Hirvonen, H., Leppämäki, T., Rinne, J., Muukkonen, P. & Fink, C. (2020). Modifying and analyzing Flickr data for wildlife conservation. *In* Muukkonen, P. (Ed.): *Examples and progress in geodata science*, pp. 66–90. *Department of Geosciences and Geography* C19. Helsinki: University of Helsinki.

Chapter VII

Modifying and analyzing Flickr data for wildlife conservation

Hirvonen, H., Leppämäki, T., Rinne, J., Muukkonen, P. & Fink, C.

hanna.hirvonen@helsinki.fi, University of Helsinki
tatu.leppamaki@helsinki.fi, University of Helsinki
jooel.rinne@helsinki.fi, University of Helsinki
petteri.muukkonen@helsinki.fi, University of Helssinki
christoph.fink@helsinki.fi, University of Helsinki

** Corresponding author christoph.fink@helsinki.fi*

Abstract

Applying social media data in researching protected area visitors can be useful in minimizing their impact on the biodiversity. An increased social media activity can be expected in national parks due to the growing nature-based tourism and the increasing use of social media. The tourism in national parks can lead to an impact on the area's biodiversity, resources, and environment. In this work, we study the possibilities Flickr data offers for conservation science, while aiming to provide methods for further research. We describe the dataset in multiple ways and examine the link between the accessibility and the frequency of social media posts. We create and utilize a script to merge geotagged social media point data with national park polygon data and global accessibility data, calculate social media post densities in national parks, and summarize them at the national and regional levels. We study point patterns in Sub-Saharan African national parks and create a kernel density raster layer of the Flickr posts in the region. Finally, we perform a cursory analysis of the linguistic content of the Flickr posts globally. Our results do not show a clear correlation between the Flickr post density in national parks and the accessibility from the nearest population center globally, which signifies a need for a regional examination or for a more sophisticated accessibility dataset. We find clear clusters of Flickr posts inside most Sub-Saharan national parks and have examples of national parks with concentrated and dispersed Flickr post distribution, although clustering is much more prevalent. Our linguistic analysis demonstrates the dominant role of English in Flickr, which might indicate an overrepresentation of people from English speaking countries in the data.

Keywords: Accessibility; Conservation; Flickr; Geoinformatics; Linguistic analysis; National park; Social media

Introduction

Social media data is useful in conservation science. Since the discipline tends to benefit from spatially precise data, it extracts social media data from different social media platforms nowadays such as Facebook, Twitter, Instagram, and Flickr (Di Minin et al., 2015). The data can be used in e.g. systematic conservation planning (Margules & Pressey, 2000; Knight, Cowling & Campbell, 2006) and in modeling species distributions (Elith et al., 2006). The data accuracy can be higher compared to more traditional data sources (University of Helsinki, 2015), the process more cost-efficient and continuous (Hausmann et al., 2017a), and the temporal and spatial resolutions better (Richards & Friess, 2015).

National parks and other protected areas have a great significance in wildlife conservation, protecting species, and possibly in reversing the biodiversity crisis (Watson et al., 2014). National parks are areas protected by the government to preserve the natural environment (Encyclopaedia Britannica, 2020). Conservation biology can be described as a mission-oriented discipline that aims to protect and restore biodiversity. It usually focuses on the issues that need quick action and could have significantly negative consequences. Because of the growing nature-based tourism (Balmford et al., 2015) and the increasing use of social media (Kaplan & Haenlein, 2010; Mayer-Schönberger & Cukier, 2013), increased social media activity in national parks can be expected. This leads to more available user content to further utilise in conservation research.

Tourism in national parks can be a double-edged sword. According to Di Minin et al. (2013), ecotourism has a potential to generate political support for protected areas. It can also generate funding for covering the park management costs (Buckley, 2009; Buckley, Morrison & Castley, 2016; Gössling, 1999) and has been promoted as a way to support biodiversity conservation and economic development (Goodwin, 1996; Krüger, 2005). Still, it can also lead to a detrimental anthropogenic impact on the biodiversity, the resources, and the environment of an area (Buckley, Morrison & Castley., 2016; Gössling, 2002). Particularly the biodiversity of small areas can suffer from the edge effect (Woodroffe & Ginsberg, 1998).

