MONASH University

Mapping Species distribution in space and time using social network site geotagged photos

Moataz Muhammad Medhat AbdulRahman Mahmoud-ElQadi

PhD

A thesis submitted for the degree of *Doctor of Philosophy* at

Monash University in *2019*

Faculty of Information Technology

Copyright notice

Abstract

Climate change is threatening pollinator insects, their habitats, ranges, and lifecycles. As a consequence, the human food supply is also threatened, since pollination by insects is vital for 75% of crop types we consume. Understanding these problems, requires constantly updated data on global scale. The spatial and temporal requirements for these data render them expensive and often impractical to collect.

Classically, scientists have resorted to manually collecting and studying species. To tackle issues requiring large scale surveys and intensive labour, scientists would sometimes mobilise volunteers to help collect data, a collaboration known as "citizen science". Citizen science continues to prove valuable to scientific research. But it requires effort to recruit, train, and incentivise volunteers, as well as to design data formats, and curate volunteered data.

"Incidental citizen science" is a new term I use here to denote the collection of data from online sources to be applied to scientific research, where the data is only incidentally related to the citizen science project being undertaken.

This thesis explores the possibilities and challenges involved in incidental citizen science. I propose that by collecting data from social network sites, and applying computer vision and classification techniques, we can find insights into ecological research questions that may help to understand the effects of climate change on insect pollination. The research case studies include Australia, seven African countries, and Japan.

A recurrent problem hindering the use of social network site data in my research is that the visual content of the retrieved photos may be irrelevant to the research questions. I have developed a pipeline process for checking photos' visual content using computer vision tools and classification techniques. Another problem I encountered is the spatiotemporal bias in data posted on social network sites by users. I address this problem by demonstrating techniques to see through the bias.

My results show that data on social network sites not only reflect social connections between human users, but can also reflect changes in ecological phenomena on Earth's surface over time. Among the

phenomena I look at is land cover type, which is a major determinant of ecological systems. From social network site data, I could obtain high quality in situ imagery that can aid in land cover classification. I also show how to augment species distribution data existing in specialised databases using geo-tagged photos from social network sites. I also show that social network data can reflect flowering time dynamics, despite spatial and temporal bias in photos posted online, and that even subtle out-of-season blooms can be detected using my methods.

Overall, in this thesis I present cost and time-effective methods to obtain ecological insights from free, public, geo-tagged photographic data available on social network sites using commercially available computer vision and machine learning tools. The data available looks likely to accumulate over time, and the machine intelligence capabilities are getting better, and more ubiquitous. Hence the approaches outlined in this thesis look set to become even more valuable in the future than they already are today.

Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature: Moataz Medhat ElQadi

Print Name: Moataz Muhammad Medhat Abdulrahman Mahmoud-ElQadi

Date: 6 December 2019

Publications during enrolment

ElQadi, Moataz Medhat, Alan Dorin, Adrian Dyer, Martin Burd, Zoë Bukovac, and Mani Shrestha. "Mapping Species Distributions with Social Media Geo-Tagged Images: Case Studies of Bees and Flowering Plants in Australia." Ecological Informatics 39 (5// 2017): 23-31. https://doi.org/http://dx.doi.org/10.1016/j.ecoinf.2017.02.006.

ElQadi, Moataz Medhat, Myroslava Lesiv, Adrian G. Dyer, and Alan Dorin. "Computer Vision-Enhanced Selection of Geo-Tagged Photos on Social Network Sites for Land Cover Classification." Environmental Modelling & Software 128 (2020/06/01/ 2020): 104696. https://doi.org/https://doi.org/10.1016/j.envsoft.2020.104696.

Acknowledgements

I am grateful to my parents for building the person I am. Thanks to my mother, Prof Ebaa ElShamy, for setting the example of life-long learning and research, and thanks to my father, Mr Medhat ElQadi, for being a role model of hard work, determination, and love for computers.

Thanks to my wife, Salma, for our life journey, for her patience, love and care. I could not imagine going through this PhD without her.

I would like to thank Assoc Prof Alan Dorin, my main supervisor, for guiding me throughout my research and for all the time, effort, and advice he gave. I am also grateful to Assoc Prof Adrian Dyer, my second supervisor, for his valuable time, discussions, and expertise.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship. The research was possible to start thanks to Australian Research Council grant DP 160100161 awarded to my supervisors, and I could continue my candidature thanks to a Monash faculty of Information Technology stipend.

I would like to thank Microsoft for awarding me an “AI for Earth” grant that allowed me to use Microsoft Azure services in my experimental work. I would also like to thank Potsdam Institute for Climate Impact Research for funding my participation in IIASA’s Young Scientist Summer Program (YSSP) where I completed parts of my research.

Table of Contents

1 Thesis Chapter 1 – Introduction

1.1 Overview

1.1.1 Motivation

Climate change is threatening pollinator insects, their habitats, ranges, and lifecycles (Kjøhl, Nielsen, & Stenseth, 2011). Habitats can be subject to drought which may result from changes to rainfall patterns caused by climate variation. Wildfires too are potentially a problematic outcome in new regions as the climate shifts (Abatzoglou & Williams, 2016). Likewise, in low lying habitats, sea level rise is a potential problem emerging from climate change (Nerem et al., 2018) that some insects may encounter. Perhaps more importantly, even than the above potentially catastrophic events, as the climate shifts, and local temperatures are perturbed from their historical ranges, insects may find themselves in regions no longer thermally suited to their physiology (Deutsch et al., 2008). Hence, their viable ranges can change in response to the climate also. Local weather conditions, such as temperature, may also act as ecological triggers for lifecycle events like insects' emergence from underground burrows and nests, for eggs hatching etc. These too can be disrupted by changes in climate. Lastly, insects depend on other animal species and plants for their food supplies (Memmott, Craze, Waser, & Price, 2007). If these are also disrupted by climate change in ways different to the disruptions the insects suffer, then mismatches between the timing of the insects' needs and the availability of the resources in the environment they need to survive, may occur. This too can have potentially fatal consequences for entire populations of insects.

These threats to pollinator insects directly endanger the human food supply (Klein et al., 2007). Much of the food consumed by humans, including fruits and nuts for instance, is dependent on pollination by insects. In addition, the seeds grown on farms are often produced using techniques of hybrid seed production. This requires insects to carry pollen from a male line of crops to a female line that then produces the fertilised seed. These seeds are later the source of the food plants that farmers grow to allow human food production to keep pace with human population growth. In short, without insects, humans would have to either find an alternative (and likely very expensive and labour intensive) way to pollinate their crops, or we have nothing to eat.

Problems of pollination in a changing climate can be potentially mitigated by, for example, investing in managed pollination services, boosting feral pollinators' populations, or using more greenhouses. To inform mitigation strategies we require rapidly updated information on global scale about pollinators' ranges and abundance. While climate, and changes affecting it, are global phenomena, studying the effects on pollination sometimes requires fine-grained spatial data. For example, a farmer needs to know whether there are enough pollinators in his farm. But a country such as Australia may need data that

covers an entire continent to assess the spread of an invasive insect species, or the change in range of a valued native pollinator.

In addition to the need for fine-grained through to coarse grained spatial data, we have a similar need for temporal data. Climate change-induced change in habitats may take many years to be evident. But a farmer interested in an individual species of pollinator and its lifecycle on his farm may need data across a single growing season of a few weeks. The range of temporal data therefore spans a week or two, right through to decades for long term trends.

The potentially high temporal and spatial resolution of data we require to address the needs outlined above, is impractical to collect using trained ecologists. There just aren't enough of them and they are a relatively expensive resource. However, today, social network sites have massive amounts of data that, although not necessarily geared towards scientific research, may still offer valuable insights to understand insect pollinators. Finding our proverbial needle in the haystack, valuable ecological sightings amongst massive public online data in social network sites, is a task that calls for the help of intelligent machines. This thesis introduces methods and tests them to employ data from social network sites for solving ecological problems. These frameworks use Computer Vision and classification models to find relevant data in the chaos.

In the rest of this overview, I show the importance of pollinator insects, and how they are affected by climate change (1.1.2). I then discuss the spatial and temporal gaps in available information on pollinator insects (1.1.3).

1.1.2 Pollinators and Climate Change

Pollination by insects, especially bees, is vital for 75% of crop types consumed by humans, which represent 35% of food volume consumed worldwide (Klein et al., 2007). The remaining crops do not depend on animal pollination, but are wind-pollinated (anemophilous), self-pollinated, or produce fruits without pollination (parthenocarpic).

Given the high percentage of food that depend on pollination, it comes as little surprise that the annual economic value of pollination in global agriculture has previously been estimated at €153 Billion (Gallai, Salles, Settele, & Vaissière, 2009). When a plant receives pollination services that are less than optimal, also known as having a pollination deficit, then the resulting produce is sub-optimal in quantity, quality, or both (Clarke, Gillespie, & Cunningham, 2017). In Australia, this pollination gap is estimated at 1 Billion dollars (Clarke et al., 2017). It thus comes as little surprise that the Australian government include in its research and development priorities: "[To] establish practices to maintain, or increase the level of free pollination from wild insects", and "[To] develop systems for managing and using alternative pollinators (such as stingless, blue banded and leaf-cutter bees) for specialised production

environments"(*Honey Bee and Pollination Program Five Year Research, Development & Extension Plan 2014/15 – 2018/19*, 2015).

Alarmingly, given insect pollinators' economic and agricultural importance, they are under threat on many fronts. Threats include habitat alteration, climate change, pesticides overuse, and invasive species infestation (Kjøhl et al., 2011). Habitat alteration is primarily caused by land use change where, for instance, land used by insect pollinators for nesting and foraging is lost to human activities of logging, farming, or building. Habitat alteration is considered the first driver of species extinction (Díaz et al., 2019), where one million species are threatened by extinction (Díaz et al., 2019).

Of these threats, climate change is particularly devastating. It threatens managed and wild insect pollinator population, as well as the crop plants. Nevertheless, studies focusing on the effect of climate change on crop or other plant pollination are very scarce (Kjøhl et al., 2011). The Intergovernmental Panel on Climate Change (IPCC) documented increased global temperature, among other changes. Kjøhl et al. (2011) argue that the raised temperature is the most important effect of climate change on plant-pollinator interactions, and that the species' response to large-scale climate change would be emigration to cooler latitudes. However, although crop species may be easy to move, pollinators might not be able to follow. The study of these potential responses to climate change requires determining pollinators' presence. This is a complex task as it depends on a combination of biotic and abiotic conditions that are variable across the landscape like habitats, floral resources, and climate (Nogué et al., 2016). Information on the geographic ranges and abundances of insects is important to understand, predict, and manage pollinator services (Biesmeijer et al., 2006; Moritz, Kraus, Kryger, & Crewe, 2007).

1.1.3 Information gap and how citizen science may help

The UN development goals of "zero hunger" and "life on land" ("United Nations sustainable development goals," 2015), Figure 1, require adequately understanding and mitigating the threats to pollinators, due to their importance and vulnerability. However, the global scale and fast rate of these ecological threats demand new ways of collecting data. Classically, scientists resorted to manual collection and study of species, a practice still relevant today (Kuiter, 2013). For example, to establish the existence of insects, traps or nets could be used to catch insects for investigation. To determine a species' distribution, scientists specialised in insect collection and identification have usually had to endure surveying expeditions that mandated the organisation of sometimes expensive and labour-intensive campaigns that are impractical, arguably impossible, to organise on the required large, highly detailed, scale (Garbarino & Mason, 2016).

Figure 1 United Nations sustainable development goals

To tackle issues of large scale and intensive labour in the study of species distribution, scientists would sometimes mobilise volunteers to help in data collection, a collaboration known as "citizen science" (Follett & Strezov, 2015; Theobald et al., 2015). The World Wide Web provided ways of publishing information that were not available before. With the advent of social network sites such as Facebook, twitter, and Flickr, and other web 2.0 technologies, the information sharing turned into a bi-directional process whereby users can not only consume published information, but also be a source of publishing and "sharing" information. This evolution provided an opportunity to more easily conduct surveys and collecting information from the population.

In summary, pollinator insects are responsible for a third of the food humans consume. These pollinator insects are endangered, unfortunately, by many factors. Mitigating these factors, and their effects, needs better understanding of these insects' distribution in space and time, as well as their habitats. However, undertaking this globally massive research is beyond the capabilities of natural scientists and their funding organisations. A form of help career scientists are increasingly recognising and seeking is that of the general public, a form of volunteering to further science usually termed "Citizen science", which we discuss in the next section.

1.2 Citizen science

1.2.1 Early citizen science

Citizen science is the practice of science by volunteers, especially in collecting and analysing data as a part of a scientific research project (Gura, 2013). Although the term "citizen science" is relatively recent (Silvertown, 2009), citizen science has always, arguably, been a main tributary to science.

A long running example of citizen science is the Christmas bird count in the USA that has been run every year since 1900 (LeBaron et al., 2004). This bird count has helped to understand bird ecology and natural history (Garbarino & Mason, 2016). Another historical example of citizen science, in 1930s Britain, an extensive land utilisation survey that recorded land use of every field in the country relied on school children guided by their teachers (Southall, Baily, & Aucott, 2007).

Citizen science continues to be used at an accelerating pace since it is proving valuable to scientific research. For example, data from the Christmas bird count had generated about 350 papers by 2009 (Silvertown, 2009). Financially, citizen scientists are estimated to contribute $2.5 billion in-kind to research (Theobald et al., 2015). In addition, the involvement of volunteers in citizen science helps the general public understand scientific issues and better appreciate science (Garbarino & Mason, 2016).

The success of citizen science comes with its drawbacks. Among these are the effort required to recruit, train, and incentivise volunteers, as well as the need to design the data formats and curate volunteered data. Technology advancement can help with some of these challenges in the new generation of citizen science projects.

1.2.2 Citizen science today

Today, new means of communication are allowing people to interact in new ways privately and publicly. A user can hold audio or video calls with another person, or even a small group of friends or family in real time or publicly broadcast his life in real time. Everyone can now be a "citizen journalist" sharing their opinion, photography, and other media.

These advances have affected citizen science. In this section we show how the web, smart phones, and social network sites have all changed citizen science.

1.2.2.1 Internet and smart phones

Since its creation in the 1990s, the world wide web continued to evolve. Around 2003~2004, the web was said to have evolved into Web 2.0. (Cormode & Krishnamurthy, 2008), there is not always a clear distinction between Web 1.0 and 2.0, but generally, in web 2.0 users were able to not only receive web pages, but also edit web pages by commenting and replying. Not only using text, but also other types of media such as photos, audio, and videos. Social network sites are a distinctive feature of Web 2.0 which brought to the web terms like "friends" and "groups" (Cormode & Krishnamurthy, 2008)

As mobile phones got more powerful hardware, they gained more features other than just making calls and text. These smartphones, as they became to be known in the 2010s, allow their users to not only access the internet away from their desktop computers, but to also run applications "apps" that leveraged

the powerful and versatile hardware of these phones. Smartphones allowed users, for example, to instantly send photos with location information, thanks to the phone GPS and camera. The combination of social network sites and smartphones understandably changed citizen science projects (Silvertown, 2009).

As we have discussed, social network sites are a key feature of Web 2.0 that is changing Citizen Science, and therefore deserve a further discussion which we include in the next section.

1.2.2.2 Social network sites

Social network sites can be defined as "web - based services that allow individuals to (1) construct a public or semi - public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site." (Boyd & Ellison, 2007). Examples of these websites include, at the time of writing, Facebook, Twitter, Flickr, and others.

In the context of citizen science, we note that a common use of social network sites is to allow people to communicate, it thus comes as little surprise that social network sites facilitate the communication process between scientists and volunteers in citizen science projects (Stafford et al., 2010). For example, some scientists have set up Facebook interest groups to collect data (Deng et al., 2012), others asked volunteers to post their contributions to a special Flickr group (Kirkhope et al., 2010; Stafford et al., 2010), or create a special platform for users to contribute (Fritz et al., 2017; Marchante, Morais, Gamela, & Marchante, 2017).

In all these cases, social network sites serve as a communication facilitator and citizen scientists were actively engaged to contribute to the research project. In fact, engaging citizen scientists is arguably a challenge when running citizen science projects (Gura, 2013). Consequently, different methods to motivate a large number of individuals to contribute to a project as crowd sensors were discussed in the literature. Jaimes, Vergara-Laurens, and Raij (2015) surveyed many possible ways to incentivise participation implemented in research papers from different fields.

Social network sites, SNS, have been recently gaining attention from the research community. This is not surprising given how social network sites are an emerging aspect of the human life that is directly affecting people living today. It is estimated that more than one billion people use social network sites regularly (Andreassen, Torsheim, & Pallesen, 2014). Two thirds of college students surveyed in one study spent 6 hours or more every day on social network sites (Wang, Chen, & Liang, 2011). Studies suggest that excessive use of SNS is considered a form of behavioural addiction (Andreassen, 2015).

At the time of writing, searching google scholar (http://scholar.google.com) for literature with the keyword “Twitter” in the title returns about 79,000 results, slightly behind the 92,000 results returned for the keyword “Facebook” in research title. The more general term “social network site” returns “only” 500 results. This is a significant interest given that these sites have come to being only recently. Facebook, for instance, was founded in 2004 (Facebook, 2019).

Of particular relevance to this thesis is the use of social network sites to conduct biological surveys, which is an emerging field of interest. A systematic review of volunteer involvement in biological surveys revealed that data collected by professionals was more accurate only in 4 out of 7 cases (Lewandowski & Specht, 2015).

Social network sites have been used in many fields of natural science research. For example, in epidemiology, hospital surveys are the standard means for identifying epidemics. These surveys are expensive and time consuming. Instead, Culotta (2010) investigated using Twitter to detect influenza epidemics through automatically classifying messages related to influenza. In environmental monitoring, which currently also depends on field surveys, social network sites were used instead to predict hurricane damage (Bohannon, 2016) and flooding impact (Barker & Macleod, 2018), to detect invasive species (Daume, 2016), and complement monitoring data on the Great Barrier Reef in Australia (Becken, Stantic, Chen, Alaei, & Connolly, 2017).

In ecological research, Hampton et al. (2013) linked the future of ecology to big data. However, they were referring to datasets of massive size, arguing that ecologists need share their data collections, and mentor the citizen science projects carried out by volunteers. It seems that the massive size, and broad spatial coverage, of the datasets are what appeal to the ecology community. The reason behind this interest could be that there are massive amounts of data available in social network sites so that even if a small fraction of the data is relevant, to a given research question, these relevant data may prove valuable. I discuss this issue in the next chapters.

In the field of biodiversity studies, the BeeID project (Kirkhope et al., 2010; Stafford et al., 2010) is an example of the application of social network site Flickr as a medium for species identification. This application allows volunteers to post bee photos, and experts to identify the bee species the images depict. The popularity of social network sites and applications gave rise to several specialist social networks dedicated to species identification. For example, iSpot connects amateurs with experts in species identification, resulting in accurate crowd-sourced species identification of 390,000 observations by 2015 (Silvertown et al., 2015). Barve (2014) also used Flickr for biodiversity research. Instead of asking volunteers to post photos, he explored the geotagged locations of previously posted photos and investigated whether they can serve as a source of biodiversity data. His study focussed on the Snowy Owl (*Buboscandiacus*) and the Monarch Butterfly (*Danaus plexippus*) and he decided that

the previously posted geotagged data was valuable as long as the species photographed are recognisable by social network site users.

Another application of SNS to research is in the area of land cover research. Land cover research requires field observations (See et al., 2013), which are often hard to obtain, especially with the large scale of research areas that might cover anything from the size of a farm, to an entire continent. This makes land cover research a primary application for photos posted on social network sites. For one example, Iwao et al. (2011) used photos from the Confluence Project (http://confluence.org/); a website that encourage volunteers to submit photos from integer coordinates[1], to validate global land cover maps they created by merging disparate land cover maps. Similarly, Estima and Painho (2013), (2014) and (Estima, Fonte, & Painho, 2014) explored the adequacy of Flickr geotagged photos for land cover classification by analysing the spatial and temporal distribution of photos and comparing them to the European Corine Land Cover database as a reference. They found that Flickr geotagged photos might present a valuable information source but suffers of biased spatial distribution of photos.

Oba, Hirota, Chbeir, Ishikawa, and Yokoyama (2014) used Flickr photos' location, text attributes, and text keywords corresponding to image features to classify land cover maps. Xu, Zhu, Fu, Dong, and Xiao (2017) and Xing, Meng, Wang, Fan, and Hou (2018) used deep learning to classify Flickr image contents to land cover classes.

In fact, social network sites have also been used in other research fields. Examples include public opinion measurement using twitter text sentiment analysis (O'Connor, Balasubramanyan, Routledge, & Smith, 2010), predicting elections results based on number of tweets mentioning a political party or a candidate (Tumasjan, Sprenger, Sandner, & Welpe, 2010; White, 2016), and studying stigma around psychological illness, where the word "Schizophrenia" was found to be misused (Joseph et al., 2015).

Citizen science continues to be a source of help to scientists to carry out tasks that are otherwise impractical. Volunteers are contributing billions of dollars' worth of effort. In their review of 388 citizen science projects, Theobald et al. (2015) estimated that among these projects, collectively, 1.3 million volunteers contributed up to $2.5 billion in-kind annually. Citizen science is becoming more ubiquitous as technology facilitates the participation of volunteers. However, there remain many open research questions pertaining to the use of social network sites in citizen science. These issues are discussed in section 1.2.3.2.

[1] A common method of specifying locations on Earth's surface is to use a spherical coordinate system where latitudes and longitudes are angles measured from the centre of a spherical model of Earth.

In many citizen science projects, social network sites are used as a medium where volunteer citizen scientists are asked to contribute data through SNS, which are then collected by the recruiting scientists.

Researchers observed that data created by users on social network sites, not necessarily during a managed project, could be a valuable source of information (Di Minin, Tenkanen, & Toivonen, 2015)

Ultimately, data that were posted to a social network site by users willing to answer a call by a scientist are, in many cases, not much different from other data posted by users without anyone asking them. These data can arguably present a valuable source of insights on natural phenomena. E.g. (Becken et al., 2017; Daume, 2016)

I term this utilisation of data, for a scientific research project, created by self-motivated users, not necessarily in course of a directed survey or project, "incidental citizen science". This stands in contrast to the classical "traditional citizen science".

1.2.3 Incidental Citizen Science

1.2.3.1 Definition

Incidental Citizen Science is a new term I use here to denote the collection of data from online sources to be applied to scientific research in the case where this data was not originally posted/published with this intention. That is, the data is only "incidentally" related to the citizen science project being undertaken and its relevance is determined by the researcher without any communication between the researcher and the data provider and no attempt to manage the process of data collection, or the behaviour of the data providers. social media and other repositories of volunteered and open data all potentially form a part of such research, but only recently has it become feasible to use these sources in this way.

This thesis explores the possibilities offered by the new availability of incidentally volunteered and open "big data" on social network sites, and the challenges involved in making sound scientific progress in ecology with this resource. Figure 2 shows sources of data used in what I identify as incidental citizen science, and its use in ecological research.

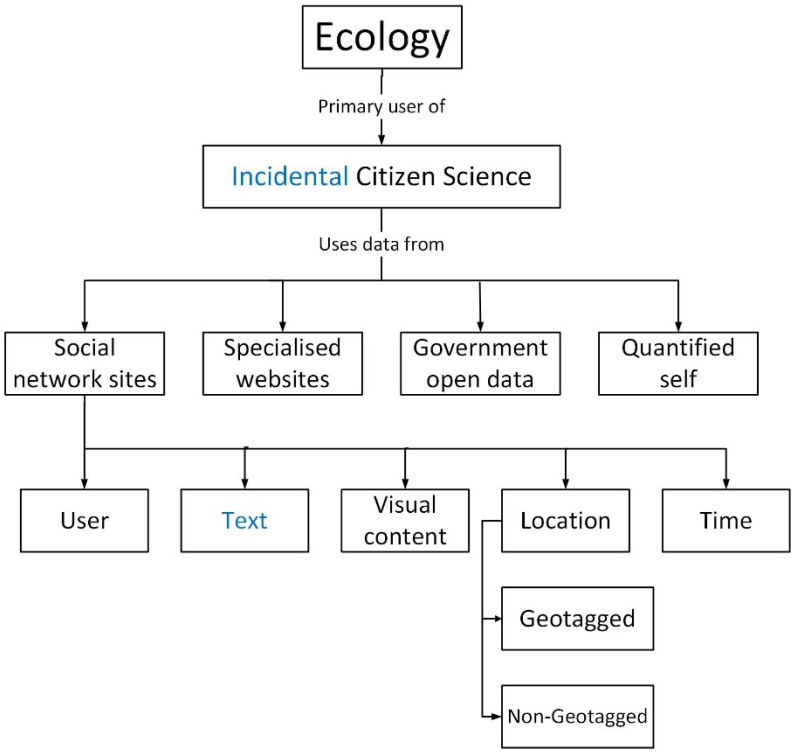


Figure 2 Incidental citizen science and its data sources. Components emphasised in the thesis are shown in Blue.

