DEPARTMENT OF GEOSCIENCES AND GEOGRAPHY A89

User-Generated Geographic Information for Understanding Human Activities in Nature

VUOKKO HEIKINHEIMO

UNIVERSITY OF HELSINKI
FACULTY OF SCIENCE

User-Generated Geographic Information for Understanding Human Activities in Nature

VUOKKO HEIKINHEIMO

ACADEMIC DISSERTATION
To be presented for public discussion with the permission of the Faculty of Science
of the University of Helsinki, in Athena, Hall 107, Siltavuorenpenger 3 A,
on the 9th of December, 2020 at 10 o'clock.

DEPARTMENT OF GEOSCIENCES AND GEOGRAPHY A89 / HELSINKI 2020

Author: Vuokko Heikinheimo
Department of Geosciences and Geography,
University of Helsinki, Finland

Supervisors: Professor Tuuli Toivonen
Department of Geosciences and Geography,
University of Helsinki, Finland

Associate Professor Enrico Di Minin
Department of Geosciences and Geography,
University of Helsinki, Finland

Pre-examiners: Professor Alexander Zipf
Institute of Geography,
Heidelberg University, Germany

Professor Kate Sherren
School for Resource and Environmental Studies,
Dalhousie University, Canada

Opponent: Professor Catherine Pickering
School of Environment and Science,
Environment Futures Research Institute,
Griffith University, Australia

Front cover photo by The Author from *Tunturiaapa*, Pyhä-Luosto National Park 2018.
Back cover photo by Aaro Keipi from *Vanhankaupunginlahti*, Helsinki 2020.

The Faculty of Science uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

ISSN-L 1798-7911
ISSN 1798-7911 (print)
ISBN 978-951-51-6580-0 paperback
ISBN 978-951-51-6581-7 pdf
http://ethesis.helsinki.fi
Painosalama, Turku 2020

Heikinheimo V. 2020. *User-Generated Geographic Information for Understanding Human Activities in Nature*. Department of Geosciences and Geography A89. 48 pages and 7 figures.

ABSTRACT

In this thesis I have investigated how user-generated data can be applied to studying human-nature interactions on different spatial and temporal scales. User-generated geographic information refers to spatial data sets generated by and about people, such as social media data, sports tracking data, mobile phone data and participatory geographic information. Users of various digital platforms and mobile devices generate considerable amounts of data about their observations, activities and preferences in different environments. These data can potentially be used to fill information gaps about spatial and temporal patterns of human activities in nature.

The aim with this thesis is to gain improved understanding of human-nature interactions based on user-generated geographic information with a focus on social media data from national parks and green spaces. The main objectives are to gain 1) a novel understanding about user-generated data, and 2) insights about human activities in nature on different scales through these questions: Where and when are people visiting nature? What are people doing and valuing in nature? Which users have shared their data from national parks and green spaces?

This thesis consists of four articles and an introductory section. Article I provides an overview of social media data sources and analysis methods relevant for nature conservation, and highlights that most of the analytical opportunities are still unexplored in the growing body of literature using social media data in conservation science. Article II compares social media data with national park visitor survey and finds similar trends in both data sources regarding popular activities and visited places. Article III compares methods for detecting national park visitors' place of residence from geotagged social media and assesses biases that affect the analysis. Article IV compares the use of social media data, sports application data, mobile phone data and participatory geographic information for understanding the use of urban green spaces and suggests that combining information from several sources provides a more comprehensive understanding of green space use and preferences.

Overall, user-generated geographic information offers valuable insights about where, when and how people use and value nature, especially from areas that are otherwise difficult to monitor. There are several issues related to data access, bias and privacy in these data. Despite evident limitations, these data contribute to a better understanding of human activities in nature and complement traditional data sources with new and dynamic perspectives. In some areas, user-generated data might be the best available information about human activities. Data comparisons from national parks and green areas presented in this thesis also feed into other fields of research using social media and other user-generated data for studying human spatial behaviour.

TIIVISTELMÄ

Tämän väitöskirjan tavoitteena on hankkia uutta tietoa ihmisen ja luonnon vuorovaikutussuhteesta sosiaalisen median ja muiden uusien käyttäjälähtöisten paikkatietoaineistojen pohjalta. Tutkimus keskittyy viheralueille ja kansallispuistoihin. Hyödynnän sosiaalisen median aineistoja, sekä muita mobiililaitteiden käytöstä syntyviä aineistoja viheralueiden ja kansallispuistojen käytön tutkimisessa, ja arvioin näiden aineistojen käytettävyyttä maantieteellisen tiedon lähteenä. Tutkimuksen tavoitteena on tarjota menetelmällistä ymmärrystä käyttäjien tuottamien paikkatietoaineistojen hyödyntämisestä luonnonsuojelututkimuksessa, sekä tuottaa tietoa luontovirkistyksen alueellisesta ja ajallisesta vaihtelusta eri mittakaavatasoilla. Tarkastelen tavoitteita seuraavien kysymysten kautta: Missä ja milloin ihmiset viettävät aikaa luonnossa? Mitä ihmiset tekevät viheralueilla ja kansallispuistoissa, ja mitä he näillä alueilla arvostavat? Ketkä jakavat maantieteellistä tietoa luontovierailuistaan?

Väitöskirja koostuu johdanto-osiosta ja neljästä osatyöstä. Artikkeli I luo katsauksen sosiaalisen median aineistojen hyödyntämiseen luonnonsuojelututkimuksessa, ja kuvailee keskeiset aineistolähteet ja analyysimenetelmät. Artikkelissa tunnistetaan lähestymistapoja, joiden mahdollisuuksia ei vielä ole täysin hyödynnetty luonnon ja ihmisen vuorovaikutuksen tutkimisessa. Artikkeli II vertailee sosiaalisen median aineistoja kyselytutkimukseen ja kävijätilastoihin Pallas-Yllästunturin kansallispuistosta. Suosituimmat aktiviteetit ja vierailukohteet toistuvat molemmissa aineistoissa. Artikkeli III vertailee aika- ja paikkatietoon pohjautuvia menetelmiä sosiaalisen median käyttäjien kotimaan tunnistamiseen ja arvioi analyysiin liittyviä rajoitteita. Artikkeli IV vertailee sosiaalista mediaa, matkapuhelinaineistoja, urheilusovellusdataa, ja osallistavan paikkatietokyselyn tuloksia kaupungin viheralueiden käytön tutkimisessa. Aineistot tarjoavat toisiaan täydentävää tietoa viheralueiden käytöstä ja arvostuksesta.

Käyttäjälähtöiset paikkatietoaineistot auttavat ymmärtämään missä, milloin ja miten ihmiset käyttävät ja arvostavat kansallispuistoja ja viheralueita, erityisesti alueilla joita on muuten hankala monitoroida. Aineistojen epävarma saatavuus kuitenkin rajoittaa näiden aineistojen käyttöä tutkimuksessa. Lisäksi käyttäjäryhmiin ja aineistojen maantieteelliseen kattavuuteen liittyvät vinoumat sekä yksityisyyden suojaan liittyvät kysymykset rajoittavat käytännön sovelluksia. Rajoitteista huolimatta ihmisten itse tuottamat paikkatietoaineistot tarjoavat arvokasta lisätietoa kansallispuistojen ja viheralueiden suunnittelun ja kestävän hallinnan tueksi. Kansallispuistoista ja viheralueilta tuotetut analyysit ja aineistovertailut tarjoavat uutta tietoa myös muille sovellusaloille joilla hyödynnetään uusia aineistoja ihmisten liikkumisen ja aktiviteettien tutkimiseen.

ACKNOWLEDGEMENTS

In spring 2015, I was in Tuuli's office explaining to her how I had tried to take a photograph of a butterfly with my smart phone, but it had flown away. Tuuli replied by introducing me to the idea of using social media data in nature conservation research; a topic they had recently been brainstorming about with Enrico and Henrikki. In spring 2016. I started my PhD in the project Social Media Data for Conservation Science (SoMeCon), and here we are now! Year 2020 has been quite exceptional, and I am happy to be able to finalize this work with (remote) support from my supervisors and peers.

I greatly appreciate having Professor Catherine Pickering as my Opponent in the public examination of this thesis and I am very much looking forward to our discussions. I am grateful to Professor Alexander Zipf and Professor Kate Sherren for investing their time in pre-examining this work and for their insightful comments. I would also like to express my gratitude to Professor Niina Käyhkö and Dr Riikka Paloniemi for the positive and encouraging comments in the thesis committee meetings.

I am very thankful to my supervisors Tuuli Toivonen and Enrico Di Minin for taking me onboard in the SoMeCon project and for encouraging and supporting me along the way. Tuuli is an exceptional role model as an academic and a human being. Tuuli has constantly helped me to believe in myself and to carry on despite doubts and delays. Special thanks to Tuuli and her family for hosting me in Cambrige, UK in February 2020. Enrico is an outstanding scholar and I have been happy to follow the big leaps in his career advance during these years. I would like to thank Enrico for involving me in interesting research projects beyond this thesis during the years and of course for introducing me to the world of rhinoceroses.

In addition to my supervisors, publications in this thesis have been co-authored by Claudia Bergroth, Joel Erkkonen, Christoph Fink, Anna Hausmann, Tuomo Hiippala, Olle Järv, and Henrikki Tenkanen. I want to thank all co-authors for their input, patience and quick reactions during manuscript preparations. Furthermore, I want to thank all anonymous reviewers for their constructive comments on the four articles in this thesis. I want to thank Liisa Kajala and colleagues at Parks and Wildlife Finland, Metsähallitus and Marna Herbst and colleagues at the South African National Parks for collaboration in workshops, and for contributing data about national park visits. Big thanks also to Reetta Välimäki, Laura Lipasti and Emil Ehnström for assisting with data classification, and Anni Virolainen for graphics design. I would like to acknowledge Ian Dobson for his language revision recommendations in the synopsis.

I am very thankful to the KONE foundation for funding the SoMeCon project, and to Tuuli for supporting my research through this project grant. I am thankful for the travel grant from the Doctoral Programme in Interdisciplinary Environmental Sciences (DENVI) and for the

facilities and support provided by the Faculty of Science in Kumpula. Special thanks to the Dean Kai Nordlund and the Kumpula campus PhD services for support during the pre-examination of this thesis. I would like to thank Karen Sims-Huopaniemi from DENVI for invaluable support with all administrative matters regarding the PhD.

I would like to thank all fellow DENVI students for great discussions and good company over the years in various workshops, writing retreats and excursions. Sincere thanks to students and staff in the geography corridors in Kumpula for the relaxing coffee breaks and good times in the GIS labs. Special thanks to Arttu Paarlahti who originally introduced me to the magical world of GIS during my freshman year.

I wish to thank current and former members of the Digital Geography Lab for creating such a supportive and positive atmosphere in the group. Thank you Maria, Henkka, Olle, Kerli, Aija, Elias, Claudia, Joel, Henna, Johanna, Enrico, Anna, Anna, Christoph, Gonza, Tuomo, Väiski, Age, Ainokaisa, Ludovic, Jeison, A-P, Joose, and Bryan for being part of the team, and Tuuli for bringing all these awesome people together. When joining the "small but efficient" accessibility research group as a research assistant, I was a bit worried that I would be the only one left in a couple of years after Henkka graduates. Now, five years later, the group has evolved and transformed into something very special without losing the essence of being *open and happy*.

Finally, I would like to thank my family and friends for unconditional support.

In Helsinki, Finland, November 2020,
Vuokko V. Heikinheimo

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications, which are referred to in the text by their roman numerals:

I Toivonen, T., **Heikinheimo, V.**, Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H., & Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation*, 233, 298–315. https://doi.org/10.1016/j.biocon.2019.01.023

II **Heikinheimo, V.**, Di Minin, E., Tenkanen, H., Hausmann, A., Erkkonen, J., & Toivonen, T. (2017). 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), 85. https://doi.org/10.3390/ijgi6030085

III **Heikinheimo, V.**, Järv. O., Tenkanen, T., Hiippala, T., & Toivonen, T. (under review). Detecting place of residence from social media data: a comparison of methods. Submitted manuscript.

IV **Heikinheimo, V.**, Tenkanen, H., Bergroth, C., Järv, O., Hiippala, T., & Toivonen, T. (2020). Understanding the use of urban green spaces from user-generated geographic information. *Landscape and Urban Planning*, 201, [103845]. https://doi.org/10.1016/j.landurbplan.2020.103845

Table of contributions

	I	II	III	IV
Idea	TT, EDM, AH, HT, VH, CF	VH, TT	OJ, HT, TH	VH, TT
Literature review	VH	VH	VH	VH
Data collection and preparation	VH, HT, CF	VH, HT, JE	HT, VH	VH, HT, CB, TH
Analysis	VH, HT, TH, CF	VH	VH	VH
Visualizations	VH, HT, TH, CF	VH	VH	VH
Writing – original draft	VH, TT, HT, OJ, TH, CF	VH	VH, OJ	VH
Writing – review and editing	VH, TT, HT, OJ, TH, CF, AH, EDM	VH, TT, EDM, HT, AH, JE, TT	VH, OJ, TH, HT, TT	VH, TT, HT, CB, OJ, TH

AH: Anna Hausmann
CB: Claudia Bergroth
CF: Christoph Fink
EDM: Enrico Di Minin
HT: Henrikki Tenkanen
JE: Joel Erkkonen
OJ: Olle Järv
TH: Tuomo Hiippala
TT: Tuuli Toivonen
VH: Vuokko Heikinheimo

1. INTRODUCTION

1.1. MOTIVATION

> *... even imperfect measures of their value, if understood as such, are better than simply ignoring ecosystem services altogether as is generally done in decision making today.*
>
> *Daily et al. 1997, p. 8*

Advances in information technology have revolutionized the collection of geographic information (Kitchin, 2014a). Remotely sensed data, such as satellite images, provide near-real time information about the environment, and mobile sensors and online platforms provide continuous information about people's observations, activities and movements (Pimm et al., 2015; Toth & Jóźków, 2016). We are living in a so-called information age (Castells, 2000) - an era characterized by considerable amounts of digital data.

At the same time, we are living in the Anthropocene Epoch - a new geological era characterized by significant human impact on the global environment and biodiversity (Steffen et al., 2011). Biodiversity and ecosystems are fundamentally important to human life (Daily et al., 1997; Haines-Young & Potschin, 2010; MA, 2005), but human activities continue to cause declines in nature (Díaz et al., 2019). Understanding the diversity of ways in which nature contributes to people's well-being is crucial for successful nature conservation and sustainable decision-making (Díaz et al., 2018).