One of the top tourist destinations is Sub-Saharan Africa (World Tourism Organization, 2015). We choose to focus on Sub-Saharan Africa due to its density of well-known tourism-oriented national parks, which are typically safari parks (Africa Sun News, 2003; Crush, 1980; Siegfried, Benn & Gelderblom, 1998). Sub-Saharan Africa can be defined as the Africa south of the Sahara Desert, consisting of countries such as Ethiopia, Ghana, Kenya, Tanzania, Namibia, and Botswana. The African parks can support wildlife conservation while the potential use of social media data to inform conservation may increase in the future (Tenkanen et al., 2017; Willemen et al., 2015). The tourists are attracted to the African protected areas mainly by their charismatic megafauna. Hausmann et al (2017b) discuss the other important characteristics of nature-based tourism in Africa, the most essential of which are the biodiversity and the landscape aesthetics. Also, particularly when studying tourism in these areas, geographical factors, such as accessibility and human influence, can be very important. Consequently, it can be deduced that accessibility can be utilised in conservation science.

The indicators of accessibility often are different distance measures and travel times (Frank et al., 2008; Mavoa et al., 2012). It can also have a great effect on the post activity and the number of park visitors. Hausmann et al concluded in their study (2017b) that accessibility was a strong predictor of the user and the post activities, meaning that accessible areas tend to have more social media posts and active users. The study also revealed that the richness of charismatic species did not influence the social media use in the protected areas of Africa but rather of importance were the socio-economic conditions of the countries and their geographical characteristics.

Hausmann et al (2017) also note that the biodiversity and the environment of accessible areas can be threatened by a high human pressure. The disturbance on the area’s biodiversity can include stamping down the vegetation (Pickering & Hill, 2007), disrupting the feeding and breeding of the fauna (Bouton et al., 2009; Ranaweerage, Ranjeewa & Sugimoto, 2015), and decreasing the successful reproduction (Steven, Pickering & Castley, 2011). Overall, the sustainability of nature-based tourism is indeed challenged (Buckley, 2011).

Tenkanen et al (2017) state that evaluating the benefits of the recreational value of national parks is often a crucial part of justifying the existence of these parks, which creates a firmer base for maintaining these areas for biodiversity conservation. Monitoring the visitor rates from social media data can be used in assessing the area’s recreational value. Based on this it is possible to conclude that applying social media data in researching protected area visitors can serve as a justification for conservation.

Social media data usually contains text, images, videos, and tags. When using the data for e.g. conservation studies, it can be restricted by search parameters, such as keywords (Di Minin et al., 2015). The posts also contain the time stamp and possibly the location data. Because of these features, social media data has uniquely great spatial and temporal resolutions of populations (Longley, Adnan & Lansley, 2015), which makes it a very suitable data source for conservation science, although the use is still limited (Di Minin et al., 2015). To access the content, ready-made application programming interfaces can be used (University of Helsinki, 2015), and in publishing the data, the user privacy has to be taken into account.

The social media posts can tell about the preferences and the engagement of the national park visitors (Hausmann et al., 2017a; Levin, Kark & Crandall, 2015; Su et al., 2016). The data can be useful in the national park management as information for minimising the impact of the visitors on the area's biodiversity (Cessford & Muhar, 2003) and for understanding the interests of the visitors for promotional purposes (Hausmann et al., 2016), along with marketing purposes (Buckley, 2009; Smith, Verissimo & Macmillan, 2010; Tenkanen et al., 2017). It may be profitable for the park management to use data from these kinds of novel sources instead of carrying out the surveys themselves, which can be comparatively time-consuming and costly (Hausmann et al., 2017a). For example, the data might reveal the species the visitors have spotted or their favourite species and landscapes (University of Helsinki, 2015). There are still some weaknesses in studying social media data from national parks. For instance, the data tends to perform better in the parks with more visitors, and sometimes the visitor statistics and the user activity do not match (Tenkanen et al., 2017). Social media data also tends to be biased to the developed countries (Di Minin et al., 2015).