1.2.3.2 Incidental versus traditional citizen science

I summarise the differences between traditional and incidental citizen science in Table 1, and the following list.

Table 1 Comparison between traditional and incidental citizen science projects

			Traditional	Incidental
Data Sources	1.	Public websites	✓	✓
	2.	Social networks	✓	✓
	3.	Quantified self	×	✓
	4.	Open government data	×	✓
	5.	IoT sensor logs	×	✓
People Involved	6.	Research (project) managers	✓	×

	7.	Researchers	✓	✓
	8.	Volunteers (citizen scientists)	✓	×
Project Operations	9.	Project scoping	✓	✓
	10.	Project marketing to volunteers	✓	×
	11.	Incentivising volunteers	✓	×
	12.	Supporting volunteers	✓	×
	13.	Communication with volunteers	✓	×
Data attributes	14.	Designed formats	✓	×
	15.	Usage consented by provider	✓	×
Data Operations	16.	Data acquisition	✓	✓
	17.	Data wrangling	×	✓
	18.	Checking and quality review	✓	✓
	19.	Data processing and analysis	✓	✓

The following numbered points address their corresponding entries in the table above:

1. Many traditional citizen science projects, and interest groups use websites to collect and present data which can be used in incidental citizen science projects, which may or may not be related to the original traditional projects. For example, iSpot (ispotnature.org, ispot.org.za) are used for crowdsource identification of organisms.(Silvertown et al., 2015). The data available on these websites can be "recycled" in incidental citizen science where it is used for scientific purposes other than originally intended.

2. Social networks have been used in traditional ecological citizen science as a medium of communication, public engagement (Kirkhope et al., 2010; Stafford et al., 2010), and as a free data

repository. The data hosted on social network sites as a result of these projects, or for other reasons, can be leveraged also in incidental citizen science projects.

3. People are becoming increasingly keen to monitor themselves and share personal information, from sleep and food consumption, physical activity and heart rate, to reading habits and media consumption (Swan, 2013).

4. Governments are making their data open in hope of enhancing transparency, and citizen well-being (Kim, Trimi, & Chung, 2014). Scientific research may not be among the immediate objectives of government open data, but it can be leveraged in incidental citizen science projects in ecology. For example, fire outbreaks data meant for evaluating emergency services can provide valuable insights into wild life disruption by fire.

5. Internet of Things (IoT) is making available extraordinary amounts of sensor data that will see more ubiquitous applications.

6. Non-technical aspects of a traditional project require a distinct project management role which is rarely the case with incidental citizen science projects.

7. Understandably, researchers are involved in both kind of citizen science projects. However, I note that the operations and skills required are different in either types. (See data attributes and operations)

8. Volunteers knowingly participate in traditional citizen science projects, while in the incidental counterpart, they might not know that there data is being used. (See data attributes)

9. Scoping is a crucial step in any project. However, in an incidental citizen science project this could only entail posing research questions while in traditional citizen science projects that would involve setting the project scope.

10 & 11. Since traditional citizen science projects depend on volunteer citizens, much interest is understandably vested in marketing the project to volunteers, and incentivise them to participate. For example, Jaimes et al. (2015) surveyed possible methods to motivate a large number of individuals to contribute to a project as crowd sensors. The survey discussed ways to incentivise participation implemented in research papers from different fields.

12 & 13. Volunteers submitting data in a traditional project need mentorship and training to perform the required tasks e.g. (Kirkhope et al., 2010). Contrarily, incidental citizen science does not require data provider training or support. In fact, this kind of project does not involve the data provider to begin with.

14. Aside from a few examples, e.g. iSpot (Silvertown et al., 2015), social network sites are built for social, rather than scientific, purposes. As such, the structure of data in citizen science projects is a challenge (Daume, 2016) that requires planning and management during the project. The nature of incidental citizen science projects clearly exacerbates the data formats problem.

15. In traditional citizen science projects, volunteers purposefully contribute to the project, which is not the case with the incidental citizen science. Content created by users that researchers could incidentally find useful was not originally created for this purpose, and the users, who are not contacted, might not consent to the usage being carried out. This remains an ethical area of concern warranting investigation.

16. Data acquisition is common in traditional and incidental citizen science. It may however take different forms (See data attributes).

17. Incidental data in raw, unstructured, formats (see point 14) need to be mapped and transformed to structures that are fit for the research in question.

18. Part of the volunteer mentorship and management in traditional citizen science is the data checking and quality review by researchers or other experienced reviewers. On the other hand, not all data available in the possible sources of incidental citizen science are related or adequate the research undertaken, rendering filtration an elaborate task to perform in this kind of projects. Consequently, despite the massive nature of these data sources, for a given research question we might be left with little to no adequate data (see challenges). This problem is virtually non-existent in traditional projects thanks to the directed nature of the project.

19. Data processing is a common step in all kinds of modern research. However, owing to the unstructured nature of the data in incidental citizen science projects (see Data Attributes in table), it is far more challenging to perform the analysis. A fact that was reported by some scholars who tried to use social networks data creatively in what I term incidental citizen science. E.g. (Daume, 2016) manually analysed twitter feeds.

1.2.3.3 Challenges and gaps in Natural science research applications

Having characterised incidental citizen science in the previous section. In the following list, I ponder challenges and gaps in incidental citizen science based on existing literature where the used methodologies conform to the characteristics of incidental citizen science.

A. Spatial and Temporal bias

The content posted generally on social networks is spatially biased (Estima & Painho, 2014); users tend to inhabit urban areas and therefore they post from these locations most frequently, Figure 3. Content is also biased temporally where users' activity change by time of day and year, and also by content subject as people tend to write or take pictures of objects deemed interesting. For example, Christmas trees aren't often discussed in March or April, and firecrackers may be a popular subject of discussion around the New Year's celebrations of different cultures. This bias further complicates the data analysis task for a researcher using SNS to attempt to obtain unbiased data. and should be taken into consideration before drawing conclusions. Bias can also be magnified by massive data set sizes, for example, the population of a city, and user content generated there, could be an order of magnitude higher than a town's. Figure 3 shows, for instance, how the capital cities are hotspots of high image density.

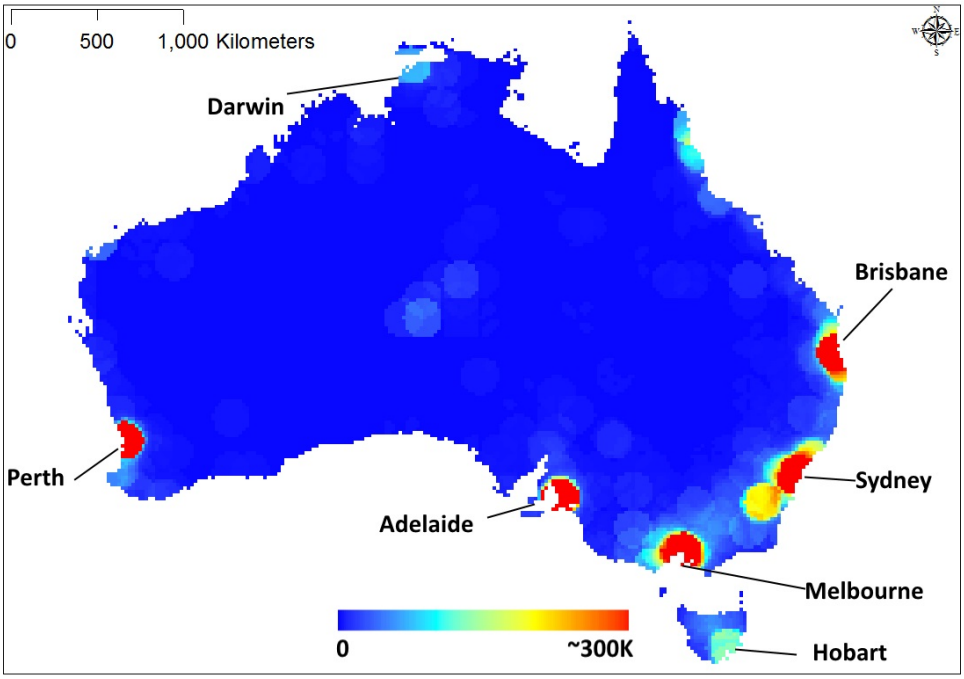


Figure 3. Heat map of all geotagged Flickr images in Australia. Capital cities show a much higher density, Oct. 2016

B. Questionable data quality

Although See et al. (2013) have shown that the quality of data provided by volunteers can be close to that of experts, the non-specialised nature of the content creators still raises questions about the data quality and trustworthiness.(Barve, 2015) also note that data quality remains a concern, especially with taxonomic accuracy, and confidence in Metadata. In fact, the analysis of incidental, unstructured, and

unevaluated content is challenging; interested users posting photos of “honey bees” (*Apis mellifera*) may well have misidentified the species.

C. Inadequate data quantity

The specialised nature of science projects lends itself to the incidental citizen science projects, creating unique challenges; Becken et al. (2017) who analysed sentiment of tweets about the great barrier reef noted that, despite the sheer volume of tweet flow, there are not enough tweets satisfying the location and content conditions. After they collected more than 200,000 tweets located around the Great Barrier Reef, the number of relevant tweets discussing the reef was much smaller, where only 0.6% mentioned “water” or “coral”.

D. Unstructured data formats

Social network sites are not generally meant to be used as a scientific data collection tool, and as such, content created in it is unstructured data not readily usable in scientific research (Deng et al., 2012). This is particularly challenging when we aim to extract knowledge from data that is incidentally relevant to our research questions, and more so when we seek to automate the process.

E. Automation

Some of the innovative research using SNS depend on manual implementation of methodology. A fact that does not affect the adequacy of the concept or the result validity, but cripples operationalising these usage models. For example, Daume (2016) uses manual checking to determine whether a tweet shows an invasive species, and concluded that challenges in using twitter for ecological monitoring are technical, not conceptual, especially with aspects like image recognition. Similarly, the Great Barrier Reef Marine Park Authority, GBRMPA, manually screen Instagram to detect extent of bleaching (Becken et al., 2017).

F. Geotag scarcity

Most citizen science applications, and virtually all of these applications related to natural research require geotagged content, i.e. the availability of geographic location in meta-data. However, only a small proportion of such content is geotagged. For example, it is estimated that only between 1.5% and 3% of twitter data is geotagged (Palpanas & Paraskevopoulos, 2015).

G. Ethical Considerations

Usage of social network sites data in some human research has been understandably the centre of fierce controversy. Especially when interfering with political processes and national sovereignty, e.g. Cambridge Analytica, a data analytics company, collected and exploited unauthorised personal data of Millions of Facebook users to influence American voters (Cadwalladr & Graham-Harrison, 2018).

Natural science research using social network site data, however, do not seem to attract such public attention, and the whether such research is ethical, is not currently subject to much deliberation. There are many potential arguments supporting this stance; this kind of natural research does not seek to experiment with humans or animals, does not alter natural environment, and more importantly in this context, does not seek to collect personal information on humans, and the data in question is publicly available and have been provided wilfully. The last points are a common denominator in laws and regulations organising data privacy and lawful use of data (e.g. GDPR (EU, 2016))

I argue here, however, that while these arguments may be valid reasons in favour of natural research using social network sites, there may still be unintended consequences to such research that ought to be taken into consideration. Firstly, although the data were consensually posted in public, the user posting this data has not always explicitly approved a certain research application. Especially in incidental citizen science uses where SNS is a data source, not only a medium of communication between researchers and citizen scientists. Secondly, it is true that data collected for natural research are normally related to natural phenomena, not human behaviour, but such behaviour can be sometimes revealed from metadata (e.g. tracing a user in space and time from their posts).

H. Generalisability

I show in the chapter 2 how different countries, even if they share a border, exhibit a different photography landscape. Hence, as much a rich data source SNS can be, transferring results from one geographical area to another, or even to global extent, is rather risky.

1.3 Thesis

In summary, pollinator insects are threatened globally by a variety of biotic and abiotic factors, this in turn threatens the human food supply. Facing this threat requires a data stream of global coverage. social network sites offer such a stream, but they are not built or typically used for this specific purpose. In this thesis, I set to innovate new ways that SNS can help in ecological research; more specifically, in mapping flowering plant and insect pollinators. I take an innovative approach to using computer vision and classification models to analyse information and metadata of geotagged photos posted on SNS, made available online by users not necessarily involved in scientific research, to solve these pressing ecological problems.

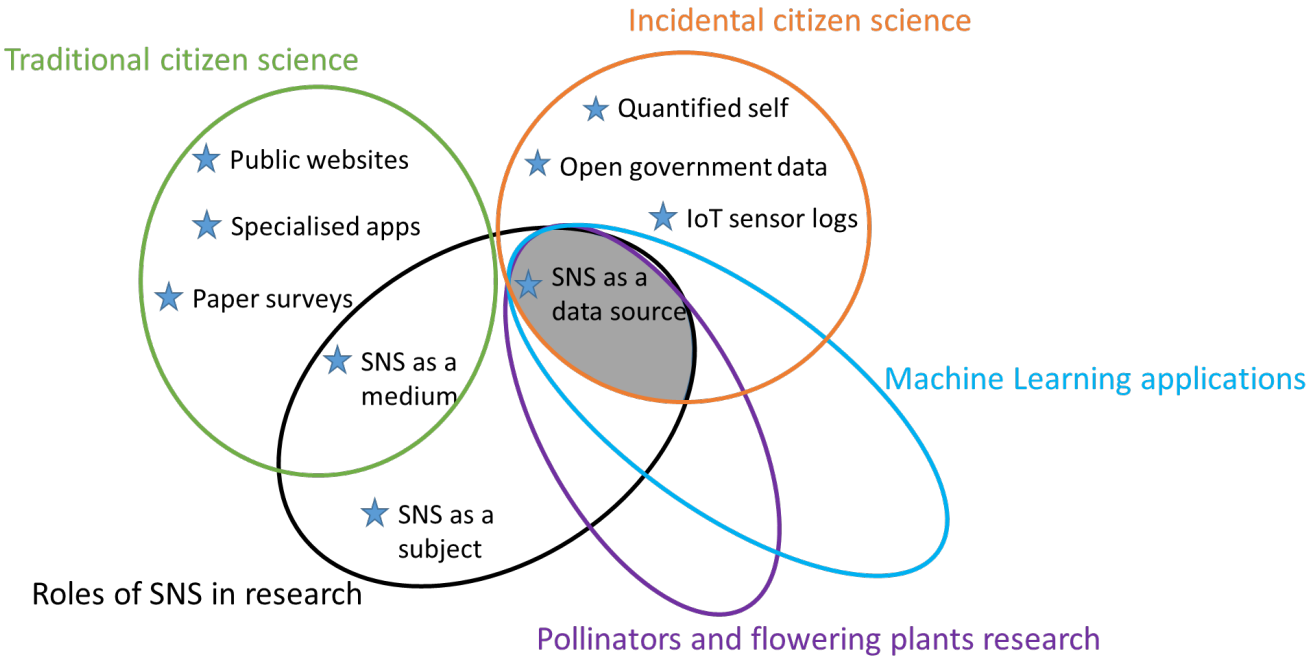


Figure 4 The grey shaded area represents the thesis; applying machine learning to social network sites (SNS) as an incidental citizen science source to study pollinators and flowering plants in space and time.

1.3.1 Scope

This work sets out to use social network sites in solving some of the pressing problems in ecological research. I introduce tools and frameworks that can help determine the spatial and temporal distribution of bee pollinators, flowering plants, and their habitats in order to inform decisions on mitigating the pollination gap. In doing this, I address shortcomings in the existing incidental citizen science literature discussed in 1.2.3.3.

In chapter 2, I suggest and implement a computer system to aid scientists to find photos relevant to land cover classification. Land cover affects the climate, and species habitat, and is therefore important to understanding pollinator and plant habitats.

In chapter 3, I contribute a method to filter-out SNS geotagged photos irrelevant to a species of interest, and show how the filtered results can effectively augment primary biodiversity data. I apply my method to pollinator insects and flowering plants in Australia.

In chapter 4, I hypothesise that the time signature of phenological activities can be inferred from temporal distribution of SNS geotagged photos of a plant. I prove this hypothesis using data of flowering cherry blossoms from Japan. I then show how powerful the method is that it can uncover previously understudied real-world phenomena, such as a subtle autumn cherry bloom.

1.3.2 Technology used

This interdisciplinary thesis is arranged by ecological topics, the below table shows the corresponding data science and information technology components in each chapter.

	Ecological topic	Dimension	Data science topics	Technology
Chapter 2	Land cover	Space	Computer Vision Binary Classification Artificial Neural Networks	Azure machine learning Azure cognitive services Web services WinForms Python
Chapter 3	Species distribution	Space	Computer Vision Statistical Distribution Spatial Analysis	GIS Python
Chapter 4	Phenological cycles	Space and Time	Computer Vision Spatial Analysis	Google vision API GIS Docker Python

1.3.3 Data processed

The social network site used through this thesis as a data source is Flickr. Flickr (www.Flickr.com) is a photo sharing social network site specialised in sharing photography, along with information on these photos, including the posting user and their bio information, the time and location of the photo, time of upload, and descriptive keywords used to match a user query. Between 2013 and 2016, Flickr had an average of two million public photos added every day worldwide (Michel, 2016).

While my methodology is social network site agnostic, I chose Flickr for some favourable characteristics; its Application Programming Interface, API, allows for queries simultaneously

constrained by textual keywords, spatial bounding box, and temporal limits. The API also allows for pagination of search results. In this work, I used the python Flickr API by Mignon (2016). Retrieved data was saved in an SQL Server Database, subsequent processing results (e.g. labelling of photos) were also saved in the same database. Spatial analysis was performed using ESRI ArcGIS and arcpy. I enabled labelling photos by experts in land cover (Chapter 2) and botany (Chapter 4) through a .NET desktop application and a python web application I created especially.

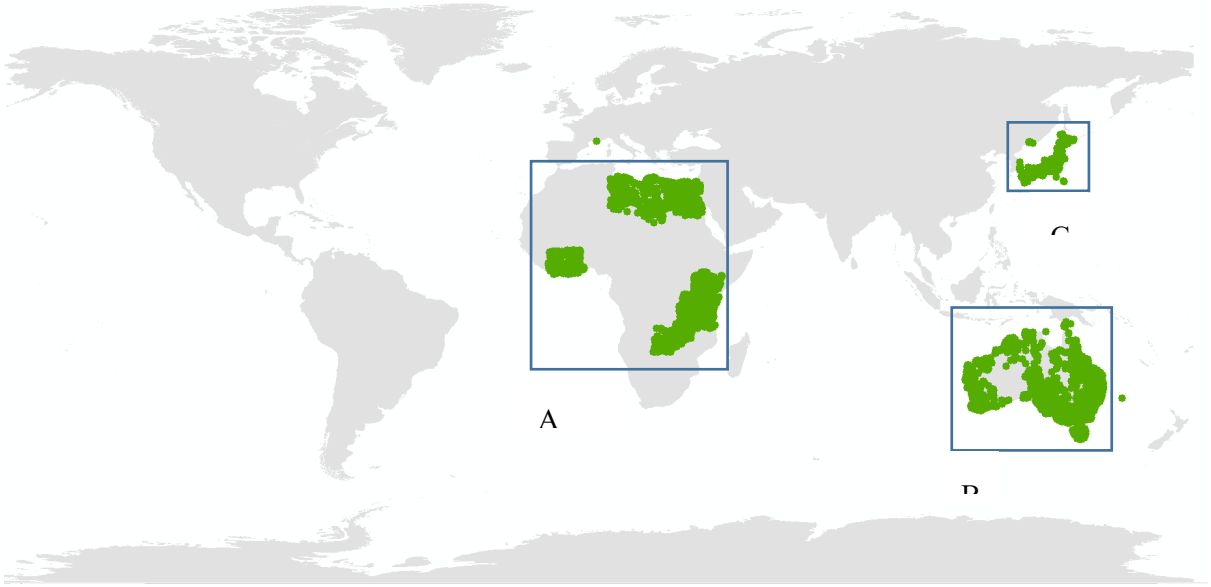


Figure 5 social network sites provide a global coverage, a property much needed in ecological research. 500,000 geo-located photos covering nine countries in three continents were used throughout this work. A) Land cover classification in Chapter 2 uses data from seven African countries. B) Australian data are used in Chapter 3 for spatial mapping of species. C) Data from Japan is used in Chapter 4 for natural phenomena temporal analysis.

2 Thesis Chapter 2 – Spatial analysis: Land cover[2]

2.1 Introduction

Land cover impacts global climate by changing biogeochemical cycles and consequently the composition of the atmosphere, as well as changing the biogeophysical processes that affect energy absorption at the Earth's surface (Feddema et al., 2005). An understanding of changes in land cover and the associated monitoring of human land use such as agriculture, mining, and urban development, also enables us to better grasp human encroachment on natural ecosystems and habitats of pollinator insects. Mapping land cover is therefore essential for understanding anthropogenic climate change and our impact on Earth's ecosystems (Feddema et al., 2005).

Land cover maps are usually created by automatic classification of satellite imagery, a process that results in discrepancies between global land cover products (Fritz et al., 2011; McCallum, Obersteiner, Nilsson, & Shvidenko, 2006). To aid in solving discrepancies, citizen science is a high value resource. For example, the International Institute for Applied Systems Analysis (IIASA) has created Geo-Wiki, an information portal that allows volunteers to improve data on land cover using satellite imagery from Google Earth (Fritz et al., 2017; See et al., 2015). Results show volunteer-contributed data to be generally equivalent to expert data (See et al., 2013).

Satellite imagery interpretation frequently depends on the existence of images taken in situ (i.e. in position), requiring researchers to seek alternative sources of visual data to build a robust model of the environment. Estima and Painho (2013) explored the adequacy of Flickr images to help the quality control of the CORINE land cover (CLC). They concluded there is potential for such use, admitting however that their study has not looked into the content of the images and their adequacy for this purpose. Nevertheless, since Flickr contains such a massive amount of photos, even if only a small fraction are relevant, there is potential to achieve high value outcomes from open data on social network sites by careful filtering, which avoids some of the issues associated with raw image data (Barve, 2014).

Social network sites' geotagged photographs were used in previous research to determine land cover classes. For instance, Estima et al. (2014) compared CORINE Land Cover information obtained from Flickr geotagged photos against classification based on satellite imagery. They concluded that social network site geotagged photos are a valuable supplementary data source. Oba et al. (2014) retrieved

[2] The following paper has been published based on this chapter.

ElQadi, Moataz Medhat, Myroslava Lesiv, Adrian G. Dyer, and Alan Dorin. "Computer Vision-Enhanced Selection of Geo-Tagged Photos on Social Network Sites for Land Cover Classification." *Environmental Modelling & Software* 128 (2020/06/01/ 2020): 104696. https://doi.org/https://doi.org/10.1016/j.envsoft.2020.104696.

photos from Flickr using text tags corresponding to land cover classes. They then classified regions to land cover types based on image feature classification using a support vector machine (SVM), as well as photos' titles and tags. Xu et al. (2017) and Xing et al. (2018) used Convolutional Neural Networks (CNN) to classify geotagged photos from the Global Geo-Referenced Field Photo Library and Flickr into land cover classes.

In this chapter, I aim to support remote sensing scientists' decision making on land cover type, rather than to automate the decision process. This distinction is made primarily because my research problem stems from discrepancies in land cover satellite imagery. Consequently, this chapter provides a practical framework to solve existing problems in land cover classification in understudied regions, rather than suggesting theory to be validated in well-studied regions as was the focus of previous work.

. To achieve these objectives, I developed a framework that uses computer vision to describe image visual content, and an artificial neural network classification model to decide whether or not the image is relevant based on this description.

The main outcome of my study is a reusable framework to find and filter imagery that can help determine land cover types. Therefore, I investigate whether particular countries should have customised models, or whether data from all countries can be used to build one generalised model. To build the suggested framework, human labour is required to label images for training the machine learning algorithm. It is therefore worthwhile testing whether models can be reused in countries other than those from where the training data originates. I developed country-based and generalised models for Africa, and compared the results to establish the most appropriate ways to use them. My framework was integrated into Geo-wiki to empower its scientist users to find relevant imagery that can help them determine land cover type.

Although my framework is independent of the social network site data source, I chose Flickr for the reasons discussed in chapter 1. In the next sections I present my methodology, and discuss the results of the different models I developed. Finally, I present my publicly available API that is allowing other users to invoke the models.

2.2 Methodology

2.2.1 Study area

The chosen study area consists of four African countries with varied geographic, demographic, and economic characteristics: Egypt, Kenya, Zambia, and Côte d'Ivoire (Figure 6). These are all subjects of a practical interest in obtaining land cover photos because although they fall within regions where high resolution data is available, it has previously been noted (Lesiv et al., 2017) that the accuracy of

this data is as low as 65%. Also, these four countries offer valuable case studies due to their variations in climate and geography.