In particular, understanding human dimensions of nature conservation is crucial for finding successful solutions for protecting biodiversity and ecosystems (Bennett et al., 2017; Venter et al., 2016). Despite the increasing availability of digital data, there is still a lack of spatially and temporally accurate information about threats to biodiversity (Joppa et al., 2016), and the benefits people get from nature (Beeco & Brown, 2013). Collecting data on human-nature interactions is resource-intesive, and funding for such data collection efforts is often lacking (Waldron et al., 2013).

Different types of green spaces and protected areas are important for biodiversity and the well-being of people (Niemelä et al., 2010; Watson et al., 2014). Protected areas attract over eight billion visitors per year globally (Balmford et al., 2015), and nature-based tourism can provide funding and political support for maintaining protected areas and natural sites in place (Buckley, 2009). National parks are protected areas particularly designed for both protecting nature and providing recreational and other cultural opportunities for people[1]. Park management in different contexts benefits from spatially explicit data on social phenomena, such as the type and intensity of different activities (Beeco & Brown, 2013; Pickering et al., 2018). For example, information about how people use and value parks and green spaces can support sustainable

[1] https://www.iucn.org/theme/protected-areas/about/protected-areas-categories/category-ii-national-park

and equitable land use planning while also supporting biodiversity conservation (Burkhard et al., 2012; Haaland & van den Bosch, 2015). However, information about human activities and values have been difficult to acquire, especially in a spatially explicit form (Beeco & Brown, 2013; McIntyre et al., 2008).

Emerging data sources from crowdsourcing initiatives, citizen science projects and big data feeds have become a recognized source of geographic information alongside official data sets produced by scientists and government authorities (Elwood et al., 2012; Goodchild, 2007; See et al., 2016). Users of various online platforms and mobile devices generate considerable amounts of data about their observations, activities and preferences in a range of environments. Web-based social media services allow users to generate and share different types of online content such as text, images and video, and users may also link their content to a specific location using place names and geotags (McCay-Peet & Quan-Haase, 2017).

Location-based social media platforms, such as Flickr, Instagram and Twitter contain information particularly about people's leisure time and positive experiences, which makes them an appealing data source for studying nature-based tourism (Di Minin et al., 2015; Hausmann et al., 2018). Rich content on social media allows for human activities to be understood "beyond the geotag" (Crampton et al., 2013); not only the time and location, but also the networks, observations, activities and motivations of people. It has been suggested that user-generated data, such as social media, could help fill in information gaps about both threats to biodiversity and opportunities for nature conservation (Di Minin et al., 2015).

This work builds on an emerging body of literature that examines human-nature interactions through user-generated data sets, such as social media data. A growing number of studies has used social media data in environmental research in the past decade (Ghermandi & Sinclair, 2019; Teles da Mota & Pickering, 2020). For example, previous research has brought forward promising results on using social media for detecting visitor rates in protected areas (Levin et al., 2015; Sessions et al., 2016; Tenkanen et al., 2017; Wood et al., 2013), and in-situ activities, preferences and values (Hausmann et al., 2018; Richards & Friess, 2015; van Zanten et al., 2016).

There are many challenges with social media data analysis not tackled in the current literature. The main limitations to applying social media data analysis in any field of research are related to data quality, biases, data availability and ethical use (Di Minin et al., in press; Ruths & Pfeffer, 2014; Senaratne et al., 2017; Zook et al., 2017). In this thesis, I have addressed some of these gaps in the context of studying human-nature interactions and have extended the discussion to other types of user-generated geographic information beyond social media. This thesis combines data from a range of sources and takes a holistic perspective by incorporating several elements of user-generated geographic information for studying the spatial and temporal patterns of human activities in nature (Figure 1).

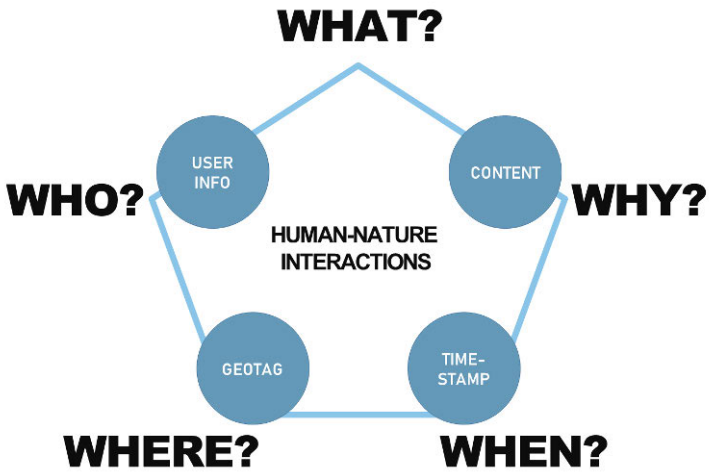

Figure 1. User-generated geographic information provide insights about different aspects of human activities in nature: Where and when are people visiting nature? What are people doing and valuing in nature? Who are those that have shared their data from nature destinations? (Adapted from Articles II, IV & Di Minin et al., 2015).

1.2. OBJECTIVES

The aim of this thesis is to lead to a better understanding of the use of novel data sources for studying human-nature interactions with a particular focus on human activities and preferences in national parks and urban green spaces. This work included data from social media platforms, mobile network operators, sports tracking applications and map-based surveys (PPGIS data). The overall objectives of this thesis are to:

1. Describe and understand how user-generated data can be used as a source of geographic information about the use of national parks and green spaces.
2. Discover spatial and temporal patterns of human activities in nature on different scales of observation from user-generated geographic information.

To achieve these objectives, I looked at elements of user-generated geographic information and related analysis methods using the framework presented in Figure 1: Where and when are people visiting national parks and green spaces? What are people doing and valuing in national parks and green spaces? Who are the users who have shared their data from national parks and green spaces?

I addressed the research objectives in the four articles and summarized the main findings in this summary report (the *synopsis*). Articles I-III focused on social media data, and Article IV also included other sources of user-generated geographic information. Each article is related to a combination of the questions about where, when, what, who and why (Figure 1), and contributes to both of the main objectives through methodological understanding (objective 1), and empirical observations (objective 2).

Article **I** contains a review of current literature using social media data in conservation science and an overview of relevant data sources and analysis methods. The main objectives were to describe the kind of information available from different social media platforms, to provide a detailed overview of social media analysis methods, to provide practical examples of the applicability of these methods to conservation science, and to highlight the limitations of using social media data for studying human-nature interactions.

In Article **II** social media data were compared with visitor survey data from Pallas-Yllästunturi National Park which is the most popular national park in Finland. The main objective was to evaluate how insights derived from social media data correspond to official visitor information, and to improve understanding of social media usage patterns in a national park context. Data comparisons presented in this article provide insights for using social media data to complement existing sources of visitor information regarding the questions about where, when, what, why and who.

Article **III** compares spatial and temporal measuring techniques for identifying the place of residence of national park visitors from social media data. The main objectives were to propose the most reliable approach to detect country of residence from geotagged social media data and to identify aspects in the data that affect the analysis. This article provides a methodological understanding for analysing who the users are and offers insights about the limitations of social media as a data source.

Article **IV** extends the focus from social media data to include other types of user-generated geographic information. The main objective was to compare how social media data, sports application data, mobile phone data and PPGIS data can be used to study where, when and how people use and value urban green spaces. This article highlights the complementarity of the different data sources in answering questions about urban green space use and values. This article provides practical insights about using new data sources for understanding visits to areas where systematic visitor monitoring rarely takes place.

Overall, this thesis is linked to the fields of geography, GIScience, conservation science and environmental studies in the broader context of studying human-nature interactions through novel data sources. Articles in this thesis are relevant to scientists and practitioners aiming to improve their understanding of the dynamics of human-nature interactions on different scales, and to those aiming to understand the potential and limitations of user-generated geographic information better.

2. BACKGROUND

2.1. HUMAN-NATURE INTERACTIONS

Human-nature interactions is a broad concept that captures interlinkages between human and natural systems (Liu et al., 2007). Human impact on the environment has accelerated profoundly since the Industrial Revolution (Steffen et al., 2011), while human life and well-being continue to rely on nature (MA, 2005). From a systems-thinking perspective, social and ecological processes are integrated in *social-ecological systems* (Figure 2) on different scales ranging from local, regional to global (Berkes & Folke, 1998). The Millennium Ecosystem Assessment (MA, 2005) used the concept of *ecosystem services* to describe the direct or indirect benefits human societies receive from nature. More recently, the Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES) has conceptualized the benefits from nature to good quality of life as *nature's contributions to people* (Díaz et al., 2018; Pascual et al., 2017). These concepts pinpoint that understanding the contributions of nature to people can help support the sustainable use and conservation of ecosystems and biodiversity.

Understanding the complexities of coupled human-nature systems requires an interdisciplinary approach (Liu et al., 2007). In this thesis, I have contributed to the topic from the viewpoint of geography, geographic information science and environmental sciences. In particular, this thesis looks into human-nature interactions with a focus on human systems (Figure 2) and direct interaction be-

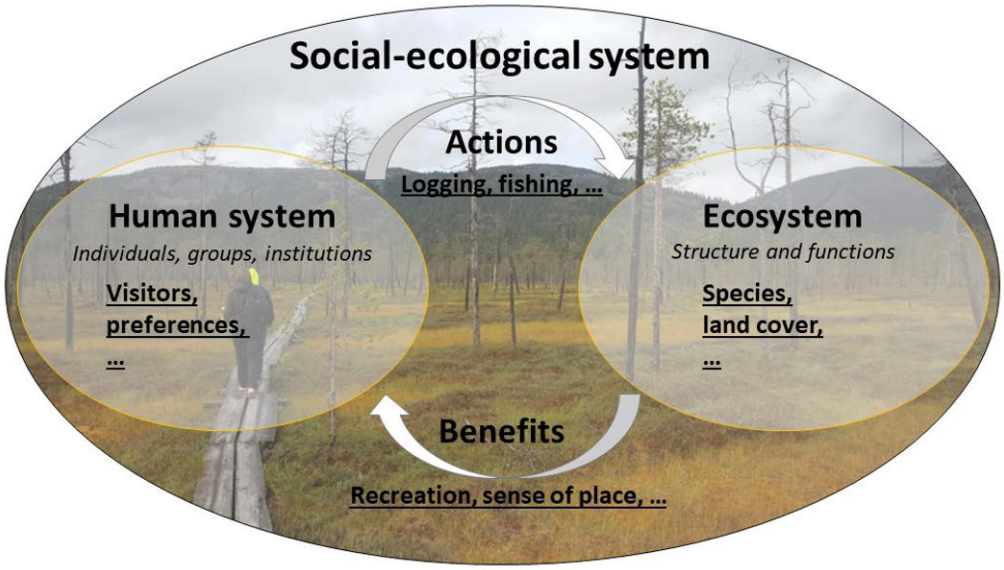

Figure 2. Elements of a social-ecological system in a conceptual diagram. The underscored text gives examples of topics that could be studied via user-generated geographic information. Diagram adapted from Resilience Alliance (2007). Photograph: Pyhä-Luosto National Park, Finland, by the author (2019).

tween humans and nature, such as in the case of recreation. In the ecosystem service literature this perspective is studied in the context of *cultural ecosystem services* (MA, 2005), in which social media and other digital data sources have gained increasing interest as a data source (see for example Richards and Friess, 2015). Methods and insights for understanding human activities in nature presented in this thesis may further feed into broader analyses and decision-making processes that also cover other aspects of social-ecological systems, such as species and land cover, as long as evident limitations of these data sources are taken into account.

2.2. DIGITAL GEOGRAPHIES

Digital geographies capture a broader disciplinary turn in geography that includes the emergence of digital tools, practices and phenomena as the tool and focus of geographic inquiry (Ash et al., 2018). This development is not unique to geography, as digital technologies are impacting research practices across different fields. This thesis is also linked to an emerging field of *digital conservation* that combines the use of digital technology and nature conservation (Arts et al., 2015), and *digital humanities*, which refers to the intersection of computer science and disciplines that study human society (Berry, 2012).

Advancements in digital technology forms the basis of using novel data sources in geographic research. First digital mapping projects and Geographic Information Systems (GIS) emerged in the 1960s in North America (Tomlinson, 1967) and since then, GIS has become an important tool for producing, storing, managing, analysing, and representing spatial data. Overall, GIS has evolved as an umbrella term for the creation, management and analysis of spatial data and has been widely used in academia and industry for solving geographic questions (Longley, 2000). As a distinction from GIS only as a tool in research, Geographic information science (GIscience) has emerged as a discipline that focuses on the theories, methods and technologies related to transforming geographic data into useful information (Goodchild, 1992; Mark, 2003) and this thesis builds upon and contributes also to this literature.

Critical geographers have criticized the use of GIS and the related quantitative approach in geography about its supposed objectivity and inability to capture human experiences (Ash et al., 2018; Kwan & Schwanen, 2009). More recently, geographic information generated by regular people might have bridged this gap to some extent (Elwood et al., 2012; Goodchild, 2007). Cultural and technological developments have allowed users of various online platforms to start generating geographic data – as opposed to authoritative data sources that were previously the main source of geographic information (Sui & Goodchild, 2011).

New data sources and analytical tools have allowed for the testing of older geographic theories in practice, including the ideas of time geography (Hägerstrand, 1970) which investigates human activities in space and time through asking the questions about where, when and who. These ideas from time geography form the basis of the questions asked in this

thesis (Figure 1). GPS devices and mobile sensors allow these spatial and temporal questions to be studied with real-world data, and social media in particular have extended the possible questions to include *what* and *why* through their rich content (Di Minin et al., 2015).

Geotags and other location references are practical means of linking digital data to the physical world in a GIS environment. Conceptually, while *cyberspace* would refer to virtual space or online communities that exist primarily in the *digital space*, *hybrid spaces* are a combination of *digital* and *physical space* connected through location-based technologies such as the smartphone (de Souza e Silva, 2006) (Figure 3). The main assumption in this conceptual approach is that objects from the digital space (such as a geotagged photo), represent something that happened in the physical space (such as a national park visit). The idea of hybrid spaces then provides the basis for analysing human activities in nature through digital data – which is the focus of this thesis.

There are several limitations to spatial big data research from an epistemological perspective (Thatcher, 2014). For example, it is relevant to ask critically if we are really capturing the physical use of space through digital data, and if so, how accurately. Furthermore, researchers using digital data sources for studying geographic phenomena should ask critically if the observed patterns exist only in the digital space and not in real life. For example, if someone has geotagged a photo into a national park, did they actually visit the place?