Established in 2004, Flickr is among the oldest social media platforms. It has some good qualities as a data source for conservation science, which is one of the main reasons we use data mined from it in this study. The site is popular among photographers and is commonly used for image sharing. The study done by Hausmann et al (2017) had results on the features of the Flickr users that were visitors in protected areas. They were described as experienced tourists and nature enthusiasts with interests towards some of the less charismatic species. In South Africa, it had the highest correlation with the official statistics. (Tenkanen et al., 2017.)

In this article, we study multiple aspects of Flickr posts in national parks. First, we inspect the relationship between the post frequency and the accessibility of a park in different spatial scales. Then we study the patterns the posts create within the parks of a chosen subregion, Sub-Saharan Africa, and look at example parks to understand where and why the posts are clustered. We focus particularly on Flickr data from Serengeti National Park and Nairobi National Park. Serengeti is a famous area of 14,763 square kilometres in Tanzania and Kenya that attracts visitors with its rich natural resources, mainly biodiversity and its highest large mammal density of the world (Eagles & Wade, 2006; Serengeti National Park). Nairobi National Park is a smaller park in Nairobi, the capital city of Kenya. Finally, we do a tentative inspection of the linguistic content of the posts. All of our research steps aim to explore the dataset and create methods, thus assisting the future research in studying accessibility and Flickr data in conservation science.

Data and methods

We employed three datasets to examine global accessibility, national parks, and social media posts. These are, respectively, the global accessibility to cities by the Malaria Atlas Project (Weiss et al., 2018), the World Database on Protected Areas (UNEP-WCMC & IUCN, 2020), and a dataset of Flickr posts represented as coordinate points. Accessibility is, in the global raster surface by Weiss et al. (2018), defined as the travel time in minutes from one raster cell to the nearest urban centre. Urban centres are areas with a high population density or a high number of built plots coinciding with at least 50,000 inhabitants. Travel time is quantified by measuring the combined effects of different highways, land features, and national borders. The data dates to the year 2015, and its spatial resolution is 30 arc seconds, or roughly 1 km^2 at the equator (Weiss et al., 2018).

The World Database on Protected Areas (WDPA) is a global collection of land and sea areas that hold high natural or cultural values and meet the standards for a protected area (UNEP-WCMC, 2019, pp. 8–9). The data consists of polygon boundaries of the protected areas, provided by various governmental and other entities. We utilised a subsection (n=2556) of the data, the areas labelled as 'natural parks'. Finally, we used Flickr posts that are geotagged coordinate points falling within the natural parks. After filtering the data for exact duplicates, we were left with over 2.3 million posts with a temporal extent from the year 2004 to January 2019. Attributes of the posts, such as the title, the textual description of the image contents, the accuracy of the positioning, and the URL of the photo were included alongside the location.

For our research, we utilised various open source geospatial software, mainly Python and QGIS (see the whole workflow in Appendix A). We began by filtering both WDPA and Flickr datasets: the first for entities marked as 'national park', and the second for duplicate posts. If a park consisted of different zones, we interpreted them as parts of the same park. Parks that lacked Flickr posts altogether were dropped. In addition, some of the national parks had overlapping boundaries which means that some Flickr posts fell within more than one national park. The exact number of these posts was 14,067. We used these datasets to calculate the average post density in each park. Density is defined here as posts per square kilometre. The areas of the parks were included in the WDPA dataset and included marine regions. We used the 'GIS_AREA' field as the indicator for the area extent. The densities were also summarised at the national and regional levels. The summaries were conducted by adding the number of the region's Flickr posts together and dividing it by the summarised area of the region. We also included some simple descriptive statistics for each level.

National parks were also combined with the global accessibility raster dataset. Different statistical values for each park were calculated by defining the park borders as zones and summarising all raster cells that fall within. This produced for example the minimum value it takes to reach the park in minutes, that is to say, how accessible the park is in the best-case scenario. The accessibility dataset was then joined with the Flickr post density dataset. Some parks are on islands where the accessibility dataset does not reach, so those parks were dropped from the accessibility statistics but were still included in the post density statistics. As with the post density statistics, accessibility statistics were summarised at the national and regional levels. We then tested whether correlation exists between the park accessibility and the post activity: the assumption being, that an easier accessibility would lead to more visitations and therefore to an increased social media posting. We used the minimum value of accessibility for each park, since the parks can be large, and some sections especially outside the road networks can be highly inaccessible. We assumed the minimum value within a park might be its entrance, since it is connected to a larger road network and thus captures the park's accessibility for the average visitor. Pearson correlation coefficient and scatter plotting were done for the minimum value and the post density both globally and regionally.