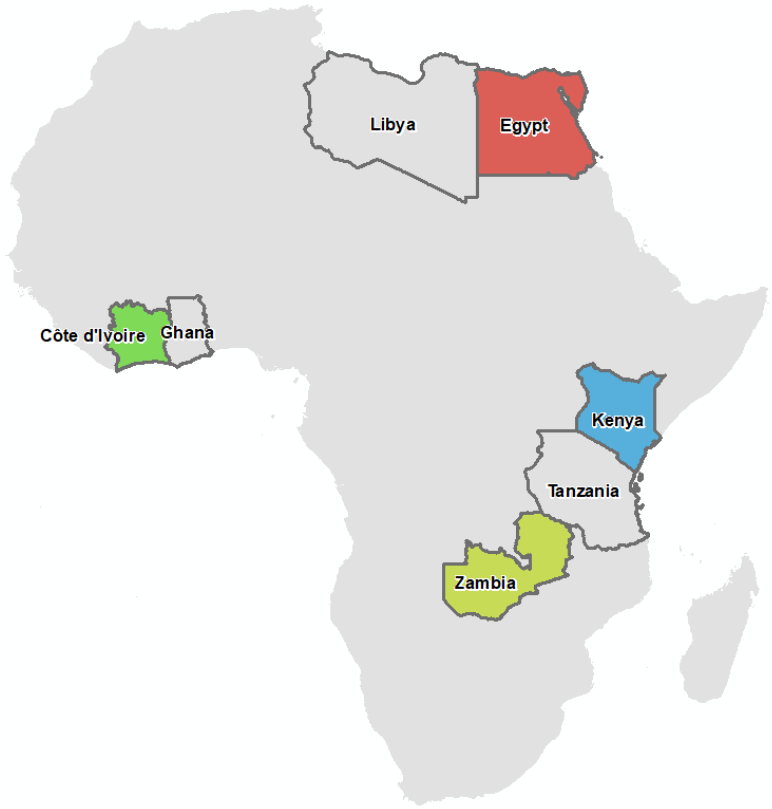


Figure 6 Four African countries with varied climate and geography are used as case studies (Egypt, Kenya, Côte d'Ivoire, and Zambia). Photo classification models from these countries were scored against data from 3 countries neighbouring the original test cases, shown in grey (Libya, Ghana, and Tanzania). Map data from gadm.org

2.2.2 Process overview

Our process uses cloud-based computer vision services to analyse the visual content of geotagged photos on social network sites and generate descriptive tags for that content. I then use these tags to train an artificial neural network to predict a photo's adequacy for land cover classification. Figure 2 demonstrates the workflow. First, I collect photos in the specified study area. I then filter the collected photos to exclude those in urban regions, since built-up land cover is already well-known and mapped, while I am primarily interested in natural and other non-urban regions.

I then select a random sample of photos from the filtered data set that I subject to automatic computer vision classification programs to generate descriptive text tags for the samples. The sample set of photos is also manually labelled by expert researchers as either relevant or irrelevant based on the photo utility to land cover determination (e.g. outdoor, outside settlement areas, containing trees, shrubs, grassland, rocks, sand, or agricultural areas, etc.). The text tags and associated relevant/irrelevant Boolean labels are used to build a classification model that can predict whether a photo described by a certain set of

tags is relevant for land cover classification. The model is later applied to filter photos based on visual content, by users classifying land cover (Figure 8).

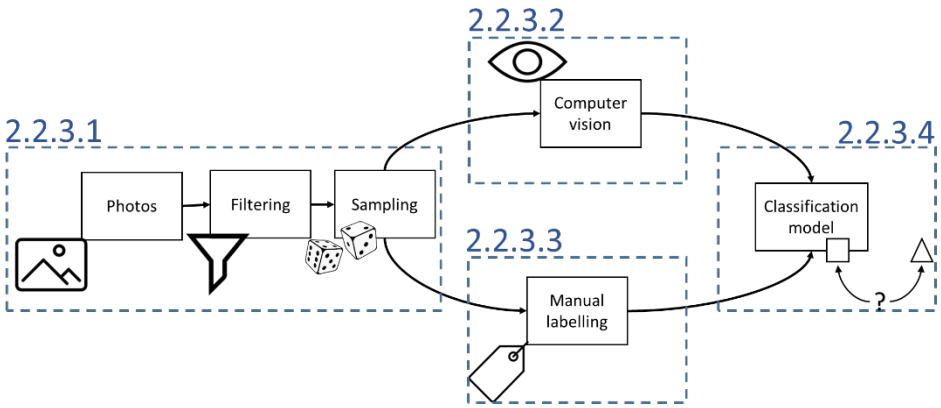


Figure 7 Methodology overview: Building a classification model to predict a photo's relevance to land cover determination. Numbers (e.g. 2.3.1) refer to text sections.

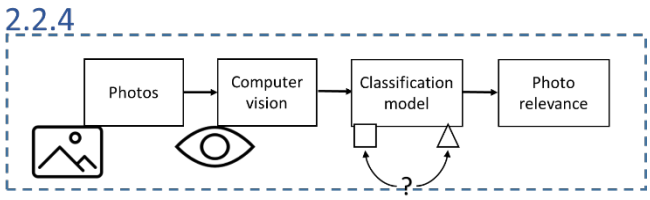


Figure 8 My reusable classification models.

2.2.3 Building the classification models

2.2.3.1 Photo collection

The method I describe can be applied to any social network site that offers an Application Programming Interface (API) to retrieve geotagged photos. In addition to the convenience of Flickr's API already noted, I chose Flickr (www.flickr.com) for its wealth of photos; with at one stage about 2 million new photos being added every day (Michel, 2016).

Flickr can be queried for photos in a particular geographic bounding box. However, for any given query, the API only returns 3600 unique results. In order to collect a larger number of photos, I divided each of my test case countries into a grid, where its cells are then used sequentially as the bounding box for photo queries. To create the grid, the outline of a country is divided into cells of side length of 1 decimal degree. The grid cells intersecting the urban centres are replaced by a smaller grid of a cell side length of 0.2 decimal degrees to allow for high-density retrieval of photos. I then delete the grid cells that are totally contained within urban areas, since I am only interested in land cover in non-urban areas. Grid

cells intersecting the numerous much smaller urban settlements were left to be used in queries. The full set of collected photos in a country is filtered using an urban settlement mask based on the Copernicus land cover dataset (© European Union, Copernicus Land Monitoring Service 2018, European Environment Agency (EEA)), leaving only photos outside registered urban settlements. A sample of the remaining photos is randomly selected for the following steps of visual tagging and data labelling (Table 3).

Although the Flickr API allows searching by location and keyword simultaneously, I didn't limit my queries by any keywords and only searched by geographic bounding box because the text tags supplied by the original image posters can sometimes be misleading (Xing et al., 2018).

2.2.3.2 Computer vision tagging

In this step, I use computer vision to assign tags to the photos based on their visual content. To achieve this, I used the computer vision cognitive services from Microsoft Azure. Microsoft Azure (https://azure.microsoft.com) is a set of cloud-based services, which include "cognitive services" (Microsoft, 2018); machine learning models, and artificial intelligence algorithms. Among these, computer vision is one of the available cognitive services.

In the cognitive services API, the "Image" function can be applied for a given image where more than one tag is returned for the image, with varied confidence levels, with unity being the highest confidence level.

I ran the "Analyse Image" service on every image in the selected sample and only chose the tags with confidence levels higher than 0.5. This is an arbitrary value that corresponds to the computer vision system being more confident than not about the generated tag. A lower threshold would have simply given more tags in which the computer vision system had poor confidence. I excluded images that had no tags with enough confidence from the machine learning model training. For each country, I compiled a list of all tag occurrences for all photos in that country. I then selected the most frequent tags that occur in at least 95% of the records. These selected tags served as my classification model features.

Preliminary analysis in this step showed that the frequent tags are different in each country, despite the fact that photos in all cases were outside urban centres. For example, in Egypt where people primarily visit to see monuments and practice water sports, the tags include: outdoor, sky, nature, person, water, building, ground, indoor, etc. (Figure 9). Conversely, in Kenya, where safari is prominent, the tags include outdoor, grass, animal, field, sky, mammal, tree, etc. (Figure 10).

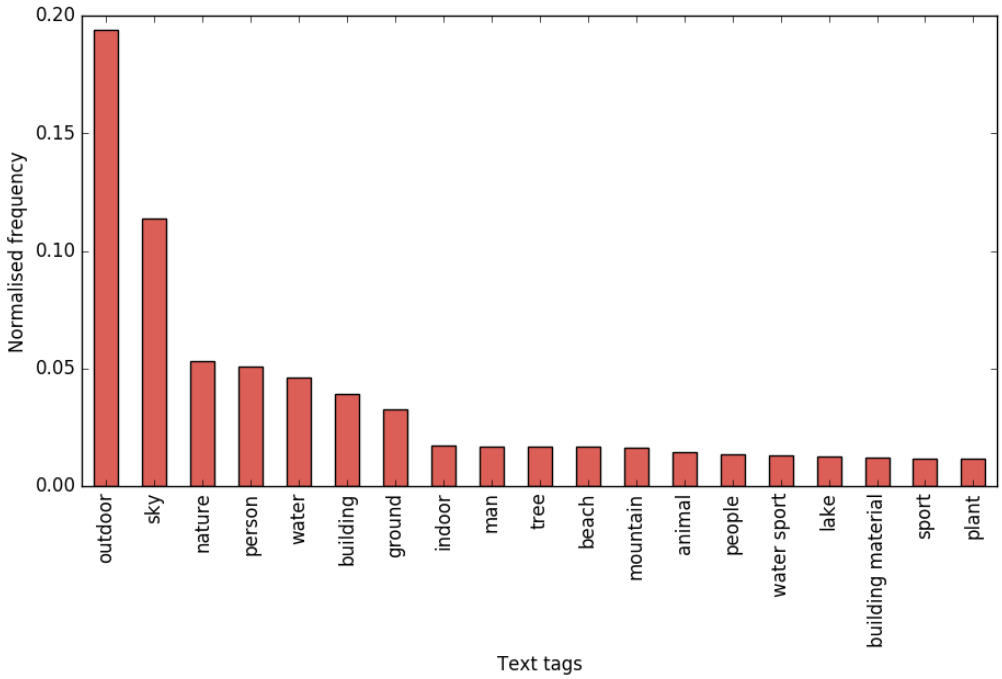


Figure 9 Normalised frequency of text tags associated with a (2%) sample of Flickr photos from Egypt outside urban areas. The tags common among 95% of the data set were selected as classification features

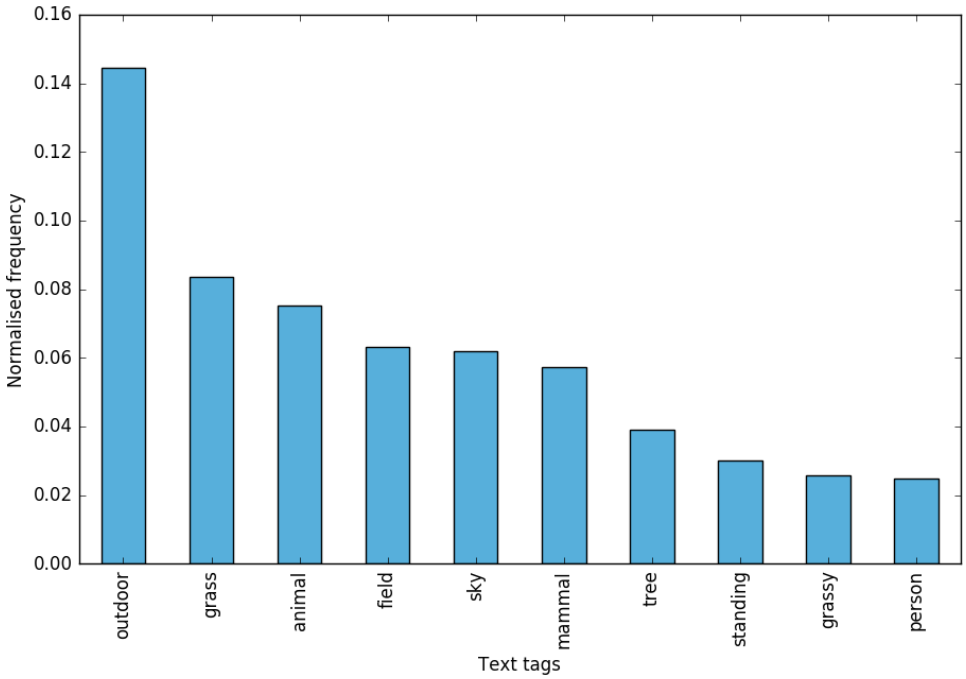


Figure 10 Normalised frequency of text tags associated with a (2%) sample of Flickr photos from Kenya outside urban areas. The tags common among 95% of the data set were selected as classification features.

2.2.3.3 Manual data labelling

The sample set of photos that have been automatically tagged based on their visual content are then manually labelled by expert researchers from IIASA as relevant if they perceived the image content presented potentially meaningful land cover information. Otherwise, the images are labelled as

irrelevant. The relevance ratio is different among countries as shown in Table 4. That these ratios differ among countries may be attributed to the different photographic activity profile of each country as discussed in 2.2.3.2.

2.2.3.4 Classification model training

Artificial neural networks (ANNs), were chosen to build the classification model. ANNs are capable of classifying patterns unseen in training data. They are tolerant to noisy data, and appropriate when little is known about the relationships in input data (Han, Pei, & Kamber, 2012). And they are ubiquitous in the remote sensing community (e.g. (Xing et al., 2018))

In order to train an ANN that can predict whether a photo is relevant based on its visual tags, I created a table containing the label I applied (i.e. whether the photo is relevant or not). Next, the tags identified as selection features in 2.2.3.2 were added to the table as binary columns showing whether each tag was present in each row (i.e. photo). See Table 2.

Table 2 Example rows from the Kenya data set. Every row corresponds to a photo, and represents the presence of the text tags selected in (section 2.2.3.2) in that photo. The first column is the manual label assigned to indicate whether a photo is relevant to land cover (section 2.2.3.3). The classification model is trained to predict the target based on the feature columns (text tags).

Model target	Model features (Text tags)									
relevant	standing	outdoor	grassy	person	Tree	sky	field	animal	mammal	grass
false	false	true	false	false	true	true	false	false	false	false
false	false	true	false	true	false	true	false	false	false	false
true	false	true	false	false	false	false	false	True	true	true
…	…	…	…	…	…	…	…	…	…	…

To build my classification models, I used fully connected neural networks with one hidden layer consisting of n neurons, where n $=\frac{Training\ set\ size}{2*(Input\ features\ +output)}$. The number of neurons in the hidden layer is based on experimentation with the rules of thumb suggested in (Heaton, 2008). My ANNs had a 0.1 learning rate, 0.1 initial learning weight, and a Min-Max normaliser. I randomly selected 70% of the data set to train the model, and the remaining 30% for scoring. The machine learning experiment was

fully implemented in Microsoft's Azure machine learning studio, a browser-based graphical user interface that runs machine learning algorithms and models in the cloud.

2.2.4 Using the classification models

The driver behind this work is the practical need for geotagged photos to help determine land cover in understudied areas. Geo-wiki users are presented with a web interface. This follows the workflow shown in Figure 8, for a given location viewed on the map, images are retrieved from SNS. Next, tags are generated for all photos using the computer vision API. The tags are then sent to my Machine Learning model that responds with a Boolean value whether the photo is relevant to land cover determination. Only relevant photos are displayed to the user.

2.2.5 Experiments

An overview of these experiments is outlined below and in figure 7.

1. For each country in set 1 (Egypt, Kenya, Côte d'Ivoire, and Zambia), I collected images (2.2.3.1), tagged them (2.2.3.2), labelled them (2.2.3.3) and trained a model (2.2.3.4).
2. A union of data, drawn from across set 1 to ensure a spread of representation from each country, was used to train and test a single generalised model through the process outlined in (2.2.3.1 to 2.2.3.4).
3. For each country in set 2 (Libya, Ghana, and Tanzania), which neighbour set 1 countries, I collected images (2.3.1), tagged them (2.2.3.2), labelled them (2.2.3.3) but did not train a model.

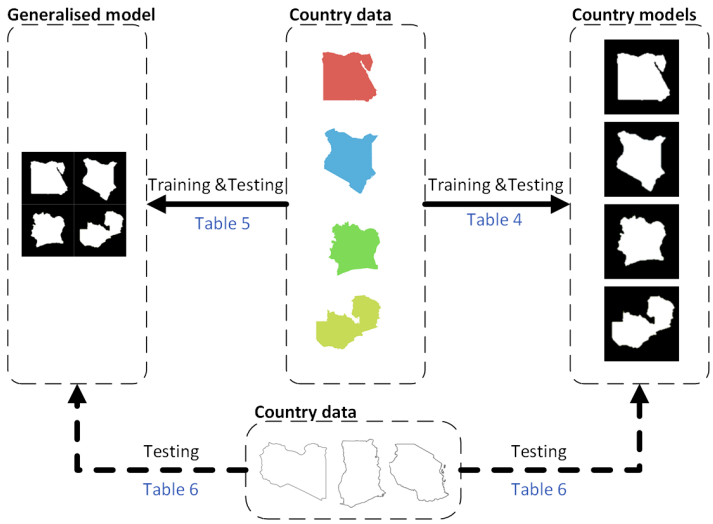


Figure 11 Data from four countries (shown in Figure 6) were used to train and test four corresponding country models. A union of data, drawn from across set 1 to ensure a spread of representation from each of the same four countries, was used to train and test a single generalised model. Data from three neighbouring countries (shown in grey in Figure 6) were tested against the country-specific models and the generalised model.

The first experiment enabled us to test my models in the countries from which the data were obtained. The second experiment tests the performance of a unified, generalised model. The third experiment enabled us to test model reusability. I.e., can a model from one country filter or predict the relevance of data from another?

2.3 Results and discussion

2.3.1 Experiment 1: Country-specific models

The results of the data collection process described in 2.2.3.1 are summarised in Table 3. Here I report "Total number of available photos" which is the count of records reported by the Flickr web interface when searching for geotagged images in a country. “Total collected” is the number of photos collected outside large urban areas using the grid cells I calculated in 2.2.3.1. The API limits the number of downloaded photos in grid cells with high photos density. These would mostly be urban areas. “Non-urban” is the count of remaining records after masking out urban areas. “Sample” is the count of records randomly selected for automatic visual tagging and manual labelling, the sample size is about 2% in Egypt and Kenya. However, since fewer data are available in Côte d’Ivoire and Zambia, the sample size was taken to be 50% and 10% respectively to allow for a representative sample from these relatively small data sets that was sufficiently large to allow for training and scoring.

Table 3 Summary of data collection activities for the four case study countries.

Flickr photos	Egypt	Kenya	Côte d'Ivoire	Zambia
Total number of available photos	580,884	261,382	8,347	38,970
Total collected	103,335	80,617	3,536	21,821
Non-Urban	50,634	69,021	1,098	11,272
Sample	985	1,273	545	1,054

The photos in the sample were tagged and labelled as described in sections 2.2.3.2 and 2.2.3.3. Some of the photos couldn’t be tagged or labelled with sufficient confidence and so were removed from the analysis; hence the difference between dataset size in Table 4 and the sample size in Table 3.

The ratio between the two classification classes (relevant and irrelevant) are shown in Table 4. The table also reports on the configuration of my ANN classification model (section 2.2.3.4), and the performance metrics for classifying the scoring dataset.

Table 4 Configuration and performance metrics for the trained neural networks. TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

		Formula	**Egypt**	**Kenya**	**Côte d'Ivoire**	**Zambia**
Metadata	Dataset size (tagged & labelled)	-	897	960	481	840
	Relevant: Total records	$\frac{Relevant\ record\ count}{Total\ record\ count}$	0.3	0.7	0.4	0.6
Configuration	No. input features	-	19	10	13	12
	No. neurons	$\frac{Training\ set\ size}{2 * (Input\ features\ + output)}$	16	31	12	23
Metrics	Specifity, selectivity	$\frac{TN}{TN + FP}$	0.83	0.72	0.90	0.79
	Accuracy, Recognition rate	$\frac{TP + TN}{TP + TN + FP + FN}$	0.84	0.83	0.84	0.85
	Precision	$\frac{TP}{TP + FP}$	0.67	0.86	0.83	0.89
	Recall, Sensitivity	$\frac{TP}{TP + FN}$	0.86	0.89	0.76	0.88
	F1 score	$2 * \frac{Precision\ * \text{Recall}}{Precision + Recall}$	0.75	0.88	0.79	0.89

2.3.2 Experiment 2: A generalised model

Although the distribution of tags differs by country as I discussed in 2.2.3.2, I checked whether there is sufficient value in spending resources on creating separate models. I thus collated records from every dataset to train a generalised model with records from each country. The number of columns, i.e. selection features, was the superset of all datasets. 70% of the data was used to train the generalised

model. The validation dataset from each country was scored against the generalised model, as well as the original (country-specific) model.

Results shown in Table 5 demonstrate that differences in model accuracy between per-country models and the generalised model are statistically significant when the per-country models perform better than the generalised model (Egypt and Côte d'Ivoire). In cases where the generalised model slightly outperformed the per-country models (Kenya and Zambia), the difference was statistically insignificant.

Table 5 For each country's test dataset: accuracy of individual and generalised models are compared. Also, the dataset results from both models are compared for statistical significance using McNemar's test.

		Egypt	Kenya	Côte d'Ivoire	Zambia
Model Accuracy	Individual	0.84	0.83	0.84	0.85
	Generalised	0.71	0.85	0.71	0.87
McNemar's test	p-value	0.01	0.13	0.004	0.5
	95% C.I.	Significant	Insignificant	Significant	Insignificant

2.3.3 Experiment 3: Set 2 countries

Here I address the question of model reusability: could I use a model from one country to filter (predict relevance of) data from another country? And, given that country-specific models are better than the generalised model, how would the generalised model perform against data from countries not used in model training?

To answer these questions, I collected new test data from a second set of countries: Libya, sharing a border with Egypt; Tanzania, sharing borders with both Kenya and Zambia; and Ghana, sharing a border with Côte d'Ivoire. Samples from these data were manually labelled (156 from Libya, 165 from Tanzania, and 162 from Ghana), then scored using all individual country models, and the generalised model. Results are shown in Table 6. The generalised model's performance with data from a new country is the best, or at least as good as, the model from a neighbouring country.

Table 6 Comparing model accuracy for data from countries not used to build the models. Models from countries sharing a border with the data country in bold.

		Model (Set 1 countries)				
		Egypt	Kenya	Côte d'Ivoire	Zambia	Generalised
Data (Set 2 Countries)	Libya	**0.77**	0.74	0.81	0.50	0.78
	Tanzania	0.55	**0.81**	0.74	**0.81**	0.81
	Ghana	0.73	0.79	**0.80**	0.82	0.83

2.3.4 Related costs and technology

The machine learning logic was implemented in the Azure machine learning cloud service from Microsoft which not only masked many of the low-level details of the neural network that are of little interest to my research, but also helped in the automatic generation of consumable web services allowing Geo-wiki users to benefit from my models as a service.

Our framework weeds out the majority of irrelevant photos (accuracy around 0.8) from the massive amounts of SNS photos. Noting that the ratio of relevant photos in samples I checked was between 0.3 and 0.7 (Egypt and Kenya respectively), my framework is saving researchers valuable time they would otherwise spend browsing many irrelevant photos.

I have opted to include text tags in a descending order of frequency (2.2.3.2), until I had included tags appearing in at least 95% of the records. However, the accuracy of my classification filter can be further enhanced by optimising the number of input features (text tags) participating in the model.

Microsoft cognitive services computer vision is implemented using deep learning (Tran et al., 2016). These are often mis-calibrated (i.e. overly confident in their own results) (Guo, Pleiss, Sun, & Weinberger, 2017). Although Microsoft cognitive services computer vision contains a dedicated module for confidence estimation (Tran et al., 2016), I arbitrarily chose to consider tags with confidence>0.5. Empirical calibration of this confidence cut-off may prove useful to the overall framework performance in the future.

2.4 Conclusions

In land cover mapping, SNS can provide much needed high resolution geotagged photos which can be manually, or automatically, classified to determine the type of land cover in a photo. However, to retrieve and classify geotagged photos from SNS, filtration methods are needed to remove irrelevant photos from the classification pipeline. In this work, I suggested, developed, and tested a framework to filter SNS photos relevant to land cover classification.

Our framework uses commercially available APIs to favour simple implementation by interdisciplinary researchers who may need to reproduce it. Technology is improving over time, I am sure the methods and frameworks I suggest in this research would be improved further when paired with more capable computer vision and machine learning in the future.

3 Thesis Chapter 3 – Spatial analysis: Pollinator and flower distribution[3]

3.1 Introduction

Databases of species occurrence, such as the Global Biodiversity Information Facility (GBIF, www.gbif.org), can play an important role in research on the effects of climate change and habitat alteration on pollinator availability. GBIF is a data source based on biodiversity records of participating institutions and governments, but there are often inconvenient gaps in its data (Robert P. Anderson, 2016). For instance, some literature (Beck, Ballesteros-Mejia, Nagel, & Kitching, 2013) specifically addresses the inventory completeness of a tropical insect, hypothesising that it is impacted by human factors including "road and tourism infrastructure, habitat encroachment, population density, conflict and colonial history". Filling such gaps is an important challenge for both biologists and information scientists.