An underlying assumption in big data approaches often is that user-generated content reveals relevant and meaningful information (Kitchin, 2014a; Thatcher, 2014). The assumption is that the data

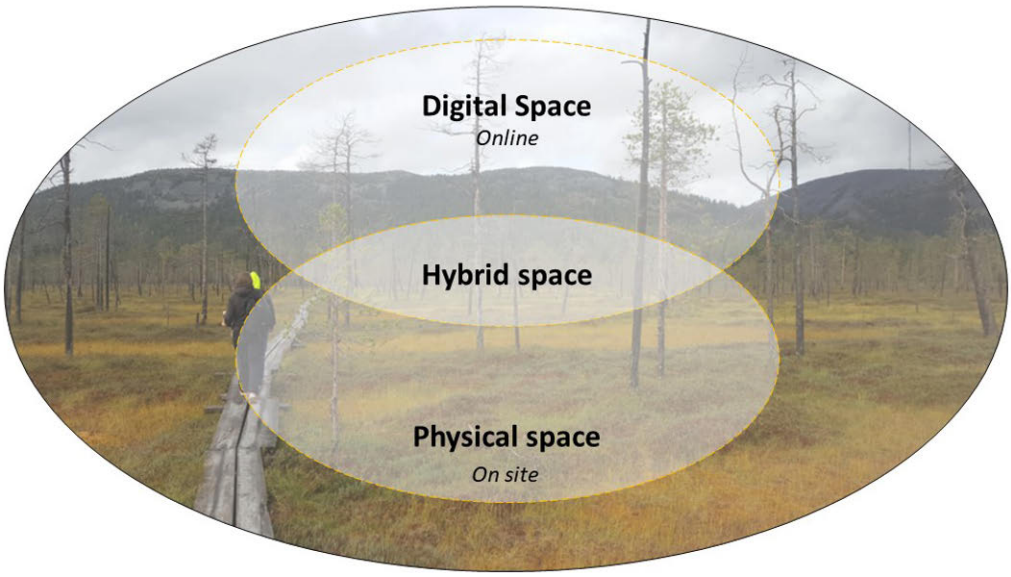

Figure 3. Illustration of the digital space, physical space and their intersection – the hybrid space. Inspired by Ash et al. (2018); de Souza e Silva (2006).

are not only noise and from the emergent patterns, we can find an explanation for the observed phenomena. With this thesis my aim is to question this assumption by investigating whether user-generated geographic information contains meaningful information for studying human-nature interactions.

When studying geographic phenomena through digital data, it is important to consider whose data we are actually analysing. The fact that social media platforms are used by specific groups of people is a major limitation for using these data to understand socio-spatial phenomena more generally (Ruths & Pfeffer, 2014). Also, not everyone geotags their social media posts, and the tendency to add a geotag varies in different cultural regions (Huang & Carley, 2019). Unequal representation across different geographic regions is a relevant issue when using social media and other online data sources. The popularity of social media platforms – and the overall access to internet varies in different regions. The concept of digital divides (Castells, 2000; Crampton, 2009) captures the fact that access to digital technology and the internet is unevenly distributed. Understanding who shared their data from national parks and green spaces is one of the main sub-objectives of this work (Figure 1).

2.3. USER-GENERATED GEOGRAPHIC INFORMATION

The concept of user-generated geographic information captures several types of data sets produced through interactions with people and location-based technologies. Examples of such data sets include social media data, sports tracking data, mobile phone data and participatory geographic information. There are several interrelated concepts that describe 'novel' sources of geographic information such as big data, crowdsourced data, volunteered geographic information (VGI), and user-generated content (Table 1). In this work, I chose to use the broadest possible concept that covers several kinds of geographic information generated through the active or passive interaction of people and location-based technologies.

The conceptual starting point for user-generated geographic information is the idea of "citizens as censors" and the notion of volunteered geographic information (VGI) (Goodchild, 2007) – a special case of user-generated content on the internet and the so-called web 2.0 (Table 1). However, as a concept, VGI puts an emphasis on the act of volunteering, which refers to active and aware contribution. On the one hand, VGI is a good definition when it comes to spatial data contributed to citizen science projects, or OpenStreetMap (www.openstreetmap.org), which is an open-source database for street networks and other geographic features that anyone can edit ("the Wikipedia of maps"). On the other hand, there are many user-generated geographic data sets, such as geotagged social media data, and mobile phone records that are stored and used for purposes other than those originally intended, and which cannot be categorized as actively volunteered data. From this perspective, the concept of *user-generated geographic information* captures both actively and passively contributed sources of data (replacing the word "vol-

unteered" with "user-generated" in VGI). The exact concept of user-generated geographic information has been used occasionally (Shelton et al., 2015) but not very often, perhaps due to its length and the plurality of related concepts available (See et al., 2016).

The concept of 'big data' aims to capture the overwhelming nature of these vast data sets and data streams. Big data don't only refer to storage space requirements. In addition to their huge volume, big data are characterized by high velocity, variety (non-uniform structure), exhaustive scope, fine-grained resolution, relational nature, flexibility and scalability (Kitchin, 2014b). Different types of big data provide new opportunities for studying people and nature and might fill in data gaps in nature conservation (Di Minin et al., 2015). In this thesis, however, the data sets used eventually become *small data* (Poorthuis & Zook, 2017) – smaller extracts of the huge volumes of publicly-shared content from different social media platforms and data sources.

Geotagged social media data is the primary example of user-generated geo-

Table 1. Key concepts related to user-generated geographic information. Finnish translation is provided in the parenthesis. Definitions adapted from: See et al. 2016; Oxford dictionary (https://www.oxfordlearnersdictionaries.com/).

Concept	Definition	Examples
Big Data (fin. *massadata*)	Large and complex data that pose challenges to computation and storage.	Data streams from social media such as Twitter
Web 2.0	Participatory form of the World Wide Web where users have a central role in generating content.	Social media platforms where users can share text, images and video; Different Wiki pages that anyone can edit, such as Wikipedia
User-generated content (fin. *käyttäjien tuottama sisältö*)	Content, such as text photographs or video created and shared by users of online platforms.	Social media posts, content on Wikipedia
Crowdsourcing (fin. *joukkoistaminen*)	The practice of obtaining information through contributions from a large number of people (paid or unpaid) typically via the Internet.	OpenStreetMap is a crowdsourced dataset where registered users can edit the map from anywhere in the world.
Citizen science (fin. *kansalaistiede*)	Engagement of the general public in scientific research through data collection, data analysis or defining the actual research problem.	eBird is a citizen science project focused on bird observations. iNaturalist is a citizen science platform for collecting species observations. iNaturalist uses crowdsourcing and machine learning for species identification.
Public Participation GIS (PPGIS) (fin. *osallistavat paikkatietomenetelmät*, "pehmoGIS")	An approach for using geospatial technologies to enable participation in public processes.	Map-based surveys, such as http://kerrokartalla.hel.fi/ by the city of Helsinki
Volunteered Geographic Information (VGI) (fin. *vapaaehtoisten tuottama paikkatieto*)	Geographic information produced by volunteer contributors.	OpenStreetMap data consists (mostly) of volunteered contributions. In iNaturalist, users can voluntarily share their nature observations.

graphic information used in this work. By definition, social media are "web-based services that allow individuals, communities and organizations to collaborate, connect, interact, and build a community by enabling them to create, co-create, modify, share, and engage with user-generated content that is easily accessible" (McCay-Peet & Quan-Haase, 2017). A social media post refers to an item such as text, image, video or combination of these that a user has shared on a social media platform. Different elements of social media data (Figure 4; Article I) provide new possibilities for studying human behaviour in different environments (Di Minin et al., 2015).

Social media data match the definition of big data; social media feeds are huge in volume, high in velocity, diverse in variety, exhaustive in scope, fine-grained in resolution relational in nature, flexible, and scalable – in line with the characterization of big data by Kitchin (2014b).

In this thesis, I used social media as a source of information about the whereabouts and activities of people. Social media data also provide other perspectives about human-nature interactions beyond in-situ visits. Social media have a role in disseminating information, raising awareness, or as a platform for discussions, but these viewpoints on social media are outside the scope of this thesis.

This work also includes other types of user-generated content in addition to social media data. **Mobile phone data** contain location information and other details about mobile devices in the mobile phone operator's network (Ahas et al., 2010). GPS tracking data collected via sports applications contain spatial and temporal information about physical activities with relatively high accura-

Figure 4. Elements of social media data as illustrated in Article I.

cy. **Public participatory GIS (PPGIS)** and participatory geographic information systems (PGIS) refer to approaches that combine participatory methods with geographic information technologies to support public processes such as planning (Brown & Kyttä, 2014).

We can further categorize user-generated geographic information according to the level of participation and engagement. Building on Arnstein's (1969) ladder of participation, Haklay (2013) identified several levels of participation in geographic citizen science projects, ranging from citizens as merely information providers (Level 1:"Crowdsourcing"), to highly collaborative participation, in which citizens are also involved in problem definition (Level 4: "Extreme citizen science"). In this typology, social media data, sports app data and mobile phone data represent crowdsourcing (the lowest level of participation), as the data producer is most likely to be unaware of the research or planning project in which their data are used. PPGIS data belongs to somewhere between Level 2: "distributed intelligence" and Level 3: "participatory science" having a higher level of participation and interaction between the citizen data producer and the end analyst in comparison with the other data sources.

However, Haklay's typology focuses on the values in specific projects, which aim to collect new information and I would argue that it does not directly apply to data sets generated for other purposes. For example, while social media data fit under "crowdsourcing" and the idea of citizens as sensors to some extent, there is still a clear distinction with crowdsourced data for citizen science, and data mined from online networks. Social media data exist regardless of research and planning initiatives (unless special promotion campaigns have taken place), and thus the researcher has less control of the structure and content of the generated data in comparison with conventional citizen science data sets for which data are collected, for example, via a custom-made application.

Overall, the literature has attempted to define, categorize and position the existing interrelated concepts in different ways (Brown & Kyttä, 2014; Mocnik et al., 2019; See et al., 2016), but there seems to be a lack of a common typology for different types of user-generated geographic data sets that persists over time. For example, previous studies have referred to social media data as volunteered geographic information (VGI), ambient geographic information, mobile big data, and (passively) crowdsourced data to name a few categorizations. The user base, content, technical structure and availability of user-generated data evolve over time, and that is why it is difficult to set one label on these novel data sources.

2.4. PREVIOUS WORK USING SOCIAL MEDIA DATA FOR STUDYING HUMAN-NATURE INTERACTIONS

An increasing number of studies on human-nature interactions based on user-generated data, social media in particular, have been analysed since the emergence of these new data sources in the two first decades of the 2000s. Article I in this thesis reviewed the relevant literature and analysis methods in the con-

text of conservation science, and other recently-published review articles have reported summaries of the use of social media data in the context of environmental research (Ghermandi & Sinclair, 2019) and nature-based tourism studies (Teles da Mota & Pickering, 2020). Work related to this thesis has often been framed under the interrelated topics of ecosystem service mapping, nature-based tourism, digital conservation, and other aspects of environmental studies more generally.

Studies focusing on ecosystem services have used a range of indicators for mapping ecosystem service demand (Wolff et al., 2015). For cultural ecosystem services such as recreation and aesthetics (MA, 2005), demand has been measured as the direct use of the service, or preferences and values attached to the service (Wolff et al., 2015). PPGIS emerged first as a new way of mapping cultural ecosystem services (Brown, Schebella, et al., 2014; Brown, Weber, et al., 2014; Ives et al., 2017; Laatikainen et al., 2015), and social media data have also been increasingly used in this context (Gliozzo et al., 2016; Oteros-Rozas et al., 2016; Richards & Friess, 2015).

Visitor monitoring in protected areas and green spaces has been traditionally conducted using surveys and on-site counters (see for example Kajala et al., 2007). User-generated data sources have provided new opportunities for complementing existing visitor monitoring schemes and understanding use patterns and values in unmonitored areas. Several studies have focused specifically on using social media for estimating visits to protected areas (Fisher et al., 2018; Hausmann et al., 2017; Levin et al., 2015; Sessions et al., 2016; Tenkanen et al., 2017; Wood et al., 2013), urban parks and trails (Donahue et al., 2018; Hamstead et al., 2018; Wu et al., 2017), in-situ activities and preferences (Hausmann et al., 2018), and landscape values (van Zanten et al., 2016). Recreational use estimates based on geotagged photos from Flickr (approach from Wood et al., 2013) are integrated in the InVest software[2] used by researchers and practitioners for ecosystem service mapping.

Studies framed under conservation culturomics have leveraged content analysis methods for understanding cultural trends related to nature conservation (Ladle et al., 2016). Culturomics studies often focus on the digital space (Figure 3), for example, focusing on online sentiment for iconic animals (Fink et al., 2020), or public awareness of the value of biodiversity (Cooper et al., 2019). While most conservation culturomics studies have focused on text content (by definition, culturomics refers to analysis of large bodies of text), there have been calls for visual content also to be leveraged in this approach (Sherren, Smit, et al., 2017).

While the articles included in this thesis focus mainly on the positive aspects of human-nature interactions such as nature recreation that can in turn promote opportunities for conservation, user-generated data can also inform scientists and practitioners about threats to biodiversity, as proposed by Di Minin et al. (2015). For example, social media offer new opportunities to study the ille-

[2] https://naturalcapitalproject.stanford.edu/software/invest

gal trade in wildlife products (Di Minin et al., 2018). Geotagged social media data may also serve as proxy for human-induced threat in protected areas, while at the same time offering valuable information to support nature-based tourism and management in vulnerable sites (Hausmann et al., 2019).

While the number of publications using social media data to study human-nature interactions is increasing rapidly, it is still uncertain how applicable this approach will be in the long term, as the data sources are relatively new. There is a need to focus on known information gaps related to human-nature interactions and it continues to be important to improve understanding of how to address the known limitations of social media data. Many studies only utilize a limited portion of the available data, for example, only focusing on the geotags, or text content, or a single data source (Article I; Ghermandi and Sinclair, 2019; Teles da Mota and Pickering, 2020). Most of the studies applying social media have been conducted in North America and Europe by North American and European researchers (Teles da Mota & Pickering, 2020). Furthermore, patterns observed from social media are difficult to validate due to the absence of ground-truth data, and limitations related to data quality, data availability and ethical use persist. Developing robust workflows that help protect personal data are needed (Di Minin et al., in press). This work aims to address some of these gaps in the four articles that build on and complement the existing body of literature.

3. MATERIAL AND METHODS

Article I presents an overview of analysing social media data following the main steps of a generic data mining workflow (Figure 5; Han et al., 2011). The main steps include defining the research question; acquiring and storing the data; filtering, enriching and analysing the data; and finally, critically assessing the results. If needed, these steps should be iterated and adjusted, for example, to collect additional data or to do further data filtering.