We then focused on Sub-Saharan African national parks and the concentration of the Flickr posts within them. Three methods to study the point patterns and their dispersion were employed: the goodness-of-fit test based on the quadrat counts, the Ripley's K Function, and the kernel density estimation (*KDE*). To increase the reliability of the point pattern analysis, we limited our scope to the parks with at least 500 posts. The first method simply tested against the null hypothesis that the points are randomly distributed across the study area by applying a uniform grid across the area and counting the points in each cell, or quadrat (Anselin, 2015). Then the probability of the pattern being random was tested with the Pearson χ^2 test, the alternative hypothesis being spatial clustering. After that requirement was satisfied, clustering and dispersion were tested in different scales using Ripley's K Function (Gillan & Gonzalez, 2012).

In the final approach, we created a kernel density raster layer of the Flickr posts in the region using the QGIS Heatmap Plugin with a quartic kernel shape, a grid cell size of 1x1 km, and a radius of 10 km. Universal guidelines for parameter selection appear sparse with case-by-case evaluations for each dataset being more common. Harth and Zandbergen (2014) propose that the grid cell size has little effect on the predictive accuracy. However, too large of a cell size creates coarser results but a smaller grid cell size can increase the processing time of the algorithm. Both Hart and Zandbergen (2014) and Garcia et al. (2015) note the importance of the bandwidth, or the radius, on the final results. We chose the radius based on the size of the national parks and the fact that Flickr posts seemed to be fairly concentrated in general. A larger radius would have saturated the highly concentrated areas, and a smaller radius would not have provided enough distinction between the areas when the pixel size is taken into consideration. We then chose two example national parks, one with a high concentration and one with a low concentration of Flickr photos and created kernel density maps of them. Because of the small size of the

Nairobi National Park, the example park with a low concentration of Flickr photos, we decided to use a grid cell size of 100x100 m and a radius of 1 km for its kernel density map, retaining the cell size to the radius ratio.

Finally, we studied the linguistic content of the Flickr posts by determining the likely language used and examining the most common words in the posts. The text was first pre-processed by discarding various non-linguistic features, such as URLs. Also, to get more accurate results in the language detection, the minimum length of the texts was set to be 15 characters. We noticed the Flickr users describe the images to varying degrees in both the title and the separate description field. Because of this, only posts where both these fields meet the requirements were used. Filtering left a total of 227,434 posts. The language of each post was determined using the Python implementation of the Langdetect library (Shuyo, 2010) which supports the detection of 55 languages. Short texts and multilingual posts created uncertainty in the detection, which is why a post was deemed identified only if the confidence of detection given by the software is over 85 %. Lastly, the posts were filtered for stopwords that occur often but have low semantic information, such as *the* or *his*. The most important words of the dataset were determined by a simple word count and by the term frequency-inverse document frequency (*tf-idf*) method. Tf-idf was used to highlight the terms that are frequent in single documents (in this case, posts) but not in the whole dataset, like the aforementioned stopwords. It is a widely used method in information retrieval and variations of it have been utilised in e.g. summarising recent events on Twitter (Alsaedi, Burnap & Rana, 2016). We employed it to attempt to highlight the infrequent words that still summarise the common topics discussed in the data.