This chapter aims to use SNS data to patch spatial and temporal gaps in specialised species occurrence databases (e.g. GBIF). In addition, SNS geotagged photos can help overcome the occasional lack of supporting photographs or video in GBIF species occurrence data since such image-based media can be very useful. For instance, images may enable ecologists to determine attributes of the specimens that are not reported in the textual data.

The potential to better inform or supplement biodiversity research with social media images may have widespread research value. Such a framework may also lead to improved or targeted use of public science for contributing to important research questions.

In this chapter, I use a novel, technically accessible approach to test the relevance of geotagged image content to search keywords (section 3.2). The filtered images, together with Atlas of Living Australia, ALA, species distribution data, contribute to the construction of land map overlays. Both sets of data are compared to investigate how the filtered Flickr images complement the available information on species occurrence obtained from ALA (section 3.3). I outline my methodology below, and then present

[3] The following paper has been published based on this chapter.

ElQadi, Moataz Medhat, Alan Dorin, Adrian Dyer, Martin Burd, Zoë Bukovac, and Mani Shrestha. "Mapping Species Distributions with Social Media Geo-Tagged Images: Case Studies of Bees and Flowering Plants in Australia." Ecological Informatics 39 (5// 2017): 23-31. https://doi.org/http://dx.doi.org/10.1016/j.ecoinf.2017.02.006.

my results for test cases, along with a discussion of the strengths and weaknesses of this new research tool (section 3.3).

3.2 Methodology

In overview, I searched for geotagged images in Flickr using species' common and scientific names. These images were subsequently fed into Google's reverse image-search to find tags that best describe the content of these images. These tags were next used to exclude images deemed irrelevant to the studied species. The scientific name of the species is used to search ALA (ala.org.au), henceforth referred to as ALA, a data source that is a participant node of GBIF. The filtered images from Flickr, and the occurrences from ALA, are overlayed on a geographical map to allow us to draw comparisons and conclusions. An outline of the methodology is depicted in figure 12 and each step is detailed below.

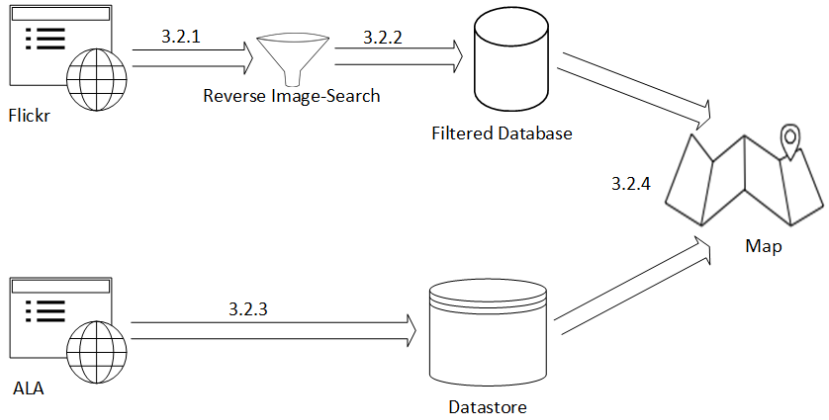


Figure 12. Overview of the process: Data from Flickr are filtered using reverse image-search. This filtered data is overlayed with the ALA data for further analysis. Numbers (e.g. 3.2.1) refer to text sections.

3.2.1 Flickr image retrieval

Flickr was selected as the social network site with which to test my study for the reasons previously discussed in chapter 1. The image URLs were saved to a database along with the specific search keywords used, and image location coordinates.

3.2.2 Image content check

Images returned from search queries in Flickr may not always be related to the species sought, for example, a search for "honey bee" could return images for honey jars in the result. For this reason, image content requires a separate stage of validation. I used Google's reverse image-search, a tool for finding images that are visually similar to an input reference image. As a result of this search, both a text-label estimate of the image subject, and a set of visually similar images are returned as potential matches. Using Google search's result, I can thus differentiate between images of Honeybees, and, for example, images of jars of honey. The relevance of each member of the set of images previously

returned from Flickr, can now be assessed by referring to its associated text tag. My filtering is therefore a useful way to remove images that may be related to the search keyword in ways irrelevant to the ecological study goals.

3.2.3 Obtaining reference data

Data on species distribution are usually collected by experts or citizen scientists, then subsequently vetted by experts, before being standardised and published. Due to their quality and the recognition by researchers, I benchmark my data against available data for Australia obtained from ALA, a node of GBIF.

3.2.4 Geographic map overlay

In order to visually compare the obtained data, ArcGIS software (ESRI, 2016) was used to create geographic maps. The locations associated with the filtered geotagged images (section 3.2.2) and ALA data (section 3.2.3) were plotted on the map. The ALA reference points were buffered to create polygons on the map extending a distance of 100 Km around species' occurrence points. The 100 km buffer is a coarse proxy for the variable maximum coordinate uncertainty in the obtained ALA reference data. It has been chosen for clarity, keeping in mind the scale of the map on which I have plotted the data. Actual uncertainty range values reported in the GBIF dataset were: 100 km Blue-banded bee, 125 km Sturt's Desert Pea, 125 km Pink Heath, 10 km Honeybee. The Flickr image locations were then overlayed as shown in figures 23, 24, 25, and 26. Results are discussed in section 3.3.

3.2.5 Case study selection

The methodology described in section 3.2 was applied to four case studies: two insect pollinators and two flowering plants. The pollinators were the Honeybee (*Apis mellifera*: Apidae), Figure 13, and the Blue-banded bee (*Amegilla cingulata*: Apidae), an Australian native pollinator, Figure 14. The flowering plants were Sturt's Desert Pea (*Swainsona formosa*: Fabaceae), Figure 15, and Pink Heath (*Epacris impressa*: Ericaceae), Figure 16, both native Australian species.

Figure 13 A Honeybee (Apis mellifera: Apidae). ©Andreas Trepte [CC BY-SA 2.5 (https://creativecommons.org/licenses/by-sa/2.5)]. https://upload.wikimedia.org/wikipedia/commons/4/4d/Apis_mellifera_Western_honey_bee.jpg

Figure 14 A Blue-banded bee (Amegilla cingulata: Apidae). ©Chiswick Chap [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)]. https://upload.wikimedia.org/wikipedia/commons/b/b5/Amegilla_cingulata_on_long_tube_of_Acanthus_ilicifolius_flower.jpg

The two insects were chosen because both are reasonably distinctive but differ in abundance, distribution, and behaviour. The Honeybee is an important pollinator of agricultural and horticultural crops and natural ecosystems around the world, and has almost iconic public recognition from school aged children to seniors in the community. The Honeybee is an introduced, but abundant and economically important pollinator on mainland Australia, the domain of my current study. Additionally, the Honeybee is relatively slow moving, making it an easy and popular subject for amateur photographers. The Blue-banded bee by contrast, although relatively common in many parts of Australia, forms solitary nests rather than large colonies and so is not nearly as abundant as the Honeybee. Perhaps consequently, it is not well recognised by the general public, and often moves fast. Also, the Blue-banded bee is not currently well known as a pollinator of global significance, although there is interest in how it might be employed as a pollinator in some circumstances (Hogendoorn, Coventry, & Keller, 2007; Switzer, Hogendoorn, Ravi, & Combes, 2016). Hence, it makes an interesting contrast against which to compare the utility of Flickr user images for the purposes of collecting species range data.

Figure 15 Sturt's Desert Pea (Swainsona formosa: Fabaceae). Photo by Blueday at English Wikipedia [Public domain].

Figure 16 Pink Heath (Epacris impressa: Ericaceae) near Great Otway national parks, Victoria, Australia. Photo courtesy of Alan Dorin.

The floral case studies were chosen from the Australian states' emblems. Each of the seven states and two territories into which Australia is divided has a floral emblem that is loosely associated with the region. The South Australian coat of arms, Figure 17, features the Sturt's Desert Pea, has a bright red flower with very unusual structure. Pink Heath, by contrast, shown in the Victorian coat of arms, Figure 18, is certainly bright and striking *en masse*, but structurally it is not as iconic as the South Australian emblem. Both are short perennials of open woodlands, but Sturt's Desert Pea occurs more frequently in the continental interior while Pink Heath is found in cooler habitats in the southeast. These flowers, perhaps surprisingly to an outsider, are not uniformly well known to the public, despite their official status as state emblems. Because the two species differ in the degree of visual spectacle they offer, Pink Heath is a less popular subject for photography than the Sturt's Desert Pea, making the combination of the two floral emblems interesting cases for this study.

Figure 17 State of South Australia coat of arms showing Sturt's desert peas at the top.

Figure 18 State of Victoria coat of arms, showing pink heath growing from a grassy mound.

3.3 Results

In section 3.3.1, I present the results of applying my image content filter (section 3.2.2) to my case studies (section 3.2.5). The findings obtained from laying the data over geographic maps (section 3.2.4) are presented in section 3.3.2.

3.3.1 Image content validation

Flickr was searched for my case studies using the scientific and common names shown in Table 7. The table also indicates the number of images returned by each search, and the percentage of these misclassified as false positives and false negatives using the image text tags returned by Google's reverse image-search. The image text tags' frequencies are plotted in detail in Figures 19, 20, 21, and 22 for each case study.

In the case of the Honeybee, the five most frequent image text tags accounted for 69% of all images returned. By visual inspection of the images themselves, these tags were determined to match the intent of the search. In the case of the Blue-banded bee, the six most frequent tags were intuitively related to the search term, except for the tag "performance" which related instead to a performance of Debbie Harry (Blondie) and her rock band. The remaining five relevant tags accounted for 68% of all images.

Table 1 highlights the extent to which the non-expert community of photo-sharers chooses to label their uploads using common name labels, rather than scientific names, as well as the variation in this practice (from 2% for Sturt's Desert Pea to 28% for the Pink Heath). Hence, at this step in the process of data collection from the SNS, it seems worthwhile to conduct a search using both types of name.

Table 7. Flickr search results for Australian case studies[4]. Search terms were species scientific and common names. Percentages are provided for the fraction of images returned using scientific names over the total number of returned images.

[4] Search dates: 1/Jun/2016 Honeybee, 16/Jun/2016 Blue-banded bee and Sturt's Desert Pea, 2/Oct/2016 Pink Heath.

"False positive %" refers to the percentage of images included in the data set after filtering using Google reverse image-search that were then determined by visual inspection not to be photographs of the target species. "False negative %" refers to the percentage of images excluded from the data set after filtering using Google reverse image-search that were then determined by visual inspection to be photographs of the target species.

Scientific name (No. of images)	Common name (No. of images)	Images returned by scientific name %	No. of images included by tag (False positive %)	No. of images excluded by tag (False negative %)
Apis mellifera (50)	honey bee (759)	6%	559 (12%)	250 (13%)
Amegilla cingulate (23)	Blue banded bee (342)	6%	248 (13%)	117 (27%)
Swainsona formosa (7)	Sturt pea (419)	2%	310 (3%)	116 (34%)
Epacris impressa (103)	Pink Heath (267)	28%	223 (43%)	147 (13%)

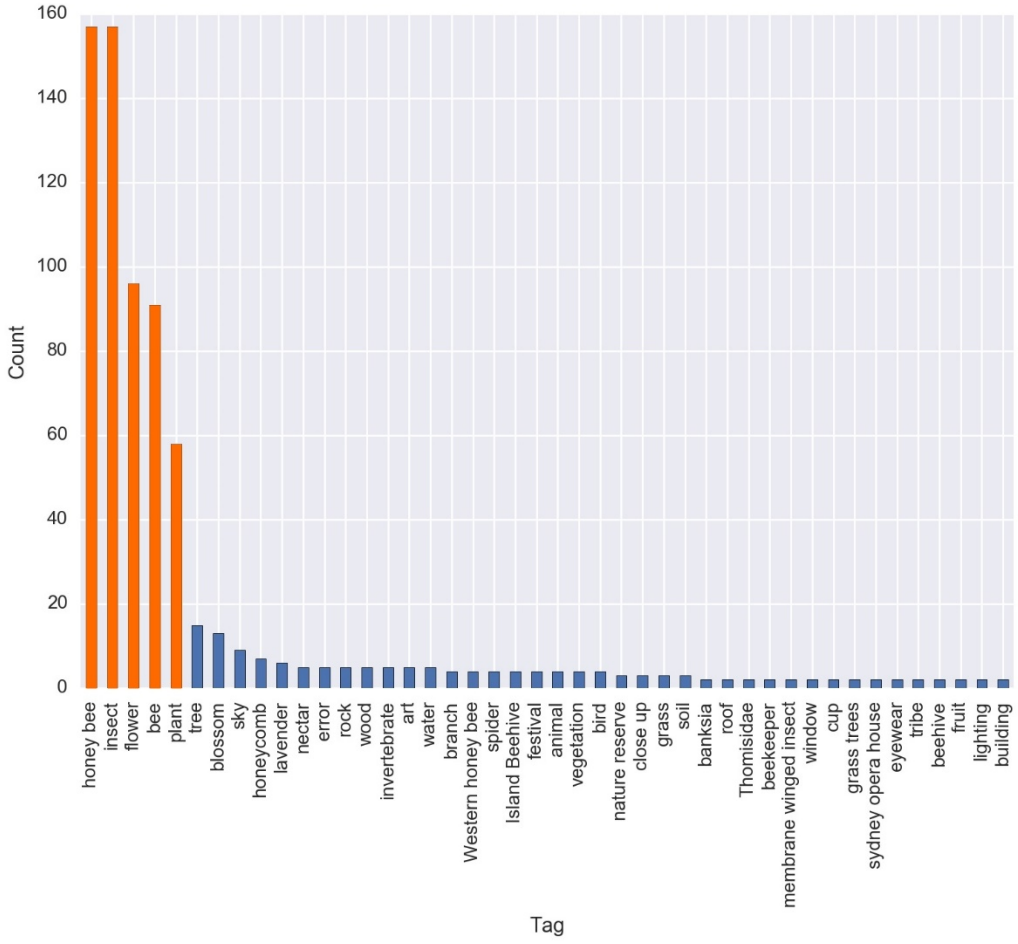


Figure 19. Flickr was searched for images of the Honeybee (Apis mellifera). The results were passed to Google reverse image-search. The frequencies of tags returned by Google from this image set are shown. The five most frequent tags (in orange) are deemed relevant by human inspection, they account for 69% of the 809 images returned. Note that a long tail of tags with frequency 1 is not shown in the graph.

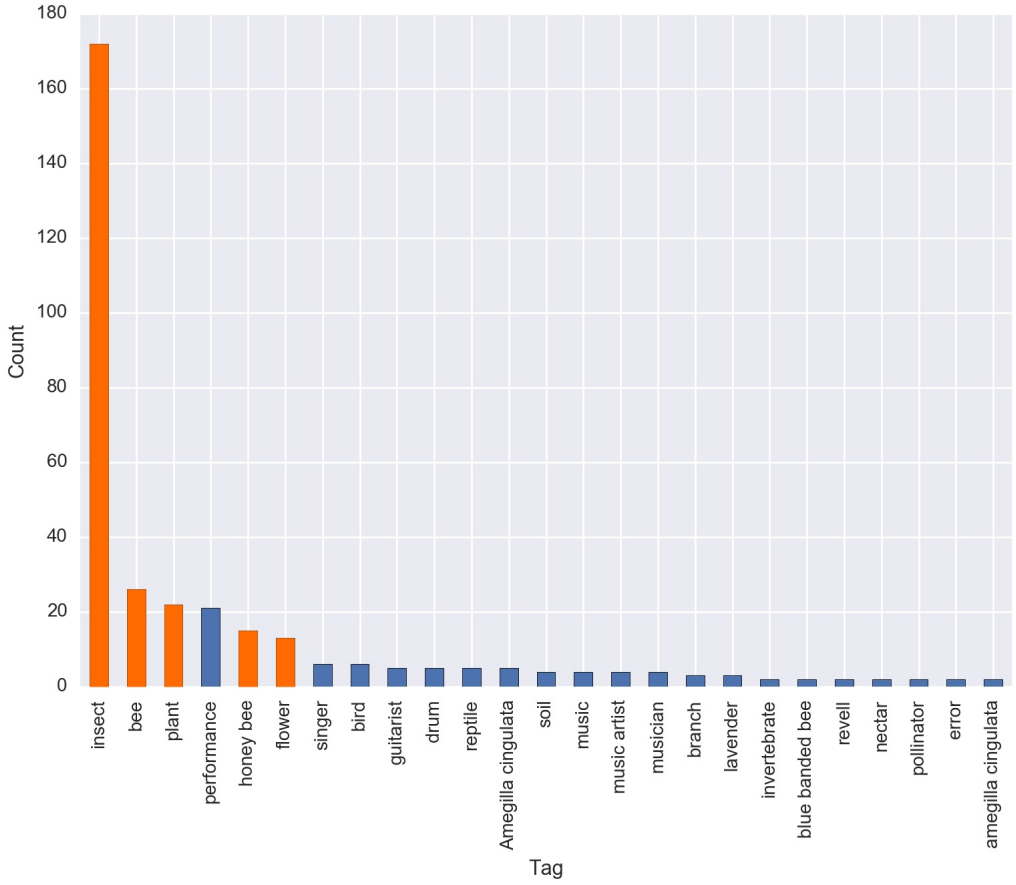


Figure 20. Flickr was searched for images of the Blue-banded bee (Amegilla cingulata). The results were passed to Google reverse image-search. The frequencies of tags returned by Google from this image set are shown. Five out of the six most frequent tags are deemed relevant by human inspection (in orange), they account for 68% of the 365 images returned while the irrelevant "Music Performance" tags successfully excluded account for ~13% of all images.

A search for Sturt's Desert Pea returned 426 images. Google reverse image-search classified a 73% majority of these into two categories, "plant" and "flower", and I visually confirmed that each category related to my search intent. Applying the same process to Pink Heath yielded 370 images, with a 60% majority divided between the four tags "plant", "flower", "Epacris impressa", and "Epacris", in decreasing order of frequency.

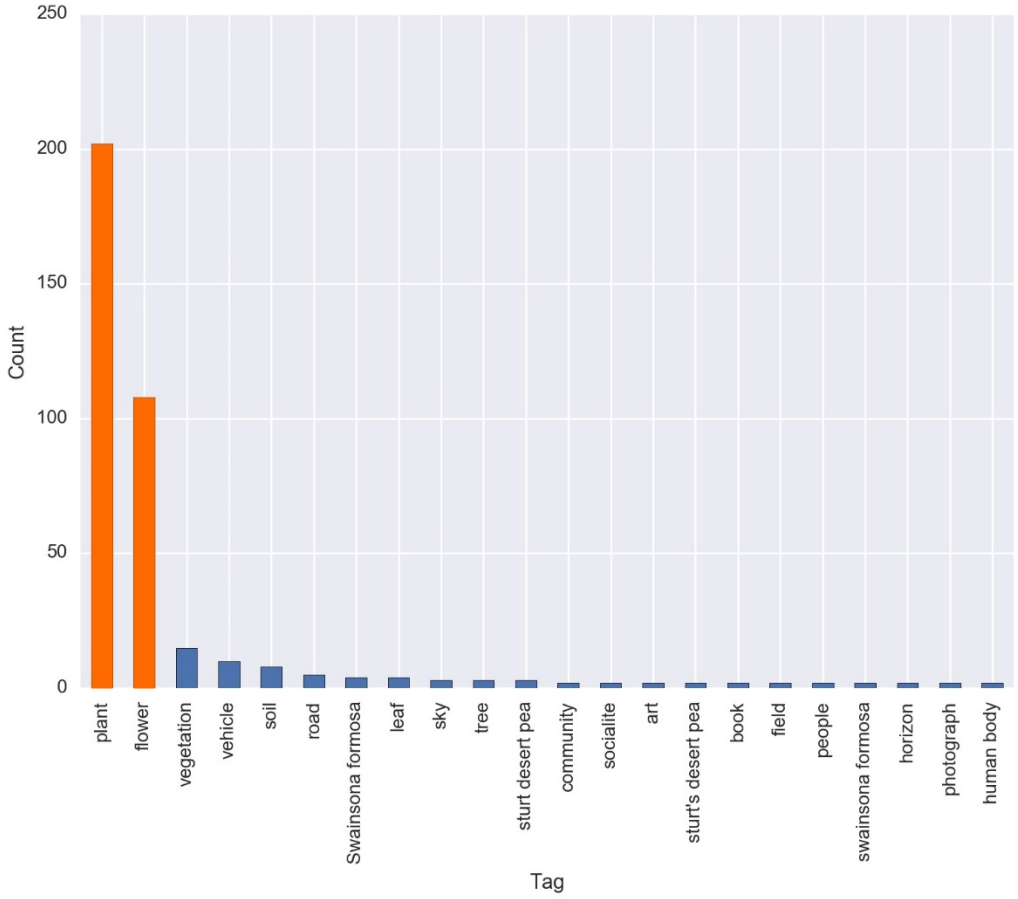


Figure 21. Flickr was searched for images of Sturt's Desert Pea (Swainsona formosa). The results were passed to Google reverse image-search. The frequencies of tags returned by Google from this image set are shown. The two most frequent tags (in orange) are deemed relevant by human inspection, they account for ~73% of the 426 images returned.

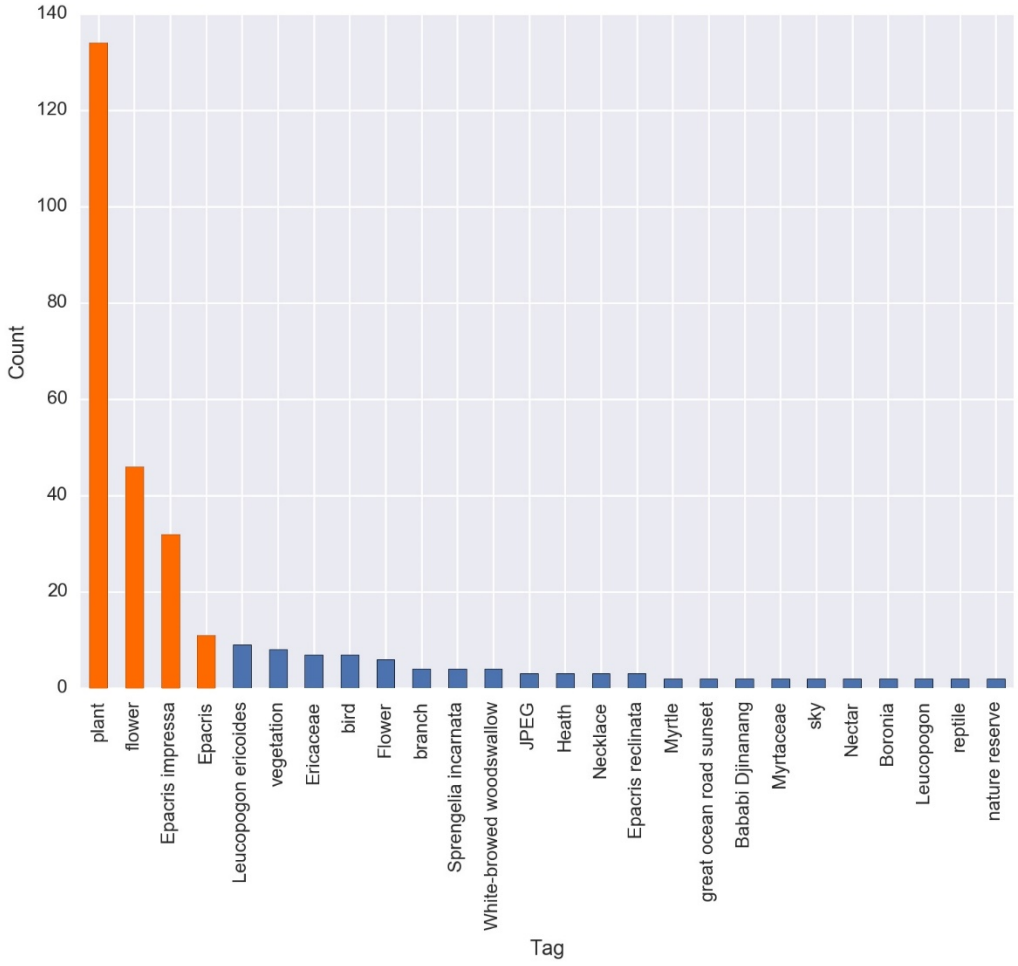


Figure 22. Flickr was searched for images of Pink Heath (Epacris impressa). The results were passed to Google reverse image-search. The frequencies of tags returned by Google from this image set are shown. The four most frequent tags (in orange) are deemed relevant by human inspection, they account for 60% of the 370 images returned.

3.3.2 Geographic Results

Data obtained from ALA and from social network site Flickr were overlaid on a map of Australia for each of the species in my case studies. Data points obtained from Flickr appear to expand the ALA-based ranges. For instance, the Honeybee map (Figure 23) shows that the Flickr data generally confirm the ALA data, but extend these data on the East coast (close to Australian urban centres) and in the centre of the continent (near Alice Springs, a popular tourist destination, remote from major cities).