In this work, the main empirical questions are related to the spatial and temporal use of national parks and green spaces, related activities and preferences, as well as understanding better who the users are, as illustrated in Figure 1. All articles include analysis of social media data, and article IV also includes other sources of user-generated geographic information. Furthermore, Articles II and III use official national park visitor data. We acquired the data from Application Programming Interfaces (APIs), open data portals, and through collaborations. To answer the questions, we used spatial and temporal analysis approaches, as well as visual and textual content analysis methods. The data collection tools were written in the Python programming language (Python versions 3.5, 3.6 and 3.7; www.python.org), and the data storage used a PostgreSQL (Versions 9.5 and 11.7) database with a PostGIS extension on a secure server. Centrographic methods in Article III used tools available in ArcMap 10.3. implemented in Python 2.7.

The main analysis scripts and supporting information for Articles I, III and IV are available online via GitHub.com[3],[4],[5]. In principle, the raw social media data sets

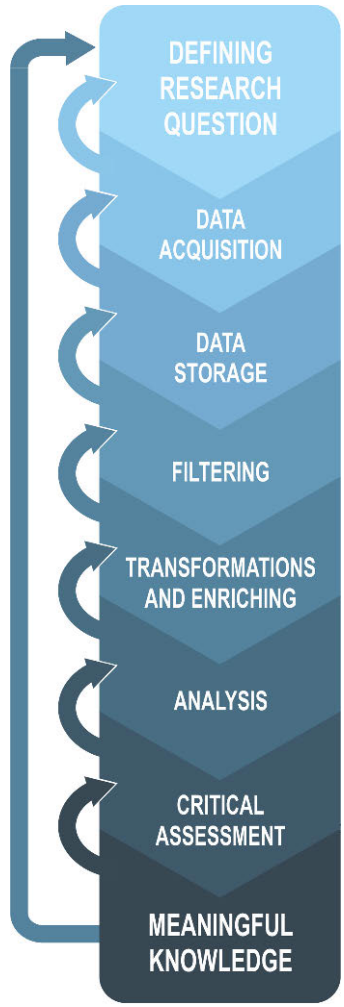

Figure 5. Main steps in a social media data analysis workflow as illustrated in Article I.

[3] Article I: https://github.com/DigitalGeographyLab/some-conservationscience

[4] Article III: https://github.com/DigitalGeographyLab/some-origins-demo

[5] Article IV: https://github.com/DigitalGeographyLab/some-urbangreens

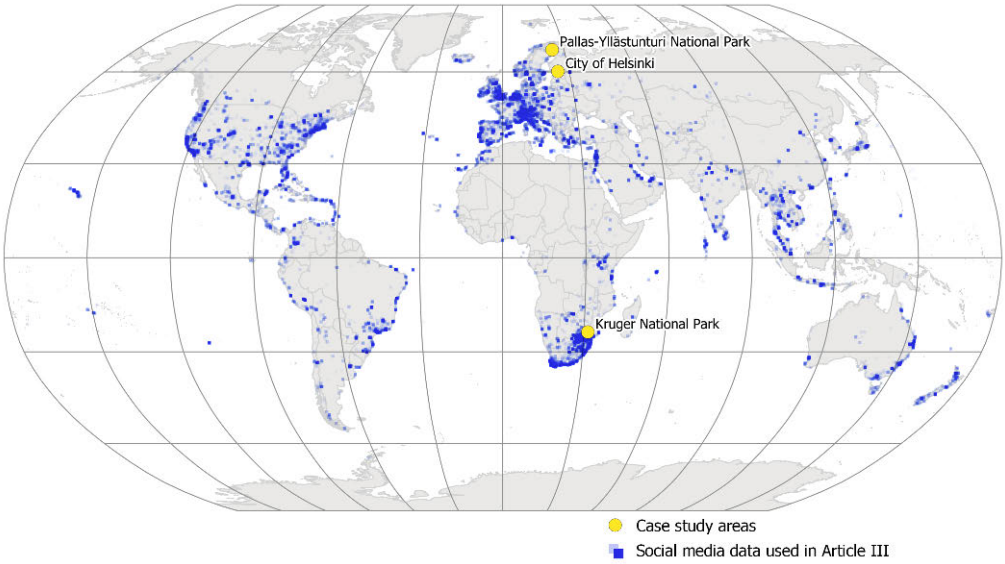

Figure 6. Study areas included in this thesis. The global map shows the input data used in Article III representing posting history of social media users who visited Kruger National Park in South Africa. The darker the shade, the more data from that location.

used in this thesis have not been made publicly-available due to the terms and conditions of the platforms and in order to protect the privacy of data subjects. However, the scripts for Article III are accompanied by a fabricated data set representing the global posting history of social media users that can be used to test how the code works and how the data should be structured.

3.1. STUDY AREAS

The scale of observation in this thesis spans from global to regional to local. Article I presents analysis examples at several scales. Article II focuses on Finland's most popular national park (Pallas-Yllästunturi National Park) and its subregions, and also looks into the users' countries of origin. Article III looks into the countries of origin of social media users who visited Kruger National Park in South Africa based on a global data set. Article IV zooms into urban green spaces in Helsinki, Finland, and uses a 250 m x 250 m statistical grid for comparing several sources of user-generated geographic information.

3.2. DATA SOURCES

Geotagged social media data aggregated into national parks (and subregions), statistical grids, green spaces, countries and regions were the starting point for all analyses in this thesis. Empirical examples in this work include data from three different social media platforms: Flickr, Instagram and Twitter. The data collection tools, developed by Tenkanen (2017) were used to access publicly shared geotagged content via the Application Programming Interfaces (APIs) of each platform. In principle, social media APIs allow programmatic access to publicly-available content, including text, media links (photographs, video) and user profiles.

Flickr is a photo-sharing platform that is popular for sharing nature-related

content. The Flickr API allows free access to publicly-available content as long as the restrictions set by the users are honoured[6]. For example, a user might have attached an additional creative commons license to their photographs. The location information on the Flickr platform is often relatively precise (Hochmair et al., 2018), as the platform allows users to upload coordinate information directly from their device. Alternatively, users might geotag their posts afterwards by zooming in on a web map.

Instagram is a photo and video sharing platform that allows users to geotag their content. The Instagram platform and the related API has changed during the execution of this thesis. Earlier, Instagram posts could contain precise geotags based on device location, as exemplified in the Article I supplements. At the time of data collection for this thesis, Instagram geotags were only attached to points-of-interest (POIs); a set of place names at various scales linked to a specific location (in coordinates). For example, a POI can represent a specific café inside a national park, and there can also be a POI for the entire national park, as illustrated in Article II.

Twitter is a social media platform for sharing short messages (currently max 270 characters, max 140 characters up until November 2017) and related content. Twitter API allows several approaches for data collection. The standard search functionality is limited to public content approximately from the past seven days, which means that the data have to be continuously collected in order to capture temporal trends from a specific area or related to a specific key word. Timeline functionality allows collecting tweet history from a single user up to 3200 tweets by the time of writing. Commercial options, such as Twitter's enterprise API offer improved access to data. Many geotagged tweets are originally from Instagram (Heikinheimo et al., 2018) potentially allowing for alternative access to Instagram content.

Social media data from Flickr, Instagram and Twitter were compared to other user-generated data sets in Article IV to investigate the spatial and temporal use patterns of urban green spaces in Helsinki, Finland.

Sports tracking data from Strava (Article IV) were acquired as a pre-processed data product (Strava Metro) through collaboration with the City of Helsinki. The data analysed covers 2015. Strava Metro is an aggregated data product; the data represent the number of athletes and trips in each road segment based on users' tracks on the Strava sports tracking application. Global heatmap visualization of Strava data is openly available online[7], and can, for example, be added as a background layer in the QGIS software.

Mobile phone data were acquired from a major Finnish mobile network operator company Elisa Oyj (Article IV). The data covered a two-and-half month period (from 28.10.2017 to 9.1.2018). The original data contained the number of hourly data use attempts (for example, browsing the internet on a mobile device) in the mobile network (Bergroth, 2019). The data set

[6] https://www.flickr.com/help/terms/api

[7] https://www.strava.com/heatmap

was anonymized by the mobile network operator and included a random error up to 200 metres. Bergroth (2019) aggregated the data from regular weekdays (Monday to Thursday) into 250 m x 250 m statistical grid squares on an hourly interval following the dasymetric interpolation method (Järv et al., 2017). Each grid square in the final data set contains information about the relative share of estimated population across the whole region. Grid squares with their centroid in the green space polygons were considered in the final analysis.

Public participatory GIS (PPGIS) data (Article IV) used in this work came from two PPGIS surveys conducted by the City of Helsinki. We downloaded the data from the Helsinki Region Infoshare open data portal[8]. The Helsinki 2050 survey was conducted in 2013 to support the preparations of the master plan for the municipality (Kahila-Tani et al., 2016). The National Urban Park Survey was conducted in 2017 to support the planning of a national urban park in Helsinki[9]. From both data sets, we used questions related to green space use and preferences, and further selected those markers that intersected the green space polygons.

National Park visitor statistics from Finland and South Africa were compared to results from social media in Articles II and III. Parks & Wildlife Finland (*Luontopalvelut, Metsähallitus*) provided visitor surveys and visitor counter information from the Pallas-Yllästunturi National Park used in Article II. In Finland, on-site visitor surveys are conducted every ~5 years in each park. Official visitor numbers are estimated using on-site counters (Kajala et al., 2007). Entrance to the parks is free, and visitors do not need to register. South African National Parks (SANParks) provided the visitor information from Kruger National Park used in Article III. In South Africa, visitors enter fenced parks through gates, and need to sign up (and pay) when entering the park.

3.3. SPATIAL AND TEMPORAL ANALYSIS

This work includes place-based and person-based approaches for spatial and temporal analysis (see Article I for full description). These approaches require direct or indirect information about the location and time related to the post. In this study, all the social media posts analysed contained a geotag added by the user and an automatically generated timestamp.

Articles I, II and III included place-based analysis of observations aggregated to national parks. Article II also focused on national park sub-regions, and Article IV focused on statistical grid squares and green space polygons in an urban area. Aggregated data sets included count of photos, count of users and count of photo-user-days per unit area. A photo-user-day counts each user only once per day they have posted content (Tenkanen et al., 2017; Wood et al., 2013). Counts of unique users or photo-user-days is often a better measure of visitor rates instead of number of posts, as one user might contribute hundreds or thousands of photos in one day.

We inspected the social media data for evident outliers, such as bots (auto-

[8] https://hri.fi/en_gb/

[9] https://maptionnaire.com/blog-list/online-public-participation-questionnaire

matically generated content) or users who would have posted a significant number of posts compared to other users, perhaps for marketing purposes. We also manually inspected the location precision of posts inside the national park (Article II) and in urban green spaces (Article IV) to exclude posts with an imprecise geotag. For example, posts tagged to "Pallas-Yllästunturi National Park" were excluded from the within-park analysis.

Articles I, II and III included person-based analysis of time and location to estimate the origins of national park visitors. Article I introduced the general approach, Article II used a simple approach commonly used in related literature for detecting visitors' origins based on the maximum number of photos per country (see eg. Hawelka et al., 2014).

Article III applied various spatial and temporal approaches for detecting the place of residence of national park visitors with the aim of suggesting the most suitable method(s) for this purpose. Spatial approaches included simple aggregation to countries and regions, spatial clustering (DBSCAN-method Ester et al., 1996; implementation following Boeing, 2018), and four centrographic measures. The centrographic measures included the calculation of mean centres (average of x and y coordinates), median centres (minimizes distance to all points), as well as the centroids of standard deviational ellipses and circles. Temporal approaches focused on the duration of stay in one place based on different thresholds. We calculated the maximum number of unique days (*max days*), weeks (*max weeks*) and months (*max months*) in one country as well as the longest distance between the first and last post (*max timedelta*) in one country. These approaches have been applied in the related literature for detecting home locations on different scales (for example, Bojic et al., 2016; Hawelka et al., 2014; Li et al., 2018), but rarely validated.

We evaluated these methods using two approaches in Article III: 1) comparing the detected countries of residence to an expert assessment at an individual level, and 2) comparing the detected countries of residence to official visitor statistics at a country level.

Two experts (myself and a research assistant) manually estimated the country of origin for a 33% sample of all users who had visited Kruger National Park in 2014 (n=430) to establish a ground truth of the users' origins. Those users whose origins the experts agreed on were included in the final ground truth (n=375). We used F1 scores (see e.g. Ajao et al., 2015 for a related application) as implemented in the Scikit-learn Python-package (Pedregosa et al., 2011) to evaluate each method against the expert assessment. F1 score is calculated as the harmonic mean of precision (proportion of correctly detected origins of users from a given country) and recall (proportion of users from a given country in the ground truth detected correctly by the method) for each label (country). In other words, precision takes into consideration the number of false positives detected by the approaches for each country, and recall considers false negatives. Finally, we considered the macro-average of F1 scores, which is calculated as the unweighted mean of F1 score, precision and recall and for each country. The macro-

average corrects the estimate for the imbalanced distribution of countries of origin in the data, where *domestic* visitors from South Africa were the biggest visitor group (n=172 in the expert assessment subset).

We compared the detected countries of residence to the official visitor statistics using spearman rank-order correlation, i.e. we compared the rank of countries detected by each approach to the rank of countries based on the number of official visitors.

For Article IV we used the Jaccard index to compare the spatial overlap between the spatial distributions of different data sets following the approach presented by Lehtomäki et al. (2015). The Jaccard index calculation divides the intersection of two sets with their union. Value 1 on the Jaccard index refers to complete overlap between the two sets, and value 0 means no overlap.

3.4. SOCIAL MEDIA CONTENT ANALYSIS

The aim of the content analysis of social media data was at enriching the data with information about observations, activities and preferences of the visitors to national parks and green spaces. Furthermore, content analysis helped to confirm that the content of the social media posts was relevant for studying human-nature interactions.

Article I describes and exemplifies analysis methods for visual and textual context analysis of social media data and sets the methodological direction for future work. Other articles in this thesis applied manual content classification following the approach presented in Hausmann et al. (2018). Detailed manual analysis was feasible due to the focused study areas with relatively small extracts of data. In Article II, we manually labelled photograph content (Instagram) from Pallas-Yllästunturi national park, and in Article IV we classified geotagged photographs (Flickr and Instagram) from urban green spaces.

Content categories in Article II were adapted from the national park visitor surveys conducted by Parks and Wildlife Finland in order to allow the two data sources to be compared. The classification also included categories related to landscape and different seasons. Content categories in Article IV were more general with the main focus of detecting if the content was overall relevant to urban green spaces, and further detecting what activities appeared in the different data sources.

In Article II, we compared the frequency distribution of activities indentified from social media content with activities reported in the visitor survey using Pearson's Chi-square test. In Article IV, we plotted venn-diagrams to compare the proportion of different content categories in Instagram and Flickr visually.