Results

Accessibility and post density

Our results on Flickr post densities in national parks show that post densities are the highest in small parks. An interesting result is that of the top 15 national parks with the highest Flickr post density, five were in the British Virgin Islands. Another noticeable statistic is that many Israeli national parks rank high on the list as well. The highest number of Flickr images was in Yosemite national park in the United States. Overall, 12 of the top 14 national parks with the highest number of posts are in the United States. On the other hand, the parks with the lowest Flickr post density, all have an area over 10,000 km² and were in Canada, South Sudan or Venezuela. At the national level both British Virgin Islands and American Virgin Islands rank on the top in regards of the post density. Overall, island nations rank high on the list. Map of the densities at the national level is presented in Figure 1. Most of the nations with the lowest Flickr post densities are in Sub-Saharan Africa. The highest density region is Eastern Asia with an average of 8.55 Flickr posts per km² (Figure 2, Table 1). The highest number of Flickr images has been posted in Northern America. Excluding Greenland, the lowest densities are in Central Asia and Northern Africa.

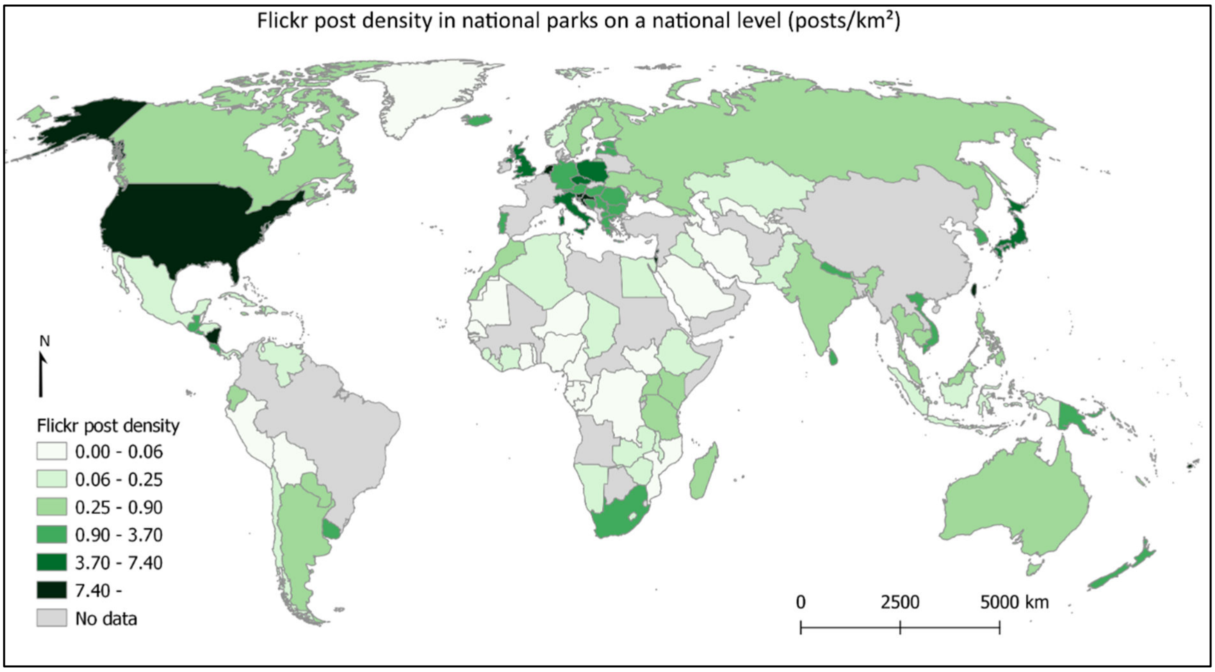


Figure 1. The Flickr posts per km² in national parks at the national level.

