people the most on Twitter? *Biodiversity and Conservation*, 23(3), 715–726.

See, L., Mooney, P., Foody, G., Bastin, L., et al. (2016) Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS International Journal of Geo-Information*, 5(5), 55.

Shi, B., Zhao, J. & Chen, P.-J. (2017) Exploring urban tourism crowding in Shanghai via crowdsourcing geospatial data. *Current Issues in Tourism*, 20(11), 1186–1209.

Stafford, R., Hart, A.G., Collins, L., Kirkhope, C.L., et al. (2010) Eu-social science: the role of internet social networks in the collection of bee biodiversity data. *PloS one*, 5(12), e14381.

Steiger, E., Westerholt, R., Resch, B. & Zipf, A. (2015) Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, 54, 255–265.

Sui, D. & Goodchild, M. (2011) The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, 25(11), 1737–1748.

Thiagarajah, J., Wong, S.K.M., Richards, D.R. & Friess, D.A. (2015) Historical and contemporary cultural ecosystem service values in the rapidly urbanizing city state of Singapore. *Ambio*, 44 (7), 666–677.

Tyrväinen, L., Mäkinen, K. & Schipperijn, J. (2007) Tools for mapping social values of urban woodlands and other green areas. *Landscape and Urban Planning*, 79(1), 5–19.

Venter, O., Sanderson, E.W., Magrach, A., Allan, J.R., et al. (2016) Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. *Nature Communications*, 7, 12558.

Wood, S. a, Guerry, A.D., Silver, J.M. & Lacayo, M. (2013) Using social media to quantify nature-based tourism and recreation. *Scientific reports*, 3, 2976.