Article IV also included language detection of the text content implemented following the approach presented in Hiippala et al. (2019). Language was detected for each sentence from each user using the FastText algorithm (Bojanowski et al., 2017). We excluded sentences less than seven characters long, as well as results with low language detection confidence. We further detected the primary language for users who had posted in multiple languages based on the count of sentences in each language, excluding English.

4. RESULTS AND DISCUSSION

4.1. DIVERSE INFORMATION ABOUT HUMAN-NATURE INTERACTIONS

Social media and other user-generated data provide diverse information about human-nature interactions. This work included a combination of elements from these data in order to answer questions about where, when and how people use and value national parks and urban green spaces, and to understand better who are represented by these data. User-generated data contain information about people; about the human aspects of social-ecological systems as illustrated in Figure 2. They also contain information about landscapes, ecosystems and species (see Article I for literature and examples) as observed and experienced by people. There are information gaps when it comes to human behaviour (Bennett et al., 2017), and human pressures on the environment (Venter et al., 2016) in the Anthropocene Era and this work suggests that several elements of user-generated data (geotags, timestamps, content, user profile) may help fill some of these gaps as long as the limitations of these data are taken into account.

The literature review in Article I highlighted that user-generated data provide multiple perspectives for understanding social-ecological systems and human-environment relationships ranging from nature-based tourism to species information and public attitudes. Three main categories of research focus emerged from the review: 1) humans in nature, 2) biodiversity monitoring, and 3) online discussions.

The focus of most of the articles reviewed was on spatial and spatio-temporal aspects of human activities in nature based on geotags and timestamps. Analysis of human activities based on images, video and textual content were less prominent in the articles. Overall, the review in Article I suggests that image content is still under-utilized when analysing location-based activities of people in nature. There is a lot of potential in applying image content analysis in conservation science to complement social impact assessments (Sherren, Parkins, et al., 2017), for example.

The reviewed studies used social media for biodiversity monitoring with the aim of gathering location-based observations of animal species from a group of enthusiasts – i.e. similar to citizen science approaches (See et al., 2016). Biodiversity information can be extracted from existing data from several social media platforms, or asking for active contributions from users, such as using a specific hashtag or in a focused online group. In a way, some citizen science applications, such as iNaturalist also match the definition of a social media platform.

Studies related to online discussions reviewed in Article I aimed at understanding the online attention on species (Jarić et al., 2016), or to understand online reactions to a controversial topic such as trophy hunting (Macdonald et al., 2016), among other examples. While the rest of this thesis focuses mainly on location-

based data, aspatial studies that focus on online discussions, such as those related to conservation culturomics (Ladle et al., 2016) also offer valuable insights into understanding the complex dynamics of human-nature interactions in the information era.

Social media and other user-generated data sets are often voluminous and diverse, matching the definition of big data (Kitchin, 2014a). The strength of using these data does not rely on one aspect, such as the geotags or timestamps, but in combined understanding of where, when, what, how, and who. At the same time, these data do not represent the whole population, and might contain data generated by bots, trolls and technical errors that increase the volume of the data with noise. Despite evident limitations, user-generated data might often be the best available information about human activities and preferences in nature. Even imperfect measures, if understood as such, can support sustainable planning decision-making (Daily et al., 1997).

Overall, this work highlights the potential of combining information from several elements of social media data and from different data sources in order to gain the most diverse and versatile understanding about human-nature interactions from different perspectives.

4.2. WHERE? – SPATIAL HOTSPOTS AND FLOWS

User-generated data contains information about visitor flows and hotspots in national parks and green spaces on different scales. Geotags and other references to location allow linking digital data to physical locations in a new way, combining the digital space to the physical space as illustrated in Figure 3. Spatially accurate information about human-nature interactions so far have been rarely available over large areas, and in this thesis I investigated how user-generated geographic information could fill in some of the data gaps. This thesis presents methods for analysing spatial patterns in user-generated data, and compares different data sources in order to learn more about their applicability to studying human presence and movements in nature.

Data comparisons between social media data and national park visitor statistics in this work (Article II) and related studies (Tenkanen et al., 2017) show that social media reflects visiting patterns in popular national parks such as the Pallas-Yllästunturi National Park in Finland and frequently visited sub-regions. At the same time, data from remote and less popular national parks and sub-regions often do not reflect the popularity-rank of the destination.

Sporadic data and the lack of digital footprints might be caused by poor mobile network connections or long travelled distances and a drained battery, in addition to low visitor numbers as such in a particular destination (Article II; Tenkanen et al., 2017). Social media platforms also allow users to post and tag content afterwards, which allows geotagging data to areas without network coverage, for example. However, posting and tagging photos afterwards might lead to coarse or imprecise geotagging.

In Article IV, the hotspots of different user-generated data sets from urban

green spaces had relatively distinct patterns and low overlap as measured by the Jaccard index. This might indicate that one single data source does not capture all popular and valuable sites.

Analysis in Articles II and IV also investigated if data geotagged to national parks and green spaces actually represents those places, i.e. in what cases the digital space was not in fact linked to the physical space even if the geotag indicated so. In both cases – Pallas-Yllästunturi National Park and urban green spaces in Helsinki, Finland – the proportion of irrelevant data was relatively low. This might indicate that geotagged data from national parks and green spaces is often relevant for studying the actual use of these areas. However, this result might be specific to national parks and green spaces particularly in Finland, while the amount of noise might be greater in larger cities and internationally-renowned destinations. Users might also intentionally falsify geotags – a practice referred to as location spoofing (Zhao & Sui, 2017). However, I did not observe instances of fake geotags in the manually assessed samples.

Enriching the spatial information with other elements is the biggest strength of social media data, i.e. asking the questions when, what and why at the same time when looking at spatial patterns. Studying visitor flows is an example of combining the questions of where, when and who. Social media data allow studying visitor movements and flows to and inside national parks if location-based data have been collected from a longer period of time for each user (Article I, Article II, Article III). Similar person-based analysis could be possible using mobile phone data, if available.

4.3. WHEN? – TEMPORAL TRENDS AND DYNAMICS

Continuously generated data sets provide unique information about temporal trends of human activities in nature. Even in areas with an established monitoring system, user-generated data may offer complementary information over a more frequent time interval in comparison with official visitor statistics, as Article II suggests. In areas that are not regularly monitored, user-generated data might be the only source of temporal information about human activities.

User-generated data sets applied in this work have relatively good temporal coverage in comparison with traditional data sources such as surveys (Article II). Passively contributed data sets allow continuous near-real time data, while data generated through active participation (such as PPGIS) are more limited in temporal extent and granularity (Article IV). Social media data reflect activities after office hours, while mobile phone data and GPS data also cover commuting in the morning (Article IV).

Social media data reflect visits over time captured in official visitor counts in popular parks such as the Pallas-Yllästunturi National Park (Article II; Tenkanen et al., 2017). Results from urban green spaces highlight that social media data capture mostly leisure time activities (Article IV), while mobile phone data and sports tracking data also capture commuting patterns.

Temporal aggregation reveals periodical trends even if data would be sporadic (Article I, Article II, Article IV). For example, combining observations over a longer period of each weekday or hour of the day, shows the general trend of sharing data on social media from green spaces during leisure time after work and on weekends (Article IV). At the same time, this might mean that social media data are not fit for near-real time monitoring of visitor flows, and only reveal the general temporal patterns when aggregating over longer periods of time.

Temporal questions are often coupled with spatial questions as was done in all articles in this thesis, but temporal trends can also be observed without specific reference to location or area. The combination of when and what is common when studying nature-related questions from digital data. For example, temporal analysis of specific topics can reveal public interest and reactions to topical issues related to biodiversity conservation (Box 4 in Article I; Fink et al., 2020).

User-generated data accumulate continuously as users post content on social media and use their mobile phones and GPS tracking applications. However, these continuous data streams are not automatically available for research. Social media APIs might change their functionality and access to data (Article I; Freelon, 2018), mobile phone companies might give out data products for limited time periods, accumulation of sports application data is dependent on user's activity similarly to social media, and PPGIS studies are often limited in time for practical reasons (Ives et al., 2017) as they require active effort from researchers and the study participants.

4.4. WHAT AND WHY? – ACTIVITIES AND PREFERENCES

Content analysis of social media texts and images allows understanding activities and preferences in national parks and green spaces from a new perspective. Studies using social media for understanding human activities in nature have previously focused mostly on analysing human presence and absence through geotags (Article I, Table 1). Earlier studies have recognized content analysis of social media photographs as a rapid way of analysing the use of natural areas (Richards & Friess, 2015), but more information is still lacking about the applicability of different data sources in different geographic contexts (Article I; Ghermandi and Sinclair, 2019; Teles da Mota and Pickering, 2020). This thesis will contribute to a better understanding how social media content reflects surveyed activities (Article II), and will highlight differences of available content on different platforms (Article IV).

Visual content analysis in this thesis was focused on identifying physical activities in social media photographs in addition to characterizing the visual information content in general, and detecting what proportion of the content was relevant to the use of national parks and green spaces. This thesis shows that content analysis of social media helps to filter out irrelevant data (Articles I, II, IV), and to enrich the spatial and temporal dimensions with information about what peo-

ple are doing and valuing in green spaces (Articles I, II, IV).

Image content analysis in Articles II and IV showed that most of the visual content shared from parks and green spaces was relevant for studying human activities in nature. Non-relevant data included advertisement and internet memes that had been geotagged to the parks and green spaces.

In the Finnish context where wildlife is relatively difficult to spot and photograph, most of the content from national parks and green spaces portrayed people, landscapes and generic nature photos (Articles II and IV). In a related study from Kruger National Park in South Africa, large-bodied mammals, such as elephants (*Loxodonta africana*) and lions (*Panthera leo*), were more frequently portrayed on social media in comparison to landscapes and human activities (Hausmann et al., 2018).

Activities detected from social media content reflected surveyed activities in Pallas-Yllästunturi National Park, Finland (Article II); hiking and cross-country skiing were the most popular activities both in the survey and in social media data. This supports previous work conducted in Kruger National Park regarding visitors' preferences for nature-based experiences (Hausmann et al., 2018). Social media photographs from Pallas-Yllästunturi National Park also contained activities not captured in the survey, such as winter biking – an activity which has gained popularity only in the recent years in the study area. In practice, social media content could serve as an indicator of emerging activities, and could provide insights when designing new surveys, for example.

Social media content is most clearly linked to the questions about what and why, as illustrated in Figure 3, but images and text can also contain relevant insights about the other questions. For example, presence or absence of snow in geotagged photographs reflects seasonal patterns in the Pallas-Yllästunturi national park (Article II), and analysing the language of text content can give hints about different user groups (Article IV).

Automated content analysis methods allow high volumes of data to be analysed (Lee et al., 2019; Richards & Tunçer, 2018; Väisänen et al., in press). In this thesis, Articles II and IV applied manual content analysis relying on expert assessment and a pre-defined classification scheme. The number of classified photographs in these studies was manageable as the geographic extent and timeframe of these studies was limited. Article I provided an introduction to and examples of automated analysis of visual and textual content setting the direction for future work. Automated methods could be especially useful for analysing continuous flows of data.

In summary, rich content in social media provides an opportunity to gain deeper understanding of what people are doing in parks and green spaces, and perhaps also why they have chosen to visit these areas. This spatially-explicit information about activities and preferences can be useful, in planning and managing the areas – information that has been previously difficult to gather (Beeco & Brown, 2013). While expert assessment of image

content (manual analysis) allows for in-depth interpretation of activities and preferences, automated methods provide opportunities to analyse large quantities of content that would otherwise be laborious or impossible to go through systematically.

4.5. WHO? – ORIGINS AND OTHER CHARACTERISTICS

Understanding who the users are is important for assessing data quality and for understanding who the visitors to national parks and green spaces are. This thesis provides new information about the visitors who share their data from national parks (Articles II, III) and green spaces (Article IV), and presents approaches for enriching user-generated data - social media in particular - with additional information about who the users are.

Social media and other user-generated data sets provide continuous information about where and when people visit nature, and even about what they are doing while visiting parks, as discussed in the previous sections, but linking this information to relevant demographic information is often a challenge, as highlighted in Article I. Even basic information about the user base is often not readily available via social media APIs. Lack of information about who produced the data is one of the key epistemological limitations of user-generated data (Ruths & Pfeffer, 2014), and GIS analysis in general (Ash et al., 2018). It is thus important to ask whose nature-based experiences are we eventually studying based on these data.

The who question can be approached from multiple perspectives including the age, gender, origins, and even preferences of people. Article I (Table 3) identified approaches for deriving information about the users based on different elements of social media data. Previous studies have used social media user profile information to derive demographic data such as age, nationality and occupation (Longley et al., 2015; Sloan et al., 2015). For example, distinguishing between national and international visitors can be an important analysis step for understanding national park visitors (Sessions et al., 2016).

National park survey reported in Article II confirms that not everyone shares their national park experience on social media. According to the survey results, the average age of social media users was lower in comparison to the average age of national park visitors in general. Findings in Article III further highlight that social media data captures both local and international visitors in a popular national park but omits completely visitors from many African and Asian countries, where other social media platforms are more popular. Furthermore, it should be acknowledged that data sets used in this thesis omit social media users who do not geotag their posts (Huang & Carley, 2019). Other related work highlights the potential gender bias in social media data; in a recent study we found that the majority (75%) of Flickr data shared from Finnish National parks was generated by Finnish men (Väisänen et al., in press).

Detecting the origins of people at an aggregate and privacy-preserving level, such as countries or cities, can help

characterize the users further. Regarding the country of origin, Article II utilized a simple approach for detecting the potential home location of social media users based on maximum number of posts, revealing similar trends as the survey regarding both national and international visitors. Also, visitor group sizes observed from social media were similar with the group sizes reported in the official visitor statistics.

Article III further explored different methods for detecting the home locations of national park visitors based on their posting history and suggests that the combination of spatial and temporal information (the maximum number of unique months in a country) yields best results for detecting the user's country of origin. Social media usage patterns, namely the number of countries visited and the temporal duration of posting history, had an impact on the reliability of the results.

The level of detail about the users varies between user-generated data sources, as highlighted in Article IV. For example, sports application data might contain metadata about the user base, at least on an aggregate level. Sports tracking data often represent active men over other groups, which should be acknowledged when using these data (Article IV; Oksanen et al., 2015). Similarly, PPGIS data used in Article IV contained aggregated metadata about the respondents. Mobile phone operators possess information about the account holders and this information has been used in research (see for example Järv et al., 2015), but this information was not available for analysis in Article IV. In the future, aggregated data products could perhaps contain some of this information without compromising the privacy of individual users.