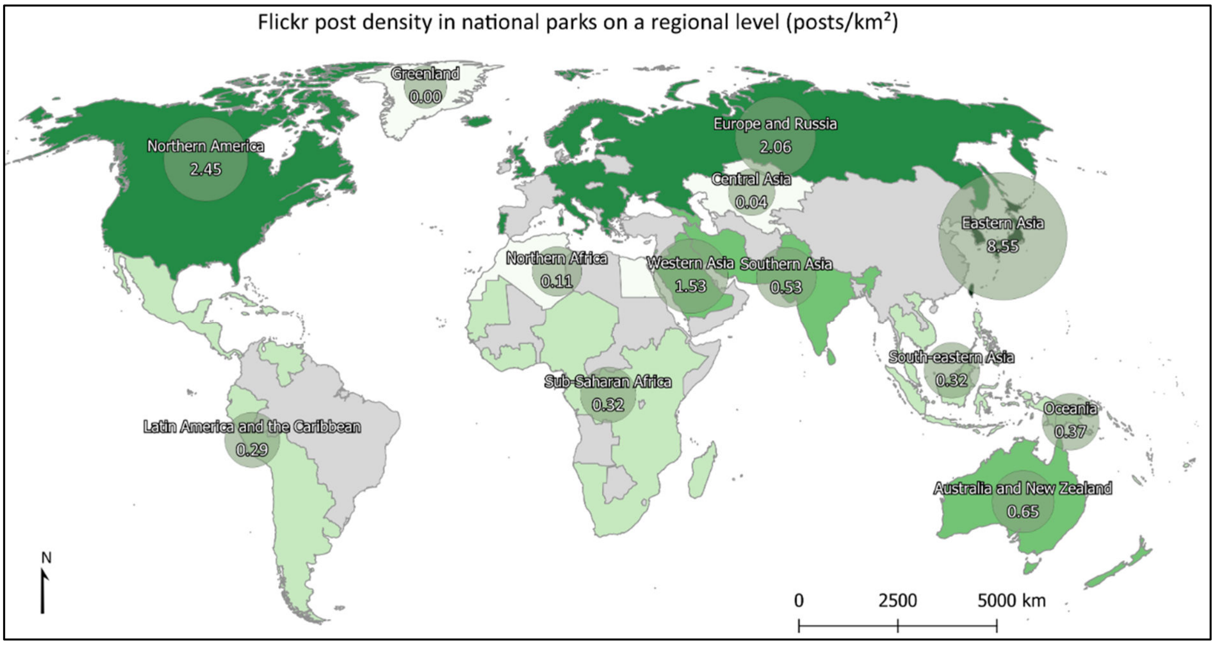


Figure 2. The Flickr posts per km² in national parks at the regional level.

Of the national parks in our dataset that have at least one Flickr post within their area, 125 have a minimum accessibility of 0 minutes. The Flickr post densities of these parks vary from 0 to over 8,500, which is the highest density of any park with accessibility data available. The average minimum accessibility of all the national parks is 162.73 minutes. At the national level, Netherlands has national parks with the highest median minimum accessibility (the lowest number) when the nations with at least five national parks are included. Within these countries, Canada is on the other end of the list with its lowest median minimum accessibility. At the regional level, Eastern Asia and Western Asia have the national parks with the highest median minimum accessibility, while Northern America has the lowest accessibility (Table 1).

Table 1. The summary statistics of the Flickr post densities and the national park accessibilities at the regional level. In the "Number of Parks" column, the number inside the parentheses indicates the number of parks that are included in calculating the park accessibility indices.

Region	Number of parks	Park area total (km²)	Number of Flickr -post total	**Post density (posts/km²)**	Accessibility index (median)
Eastern Asia	41	24,032	205,477	**8.55**	**13.00**
Northern America	99	426,094	**1,044,805**	2.45	123.00
Europe	338 (336)	198,868	410,160	2.06	43.50
Western Asia	83 (81)	15,185	23,292	1.53	16.00
Australia and New Zealand	**500 (497)**	309,351	201,723	0.65	115.00
Southern Asia	89	72,137	38,060	0.53	27.00
Oceania	13	11,024	4,025	0.37	41.00
Sub-Saharan Africa	205 (203)	533,901	170,699	0.32	66.00
South-eastern Asia	194	182,918	58,239	0.32	44.50
Latin America and the Caribbean	350 (339)	607,915	179,121	0.29	63.00
Northern Africa	25	93,183	10,068	0.11	49.00
Central Asia	4	30,685	1,258	0.04	62.50
Greenland	1	**961,673**	689	0.00	2,320.00

According to our analysis, there is no clear correlation between the density of Flickr photos and the minimum value of the accessibility index between the national parks (see Figure 3). The variation between the parks is large, and there are a great number of outliers on both axes: the parks that have either an extremely high post density or a poor accessibility. These parks lie on all regions without a clear pattern. Aggregating the values to the regional level (Figure 4) reveals a weak negative correlation (the post density increases as the time to access the parks decreases), but the data points are too scattered to call this result robust.