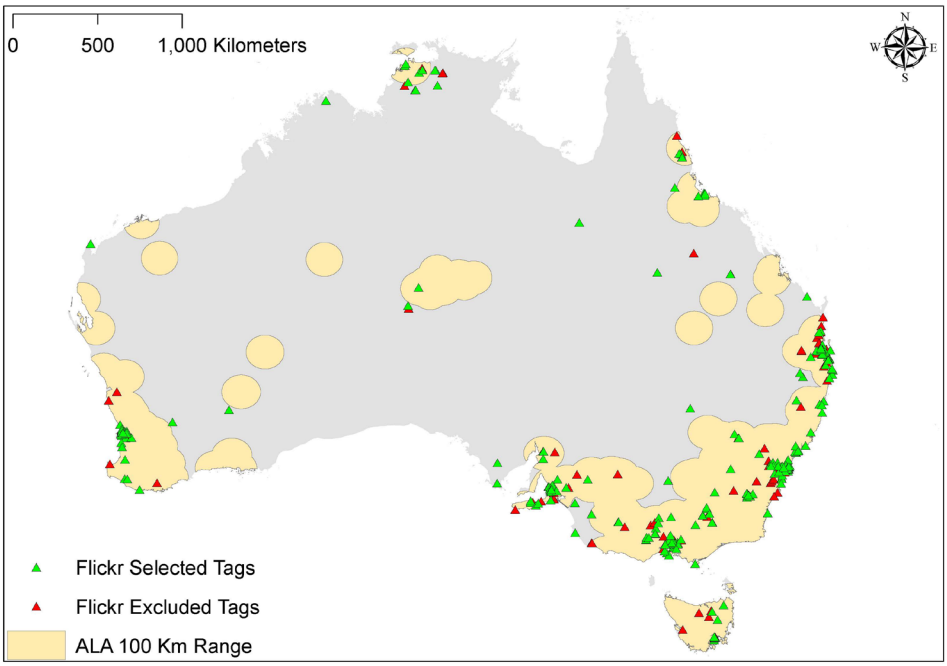


Figure 23. Honeybee (Apis mellifera) images obtained from Flickr (triangle symbols) mostly fall within a 100 km region around ALA reference points (region in yellow, ALA points not shown).

In the Blue-banded bee map (Figure 24), the Flickr points generally fall in the range known from ALA close to the coast, while suggesting new habitat in the centre, north and south of the continent. I have confirmed by visual inspection that the images in these locations warrant inclusion as insect sightings. Interestingly, the occurrence marked with a square in Figure 24, was of a Neon-cuckoo bee (*Thyreus nitidulus*) which is parasitic on the Blue-banded bee (Cardale, 1968). Hence, its appearance may point to the presence of the Blue-banded bee.

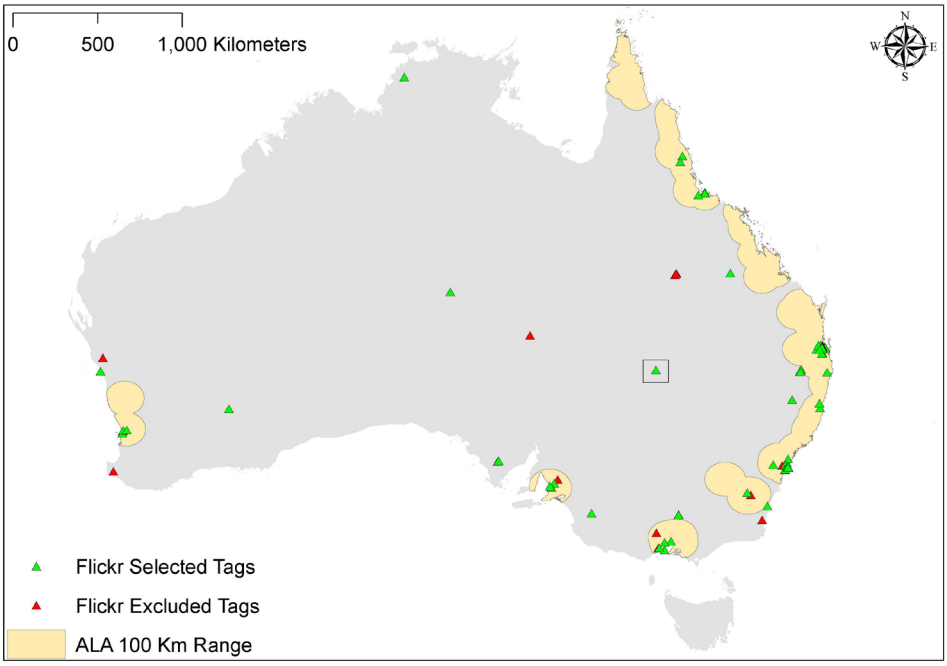


Figure 24. Blue-banded bee (Amegilla cingulata), images obtained from Flickr (triangle symbols) mostly fall within a 100 km region around ALA reference points (region in yellow, ALA points not shown). However, The Flickr filtered images showed occurrences outside the coastal regions marked in ALA (Yellow). The Flickr point inside the square is created from an image of a Neon-cuckoo bee (Thyreus nitidulus) which is parasitic on the Blue-banded bee.

Flickr data also extend the Sturt's Desert Pea range in the centre of the continent (Figure 25). However, although the data suggest the existence of the plant in Victoria, the images in Victoria were examined individually and subsequently found to be located in gardens and museums. Similarly, one specimen in the Northern Territory was found to be a potted plant at a roadside "Holiday Park". In the case of Pink Heath, Flickr data generally coincides with the ALA range in Victoria and Tasmania, but a few images were registered in locations far from the ALA ranges. These images were confirmed to be false matches that happened to have the words 'pink' and 'heath' in their Flickr tags without being specimens of Pink Heath.

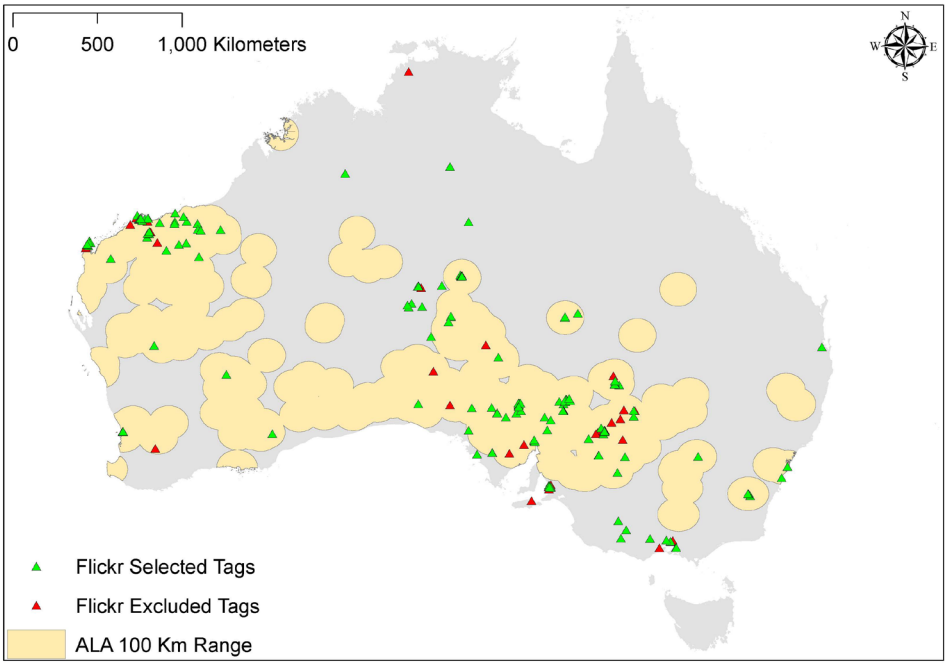


Figure 25. Sturt's Desert Pea (Swainsona formosa) images obtained from Flickr (triangle symbols) mostly fall within a 100 km region around ALA reference points (region in yellow, ALA points not shown). However, although ALA occurrences are missing in some areas in the centre of the continent, Flickr geotagged images cover those regions.

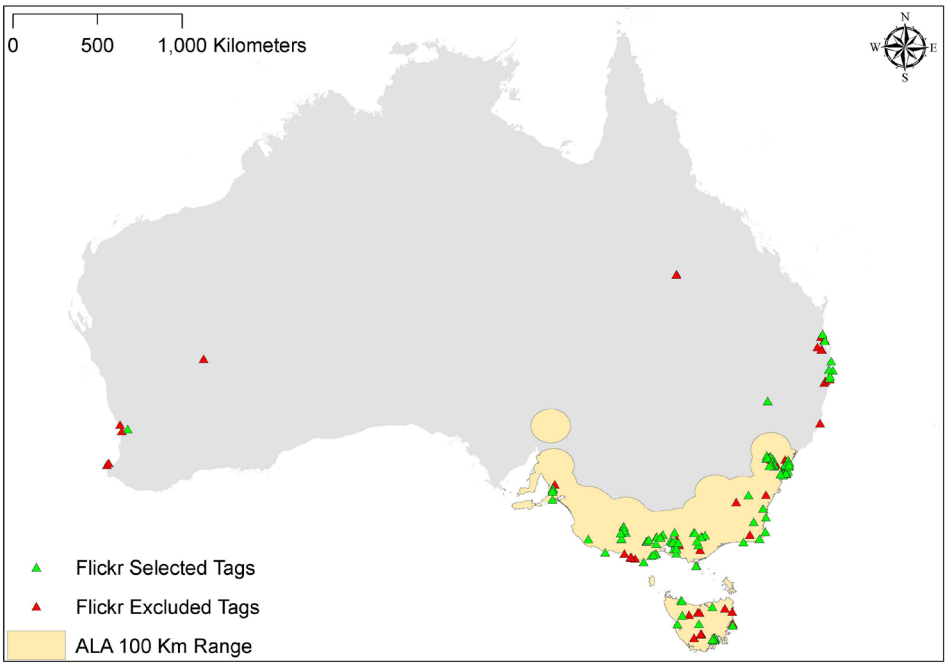


Figure 26 Pink Heath (Epacris impressa) images obtained from Flickr (triangle symbols) mostly fall within a 100 km region around ALA reference points (region in yellow, ALA points not shown). Flickr points far from ALA regions were visually confirmed to be false positives due to similar text keywords.

3.4 Discussion

Our results suggest that the SNS data can be helpful in determining the existence of species in certain regions. In particular, geotagged images can be used to detect or verify the occurrence of a species outside its previously documented or inferred range. Tools for obtaining and filtering such information have many potential uses in a world in which species ranges are shifting at unusually high rates due to climate change (Chen, Hill, Ohlemüller, Roy, & Thomas, 2011). Timely detection of expanding ranges, for example, can assist in biocontrol and other management efforts directed at invasive species (Fagan, Lewis, Neubert, & Van Den Driessche, 2002).

Our results also show that selecting images identified by the most frequent tags is generally a useful approach, with an error rate (false positive identification) that varies from as little as 3% in the case of Sturt's Desert Pea (*Swainsona formosa*) to 43% for Pink Heath (*Epacris impressa*) (Table 7). Selecting images with the most frequent tags is certain to overlook other tags that are less-frequent yet relevant. This is evident in the case of the Blue-banded bee where 27% of the images excluded based on the tag filtering were found by visual inspection to, in fact, be relevant to my search (Table 7). This is not surprising when I notice that "Ameguilla cingulata" and "blue banded bee" are among the less-frequent tags reported in Figure 20. Similarly, in the case of Sturt's Desert Pea (*Swainsona formosa*), Figure 21 shows that "Swainsona formosa" and "Sturt's desert pea" are among the infrequent, hence excluded, tags. These tags' exclusion by my simple approach accounts for the high number of false negatives reported in Table 7. My goal is to devise a method that facilitates excluding irrelevant images, based on

their tag frequency. If unnecessary exclusion of relevant images is of a concern in an application, for instance in cases where data are rare, I would advise including images tagged with the search keywords (being scientific and common names) in the "relevant" image set, even if these tags were infrequent.

Taxonomic accuracy of SNS images remains a concern (e.g. (Stafford et al., 2010)), but my experience indicates that visual validation by a specialist can greatly reduce the error rate. The pre-filtering techniques I have proposed assist to save time in this aspect of the process. The combination of machine filtering and human verification may be especially viable for targeted uses of SNS data, such as monitoring efforts directed at a single or limited number of species. In common with others (Barve, 2014; Daume, 2016) therefore, I see considerable potential in social media sources of biodiversity information and a need to develop tools to realise that potential.

SNS images can provide an inexpensive and abundant source of geographic occurrence data. Even limited data on species occurrences may, in conjunction with models of range expansion (Fagan et al., 2002) or range-abundance relationships (Gaston et al., 2000), greatly assist basic research and applications in ecology, conservation and environmental management.

I found Google reverse-image search to be a useful tool, but not at the level of distinguishing between similar insect species unless a separate validation stage was implemented. For instance, it provided the tag "honey bee" in response to an input image of a Blue-banded bee (Figure 3). I confirmed visually that the images on Flickr were in fact Blue-banded bees.

3.5 Conclusion

Social network sites provide a potential wealth of geotagged images that many researchers could use to complement existing ecological information. However, since social network sites are often used by non-specialists, they may suffer from two major problems including that common names are frequently used to describe species which might by synonymous with other objects or events; and uploaded images can be misclassified by non-specialists.

Our new findings suggest that checking the image content using Google reverse image search is useful in filtering out images broadly unrelated to the species sought. However, the method is not sufficiently fine-grained to distinguish between species. Expert human validation is still needed for reliable classifications. Despite this, the effort to manually confirm species classifications among the potentially large amounts of data I have available is relatively minor given my approach of identifying relevant tags. Future image classification systems may improve classification reliability further, and at any rate, the use of human expertise for validation of images is more cost efficient than sending experts to the field, especially in cases when distances between study sites are great.

I have shown that the filtered geotagged images from Flickr can in fact complement existing data in many cases, by providing valuable data in locations where the existing data is thin. This research method can also reveal the presence of species in areas not previously considered, allowing for improved planning for more traditionally focussed ecological methodologies.

4 Thesis Chapter 4 – Temporal analysis: Japan's cherry blossoms[5]

4.1 Introduction

There are a number of ecological climate-influenced phenomena that are highly visible and of ongoing public interest. Some of these are the subject of tourism-related activities, others are the "poster children" of television nature documentaries. Examples include the "Penguin Parade" of little penguins in southern Australia and their diurnal return to the beach after feeding (Dann & Chambers, 2013), the mass migration of the monarch butterfly (Diffendorfer et al., 2014), the tourism surrounding autumn leaf viewing in the USA, Canada and Japan ("leaf peeping", or in Japan, "momijigari") (Liu, Cheng, Jiang, & Huang, 2019), and, the spring bloom of cherry blossoms in Japan ("hanami") (Liu et al., 2019). All of these phenomena are potentially helpful events by which to assess the impact of climate change on organism behaviour. In these cases, their general popularity generates a strong footprint online consisting of photographic records and social media posts, many of which are geotagged. Such postings can in fact be used for incidental citizen science purposes to monitor seasonal ecological phenomena – specifically, as I detail below, I examine the phenology of cherry trees, a traditional cultural hallmark of spring in Japan, through the lens of Flickr users in Tokyo.

Climate variation affects plant phenology (Morton & Rafferty, 2017; Parmesan, 2006) and, although precise empirical data is typically difficult to collect, especially over long time frames, the timing of plant phenological events can therefore be an indicator of environmental change. One such indicator could be the cherry tree of Japan. Thanks to their cultural significance, records of blooming activities of Japan's cherry trees have been preserved for centuries in diaries and chronicles (Aono & Kazui, 2008). Consequently, this long history of written records of cherry tree blooms in Japan was used to reconstruct seasonal temperature variation over several centuries in Tokyo (Aono, 1998), (Aono, 2015) and Kyoto (Aono & Kazui, 2008).

Today, the daily thoughts and pictorial mementos of citizens today are documented online more abundantly than they were in hand-written chronicles and diaries of years gone by. Consequently, I would expect to uncover a mass of SNS data on highly popular ecological events; Japan's cherry tree blooms being a case in point due to their cultural significance, visibility, and status as a national and international tourist attraction. The potential of SNS data for analysing current bloom activity has previously been demonstrated by Endo, Hirota, and Ishikawa (2018), who used geotagged tweets and data interpolation methods to provide real-time information on the best times and locations to observe cherry blossoms in Japan. Instead of tracking blooming events in real time, I track the historical blooming patterns of the trees. To generate this information, I utilise the flowering record encapsulated

[5] A paper based on this chapter is under preparation.

in Flickr photos 2008-2018, and compare my estimates with the full bloom dates published by the Japan National Tourism Organisation (JNTO), sourced from the Japan Meteorological Corporation (JMC). I build on my method from previous chapters which incorporates both temporal and spatial factors, producing more direct assessments of flowering plant phenology. After establishing the fidelity of SNS data to the real-world blooming phenomenon, I investigate a data-anomaly that suggests cherry blossoms have been consistently blooming in both spring, and surprisingly, autumn, over the past decade.

4.2 Methodology

Our approach to collecting cherry blossom flowering event data can be divided into two steps 1) data extraction from the social network site, entailing several distinct sequential stages, and then 2) filtration of data for relevance using increasingly finer resolutions. I report in this section on the results of each intermediate stage in order to simplify the explanation of my filtration process. Results obtained from analysis of the final filtered data set are provided in the Results section below (section 4.3).

4.2.1 Initial SNS site search.

Although my methodology is agnostic to the data source, as noted in earlier chapters, I chose Flickr for my initial data collection process by searching for the keyword "cherry blossom" using the python API client (Mignon, 2016). My search was confined to photos taken in Japan, geo-located within a bounding box (lat. 31.186°– 46.178°, and long. 129.173°– 145.859°), in the time span 1 January 2008 to 31 December 2018. This data was then filtered by masking against the geographic boundary of Japan obtained from gadm.org. A total of 80,915 photos remained, their distribution in space and time are shown (*Figure 27*).

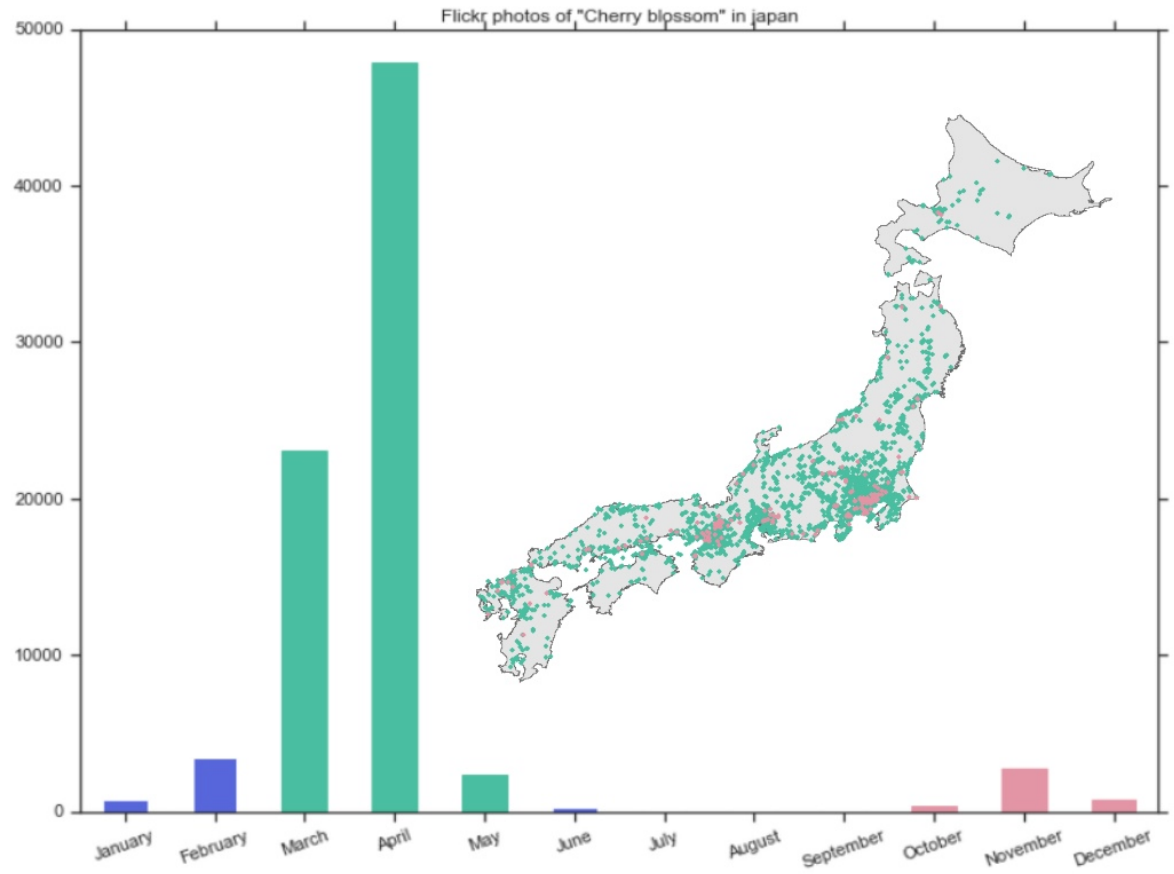


Figure 27. Total count of search results of "cherry blossom" photos from Flickr in Japan from 1 January 2008 to 31 December 2018 are concentrated in spring, but show an interesting small spike in autumn, peaking in November.

4.2.2 Computer vision-generated text tags and irrelevant text tag filtration

Although the Flickr photos were constrained using the search term "cherry blossom", SNS content is frequently falsely (or only loosely) associated with search terms (ElQadi et al., 2017; Xing et al., 2018). Hence, I subjected all photos to a computer vision API that generates descriptive text tags for each photo based on its visual content, as a means to automatically double-check the relevance of individual data points. I used Google's computer vision API (cloud.google.com/vision/) for this purpose. This API uses pre-trained machine learning models to assign labels to images based on predefined categories. It is thus an improvement over the simple reverse image search I used in chapter 3 which returned only a "best guess" for a given image.

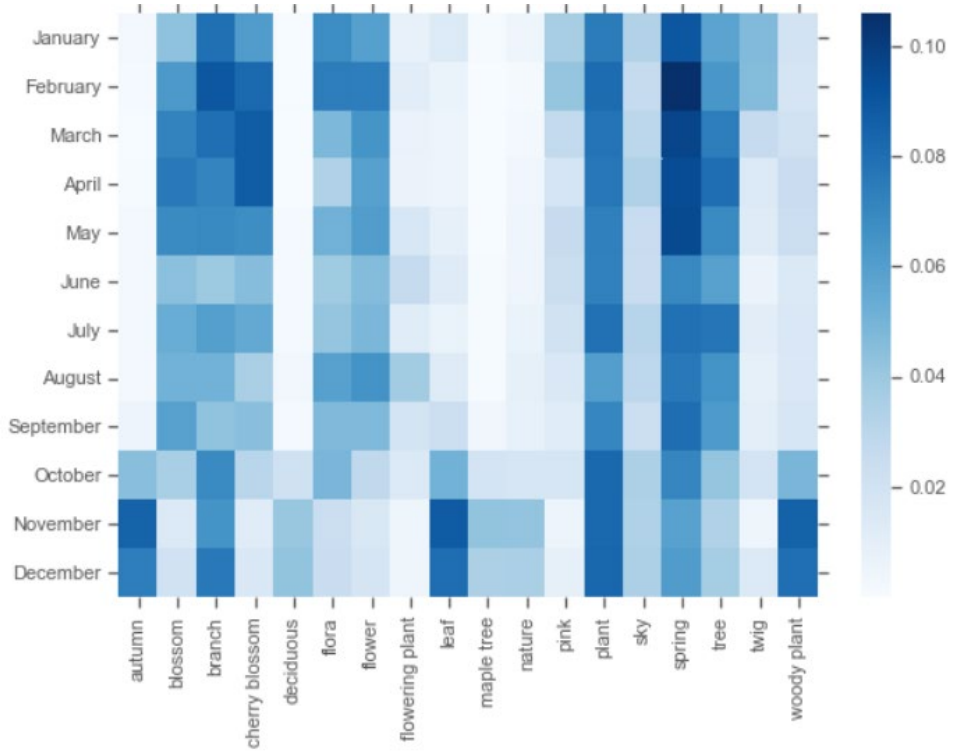


Figure 28 Normalised frequency of computer vision tags assigned to photos returned by the Flickr search "cherry blossom", Japan-wide for the period January 2008 to December 2018. Images within tags are not mutually exclusive; an image may be labelled simultaneously by several tags. To generate this plot, the top-10 most frequent text tags returned for images in each month were collected into a set of 18 tags. These appear alphabetically along the x-axis. The square marked for each month against a tag is coloured according to the relative frequency of its prevalence in that month. Note that "autumn" is a frequent tag in October, November, and December. These photos were often found to contain autumn leaves rather than blossoms.