Overall, the level of user details varies between data sources and self-selection bias is an inherent property of user-generated data sets. Population biases are often neglected in studies based on social media data, due to the lack of readily-available information. Further analysis of social media user profiles and posting history can help understanding different user groups in the data, and integrating different data sources into the analysis allows different groups of people to be taken into account, as highlighted in Article IV. For example, combining user-generated data, such as social media posts or mobile phone records from recruited participants to further survey questions would allow for deeper understanding of the sample.

4.6. SCALE AND CONTEXT MATTER

Spatial and temporal context matter when selecting the data source and analysis methods. Spatial accuracy and extent of user-generated data varies between platforms and data sources affecting the selection of applicable data as highlighted in Article IV. Also, the time range of the study affects the selection of the appropriate data source, and seasonal phenomena and events can influence the accumulation of data.

The advantage of social media data in general is that they are available over large areas across national borders and even globally. Social media data also ac-

cumulate continuously as users post new content. This work included social media data from Flickr, Twitter and Instagram - platforms that are popular in the Global North and applicable to protected areas popular with visitors from these regions (Tenkanen et al., 2017). For example, in Asia, other platforms such as Sina Weibo would be a more optimal data source for capturing visitors from that region, as highlighted in Article I.

The varying scale of points-of-interest (POIs) in social media data should be taken into account in subsequent analysis (Hochmair et al., 2018). For example, the data set used in Article II contained posts geotagged to a POI called "Pallas-Yllästunturi National Park", and this POI was located outside the park borders. These points were included in the park-level analysis but discarded in the sub-region analysis. Latitude and longitude coordinates in social media geotags might give the impression of high spatial accuracy, but analysis is often more meaningful at a coarser aggregate level.

Articles II and III included the global posting history of users who visited the national parks. Even if these are global data sets, they are related to the local context of the national parks and capturing typical visitor groups for those destinations. Results from Article III revealed the evident spatial bias in a global social media data set reflecting digital divides at least when it comes to the use of Instagram. Even the best method in Article III did not capture visitors from most Asian or sub-Saharan African countries indicating that national park visitors from these countries are not represented in the social media platform that was used.

For studying nature-based tourism, social media continue to be an interesting source of data due to the global scope and diverse content. Instagram would be the most versatile data source in terms of activities and users, but access to data is limited. While Flickr data continue to be accessible via the API, and it is a popular data source among nature enthusiasts (Di Minin et al., 2015), its user base is narrow in comparison to Instagram. Twitter is another data source for a destination with visitors from western countries, but it is less optimal than Flickr and Instagram for studying national park visitors (Tenkanen et al., 2017) and urban green spaces (Article IV).

On local to regional scales, GPS tracks and mobile network data may provide more meaningful information about spatial patterns if available, and PPGIS studies and other surveys offer tools for gaining insights about activities and preferences in finer detail.

Meaningful scale of analysis for mobile phone data is determined by the antenna network, and potential data aggregation by the operator. The accuracy of mobile phone data can be enhanced through dasymetric interpolation (Järv et al., 2017) but in principle, mobile phone data are not the optimal source for studying small area units such as smaller urban parks (Article IV). Also, identifying national park visits from mobile network data can be challenging, for example, if there is a big road or urban area right next to the park. Mobile phone data is often available on a national scale, sometimes

also containing information about international visits (Ahas et al., 2008). In this work, the acquired mobile phone data was specific to the Helsinki region. The time-range of the mobile phone data used in Article IV was not optimal, capturing only winter months.

GPS data from sports tracking applications offer high spatial accuracy, but often to a limited spatial extent in comparison with social media data or mobile phone data. For example, if interested in studying specific route choices inside national parks or green area networks, then GPS based data are often the best option.

Spatial accuracy of PPGIS data depends on the technique used, and local knowledge of the respondent. On the other hand, other types of data, such as GPS tracks can also be contributed through a participatory campaign (Korpilo et al., 2017), in which case the spatial accuracy would be higher, as suggested in Article IV.

Season and events evidently impact observed patterns in different data sources. While social media data reflect temporal fluctuations of visitors in popular national parks (Article II; Tenkanen et al., 2017) social and natural events might lead to an increase or decrease in social media activity. One user might post a significant number of photos from a single sports event (Väisänen et al. under review) and natural events such as flower blooming might show up in increased social media activity (Tenkanen et al., 2017). Overall, selecting the appropriate data sources is often a compromise.

4.7. MAIN CHALLENGES

Main challenges related to using user-generated geographic information include data quality, limited access to data and privacy issues. Uncertainty about data reliability, changes in data access and concerns over protecting personal information hinder the use of these data in research and practice (boyd & Crawford, 2012; Ruths & Pfeffer, 2014; Zook et al., 2017).

User-generated data sets are often heterogenous and of varying quality (Senaratne et al., 2017). As discussed in Article IV, the data sources used in this thesis differ in positional and temporal accuracy, thematic content and information about the users. When possible, gaps in these aspects can be complemented with other available data. For example, different social media platforms can be combined for an improved understanding of human activities in nature (Article IV; Tenkanen et al., 2017). Thorough data exploration is an important part in a data analysis workflow (Article I) that helps identify shortcomings in the data and select appropriate analysis methods. For example, if the positional accuracy does not allow fine-grained analysis inside a national park, meaningful insights can still be acquired by aggregating the data. Overall, the relative shortcomings of data sources are pointed out in all articles in this thesis.

Limited access to data affects the repeatability of the analysis and the execution of longitudinal research projects (Lomborg and Bechmann, 2014; Article I). To a large extent, this work was based on Instagram data collected

in spring 2016 and the applied data collection approach is no longer feasible. Future studies in which Instagram data would be the best data source might need to rely on alternative means of access, such as web scraping or purchasing an access or a data product as discussed in Article I. However, web scraping is technically more challenging in comparison with data access through APIs and remains legally and ethically in the grey area (Freelon, 2018).

Acquiring other types of user-generated data can also be challenging. For example, mobile phone data from national parks was unfortunately not available for this study, and negotiations for acquiring the data for Article IV were lengthy. Data collection from active participants is possible through designing and implementing citizen science campaigns and participatory surveys. In Finland, PPGIS surveys conducted by the City of Helsinki have been released as open data (as those used in Article IV), and citizens' biodiversity observations from iNaturalist feed into the global GBIF[1] biodiversity database. However, citizen science campaigns are often geared towards the natural system, and active data collection approaches about the human system come with their own challenges, such as low response rates (Brown, 2017) and limited duration and extent for practical reasons (Ives et al., 2017), and thus do not fully complement the information available continuously from big crowds through other types of data.

Collecting, storing, analysing and publishing results based on user-generated data sources require special consideration of privacy issues. Even if using publicly shared content, or purchased data products, researchers still need use the data responsibly, protect privacy, and minimize potential harm (Zook et al., 2017; Di Minin et al. in press). The European Union (EU) general data protection regulation GDPR (2016/679) that came into effect in 2018 aims to protect an individual's privacy rights and sets restrictions on the processing of personal data in the EU. The GDPR also applies to user-generated information such as social media data that contain any personal data, i.e. any information that can be linked to an identifiable person (Di Minin et al. in press). According to the GDPR, personal data must be processed fairly for specified purposes on a legal basis. Processing of personal data can be considered lawful if the data subject has given consent, or if the task is carried out in the public interest (such as research), among other potential legal bases (Article 6 in the GDPR[2]). In the context of user-generated data, informed consent is often feasible to obtain when collecting actively contributed data through PPGIS (for example, Brown et al., 2014b). However, with mobile big data from mobile phone operators or social media platforms, acquiring consent from study participants is often not feasible (Ayers et al., 2018; Di Minin et al. in press).

Data minimization (collecting and storing only a minimum amount of data

1 https://www.gbif.org/

2 http://www.privacy-regulation.eu/en/article-6-lawfulness-of-processing-GDPR.htm

needed) and pseudonymization (removing direct personal identifiers) can help protect users' privacy while allowing for responsible use of social media data in research (Di Minin et al., in press). Pseudonymization/de-identification of individuals is also possible through aggregating the data into larger groups or area units following the area-based approach presented in Article I. From a legal perspective, if data are properly de-identified (for example through aggregation) they are no longer considered to be personal data. However, simply removing personal identifiers might leave the data re-identifiable. For example, even a single quote can be enough to re-identify pseudonymized social media users (Ayers et al., 2018). Protecting personal data is important at different stages of the data analysis workflow and, if possible, should be done at the data collection stage (Di Minin et al. in press). Results of a person-based analysis can also be further aggregated into area units in order to preserve the privacy of individual users, even if analysis needs to be done at an individual level as was done in Article III.

Terminology plays a key role when considering privacy issues. It is important to distinguish the level of participation by the data producer (Haklay, 2013; See et al., 2016) in order to differentiate between actively and passively contributed data. If social media data are referred to as *volunteered* data (Goodchild, 2007) this might give the impression that users have directly given their consent for use of their data. I would argue that in addition to data sources like the OpenStreetMap, the concept of volunteered geographic information better suits location-based crowdsourcing campaigns such as the iNaturalist application through which volunteers contribute biodiversity observations, whereas examples of people volunteering data about themselves are scarce. The idea of "citizens as censors" by Goodchild (2007) still holds with currently available sources of user-generated geographic information and "crowdsensing" is also recognized as an emerging trend in remote sensing literature (Toth & Jóźków, 2016). Overall, careful consideration is needed if the data have been actively volunteered or not and if the data are about people themselves, the environment or both (Figure 2) when talking about privacy concerns.

4.8. COMPLEMENTARY INFORMATION FROM DIFFERENT DATA SOURCES

Different sources of user-generated geographic information can complement each other's gaps. All articles in this thesis highlight the potential and challenges of different social media platforms, and Article IV extended the focus to sports application data, mobile phone data and PPGIS data. Different data sources capture different user bases from different spatial and temporal extents, as highlighted in Article IV. Social media captures *being* in the park, while sports tracking data and mobile phone data reflect *moving* through the park during the daily commute. PPGIS data reflects *valuing* specific places. Relying only on one source of data evidently provides a limited view of the use patterns and values related to human activities in nature.

This thesis highlights that social media platforms such as Instagram and Flickr, contain information particularly about people's leisure time activities and experiences, which makes them an appealing data source for studying human activities in nature. The proportion of information about activities and landscapes was similar in both data sets (Article IV), but their user bases and usage patterns differed. Instagram is a popular social media platform that also contains relevant content from national parks and green spaces as highlighted in this thesis (Article II; Article IV) and other relevant work (Hausmann et al., 2018). However, changes in the Instagram API have restricted access to data limiting their use in research. Flickr is popular among nature enthusiasts (Di Minin et al., 2015), and thus an optimal source of data for looking at biodiversity-related content. At the same time, the user base of Flickr is narrow in comparison to other platforms. Twitter is widely used in geographical research due to its popularity and continuous access to data through the platform API. However, Twitter is the least useful platform for studying human nature interactions among the data sets included in this thesis according to the findings from Finland and South Africa (Article IV; Tenkanen et al., 2017). At the same time, Twitter allows secondary access to geotagged Instagram posts, as users tend to share the same content across different platforms (Heikinheimo et al., 2018; Juhász & Hochmair, 2018).

Sports tracking data and mobile phone data would best fit the purpose of answering detailed questions about the spatial and temporal patterns of human activities in nature. For example: Where and when are people moving in national parks and green spaces?

The COVID-19 coronavirus pandemic has motivated technology companies to release mobility data (data about where, when and what) to support the global health response during 2020. For example, Google[3] and Apple[4] have released aggregated and anonymized information about the temporal mobility patterns in different land uses and different travel modes. Mobile phone operators have also released new data products for research and practice during spring 2020. The Google mobility index for parks is particularly interesting for understanding the recreational demand of national parks and green spaces, and the topic is starting to be further investigated at the time of writing this synopsis[5]. In contrast to increased use of green spaces in countries like Finland and Sweden, The Google mobility index for parks also indicates a decline in visitor rates in the Global South. The lack of income from tourism might lead to complex issues and land use pressures in many regions[6]. Combining information from aggregated and anonymized mobile big data sources might be one way forward to capture the rapidly changing patterns of human activities in nature globally. However, the long-term avail-

3 https://www.google.com/covid19/mobility/

4 https://www.apple.com/covid19/mobility

5 pre-prints https://osf.io/3wx5a/ and https://osf.io/preprints/socarxiv/97qa4/

6 https://blogs.helsinki.fi/digital-geography/2020/06/01/covid-madagascar-protectedareas/

ability of these data is unclear. Main limitations for using these new data sources are mostly the same as highlighted in Article IV for social media data, mobile phone data and sports tracking data.

Overall, different sources of user-generated geographic information complement each other in answering questions related to where, when, what, why and who (Article IV). For example, sports application data mobile phone data might provide the best information about the temporal use patterns of a specific area, while social media data provide hints about what people are doing in that area. PPGIS could further provide in-depth understanding of the motivations of different visitor groups in a focus area. I see user-generated geographic information such as social media data best used in so called pre-emptive mapping that can then inform about the need for more in-depth data collection efforts such as surveys. Overall, user-generated data complement, but do not replace authoritative and traditional data sources.

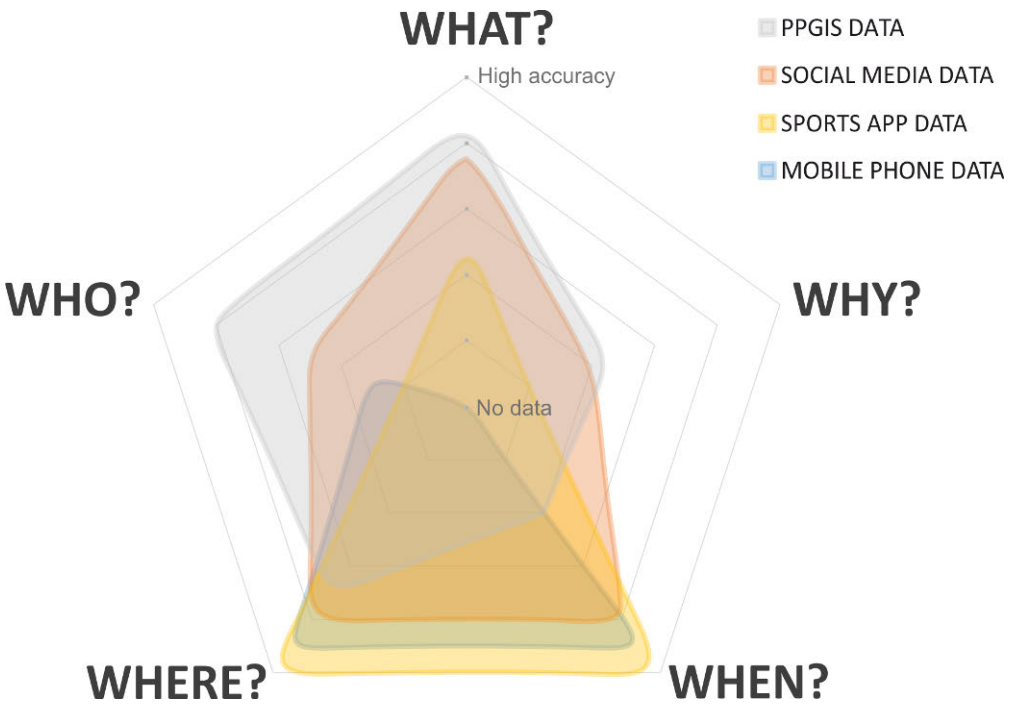

Figure 7. An illustration of the potential applicability of different sources of user-generated geographic information for answering the questions where, when, what, why and who. Article IV.