As anticipated, the computer vision API returned the text tag "cherry blossom" for most photos. Human analysis of the other returned tags revealed that most were conceptually related to cherry blossoms, except perhaps for the tags "autumn" and "maple tree", that appeared in the last months of the year associated with some photographs (*Figure 28*). I might expect these tags to be associated by the computer vision algorithm with autumn leaves and Japanese Maples. Hence, I conducted a cursory visual inspection of a sample of the image content corresponding to these tags. I readily discovered that photos tagged "autumn" or "maple tree" usually contained autumn leaves, not cherry blossoms. This confirmed both that the computer vision API was correct in its assignment of the tag, but also that the data originally downloaded from Flickr using the search term "cherry blossom" contained images that were not relevant to my particular goal. To refine the data set, I therefore chose to automatically maintain only photos with the tag "cherry blossom" and without tags "autumn" or "maple tree".

4.2.3 Human-expert validation of automatic tag-based image filtration

To manually assess the effectiveness of my text-tag based filter, a botanist visually determined whether or not photos remaining in the data set depicted cherry blossoms. To achieve this, I asked them to check

in sequence whether or not each photo: 1) included a plant/tree; 2) if the plant/tree held flowers 3) if the flowers were identifiable as cherry blossoms (as distinct, for instance, from plum blossoms). I dismissed any photographs that did not return "true" for any of the steps, or which were of such poor quality or low resolution, that I couldn't accurately identify their content. These were omitted from the dataset to ensure what remained was indeed human-validated evidence of cherry blossom activity (see Table 8).

The visual similarity between cherry blossoms and other trees in the same genus, and also the inevitable lack of familiarity that international tourists may have with the cherry blossoms, can both possibly account for the misidentification of many cherry as plum blossoms. In each photograph, I distinguished cherry blossoms (Prunus subgenus Cerasus; Kato et al. 2014) from plum blossoms (P. mume) by their retuse petal apices, pedicellate flowers, horizontal bark lenticels and green leaves (CHANG, CHANG, Park, & Roh, 2007; Ohwi, 1965) .Plants were identified as plum blossoms (and therefore not validated as cherry blossoms) by their rounded petal apices, round buds and nearly sessile flowers (eFloras, 2008; Ohwi, 1965). A small number of other plants also misidentified as cherry blossoms were easily distinguishable as being from genera other than *Prunus* (e.g. *Wisteria*).

After the initial data collection and filtration of photos sourced from across Japan (Figure 27), I used the geographic data associated with each image in the set to focus specifically on a single region, Tokyo, as defined by its district boundaries as defined by gadm.org. This restricted my final analysis to a built environment, which, along with climate generally, is important in considering tree phenology since temperature differences between rural areas and built environments can be several degrees due to the heat-island effect (Shimoda, 2003). Tokyo represents a single climate zone and region of touristic interest, climate generally being a key determinant of cherry tree phenology (Sakurai et al., 2011), and tourism being a key determinant of photographic and social media activity (see Discussion, section 4.4).

4.2.4 Full bloom date estimation

To estimate full bloom dates, I generated a one-day resolution time series for all photos within the final image data set. To this I applied a triangular rolling average of 7-day width (centroid±3 days, $w(n) = 1 - \left|(n - \frac{N-1}{2})/\frac{N+1}{2}\right|$) to smooth human photographic activity fluctuations occurring for Japanese local tourist activity between weekdays (typical working periods) and weekends (typical recreational periods), *Figure 29 and Figure 30*. The crests of photographic activity on the resulting graphs were identified as the peak bloom times according to my model.

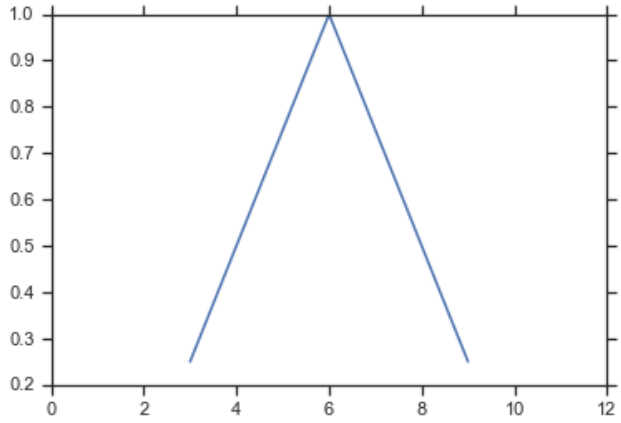


Figure 29 A triangular window of width 7, centred at 6

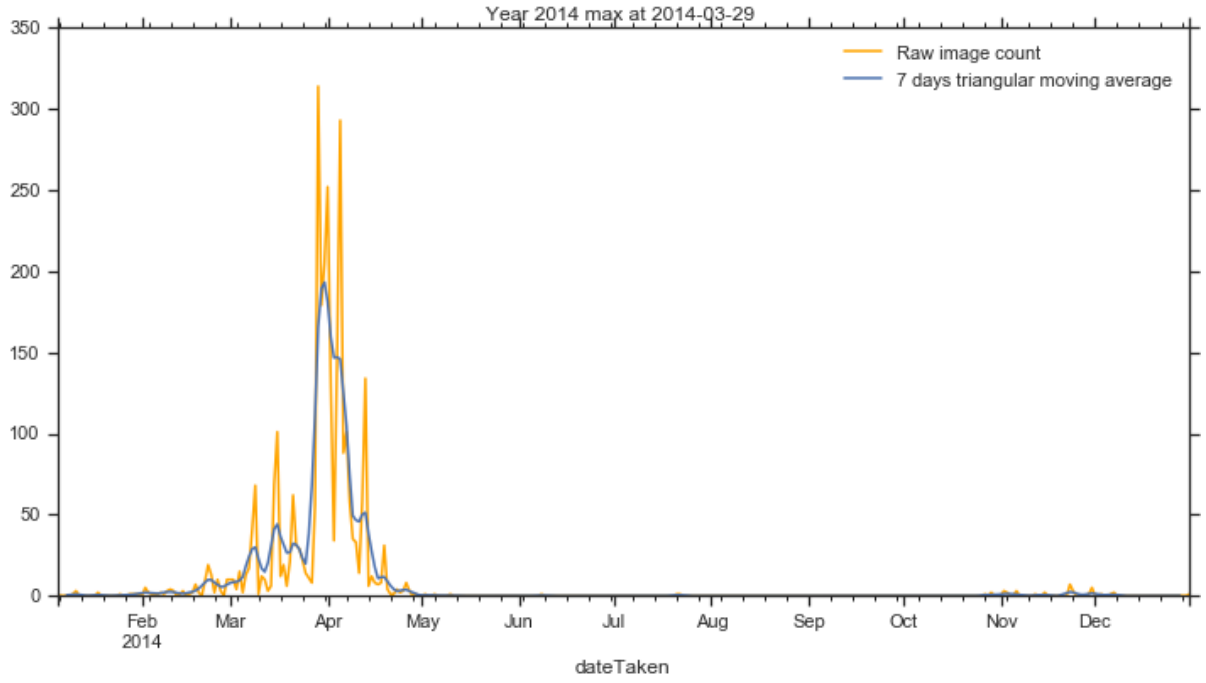


Figure 30 Estimating Cherry full bloom date in Tokyo 2014 from Flickr image count. Raw image count is smoothed using triangular window shown in Figure 29.

4.2.5 Kernel density maps

To show the spatial density distribution of cherry blossom photos, I created kernel density maps using ArcGIS 10.5 (ESRI, 2016) Kernel Density geoprocessing tool based on the quartic kernel function described in (Silverman, 1986).

4.3 Results

Table 8 Number of image data in the filtered dataset at each stage of my process.

A	B	C	D	E	F	G
Month	**Flickr "cherry blossom"**	**CV+tag filter**	**Human checked**	**Human confirmed**	**Accuracy (E/D)**	**Total cherry blossoms**
January	243	130	129	59	0.46	60
February	1211	944	221	71	0.32	*302
March	11280	8888	207	160	0.77	*6844
April	14804	11320	210	147	0.7	*7924
May	209	100	99	56	0.57	57
June	77	24	24	18	0.75	18
July	45	25	25	18	0.72	18
August	47	12	10	3	0.3	4
September	24	10	9	3	0.33	3
October	83	47	43	31	0.72	34
November	562	94	92	63	0.68	64
December	290	39	39	30	0.77	30
Total	**28875**	**21633**	**1108**	**659**		**15358**

* Calculated from a sample

Our Flickr search for “cherry blossom” (section 4.2.1) returned 28,875 photos, geo-located within the district boundaries I used to identify the Tokyo region of Japan (Table 1). The text tags, and their relative frequencies, returned by the machine vision API for this data set are reported in Figure 31. Out of the images returned from the text tag filter, 21,908 were subsequently tagged “cherry blossom” by the computer vision API (section 4.2.2), but some were eliminated due to them being tagged also “autumn” or “maple tree”, resulting in a set of 21,633 images.

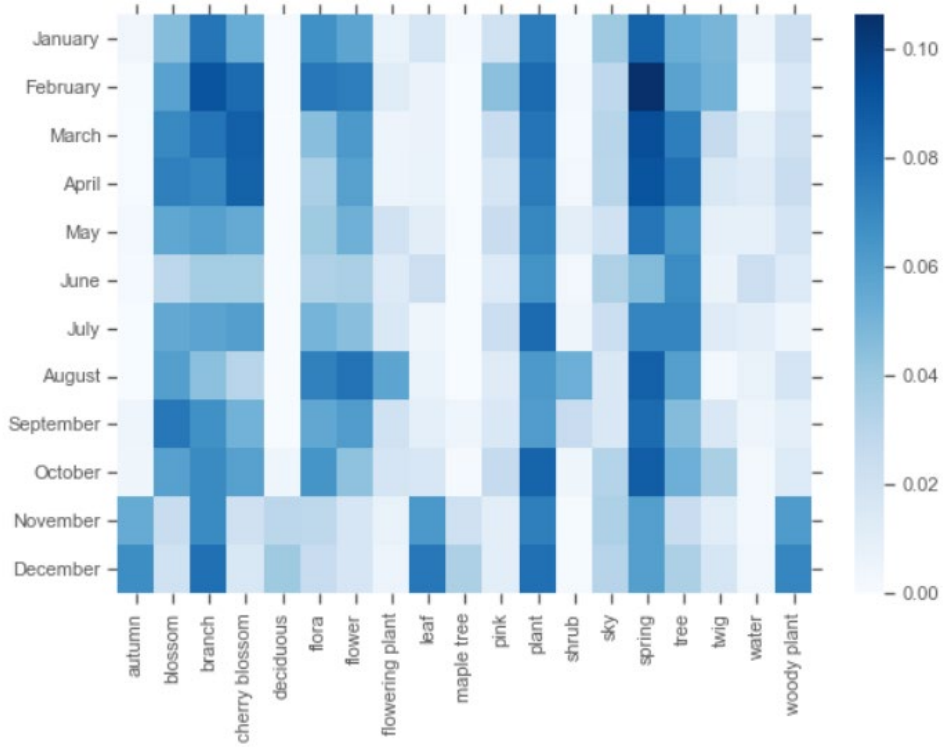


Figure 31 Normalised frequency of computer vision tags assigned to photos returned by the Flickr search “cherry blossom”, in the Tokyo region for the period (inclusive) January 2008 to December 2018. Images within tags are not mutually exclusive; an image may be labelled simultaneously by several tags. To generate this plot, the top-10 most frequent text tags returned for images in each month were collected into a set of 18 tags. These appear alphabetically along the x-axis. The square marked for each month against a tag is coloured according to the relative frequency of its prevalence in that month. Note that “autumn” is a frequent tag in November, and December. These photos were often found to contain autumn leaves rather than blossoms.

4.3.1 Spring bloom

The crests of the time series of my filtered and smoothed SNS cherry blossom photographic activity are plotted for each year of the study period in Figure 32 . To validate these results, my estimated dates are compared against the published JNTO full bloom dates ("The bloom of Cherry Blossoms 2018 Tokyo," 2018). The Root Mean Squared Error (RMSE) in my date = 3.21.

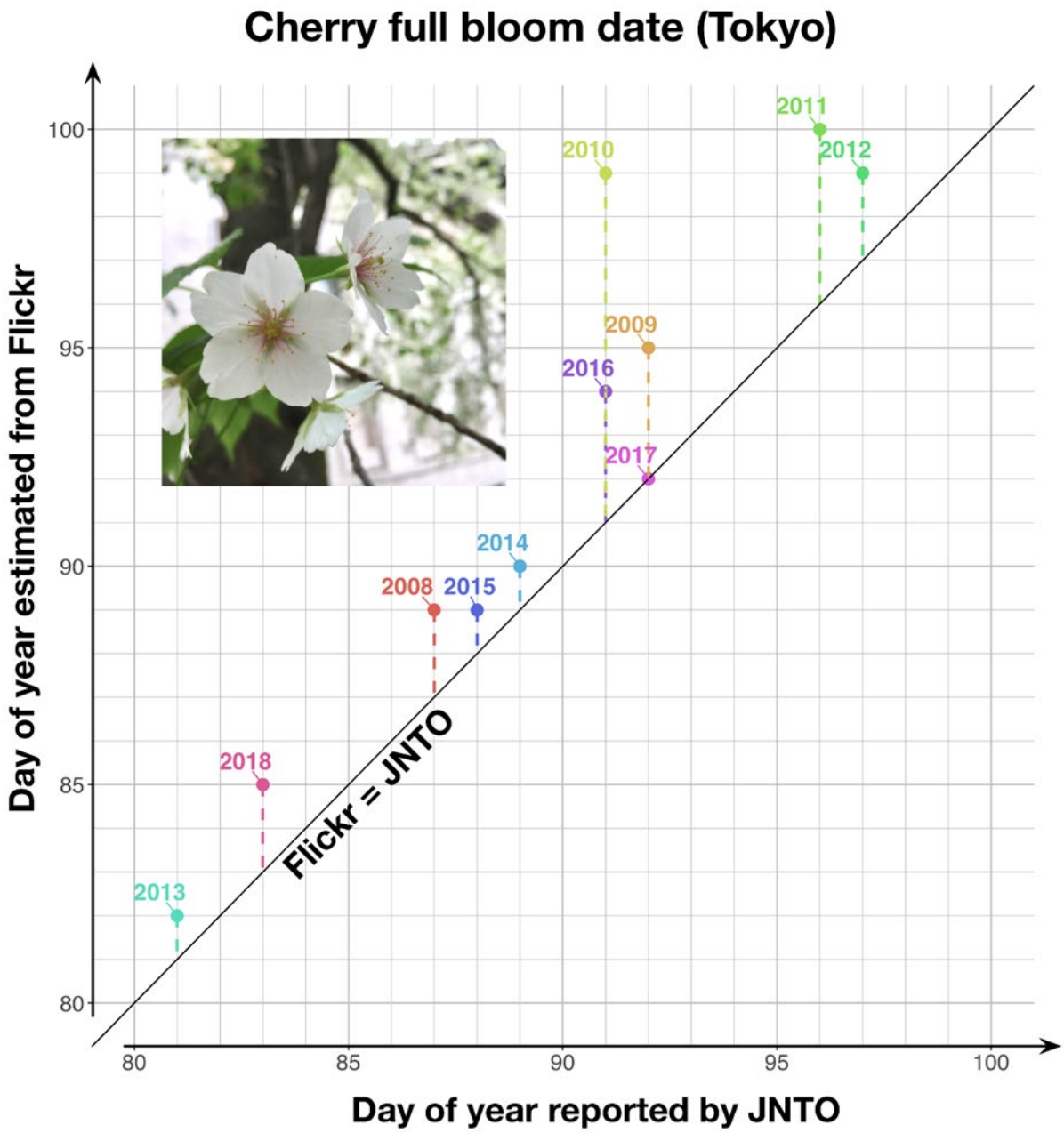
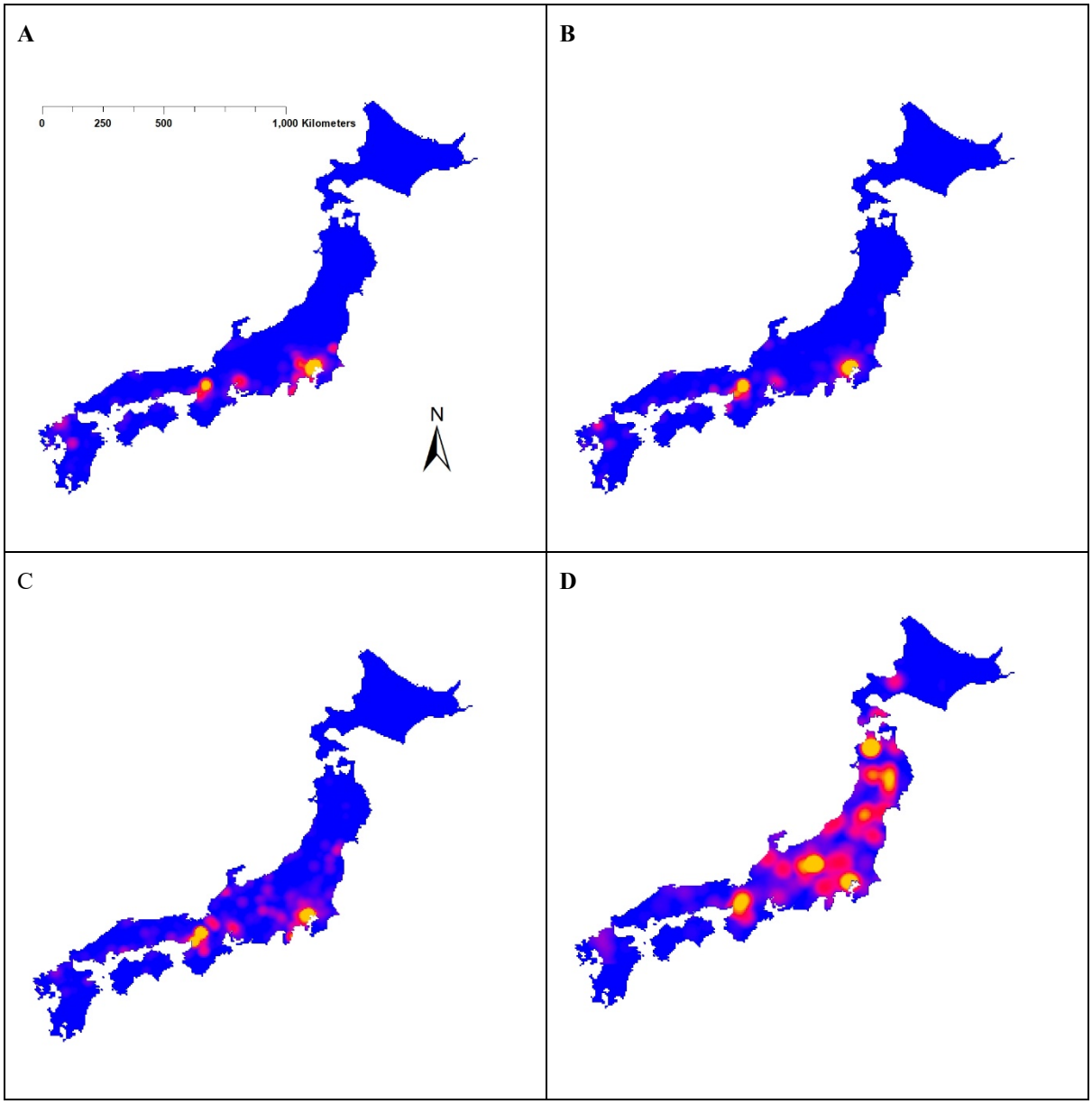


Figure 32. Bloom date estimates plotted as day number from 1st January of each respective year during the study period 2008-2018. The black diagonal line represents the front along which JNTO and Flickr estimates are equal. The dashed lines extending from this black diagonal line represent the difference in number of days between my estimates and the date specified by JNTO. Note that all SNS-based estimates fall above or on the diagonal line, meaning that SNS-based activity lags or coincides with JNTO posted dates for reasons discussed below. Cherry blossom photo courtesy of Alan Dorin.

The filtered dataset of cherry blossom geotagged photos in spring months were overlayed on a set of maps, each showing coordinates of photos taken within a 2-week interval. Heat maps were created for each period as described in 4.2.5. These heat maps show a clear pattern of blooming starting in the warmer south and advancing northward as weather gets warmer. Figure 33.


A
0
250
500
1,000 Kilometers
N
B
C
D

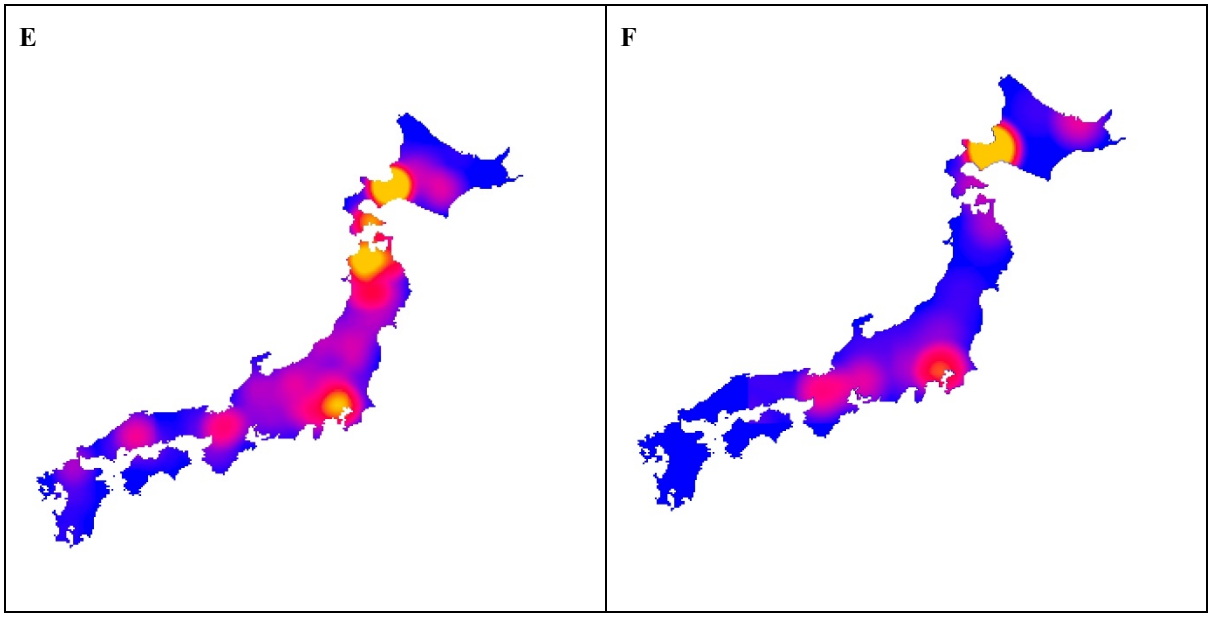


Figure 33 Hot spots of cherry blossom photos in Japan selected by my framework from aggregate data on Flickr from 2008 to 2018. Each frame corresponds to 14 days starting from day 67 (early March) of each year, the last frame thus ends at day 151. A) and B) Cherry blossom photos appar in the warmer southern part of Japan, a higher concentration of photos is shows in the urban centres of Tokyo and Kyoto. C) Cherry blossom photos stretch northwards in the main island of Honshū. D) Cherry blossoms photos continue to march north, starting to appear in the northern island of Sapporo. Tokyo and Kyoto are not the only cherry blossom photo hotspots despite their high urban population. E) Cherry blossom photos intensify in the southern part of Sapporo and northern part of Honshū. F) Finally, cherry blossom photos creep further north, appearing in the northern part of Sapporo; the main island of Honshū shows only a low density of cherry blossom images concentrated around Tokyo and Kyoto.

4.3.2 Autumn bloom

The temporal distribution of cherry blossom search results in Tokyo (Figure 35) shares some characteristics with the Japan-wide distribution (Figure 27), although of course it contains less images than its superset. In both the Japanese national and smaller Tokyo-restricted datasets, a secondary peak of SNS site images of cherry blossoms is apparent in November (Figure 34).

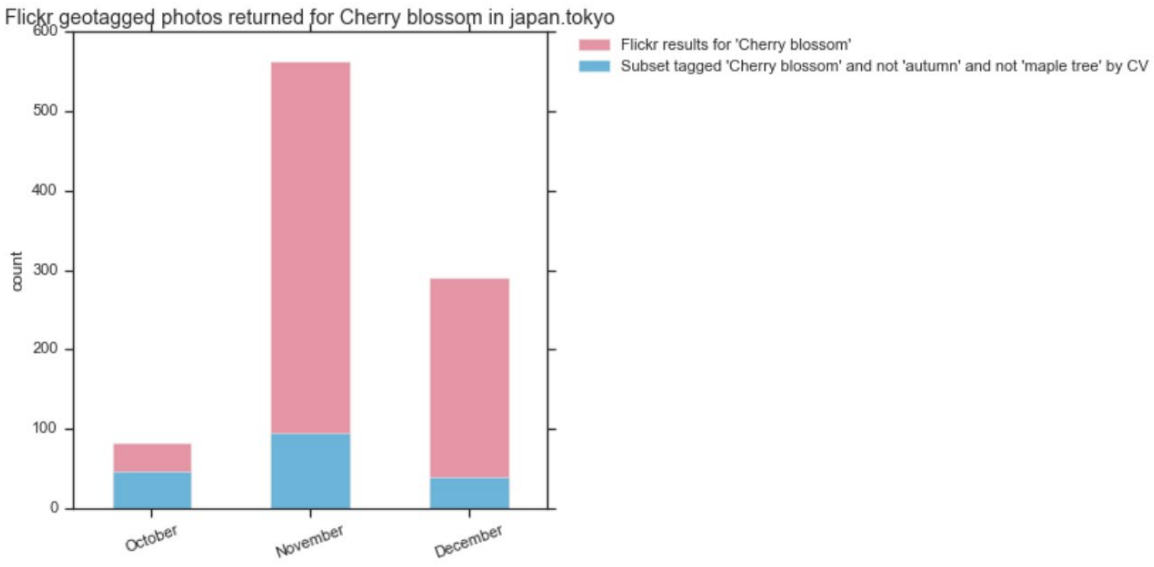


Figure 34. Total count (pink + blue bars) of search results from Flickr tagged "cherry blossom" in Tokyo, Oct-Dec, 2008 to 2018. A subset of these photos (blue bars) was also tagged “cherry blossom” and *not* “autumn” or “maple tree”, by the computer vision API.

The monthly mean counts of photos in my filtered data set over the study period (2008 to 2018) are plotted in Figure 35 revealing the extent of seasonal variation. In particular, an obvious spring peak corresponding to the standard bloom time plotted in Figure 4 is evident. But so is a secondary bloom centred around November evident, as it was for the Japan national data set (Figure 1).

The evidence for the November-centred flowering period was unexpected. Hence, to ensure that the photographs were not simply misclassified in this region due to an error in my method, I manually verified the content of the photos for months June to December for all years in my dataset. The monthly count of images I could readily confirm to be cherry blossoms are plotted in Figure 35, demonstrating the veracity of the observed November-centred flowering activity.

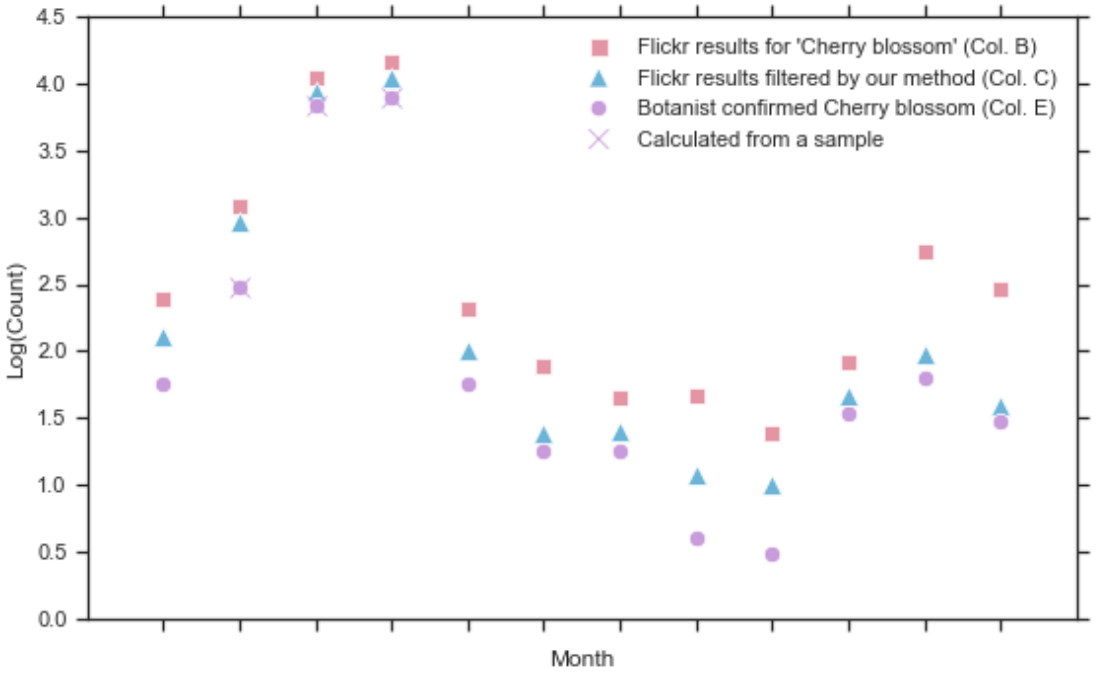


Figure 35. (a) Monthly total count distribution of results returned from a search for “cherry blossom” photos on Flickr over the years 2008-2018 in Tokyo. Legend shows the corresponding columns in Table 8 (b) Method accuracy, calculated as the ratio between number of photos confirmed by an expert, and number of photos resulting from my filtration method (column F in Table 8).

4.4 Discussion

There are several reasons why Tokyo is a suitable study site for the application of SNS data to ecological phenomena, whereas other study sites, and other high-profile ecological phenomena besides cherry blooms, might be more difficult subjects for study. Firstly, Tokyo’s status as a popular tourist destination, the eighth most-visited city in the world (MasterCard, 2018), ensures that many geotagged photos of cherry blossoms are shared by visitors on popular SNS. Additionally, the Japanese National

Tourism Office (JNTO) publishes the annual cherry blossom full bloom dates for the city ("The bloom of Cherry Blossoms 2018 Tokyo," 2018), against which it was possible to validate my method. Finally, the value Japan's own citizens place on cherry blossoms culturally, and on photography as a means to share cultural experiences, has long been documented (Mok & Lam, 2000). This potentially assists in enlisting the local population, including Tokyo's 9.27M people, (UN Data, 2015) as great incidental citizen scientists for research of the nature I describe. It is worth considering the value of my method when a smaller Japanese city, such as Kyoto, is the subject of the study.

Kyoto is another of Japan's renowned hanami destinations, but its local population is only 1.47M people (UN Data, 2015), around a sixth Tokyo's size. My search for geotagged photos tagged "cherry blossoms" in the geographic boundary of Kyoto (gadm.org) during the study period January 2008 to December 2018 returned 14,516 results. Of these, 10,425 were tagged "cherry blossom", and not "autumn", and not "maple trees" by the computer vision API. As shown in Table 9, my reported Spring peaks closely match the posted JNTO dates ("The bloom of Cherry Blossoms 2018 Kyoto," 2018), albeit with a slightly larger error (Root Mean Squared Error (RMSE) = 3.32) than for my Tokyo data.

Table 9 For Kyoto, cherry blossom full bloom JNTO-published dates ("The bloom of Cherry Blossoms 2018 Kyoto," 2018) are compared against the date of maximum geotagged photo count in Flickr based on a 7-day triangular rolling average.

Year	JNTO full bloom	Our method	Error (days)
2018	28-Mar	29-Mar	1
2017	7-Apr	7-Apr	0
2016	2-Apr	9-Apr	7
2015	1-Apr	4-Apr	3
2014	2-Apr	7-Apr	5
2013	30-Mar	31-Mar	1
2012	9-Apr	10-Apr	1
2011	7-Apr	8-Apr	1

2010	1-Apr	5-Apr	4
2009	5-Apr	8-Apr	3
2008	1-Apr	4-Apr	3

Applying my methodology, then, the SNS data can be used to regenerate the dates of the flowering events to within a few days of the JNTO-published dates in each year over a period of ten years, in both Tokyo, and Kyoto. I note that my estimate each year for both regions consistently lagged or matched, but never foreshadowed, the date reported by JNTO (Tokyo: Figure 32; Kyoto: Table 9). There may be a number of reasons for this. For instance, the SNS data peak may be generated as part of a self-actualising prophecy in which it follows the JNTO-published date. For this to be the case, for some reason a majority of visitors to the cherry blossoms would choose to visit only after published full bloom dates, biasing the SNS data in the way I observed. An ethnographic survey might elucidate the relevance of this effect on my project, but that is well beyond my present scope. However, I comment below on the extent to which my method stands independently of the JNTO-posted dates.

An alternative reason for the lag in SNS data with respect to the published dates might derive from some characteristic of the timestamping of images uploaded. Perhaps this might be biasing the SNS data in such a way that, for instance, visitors to the site were more likely to have their phones set to a time zone preceding Tokyo's, ensuring that the images uploaded are offset as I observed. I find this argument difficult to justify given the auto-update of mobile phone time zones connecting to the cellular network. Lastly, it is also possible that photographic peaks are skewed by the weather – something that has been discussed elsewhere in the literature (ElQadi et al., 2017). The first sunny day following the advertised bloom date might be the favoured time for visiting and photographing the cherry blossoms

Surprisingly, my analysis of the Flickr photos in Tokyo revealed evidence of a bloom occurring in autumn. The peak in the data at this time was small but confirmed by manual inspection. The size of the peak could be indicative of several factors. Perhaps there were few photographs of the autumn bloom because there were few trees in bloom. Perhaps there were few photographs because few people knew about the bloom and/or few visited to photograph them. Autumn is not the traditional hanami period and so, even if people did know through media and social networks about the bloom, they would have been unlikely to have made plans to visit and photograph the trees. In short, the evidence of the late autumn bloom is likely to be doubly incidental in the sense that the "citizen scientists" were incidentally contributing ecological data, but also, they were unlikely to have deliberately set out even as tourists to witness this phenomenon.

The autumn bloom has, it turns out, recently gained mainstream media attention: “for the first time in memory”(Livingston, 2018; Wamsley), “premature” (Victor, 2018), and “unexpected” ("Cherry blossoms bloom unexpectedly in Japan," 2018). Yet my data shows it to have been evident in Tokyo for many years. This is certainly interesting in and of itself, but also, it is evidence that my method is not simply a demonstration of the self-fulfilling prophecy generated by the JNTO-published dates –this secondary peak is not officially noted by JNTO yet it appears within the data.

One possibility is that the extra bloom is due to the presence of cherries of species *P. Subhirtella*, known to flower in colder weather(Watson, 1994). However, the cited news outlets have proposed the event to be a possible outcome of climate change. The merit of my SNS image analyses methodology is that it can detect such fine-grained shifts in flowering plant behaviour and spark further investigation.

Figure 33 shows the continued shift in cherry blossoms from southern to northern Japan over 12 weeks. This change can now be mapped with at least 2-week resolution, allowing a level of precise mapping with big data, with no costly formal survey; something unimaginable until now. With growing concerns about the effects of climate change on plant phenology, and on changes to the spatiotemporal distributions of flowering plants and their pollinators(Hegland, Nielsen, Lázaro, Bjerknes, & Totland, 2009; Scranton & Amarasekare, 2017), incidental citizen science of the type I apply here is a powerful new tool to map the climate’s effect on our environment.

Our results show that in the photo set I collected from Flickr searching for “cherry blossoms”, photos in January and February were falsely interpreted by my method as cherry blossoms (only 44% and 32% were correct respectively). This may be attributed to photos of plums, which precede cherries as they bloom in colder weather ("Japan's plum blossoms are already in the pink," 2014), often whilst snow is still present. So plums are concentrated in this time of the year and may be confused with cherry blossoms. Accuracy is also low in July and August. In these months, cherries are not in bloom, and very few photos exist for these months during the entire 11-year period of my study. It is worth noting that the distance between the phenological front of plum and cherry bloom times varies across the length of Japan (Yoshino & Ono, 1996). Therefore, the classification error may be expected to change depending on the region under study. For instance, plums and cherries may flower simultaneously in Hokkaido in the North, while in Southern Kyushu the difference in blooming dates may be wide.

Computer vision algorithms were usually focused on recognising generic object categories (e.g. plant, car,…), only recently, fine-grained visual recognition competitions (FGVCs) are gaining traction, where the objective is to differentiate between sub-categories (Kaeser-Chen, 2019), for example, different species, specific car models, or artistic genres. Computer vision API providers, like Google, are taking part in workshops such as the annual workshop on Fine-Grained Visual Categorization ("FGVC6,"). The aim in 2019 was to arrive at algorithms capable of differentiating between bird species

(URL). Such development would certainly improve research workflows similar to ours. In addition, more specialised, purpose-trained, computer vision APIs for species classification are worth developing the future in order to correctly identify cherry blossoms, and to distinguish them from plum blossoms.

Our results show that the SNS data can give valuable insights into phenological activities. However, one has to be careful when drawing conclusions about photos' relative densities in space and time; ElQadi et al. (2017) showed that SNS geotagged photo content may be spatially and temporally biased, not only by the underlying physical subject of the content, but also by the human photographical activity. In this work, I localised my temporal analysis of cherry blossoms to Tokyo in order to minimise the variation in human population density (and the ensuing photo density) as well as minimise the physical climate variation. Also, the popularity of Tokyo as a global tourist destination can potentially add more data points that are independent of the local population variation.

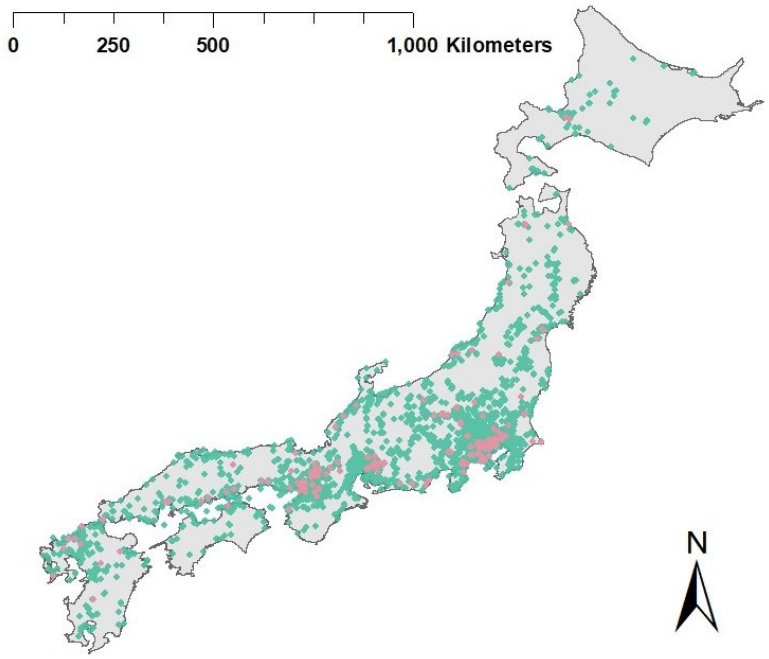


Figure 36 Main islands of Japan. Flickr photos tagged "cherry blossom" by computer vision in Spring Months (March, April, May) in green, while manually confirmed cherry blossom photos in autumn (October, November, December) are in mauve.

My data collection was based on searching for photos labelled "cherry blossom" on Flickr, and geo-localised to Japan. The Japanese character for Cherry blossoms 桜 (Sakura) was also tested on Flickr website and returned a relatively smaller result set. However, it is technically more challenging to use the Japanese character in the search API since it is a Unicode, not an ASCII character. Therefore, I opted for the English search keyword for my cultural and technical convenience. It is worthwhile to collect the cherry blossom photos that were text-tagged in Japanese. While the research community is learning to better understand how to use SNS data in scientific research and what can be achieved, some authors have discussed the use of scientific as well as common names to search for species (e.g. (Barve,

2014)). The human language used by the SNS users creating the researched content may also need be considered.

While analysing photo content, I excluded photos that were hard to confirm whether they were cherry blossoms. In the future, this decision can be enhanced by exploiting the location information and looking up other sources (e.g. Google street view in the same location, or photos from other social networks) especially in urban areas where there is a high photographic activity.

4.5 Conclusion

Studying SNS data on cherry blossoms in Japan, I was able to calculate the full bloom date from SNS geotagged photos. I have also showed that this data source has the power of revealing anomalies in phenological phenomena (e.g. flowering times). These findings suggest that my methods have the potential to help us understand changing ecological phenomena in a world that is increasingly putting its data online for all to see. This will help us understand the impact of the changing climate on our ecosystems in cases where the phenomena are popularly photographed and shared online

5 Thesis Chapter 5 – Future work, Discussion, and Conclusion

There are both direct and indirect effects of climate change. Direct effects include change in temperature, rainfall, and sea levels. An indirect effect of relevance to this thesis is the possibility of food chain disruption as a result of pollinator insect population declines (Kjøhl et al., 2011).

A new industrial revolution characterised by the unprecedented proliferation of data and the availability of computing power and algorithms to process that data is arguably taking place (Schwab, 2017). I see that the current industrial revolution can help us solve problems caused by the first industrial revolution. In other words, that using data available on social network sites with machine vision and classification, we can better find insights into ecological research questions that might, if applied, help us to mitigate the effects of climate change on insect pollination.

The two parts of the solution, source data and algorithms for processing that data, each present their own challenges. On the data side, a unique challenge is that social network site data are created for a variety of reasons by different people of diverse backgrounds. These data thus suffer from quality issues if we attempt to use them in research. Among the examples we have encountered in this research is misclassification of a species by the user. For instance, blossoms of plum and cherry are often confused by non-experts. Another example of data quality issues we faced is that the data returned from search queries on social network site can also be misleading when search word order is important for its potential to change semantics. Hence, a search query for "honey bee" could return results containing "bee honey" which has the same words, but different meaning.

Another issue with data collected from social network site is spatiotemporal and subject matter bias. These biases are a complication introduced by the "social" aspect of these sites. Human social behaviour is complex and dependent on many factors. For example, people tend to live around other people which gives rise to cities and urban centres. Consequently, data posted to social network site is mostly from these urban centres. Also, people may go outdoors in favourable weather more than in rainy, windy, hot or cold days. Then again, even in good weather, people may stay indoors, say, for work, study, or another event. This complex behaviour confounds the data generated by humans on social network site. This data is spatially biased, i.e. clustered around urban centres, and temporally biased, i.e. it varies with weather, holiday schedules, and other factors.

Subject matter bias in the data is also evident in social network site since the act of posting content is in itself a social activity. People generally tend to post content that they deem interesting to them, and/or, to their followers or friends. The existence of photos of a certain species, or lack thereof, on social network site could thus depend on whether this species is deemed interesting, catchy, or photogenic.

The sheer size of social network site data is also challenging. In fact, the universal explosion of data available on the web, including social network site, gave rise to the term "big data" (Lohr, 2012). Facebook alone in 2014 had 300 petabytes in their data warehouse, and generated 4 new petabytes every day (Wiener, 2014).The massive size of data on social network site means that manual processing of this data is beyond the capacity of human manual labour, and that there probably is data that may help address many research questions, if one could "look" carefully and quickly.

As noted above, the ideas of using social network site data in conjunction with algorithms drawn from the field of "Artificial Intelligence" to gain insights to ecological problems is challenging on both sides of the solution, data and algorithms. Machine vision and classification algorithms are today being applied in an ever-increasing variety of different applications. One reason why this might be the case is the increasing availability of these algorithms as services usable without specialised training. For example, programs that interact with users in natural language can be created without requiring specialised training in Natural Language processing (NLP) (e.g. http://luis.ai), similarly, computer vision models capable of reading and translating text, recognising people or landmarks (Tran et al., 2016) (Microsoft, 2018) are also available as services ready for consumption without the need to invest in building and training these AI systems. This democratisation is probably helped by the advancement in computing power and the availability of enough data examples to train intelligent models. For example, the system (Tran et al., 2016) behind Microsoft Cognitive services (Microsoft, 2018) was trained using a dataset of 2.5 million labelled instances in 328k images that were obtained from Flickr (Lin et al., 2014).

These AI algorithms offer solutions to many of the data problems we have discussed. For instance, computer vision can "see" what the visual content of a photo actually is. So, it could tell, in our previous example, if a photo returned from the query "honey bee" contains a bee or a jar of honey. Also, by training a classification model, we could filter photos that are relevant to a given research question. This is particularly valuable in the case of massive data sets.

Using AI algorithms presents its own set of challenges too. At the forefront is the need of human supervision; whether to teach a classification model, to review results of classification or selection, or to decide which categories of photos to keep. So, I tried throughout the thesis to save human labour needed to train these algorithms. For example, in land cover classification I opted for a generalised model that can classify photos from countries not used in training. This choice, although affecting accuracy, can help save human effort needed to train new models for other countries.

Another challenge is the "Hidden Technical Debt in Machine Learning Systems" (Sculley et al., 2015) where there usually is a great maintenance cost to real-world ML systems. A source of these costs is the glue code needed to achieve domain-specific goals using a machine learning system. Using common

APIs can result in a more reusable infrastructure. I chose to develop my methodology on commercially available APIs to minimize the technical debt incurred by researchers implementing these methods, while at the same time the performance of such Software-As-A-Service systems, SAAS, is set to improve over time as their backend keeps evolving in terms of underlying hardware and training.

A recurrent problem hindering the use of SNS data in our research is that the content of the retrieved photos is often irrelevant. We have developed a methodology of checking photos' visual content for relevance using commercial computer vision. The method generates a coarse textual description of these photos. These text tags can be used to filter the photos to exclude the irrelevant ones. We have shown throughout the thesis that the tag filtration can be done manually by sampling from most-frequent tags, or automatically by building a classification model, or a combination of both.

This thesis uses newly available technology and social network site data to begin to understand an accelerating problem of pollinator decline due to climate and habitat change, land use change, and flowering plant phenology change. The research case studies included Australia, Africa, and Asia (Japan). We have shown that data on social network site do not only reflect social connections between human users, but can also reflect ecological phenomena on Earth's surface and their change over time, of particular interest are bees and flowering plants' distribution, flower blooming times, and flower and bee lifecycle.

5.1 Future work

In this thesis, I presented methods to obtain insights on pollinators, flowering plants, and their habitat using social network site data, and demonstrated the efficacy of these methods. While the methodology used here is agnostic to the social network site used, this methodology was only case-studied on Flickr, due to its favourable search API, and because its content is primarily photographic. However, more research on generalising these methods to all other social network site can prove valuable and is worth considering.

To search social network sites, I used only English search terms. More research is required to internationalise this methodology. While the methodology would remain the same when searching in languages other than English, there are technical challenges involved in using Unicode characters (i.e. languages other than English) in information transmission (e.g. over SNS search API) and storage (e.g. in database). Also, more research is needed to determine the most effective search terms for a particular culture of social network site users because common names can differ regionally even in the same language (e.g. a "yellowjacket" in North America is a "wasp" in Australia).

The commercially available computer vision systems I chose to use allowed for a methodology decoupled from the technical implementation of the AI. However, these systems, being non-specialised,

are generally capable of only identifying object categories (e.g. insect), but not identifying a species, for example. Although specialised computer vision models capable of species classification do exist (e.g. https://github.com/microsoft/SpeciesClassification), a general-purpose computer vision model is, by the definition of being general, expected to lack such a specialised capability.

This shortcoming of commercial AI resulted in manually checking Flickr search results that were filtered with computer vision. A possible alternative to this manual labour is to train a specialised computer vision system to differentiate between specific species (e.g. species of bee), and use such a system in tandem with a fast, cheap, commercial AI that only does a coarse identification (e.g. whether a photo has a bee).

It is important to note that all social network site data used in this research were geotagged (i.e. had associated location information). However, users rarely share their location (Compton, Jurgens, & Allen, 2015). Therefore, only a small portion of data on social network site are actually geotagged. Cheng, Caverlee, and Lee (2010), for instance, estimated that less than 1% of tweets were geotagged. There is thus a research interest in geotagging non-geotagged content on social network site, usually using text associated with that content (Singh & Rafiei, 2016) (Kordopatis-Zilos, Papadopoulos, & Kompatsiaris, 2015), visual content (Verstockt, Gerke, & Kerle, 2015) time of posting, (Paraskevopoulos & Palpanas, 2016), and social connections of the user (Compton et al., 2015). However, many of these techniques may have limited utility when it comes to ecological research. For instance, using visual content of a photo to determine its location may work well with a photo of a landmark. It may work less well, with a bee photo since macro-level photos of bees may seldom contain unique landmarks easily associated with specific regions. More work is thus needed to tailor these methods to ecological research applications.

5.2 Summary of findings

Overall, results suggest that geotagged photos on social network site have a high potential in deriving insights on the living world. I have generally found that geotagged photos on SNS can offer a valuable addition to existing survey data, especially that records of the latter frequently lack visual media. However, photos are retrieved from SNS based on associated text tags, while the visual content of the photos might be different from the object of interest used as a search term. I found that inspecting visual content of these photos using computer vision is a promising strategy.