5. CONCLUDING REMARKS

Overall, this thesis highlights the potential of using user-generated geographic information for understanding different perspectives of human activities in nature, while pointing out the evident limitations of different data sources and analysis approaches. User-generated data offer unique information that can be used to incorporate the value of nature in decision making, as long as the limitations of these data are acknowledged and taken into account.

Further studies using social media and other user-generated data in different contexts should pay attention to the strengths and limitations of the used data and if possible, complement some of the gaps using other data sources. A mixture of active and passive participation for acquiring information about the use of green spaces is one potential way towards integrating perspectives. For example, insights from social media could inform a PPGIS survey design, or researchers and practitioners could use existing online platforms to request data and observations from crowds. Furthermore, researchers should pay increasing attention to privacy and ethics. In practice, working with de-identified and aggregated data sets minimizes potential harm such as re-identification of data subjects. Developing such data packages further is one potential way forward for the continuous and systematic use of user-generated geographic information in research and practice.

Data comparisons from national parks with a systematic visitor-monitoring scheme provide insights into using these data in other environments including non-monitored natural areas and heterogenous urban regions. In Finland, national park visitor monitoring is conducted in a systematic and reliable way, and user-generated data offer mostly complementary information about visitors, their activities and movements. In contrast, other non-monitored areas in Finland and beyond can potentially gain unique data about visitors and visitation patterns through user-generated data sets. Platform, user base and available content might vary according to the region, scale and time period, and that is why thorough consideration of suitable data sources and analysis methods is needed in a new context.

Applying user-generated geographic information in practice is still under-developed, for many reasons. Access to data, and questions about data quality, ownership and ethical use are under continuous debate and change. However, even imperfect measures of nature's contributions to people are better than ignoring them completely, as pointed out by Daily et al. (1997). The same applies to incorporating understanding about mobility patterns and place-based experiences and preferences into decision-making in general. Despite challenges, there is a lot of potential to analyse user-generated data sets for the benefit of people and the environment.

REFERENCES

Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, *29*(3), 469–486. https://doi.org/10.1016/j.tourman.2007.05.014

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, *17*(1), 3–27. https://doi.org/10.1080/10630731003597306

Ajao, O., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science*, *41*(6), 855–864. https://doi.org/10.1177/0165551515602847

Arts, K., van der Wal, R., & Adams, W. M. (2015). Digital technology and the conservation of nature. *Ambio*, *44*, 661–673. https://doi.org/10.1007/s13280-015-0705-1

Ash, J., Kitchin, R., & Leszczynski, A. (2018). Digital turn, digital geographies? *Progress in Human Geography*, *42*(1), 25–43. https://doi.org/10.1177/0309132516664800

Ayers, J. W., Caputi, T. L., Nebeker, C., & Dredze, M. (2018). Don't quote me: reverse identification of research participants in social media studies. *Npj Digital Medicine*, *1*(1), 30. https://doi.org/10.1038/s41746-018-0036-2

Balmford, A., Green, J. M. H., Anderson, M., Beresford, J., Huang, C., Naidoo, R., Walpole, M., & Manica, A. (2015). Walk on the Wild Side: Estimating the Global Magnitude of Visits to Protected Areas. *PLoS Biology*, *13*(2), e1002074. https://doi.org/10.1371/journal.pbio.1002074

Beeco, J. A., & Brown, G. (2013). Integrating space, spatial tools, and spatial analysis into the human dimensions of parks and outdoor recreation. *Applied Geography*, *38*(1), 76–85. https://doi.org/10.1016/j.apgeog.2012.11.013

Bennett, N. J., Roth, R., Klain, S. C., Chan, K., Christie, P., Clark, D. A., Cullman, G., Curran, D., Durbin, T. J., Epstein, G., Greenberg, A., Nelson, M. P., Sandlos, J., Stedman, R., Teel, T. L., Thomas, R., Veríssimo, D., & Wyborn, C. (2017). Conservation social science: Understanding and integrating human dimensions to improve conservation. *Biological Conservation*, *205*, 93–108. https://doi.org/10.1016/J.BIOCON.2016.10.006

Bergroth, C. (2019). *Uncovering population dynamics using mobile phone data: the case of Helsinki Metropolitan Area* [Master's thesis, University of Helsinki]. http://urn.fi/URN:NBN:fi:hulib-201905272171

Berkes, F., & Folke, C. (1998). Linking social and ecological systems for resilience and sustainability. *Linking Social and Ecological Systems: Management Practices and Social Mechanisms for Building Resilience*, *1*(4), 4.

Berry, D. M. (2012). Introduction: Understanding the Digital Humanities. In *Understanding Digital Humanities* (pp. 1–20). Palgrave Macmillan UK. https://doi.org/10.1057/9780230371934_1

Boeing, G. (2018). Clustering to Reduce Spatial Data Set Size. *SSRN Electronic Journal*. http://arxiv.org/abs/1803.08101

Bojic, I., Belyi, A., Ratti, C., & Sobolevsky, S. (2016). Scaling of foreign attractiveness for countries and states. *Applied Geography*, *73*, 47–52. https://doi.org/10.1016/j.apgeog.2016.06.006

boyd, danah, & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brown, G. (2017). A Review of Sampling Effects and Response Bias in Internet Participatory Mapping (PPGIS/PGIS/VGI). *Transactions in GIS*, *21*(1), 39–56. https://doi.org/10.1111/tgis.12207

Brown, G., & Kyttä, M. (2014). Key issues and research priorities for public participation GIS (PPGIS): A synthesis based on empirical research. *Applied Geography*, *46*, 122–136. http://www.sciencedirect.com/science/article/pii/S0143622813002531

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. https://doi.org/10.1016/j.landurbplan.2013.09.006

Brown, G., Weber, D., & De Bie, K. (2014). Assessing the value of public lands using public participation GIS (PPGIS) and social landscape metrics. *Applied Geography*, *53*, 77–89. https://doi.org/10.1016/j.apgeog.2014.06.006

Buckley, R. (2009). Parks and Tourism. *PLoS Biology*, *7*(6), e1000143. https://doi.org/10.1371/journal.pbio.1000143

Burkhard, B., Kroll, F., Nedkov, S., & Müller, F. (2012). Mapping ecosystem service supply, demand and budgets. *Ecological Indicators*, *21*, 17–29.

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

Cooper, M. W., Di Minin, E., Hausmann, A., Qin, S., Schwartz, A. J., & Correia, R. A. (2019). Developing a global indicator for Aichi Target 1 by merging online data sources to measure biodiversity awareness and engagement. *Biological Conservation*, *230*, 29–36. https://doi.org/10.1016/j.biocon.2018.12.004

Crampton, J. W. (2009). Cartography: maps 2.0. *Progress in Human Geography*, *33*(1), 91–100. https://doi.org/10.1177/0309132508094074

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, *40*(2), 130–139. https://doi.org/10.1080/15230406.2013.777137

Daily, G., Postel, S., Bawa, K., & Kaufman, L. (1997). Nature's Services: Societal Dependence On Natural Ecosystems. *Bibliovault OAI Repository, the University of Chicago Press*.

de Souza e Silva, A. (2006). From Cyber to Hybrid. *Space and Culture*, *9*(3), 261–278. https://doi.org/10.1177/1206331206289022

Di Minin, E., Fink, C., Hausmann, A., Kremen, J., & Kulkarni, R. (in press). How to address data privacy concerns when using social media data in conservation science. *Conservation Biology*.

Di Minin, E., Fink, C., Tenkanen, H., & Hiippala, T. (2018). Machine learning for tracking illegal wildlife trade on social media. *Nature Ecology & Evolution*, *2*(3), 406–407. https://doi.org/10.1038/s41559-018-0466-x

Di Minin, E., Tenkanen, H., & Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, *3*. https://doi.org/10.3389/fenvs.2015.00063

Díaz, S., Pascual, U., Stenseke, M., Martín-López, B., Watson, R. T., Molnár, Z., Hill, R., Chan, K. M. A., Baste, I. A., Brauman, K. A., Polasky, S., Church, A., Lonsdale, M., Larigauderie, A., Leadley, P. W., van Oudenhoven, A. P. E., van der Plaat, F., Schröter, M., Lavorel, S., … Shirayama, Y. (2018). Assessing nature's contributions to people. *Science*, *359*(6373), 270 LP – 272. https://doi.org/10.1126/science.aap8826

Díaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Agard, J., Arneth, A., Balvanera, P., Brauman, K. A., Butchart, S. H. M., Chan, K. M. A., Lucas, A. G., Ichii, K., Liu, J., Subramanian, S. M., Midgley, G. F., Miloslavich, P., Molnár, Z., Obura, D., Pfaff, A., … Zayas, C. N. (2019). Pervasive human-driven decline of life on Earth points to the need for transformative change. In *Science* (Vol. 366, Issue 6471). American Association for the Advancement of Science. https://doi.org/10.1126/science.aax3100

Donahue, M. L., Keeler, B. L., Wood, S. A., Fisher, D. M., Hamstead, Z. A., & McPhearson, T. (2018). Using social media to understand drivers of urban park visitation in the Twin Cities, MN. *Landscape and Urban Planning*, *175*, 1–10. https://doi.org/10.1016/J.LANDURBPLAN.2018.02.006

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers*, *102*(3), 571–590.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., & others. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, *96*(34), 226–231.

Fink, C., Hausmann, A., & Di Minin, E. (2020). Online sentiment towards iconic species. *Biological Conservation*, *241*, 108289. https://doi.org/10.1016/j.biocon.2019.108289

Fisher, D. M., Wood, S. A., White, E. M., Blahna, D. J., Lange, S., Weinberg, A., Tomco, M., & Lia, E. (2018). Recreational use in dispersed public lands measured using social media data and on-site counts. *Journal of Environmental Management*, *222*, 465–474. https://doi.org/10.1016/J.JENVMAN.2018.05.045

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *0*(0), 1–4. https://doi.org/10.1080/10584609.2018.1477506

Ghermandi, A., & Sinclair, M. (2019). Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change*, *55*, 36–47. https://doi.org/10.1016/J.GLOENVCHA.2019.02.003

Gliozzo, G., Pettorelli, N., & Muki Haklay, M. (2016). Using crowdsourced imagery to detect cultural ecosystem services: A case study in South Wales, UK. *Ecology and Society*, *21*(3), art6. https://doi.org/10.5751/ES-08436-210306

Goodchild, M. F. (1992). Geographical information science3. *International Journal of Geographical Information Systems*, *6*(1), 31–45. https://doi.org/10.1080/02693799208901893

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

Haaland, C., & van den Bosch, C. K. (2015). Challenges and strategies for urban green-space planning in cities undergoing densification: A review. In *Urban Forestry and Urban Greening* (Vol. 14, Issue 4, pp. 760–771). https://doi.org/10.1016/j.ufug.2015.07.009

Hägerstrand, T. (1970). What about people in Regional Science? *Papers of the Regional Science Association*, *24*(1), 6–21. https://doi.org/10.1007/BF01936872

Haines-Young, R., & Potschin, M. (2010). The links between biodiversity, ecosystem services and human well-being. *Ecosystem Ecology: A New Synthesis*, *1*, 110–139.

Haklay, M. (2013). Citizen science and volunteered geographic information: Overview and typology of participation. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* (Vol. 9789400745872, pp. 105–122). Springer Netherlands. https://doi.org/10.1007/978-94-007-4587-2_7

Hamstead, Z. A., Fisher, D., Ilieva, R. T., Wood, S. A., McPhearson, T., & Kremer, P. (2018). Geolocated social media as a rapid indicator of park visitation and equitable park access. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/J.COMPENVURBSYS.2018.01.007

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hausmann, A., Toivonen, T., Fink, C., Heikinheimo, V., Tenkanen, H., Butchart, S. H. M., Brooks, T. M., & Di Minin, E. (2019). Assessing global popularity and threats to Important Bird and Biodiversity Areas using social media data. *Science of The Total Environment*, *683*, 617–623. https://www.sciencedirect.com/science/article/pii/S0048969719323095?via%3Dihub

Hausmann, A., Toivonen, T., Heikinheimo, V., Tenkanen, H., Slotow, R., & Di Minin, E. (2017). Social media reveal that charismatic species are not the main attractor of ecotourists to sub-Saharan protected areas. *Scientific Reports*, *7*(1), 763. https://doi.org/10.1038/s41598-017-00858-6

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., & Di Minin, E. (2018). Social Media Data Can Be Used to Understand Tourists' Preferences for Nature-Based Experiences in Protected Areas. *Conservation Letters*, *11*(1). https://doi.org/10.1111/conl.12343

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072

Heikinheimo, V., Tenkanen, H., Hiippala, T., & Toivonen, T. (2018). Digital Imaginations of National Parks in Different Social Media: A Data Exploration. *Proceedings of PLATIAL'18: Workshop on Platial Analysis. Heidelberg, Germany, September 20--21.*, 45–52. https://doi.org/10.5281/zenodo.1472745

Hiippala, T., Hausmann, A., Tenkanen, H., & Toivonen, T. (2019). Exploring the linguistic landscape of geotagged social media content in urban environments. *Digital Scholarship in the Humanities*, *34*(2), 290–309. https://doi.org/10.1093/llc/fqy049

Hochmair, H. H., Juhász, L., & Cvetojevic, S. (2018). Data quality of points of interest in selected mapping and social media platforms. In *Progress in Location Based Services 2018 (vol. Lecture Notes in Geoinformation and Cartography)* (pp. 293–313). Springer, Cham. https://doi.org/10.1007/978-3-319-71470-7_15

Huang, B., & Carley, K. M. (2019). A large-scale empirical study of geotagging behavior on twitter. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, 365–373. https://doi.org/10.1145/3341161.3342870

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

Jarić, I., Courchamp, F., Gessner, J., & Roberts, D. L. (2016). Data mining in conservation research using Latin and vernacular species names. *PeerJ*, *4*, e2202. https://doi.org/10.7717/peerj.2202

Järv, O., Muurisepp, K., Ahas, R., Derudder, B., & Witlox, F. (2015). Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia. *Urban Studies*, *52*(14), 2680–2698. https://doi.org/10.1177/0042098014550459

Järv, O., Tenkanen, H., & Toivonen, T. (2017). Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. *International Journal of Geographical Information Science*, *31*(8), 1630–1651. https://doi.org/10.1080/13658816.2017.1287369

Joppa, L. N., O'Connor, B., Visconti, P., Smith, C., Geldmann, J., Hoffmann, M., Watson, J. E. M., Butchart, S. H. M., Virah-Sawmy, M., Halpern, B. S., & others. (2016). Filling in biodiversity threat gaps. *Science*, *352*(6284), 416–418.