In Chapter 2: Land cover in Africa, case studies from 7 African countries showed that this data source can help in determining land cover class when satellite imagery is indeterminate. The criteria to decide if an image could help determine land cover can be found after building a classification model based on descriptive text tags associated with every image, then identifying the most influential parameters in the model. More specifically, findings of that chapter were:

1. A portion (30% to 70%) of SNS geotagged photos outside urban areas can help determine land cover.
2. To classify geotagged photos based on their fitness to land cover determination
    a. A classification model trained on data from the same country as the target photo is best, then,
    b. A generalised model trained on data including the target country, then,
    c. A classification model trained on data from a neighbouring country
3. So, a generalised model offers a trade-off between performance and reusability
4. Our models save time to check 72 to 90% of irrelevant photos in case study countries

In Chapter 3: Pollinators and flowering plants in Australia, case studies, from Australia, of two flowering plants, and two bee species showed that SNS geotagged photos offered a valuable addition to existing species distribution data retrieved from the Global Biodiversity Information Facility (GBIF). Findings can be summarized as follows:

1. Geotagged SNS photos can help establish species occurrence in certain regions.
2. Many of the photos returned have false associations
3. Visual content need be checked
4. Google reverse image search is free and relatively "accurate"; can rarely differentiate between species.
5. SNS photos are spatially biased to urban centres, therefore, species density cannot be directly established from photo density

In Chapter 4: Cherry blossoms in Japan, Studying SNS data on cherry blossoms in Japan, I was able to calculate the full bloom date from SNS geotagged photos. I have also found that SNS data has the power of revealing anomalies in phenological phenomena (e.g. subtle blooming of cherry blossoms in Autumn). These result show that SNS data to be reflective of the underlying natural phenomena, despite being subject to noise from the variation in human behaviour of photography and SNS interaction. Findings of that chapter can be summarised as follows:

1. SNS photos temporal distribution can reflect temporal behaviour of underlying ecological phenomena
2. SNS photos show cyclic variation of density through each week, we suggested a moving average smoothing to account for the weekday/weekend variation in photographic activity.

In conclusion, this thesis presents cost and time-effective methods to obtain ecological insights from free, public, geotagged photographic data available on social network site using commercially available computer vision and machine learning tools. The data available looks likely to accumulate over time,

and the machine intelligence capabilities are getting better, and more ubiquitous. Hence the approaches outlined in my thesis look set to become even more valuable in the future than they already are at present.

6 References

Abatzoglou, J. T., & Williams, A. P. (2016). Impact of anthropogenic climate change on wildfire across western US forests. *Proceedings of the National Academy of Sciences, 113*(42), 11770-11775. doi:10.1073/pnas.1607171113

Andreassen, C. S. (2015). Online Social Network Site Addiction: A Comprehensive Review. *Current Addiction Reports, 2*(2), 175-184. doi:10.1007/s40429-015-0056-9

Andreassen, C. S., Torsheim, T., & Pallesen, S. (2014). Predictors of use of social network sites at work-a specific type of cyberloafing. *Journal of Computer-Mediated Communication, 19*(4), 906-921.

Aono, Y. (1998). Climatic change in March temperature deduced from phenological record for flowering of cherry tree in Tokyo since the late 18th century.

Aono, Y. (2015). Cherry blossom phenological data since the seventeenth century for Edo (Tokyo), Japan, and their application to estimation of March temperatures. *International Journal of Biometeorology, 59*(4), 427-434.

Aono, Y., & Kazui, K. (2008). Phenological data series of cherry tree flowering in Kyoto, Japan, and its application to reconstruction of springtime temperatures since the 9th century. *International Journal of Climatology: A Journal of the Royal Meteorological Society, 28*(7), 905-914.

Barker, J. L. P., & Macleod, C. J. A. (2018). Development of a national-scale real-time Twitter data mining pipeline for social geodata on the potential impacts of flooding on communities. *Environmental Modelling & Software*. doi:https://doi.org/10.1016/j.envsoft.2018.11.013

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

Barve, V. (2015). Discovering and developing primary biodiversity data from social networking sites.

Beck, J., Ballesteros-Mejia, L., Nagel, P., & Kitching, I. J. (2013). Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? *Diversity and Distributions, 19*(8), 1043-1050.

Becken, S., Stantic, B., Chen, J., Alaei, A. R., & Connolly, R. M. (2017). Monitoring the environment and human sentiment on the Great Barrier Reef: Assessing the potential of collective sensing. *Journal of Environmental Management, 203*, 87-97.

Biesmeijer, J. C., Roberts, S., Reemer, M., Ohlemüller, R., Edwards, M., Peeters, T., . . . Thomas, C. (2006). Parallel declines in pollinators and insect-pollinated plants in Britain and the Netherlands. *Science, 313*(5785), 351-354.

The bloom of Cherry Blossoms 2018 Kyoto. (2018). Retrieved from https://www.jnto.go.jp/sakura/eng/city.php?CI=29

The bloom of Cherry Blossoms 2018 Tokyo. (2018). Retrieved from https://www.jnto.go.jp/sakura/eng/city.php?CI=10

Bohannon, J. (2016). Twitter can predict hurricane damage as well as emergency agencies. *Science*. doi:10.1126/science.aaf4182

Boyd, D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210-230. doi:10.1111/j.1083-6101.2007.00393.x

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *Sat, 17*, 22.03.

Cardale, J. (1968). Parasites and other organisms associated with nests of Amegilla Friese (Hymenoptera: Anthophorinae). *Austral Entomology, 7*(1), 29-34.

CHANG, K. S., CHANG, C. S., Park, T. Y., & Roh, M. S. (2007). Reconsideration of the Prunus serrulata complex (Rosaceae) and related taxa in eastern Asia. *Botanical journal of the Linnean Society, 154*(1), 35-54.

Chen, I.-C., Hill, J. K., Ohlemüller, R., Roy, D. B., & Thomas, C. D. (2011). Rapid Range Shifts of Species Associated with High Levels of Climate Warming. *Science, 333*(6045), 1024-1026. doi:10.1126/science.1206432

Cheng, Z., Caverlee, J., & Lee, K. (2010). *You are where you tweet: a content-based approach to geo-locating twitter users*. Paper presented at the Proceedings of the 19th ACM international conference on Information and knowledge management, Toronto, ON, Canada.

Cherry blossoms bloom unexpectedly in Japan. (2018, 18 October 2018). *BBC.* Retrieved from https://www.bbc.com/news/world-asia-45898333

Clarke, M., Gillespie, R., & Cunningham, S. (2017). *Regional Economic Multiplier Impacts, Potential Pollinator Deficits across Crops*: Rural Industries Research and Development Corporation.

Compton, R., Jurgens, D., & Allen, D. (2015). *Geotagging one hundred million Twitter accounts with total variation minimization.* Paper presented at the Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014.

Cormode, G., & Krishnamurthy, B. (2008). Key differences between Web 1.0 and Web 2.0. *First Monday, 13*(6).

Culotta, A. (2010). *Towards detecting influenza epidemics by analyzing Twitter messages.* Paper presented at the Proceedings of the first workshop on social media analytics.

Dann, P., & Chambers, L. (2013). Ecological effects of climate change on Little Penguins Eudyptula minor and the potential economic impact on tourism. *Climate Research, 58*(1), 67-79.

Daume, S. (2016). Mining Twitter to monitor invasive alien species—An analytical framework and sample information topologies. *Ecological Informatics, 31*, 70-82.

Deng, D.-P., Chuang, T.-R., Shao, K.-T., Mai, G.-S., Lin, T.-E., Lemmens, R., . . . Kraak, M.-J. (2012). *Using social media for collaborative species identification and occurrence: issues, methods, and tools.* Paper presented at the Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information.

Deutsch, C. A., Tewksbury, J. J., Huey, R. B., Sheldon, K. S., Ghalambor, C. K., Haak, D. C., & Martin, P. R. (2008). Impacts of climate warming on terrestrial ectotherms across latitude. *Proceedings of the National Academy of Sciences, 105*(18), 6668-6672. doi:10.1073/pnas.0709472105

Di Minin, E., Tenkanen, H., & Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science, 3*, 63.

Díaz, S., Settele, J., Brondízio, E., Ngo, H., Guèze, M., Agard, J., . . . Butchart, S. (2019). *Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*: IPBES secretariat.

Diffendorfer, J. E., Loomis, J. B., Ries, L., Oberhauser, K., Lopez‐Hoffman, L., Semmens, D., . . . Goldstein, J. (2014). National valuation of monarch butterflies indicates an untapped potential for incentive‐based conservation. *Conservation Letters, 7*(3), 253-262.

eFloras. (2008). Flora of China. In: Missouri Botanical Garden St. Louis, MO.

ElQadi, M. M., Dorin, A., Dyer, A., Burd, M., Bukovac, Z., & Shrestha, M. (2017). Mapping species distributions with social media geo-tagged images: Case studies of bees and flowering plants in Australia. *Ecological Informatics, 39*, 23-31. doi:http://dx.doi.org/10.1016/j.ecoinf.2017.02.006

Endo, M., Hirota, M., & Ishikawa, H. (2018). *Utilization of Information Interpolation using Geotagged Tweets.* Paper presented at the Proceedings of the first workshop on User Interface for Spatial and Temporal Data Analysis (UISTDA'18). CEUR-WS.

ESRI. (2016). ArcGIS for Desktop (Version 10.5). Retrieved from http://www.esri.com/software/arcgis/arcgis-for-desktop

Estima, J., Fonte, C. C., & Painho, M. (2014, 3-6 June). *Comparative study of Land Use/Cover classification using Flickr photos, satellite imagery and Corine Land Cover database.* Paper presented at the AGILE'2014 International Conference on Geographic Information Science,, Castellón.

Estima, J., & Painho, M. (2013). Flickr geotagged and publicly available photos: Preliminary study of its adequacy for helping quality control of corine land cover. In *Computational Science and Its Applications–ICCSA 2013* (pp. 205-220): Springer.

Estima, J., & Painho, M. (2014). Photo Based Volunteered Geographic Information Initiatives: A Comparative Study of their Suitability for Helping Quality Control of Corine Land Cover. *International Journal of Agricultural and Environmental Information Systems (IJAEIS), 5*(3), 73-89. doi:10.4018/ijaeis.2014070105

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), (2016).

Facebook. (2019). Company Info. Retrieved from https://newsroom.fb.com/company-info/

Fagan, W. F., Lewis, M. A., Neubert, M. G., & Van Den Driessche, P. (2002). Invasion theory and biological control. *Ecology Letters, 5*(1), 148-157. doi:10.1046/j.1461-0248.2002.0_285.x

Feddema, J. J., Oleson, K. W., Bonan, G. B., Mearns, L. O., Buja, L. E., Meehl, G. A., & Washington, W. M. (2005). The importance of land-cover change in simulating future climates. *Science (New York, N.Y.), 310*(5754), 1674. doi:10.1126/science.1118160

FGVC6. Retrieved from https://sites.google.com/view/fgvc6/home

Follett, R., & Strezov, V. (2015). An Analysis of Citizen Science Based Research: Usage and Publication Patterns. *PloS one, 10*(11), e0143687. doi:10.1371/journal.pone.0143687

Fritz, S., See, L., McCallum, I., Schill, C., Obersteiner, M., Van der Velde, M., . . . Achard, F. (2011). Highlighting continued uncertainty in global land cover maps for the user community. *Environmental Research Letters, 6*(4), 044005.

Fritz, S., See, L., Perger, C., McCallum, I., Schill, C., Schepaschenko, D., . . . Obersteiner, M. (2017). A global dataset of crowdsourced land cover and land use reference data. *Scientific Data, 4*, 170075. doi:10.1038/sdata.2017.75

Gallai, N., Salles, J.-M., Settele, J., & Vaissière, B. E. (2009). Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecological Economics, 68*(3), 810-821.

Garbarino, J., & Mason, C. E. (2016). The power of engaging citizen scientists for scientific progress. *Journal of microbiology & biology education, 17*(1), 7.

Gaston, K. J., Blackburn, T. M., Greenwood, J. J., Gregory, R. D., Quinn, R. M., & Lawton, J. H. (2000). Abundance–occupancy relationships. *Journal of Applied Ecology, 37*(s1), 39-59.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). *On calibration of modern neural networks.* Paper presented at the Proceedings of the 34th International Conference on Machine Learning-Volume 70.

Gura, T. (2013). Citizen science: Amateur experts. *Nature, 496*(7444), 259-261. doi:10.1038/nj7444-259a

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., . . . Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment, 11*(3), 156-162.

Han, J., Pei, J., & Kamber, M. (2012). *Data mining: concepts and techniques* (Third ed.): Elsevier.

Heaton, J. (2008). *Introduction to neural networks with Java* (Second ed.): Heaton Research, Inc.

Hegland, S. J., Nielsen, A., Lázaro, A., Bjerknes, A. L., & Totland, Ø. (2009). How does climate warming affect plant - pollinator interactions? *Ecology Letters, 12*(2), 184-195.

Hogendoorn, K., Coventry, S., & Keller, M. A. (2007). Foraging behaviour of a blue banded bee, Amegilla chlorocyanea in greenhouses: implications for use as tomato pollinators. *Apidologie, 38*(1), 86-92. doi:10.1051/apido:2006060

Honey Bee and Pollination Program Five Year Research, Development & Extension Plan 2014/15 – 2018/19. (2015). Rural Industries Research and Development Corporation.

Iwao, K., Nasahara, K. N., Kinoshita, T., Yamagata, Y., Patton, D., & Tsuchida, S. (2011). Creation of new global land cover map with map integration. *Journal of Geographic Information System, 3*(02), 160.

Jaimes, L. G., Vergara-Laurens, I. J., & Raij, A. (2015). A survey of incentive techniques for mobile crowd sensing. *IEEE Internet of Things Journal, 2*(5), 370-380.

Japan's plum blossoms are already in the pink. (2014, 2014/02/07/). Brief article. *Japan Times (Tokyo, Japan)*. Retrieved from http://link.galegroup.com/apps/doc/A357950465/ITOF?u=monash&sid=ITOF&xid=bed7f8b6

Joseph, A. J., Tandon, N., Yang, L. H., Duckworth, K., Torous, J., Seidman, L. J., & Keshavan, M. S. (2015). #Schizophrenia: Use and misuse on Twitter. *Schizophrenia research, 165*(2-3), 111. doi:10.1016/j.schres.2015.04.009

Kaeser-Chen, C. (2019). Announcing the 6th Fine-Grained Visual Categorization Workshop. Retrieved from https://ai.googleblog.com/2019/04/announcing-6th-fine-grained-visual.html

Kim, G.-H., Trimi, S., & Chung, J.-H. (2014). Big-data applications in the government sector. *Communications of the ACM, 57*(3), 78-85.

Kirkhope, C. L., Williams, R. L., Catlin-Groves, C. L., Rees, S. G., Montesanti, C., Jowers, J., . . . Goodenough, A. E. (2010). *Social networking for biodiversity: the BeeID project.* Paper presented at the Information Society (i-Society), 2010 International Conference On.

Kjøhl, M., Nielsen, A., & Stenseth, N. C. (2011). *Potential effects of climate change on crop pollination*: Food and Agriculture Organization of the United Nations (FAO).

Klein, A.-M., Vaissiere, B. E., Cane, J. H., Steffan-Dewenter, I., Cunningham, S. A., Kremen, C., & Tscharntke, T. (2007). Importance of pollinators in changing landscapes for world crops. *Proceedings of the Royal Society of London B: Biological Sciences, 274*(1608), 303-313.

Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2015) Geotagging social media content with a refined language modelling approach. In*: Vol. 9074. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 21-40).

Kuiter, R. H. (2013). *Orchid pollinators of Victoria*: Aquatic Photographics.

LeBaron, G. S., Cannings, R. J., Niven, D. K., Sauer, J. R., Butcher, G. S., Link, W. A., . . . McKay, K. J. (2004). The 104th Christmas Bird Count. *American Birds, 58*, 2-7.

Lesiv, M., Fritz, S., McCallum, I., Tsendbazar, N., Herold, M., Pekel, J.-F., . . . Van De Kerchove, R. (2017). Evaluation of ESA CCI prototype land cover map at 20m.

Lewandowski, E., & Specht, H. (2015). Influence of volunteer and project characteristics on data quality of biological surveys. *Conservation Biology, 29*(3), 713-723.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). *Microsoft COCO: Common Objects in Context*, Cham.

Liu, J., Cheng, H., Jiang, D., & Huang, L. (2019). Impact of climate-related changes to the timing of autumn foliage colouration on tourism in Japan. *Tourism Management, 70*, 262-272.

Livingston, I. (2018, 23 October 2018). Cherry blossoms all over Japan have been tricked into thinking it is spring. *The Washington Post*. Retrieved from https://www.washingtonpost.com/weather/2018/10/23/cherry-blossoms-all-over-japan-have-been-tricked-into-thinking-it-is-spring

Lohr, S. (2012). The age of big data. *New York Times, 11*(2012).

Marchante, H., Morais, M. C., Gamela, A., & Marchante, E. (2017). Using a WebMapping Platform to Engage Volunteers to Collect Data on Invasive Plants Distribution. *Transactions in GIS, 21*(2), 238-252. doi:10.1111/tgis.12198

MasterCard. (2018). Big Cities, Big Business: Bangkok, London and Paris Lead the Way in Mastercard's 2018 Global Destination Cities Index [Press release]. Retrieved from https://newsroom.mastercard.com/press-releases/big-cities-big-business-bangkok-london-and-paris-lead-the-way-in-mastercards-2018-global-destination-cities-index/

McCallum, I., Obersteiner, M., Nilsson, S., & Shvidenko, A. (2006). A spatial comparison of four satellite derived 1km global land cover datasets. *International Journal of Applied Earth Observation and Geoinformation, 8*(4), 246-255. doi:https://doi.org/10.1016/j.jag.2005.12.002

Memmott, J., Craze, P. G., Waser, N. M., & Price, M. V. (2007). Global warming and the disruption of plant–pollinator interactions. *Ecology Letters, 10*(8), 710-717. doi:10.1111/j.1461-0248.2007.01061.x

Michel, F. (2016). How many public photos are uploaded to Flickr every day, month, year? Retrieved from https://www.flickr.com/photos/franckmichel/6855169886/in/photostream/

Microsoft. (2018). Computer Vision API for Microsoft Cognitive Services. Retrieved from https://docs.microsoft.com/en-us/azure/cognitive-services/

Mignon, A. (2016). python-flickr-api [Python package]: github.com. Retrieved from https://github.com/alexis-mignon/python-flickr-api

Mok, C., & Lam, T. (2000). Travel-related behavior of Japanese leisure tourists: A review and discussion. *Journal of Travel & Tourism Marketing, 9*(1-2), 171-184.

Moritz, R. F., Kraus, F. B., Kryger, P., & Crewe, R. M. (2007). The size of wild honeybee populations (Apis mellifera) and its implications for the conservation of honeybees. *Journal of Insect Conservation, 11*(4), 391-397.

Morton, E. M., & Rafferty, N. E. (2017). *Plant–Pollinator Interactions Under Climate Change: The Use of Spatial and Temporal Transplants* (Vol. 5): SPIE.

Nerem, R. S., Beckley, B. D., Fasullo, J. T., Hamlington, B. D., Masters, D., & Mitchum, G. T. (2018). Climate-change–driven accelerated sea-level rise detected in the altimeter era. *Proceedings of the National Academy of Sciences, 115*(9), 2022-2025. doi:10.1073/pnas.1717312115

Nogué, S., Long, P. R., Eycott, A. E., de Nascimento, L., Fernández-Palacios, J. M., Petrokofsky, G., . . . Willis, K. J. (2016). Pollination service delivery for European crops: Challenges and opportunities. *Ecological Economics, 128*, 1-7.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM, 11*(122-129), 1-2.

Oba, H., Hirota, M., Chbeir, R., Ishikawa, H., & Yokoyama, S. (2014). *Towards Better Land Cover Classification Using Geo-tagged Photographs.* Paper presented at the Multimedia (ISM), 2014 IEEE International Symposium on.

Ohwi, J. (1965). Flora of Japan (English edn.). *Smithsonian Institution, Washington, DC*.

Palpanas, T., & Paraskevopoulos, P. (2015). *Fine-grained geolocalisation of non-geotagged tweets.* Paper presented at the Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on.

Paraskevopoulos, P., & Palpanas, T. (2016). Where has this tweet come from? Fast and fine-grained geolocalization of non-geotagged tweets. *Social Network Analysis and Mining, 6*(1), 89.

Parmesan, C. (2006). Ecological and Evolutionary Responses to Recent Climate Change. *Annual Review of Ecology, Evolution, and Systematics, 37*(1), 637-669. doi:10.1146/annurev.ecolsys.37.091305.110100

Robert P. Anderson, M. A., Antoine Guisan, Jorge M. Lobo, Enrique Martínez-Meyer, A. Townsend Peterson, Jorge Soberón. (2016). Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling. Retrieved from http://www.gbif.org/resource/82612

Sakurai, R., Jacobson, S. K., Kobori, H., Primack, R., Oka, K., Komatsu, N., & Machida, R. (2011). Culture and climate change: Japanese cherry blossom festivals and stakeholders' knowledge and attitudes about global climate change. *Biological Conservation, 144*(1), 654-658. doi:10.1016/j.biocon.2010.09.028

Schwab, K. (2017). *The fourth industrial revolution*: Currency.

Scranton, K., & Amarasekare, P. (2017). Predicting phenological shifts in a changing climate. *Proceedings of the National Academy of Sciences, 114*(50), 13212-13217. doi:10.1073/pnas.1711221114

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., . . . Dennison, D. (2015). *Hidden technical debt in machine learning systems.* Paper presented at the Advances in neural information processing systems.

See, L., Comber, A., Salk, C., Fritz, S., van der Velde, M., Perger, C., . . . Obersteiner, M. (2013). Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts. *PloS one, 8*(7), e69958. doi:10.1371/journal.pone.0069958

See, L., Fritz, S., Perger, C., Schill, C., McCallum, I., Schepaschenko, D., . . . Obersteiner, M. (2015). Harnessing the power of volunteers, the internet and Google Earth to collect and validate global spatial information using Geo-Wiki. *Technological Forecasting and Social Change, 98*(Supplement C), 324-335. doi:https://doi.org/10.1016/j.techfore.2015.03.002

Shimoda, Y. (2003). Adaptation measures for climate change and the urban heat island in Japan's built environment. *Building Research & Information, 31*(3-4), 222-230.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London

New York: London

New York : Chapman and Hall.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in ecology & evolution, 24*(9), 467-471.

Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., . . . McConway, K. (2015). Crowdsourcing the identification of organisms: A case-study of iSpot. *ZooKeys, 480*, 125.

Singh, S. K., & Rafiei, D. (2016). *Geotagging flickr photos and videos using language models.* Paper presented at the CEUR Workshop Proceedings.

Southall, H., Baily, B., & Aucott, P. (2007). 1930s Land utilisation mapping: an improved evidence-base for policy?

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

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data, 1*(2), 85-99.

Switzer, C. M., Hogendoorn, K., Ravi, S., & Combes, S. A. (2016). Shakers and head bangers: differences in sonication behavior between Australian Amegilla murrayensis (blue-banded bees) and North American Bombus impatiens (bumblebees). *Arthropod-Plant Interactions, 10*(1), 1-8. doi:10.1007/s11829-015-9407-7

Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E., . . . Harsch, M. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation, 181*, 236-244.

Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., . . . Sienkiewicz, C. (2016). *Rich image captioning in the wild.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.

Transforming our world : the 2030 Agenda for Sustainable Development. (2015). In: UN General Assembly.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM, 10*, 178-185.

Verstockt, S., Gerke, M., & Kerle, N. (2015). Geolocalization of crowdsourced images for 3-D modeling of city points of interest. *IEEE Geoscience and Remote Sensing Letters, 12*(8), 1670-1674. doi:10.1109/LGRS.2015.2418816

Victor, D. (2018). Japan's Cherry Blossoms (Some of Them) Appear Months Early. *The New York Times*. Retrieved from https://www.nytimes.com/2018/10/18/world/asia/cherry-blossom-japan-typhoon.html

Wamsley, L. In Japan, A Strange Sight: Cherry Blossoms Blooming In The Fall. Retrieved from https://www.npr.org/2018/10/18/658484696/in-japan-a-strange-sight-cherry-blossoms-blooming-in-the-fall

Wang, Q., Chen, W., & Liang, Y. (2011). The effects of social media on college students.

Watson, E. F. G. D. G. (1994). *prunus Subhirtella 'Autumnalis'*

'Autumnalis' Higan Cherry.

White, K. (2016). Forecasting Canadian elections using Twitter. In (Vol. 9673, pp. 186-191).

Wiener, J. B., Nathan (2014). Facebook's Top Open Data Problems. Retrieved from https://research.fb.com/blog/2014/10/facebook-s-top-open-data-problems/

Xing, H., Meng, Y., Wang, Z., Fan, K., & Hou, D. (2018). Exploring geo-tagged photos for land cover validation with deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing, 141*, 237-251. doi:https://doi.org/10.1016/j.isprsjprs.2018.04.025

Xu, G., Zhu, X., Fu, D., Dong, J., & Xiao, X. (2017). Automatic land cover classification of geo-tagged field photos by deep learning. *Environmental Modelling & Software, 91*, 127-134. doi:https://doi.org/10.1016/j.envsoft.2017.02.004

Yoshino, M., & Ono, H.-S. P. (1996). Variations in the Plant Phenology Affected by Global Warming. In *Climate Change and Plants in East Asia* (pp. 93-107). Tokyo: Springer Japan.