Juhász, L., & Hochmair, H. H. (2018). Cross-checking user activities in multiple geo-social media networks. *Proceedings of the 21st AGILE Conference on Geo-Information Science*. https://agile-online.org/conference_paper/cds/agile_2018/shortpapers/160 JuhaszHochmair_AGILE2018_revised_final.pdf

Kahila-Tani, M., Broberg, A., Kyttä, M., & Tyger, T. (2016). Let the Citizens Map—Public Participation GIS as a Planning Support System in the Helsinki Master Plan Process. *Planning Practice & Research*, *31*(2), 195–214. https://doi.org/10.1080/02697459.2015.1104203

Kajala, L., Almik, A., Dahl, R., Dikšaitė, L., Erkkonen, J., Fredman, P., Jensen, F. S., Karoles, K., Sievänen, T., Skov-Petersen, H., Vistad, O. I., & Wallsten, P. (2007). *Visitor monitoring in nature areas – a manual based on experiences from the Nordic and Baltic countries*. TemaNord 2007:534. http://www.naturvardsverket.se/Documents/bokhandeln/bokhandeln.htm

Kitchin, R. (2014a). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Kitchin, R. (2014b). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. https://doi.org/10.1177/2053951714528481

Kwan, M. P., & Schwanen, T. (2009). Critical quantitative geographies 1: Beyond the Critical/Analytical binary: Quantitative revolution 2: The critical (re)turn. *Professional Geographer*, *61*(3), 283–291. https://doi.org/10.1080/00330120902931903

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. https://doi.org/10.1016/j.landurbplan.2015.08.004

Ladle, R. J., Correia, R. A., Do, Y., Joo, G.-J., Malhado, A. C. M., Proulx, R., Roberge, J.-M., & Jepson, P. (2016). Conservation culturomics. *Frontiers in Ecology and the Environment*, *14*(5), 269–275. https://doi.org/10.1002/fee.1260

Lee, H., Seo, B., Koellner, T., & Lautenbach, S. (2019). Mapping cultural ecosystem services 2.0 – Potential and shortcomings from unlabeled crowd sourced images. *Ecological Indicators*, *96*, 505–515. https://doi.org/10.1016/J.ECOLIND.2018.08.035

Lehtomäki, J., Tuominen, S., Toivonen, T., & Leinonen, A. (2015). What data to use for forest conservation planning? A comparison of coarse open and detailed proprietary forest inventory data in Finland. *PLoS ONE*, *10*(8), e0135926. https://doi.org/10.1371/journal.pone.0135926

Levin, N., Kark, S., & Crandall, D. (2015). Where have all the people gone? Enhancing global

conservation using night lights and social media. *Ecological Applications*, *25*(8), 2153–2167. https://doi.org/10.1890/15-0113.1

Li, D., Zhou, X., & Wang, M. (2018). Analyzing and visualizing the spatial interactions between tourists and locals: A Flickr study in ten US cities. *Cities*, *74*, 249–258. https://doi.org/10.1016/j.cities.2017.12.012

Liu, J., Dietz, T., Carpenter, S. R., Folke, C., Alberti, M., Redman, C. L., Schneider, S. H., Ostrom, E., Pell, A. N., Lubchenco, J., Taylor, W. W., Ouyang, Z., Deadman, P., Kratz, T., & Provencher, W. (2007). Coupled human and natural systems. In *Ambio* (Vol. 36, Issue 8, pp. 639–649). Royal Swedish Academy of Sciences. https://doi.org/10.1579/0044-7447(2007)36[639:CHANS]2.0.CO;2

Lomborg, S., & Bechmann, A. (2014). Using APIs for Data Collection on Social Media. *The Information Society*, *30*(4), 256–265. https://doi.org/10.1080/01972243.2014.915276

Longley, P. A. (2000). The academic success of GIS in geography: Problems and prospects. *Journal of Geographical Systems*, *2*(1), 37–42. https://doi.org/10.1007/s101090050027

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of twitter usage. *Environment and Planning A*, *47*(2), 465–484. https://doi.org/10.1068/a130122p

MA. (2005). *MA = Millenium Ecosystem Assesment. Ecosystems and human well-being* (Vol. 5). Island Press Washington, DC.

Macdonald, D. W., Jacobsen, K. S., Burnham, D., Johnson, P. J., & Loveridge, A. J. (2016). Cecil: A moment or a movement? Analysis of media coverage of the death of a lion, Panthera Leo. *Animals*, *6*(5). https://doi.org/10.3390/ani6050026

Mark, D. M. (2003). Geographic information science: Defining the field. *Foundations of Geographic Information Science*, *1*, 3–18.

McCay-Peet, L., & Quan-Haase, A. (2017). What is social media and what questions can social media research help us answer. In L. Sloan & A. Quan-Haase (Eds.), *The SAGE handbook of social media research methods* (pp. 13–26). SAGE.

McIntyre, N., Moore, J., & Yuan, M. (2008). A place-based, values-centered approach to managing recreation on Canadian crown lands. *Society and Natural Resources*, *21*(8), 657–670. https://doi.org/10.1080/08941920802022297

Mocnik, F.-B., Ludwig, C., Grinberger, A., Jacobs, C., Klonner, C., & Raifer, M. (2019). Shared Data Sources in the Geographical Domain—A Classification Schema and Corresponding Visualization Techniques. *ISPRS International Journal of Geo-Information*, *8*(5), 242. https://doi.org/10.3390/ijgi8050242

Niemelä, J., Saarela, S. R., Söderman, T., Kopperoinen, L., Yli-Pelkonen, V., Väre, S., & Kotze, D. J. (2010). Using the ecosystem services approach for better planning and conservation of urban green spaces: A Finland case study. *Biodiversity and Conservation*, *19*(11), 3225–3243. https://doi.org/10.1007/s10531-010-9888-8

Oksanen, J., Bergman, C., Sainio, J., & Westerholm, J. (2015). Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *Journal of Transport Geography*, *48*, 135–144. https://doi.org/10.1016/j.jtrangeo.2015.09.001

Oteros-Rozas, E., Martín-López, B., Fagerholm, N., Bieling, C., & Plieninger, T. (2016, February). Using social media photos to explore the relation between cultural ecosystem services and landscape features across five European sites. *Ecological Indicators*. https://doi.org/10.1016/j.ecolind.2017.02.009

Pascual, U., Balvanera, P., Díaz, S., Pataki, G., Roth, E., Stenseke, M., Watson, R. T., Başak Dessane, E., Islar, M., Kelemen, E., Maris, V., Quaas, M., Subramanian, S. M., Wittmer, H., Adlan, A., Ahn, S. E., Al-Hafedh, Y. S., Amankwah, E., Asah, S. T., … Yagi, N. (2017). Valuing nature's contributions to people: the IPBES approach. In *Current Opinion in Environmental Sustainability* (Vols. 26–27, pp. 7–16). Elsevier B.V. https://doi.org/10.1016/j.cosust.2016.12.006

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pickering, C., Rossi, S. D., Hernando, A., & Barros, A. (2018). Current knowledge and future research directions for the monitoring and management of visitors in recreational and protected areas. *Journal of Outdoor Recreation and Tourism*, *21*, 10–18. https://doi.org/10.1016/j.jort.2017.11.002

Pimm, S. L., Alibhai, S., Bergl, R., Dehgan, A., Giri, C., Jewell, Z., Joppa, L., Kays, R., & Loarie, S. (2015). Emerging technologies to conserve biodiversity. *Trends in Ecology & Evolution*, *30*(11), 685–696.

Poorthuis, A., & Zook, M. (2017). Making Big Data Small: Strategies to Expand Urban and Geographical Research Using Social Media. *Journal of Urban Technology*, *24*(4), 115–135. https://doi.org/10.1080/10630732.2017.1335153

Resilience Alliance; (2007). *Assessing Resilience in Social-Ecological Systems: A Scientists Workbook. Version 1. Image is available at: http://wiki.resalliance.org/index.php/Level_2_Detail_-_Bounding_the_System:_Describing_the_Present*.

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. https://doi.org/10.1016/j.ecolind.2015.01.034

Richards, D. R., & Tunçer, B. (2018, September 18). Using image recognition to automate assessment of cultural ecosystem services from social media photographs. *Ecosystem Servic-*

es, *31*, 318–325. https://doi.org/10.1016/j.ecoser.2017.09.004

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064. https://doi.org/10.1126/science.346.6213.1063

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milčinski, G., Nikšič, M., Painho, M., Pődör, A., Olteanu-Raimond, A.-M., & Rutzinger, M. (2016). Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS International Journal of Geo-Information*, *5*(5), 55. https://doi.org/10.3390/ijgi5050055

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (Muki). (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, *31*(1), 139–167. https://doi.org/10.1080/13658816.2016.1189556

Sessions, C., Wood, S. A., Rabotyagov, S., & Fisher, D. M. (2016). Measuring recreational visitation at U.S. National Parks with crowd-sourced photographs. *Journal of Environmental Management*, *183*, 703–711. https://doi.org/10.1016/j.jenvman.2016.09.018

Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, *142*, 198–211. https://doi.org/10.1016/j.landurbplan.2015.02.020

Sherren, K., Parkins, J. R., Smit, M., Holmlund, M., & Chen, Y. (2017). Digital archives, big data and image-based culturomics for social impact assessment: Opportunities and challenges. *Environmental Impact Assessment Review*, *67*, 23–30. https://doi.org/10.1016/j.eiar.2017.08.002

Sherren, K., Smit, M., Holmlund, M., Parkins, J. R., & Chen, Y. (2017). Conservation culturomics should include images and a wider range of scholars. *Frontiers in Ecology and the Environment*, *15*(6), 289–290. https://doi.org/10.1002/fee.1507

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PLoS ONE*, *10*(3), e0115545. https://doi.org/10.1371/journal.pone.0115545

Steffen, W., Grinevald, J., Crutzen, P., & McNeill, John. (2011). The Anthropocene: conceptual and historical perspectives. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1938), 842–867. https://doi.org/10.1098/rsta.2010.0327

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. https://doi.org/10.1080/13658816.2011.604636

Teles da Mota, V., & Pickering, C. (2020). Using social media to assess nature-based tourism: Current research and future trends. *Journal of Outdoor Recreation and Tourism*, *30*, 100295. https://doi.org/10.1016/j.jort.2020.100295

Tenkanen, H. (2017). *Capturing time in space: Dynamic analysis of accessibility and mobility to support spatial planning with open data and tools* [Doctoral dissertation, University of Helsinki]. http://urn.fi/URN:ISBN:978-951-51-2935-9

Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., & Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports*, *7*(1), 17615. https://doi.org/10.1038/s41598-017-18007-4

Thatcher, J. (2014). Living on fumes: Digital footprints, data fumes, and the limitations of spatial big data. *International Journal of Communication*, *8*(1), 1765–1783.

Tomlinson, R. F. (1967). An introduction to the geographic information system of the Canada Land Inventory. *Department of Forestry and Rural Development, Ottawa, Canada*.

Toth, C., & Jóźków, G. (2016). Remote sensing platforms and sensors: A survey. *{ISPRS} Journal of Photogrammetry and Remote Sensing*, *115*, 22–36. https://doi.org/http://dx.doi.org/10.1016/j.isprsjprs.2015.10.004

Väisänen, T., Heikinheimo, V., Hiippala, T., & Toivonen, T. (in press). Exploring human-nature interactions in national parks using social media photographs and computer vision. *Conservation Biology*.

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

Venter, O., Sanderson, E. W., Magrach, A., Allan, J. R., Beher, J., Jones, K. R., Possingham, H. P., Laurance, W. F., Wood, P., Fekete, B. M., Levy, M. A., & Watson, J. E. M. (2016). Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. *Nature Communications*, *7*, 12558. https://doi.org/10.1038/ncomms12558

Waldron, A., Mooers, A. O., Miller, D. C., Nibbelink, N., Redding, D., Kuhn, T. S., Roberts, J. T., & Gittleman, J. L. (2013). Targeting global conservation funding to limit immediate biodiversity declines. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 12144–12148. https://doi.org/10.1073/pnas.1221370110

Watson, J. E. M., Dudley, N., Segan, D. B., & Hockings, M. (2014). The performance and potential of protected areas. *Nature*, *515*(7525), 67–73. https://doi.org/10.1038/nature13947

Wolff, S., Schulp, C. J. E., & Verburg, P. H. (2015). Mapping ecosystem services demand: A review

of current research and future perspectives. *Ecological Indicators*, *55*, 159–171. https://doi.org/10.1016/j.ecolind.2015.03.016

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. https://doi.org/10.1038/srep02976

Wu, X., Lindsey, G., Fisher, D., & Wood, S. A. (2017). Photos, tweets, and trails: Are social media proxies for urban trail use? *Journal of Transport and Land Use*, *10*(1), 789–804. https://doi.org/10.5198/jtlu.2017.1130

Zhao, B., & Sui, D. Z. (2017). True lies in geospatial big data: detecting location spoofing in social media. *Annals of GIS*, *23*(1), 1–14. https://doi.org/10.1080/19475683.2017.1280536

Zook, M., Barocas, S., boyd, danah, Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., & Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology*, *13*(3), e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Mobile devices and digital platforms such as social media gather considerable amounts of data about our movements and activities. For a geographer, these data offer endless possibilities for discovering spatial and temporal patterns.

In this thesis, I have investigated how these new user-generated sources of geographic information can be used responsibly to fill in knowledge gaps about human activities in national parks and green spaces.

Despite various challenges and biases, user-generated data provide diverse information about where, when and how people enjoy nature. In some areas, user-generated data might even be the best available information about human activities.

Department of Geosciences and Geography A
ISSN-L 1798-7911
ISSN 1798-7911 (print)
ISBN 978-951-51-6580-0 paperback
ISBN 978-951-51-6581-7 PDF
http://ethesis.helsinki.fi/

Painosalama
Turku 2020

UNIVERSITY OF HELSINKI
FACULTY OF SCIENCE