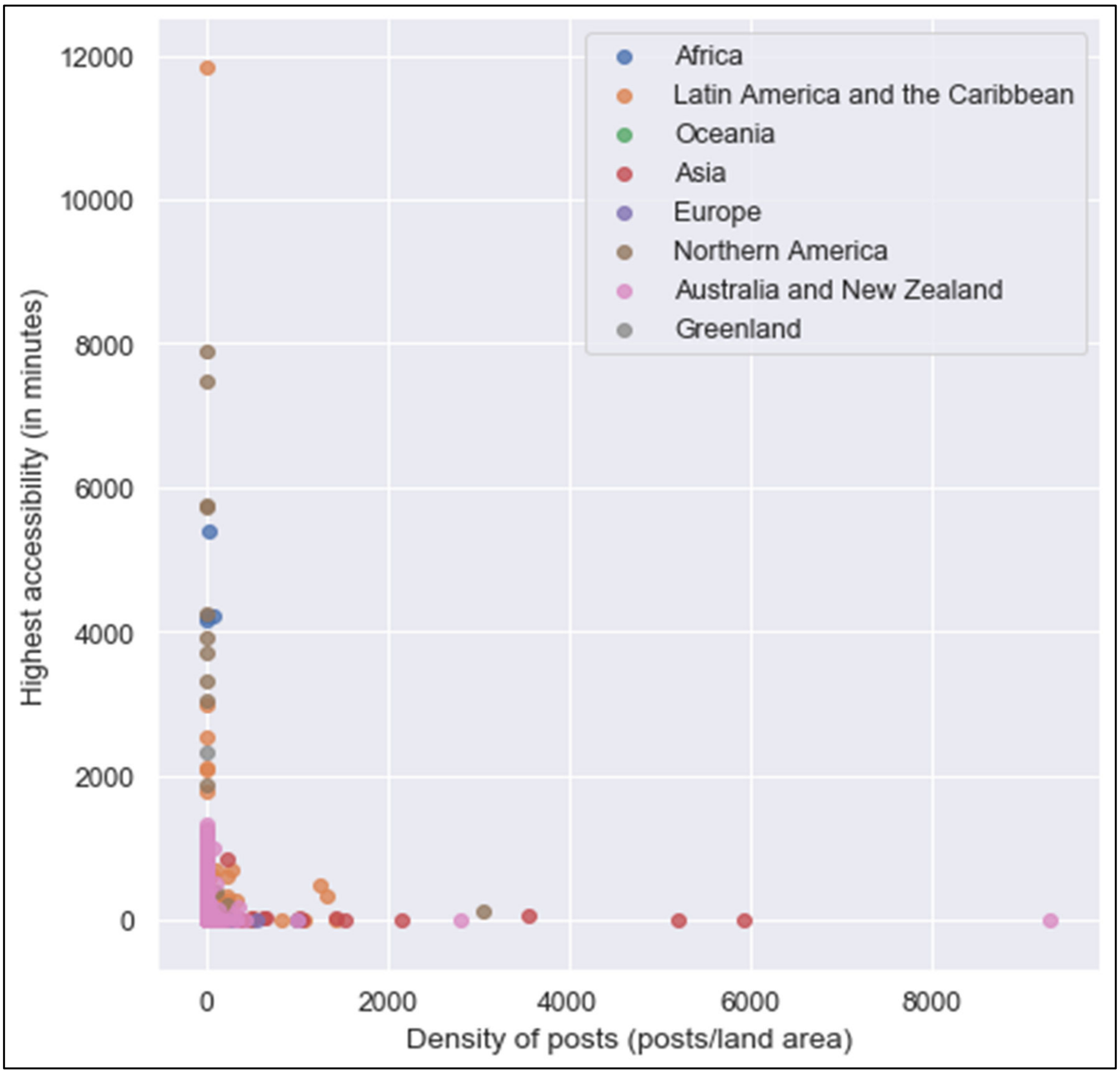


Figure 3. The most accessible (minimum value) section of the parks, and the density of Flickr posts per km². The region park *n* is in is marked by the color scheme.

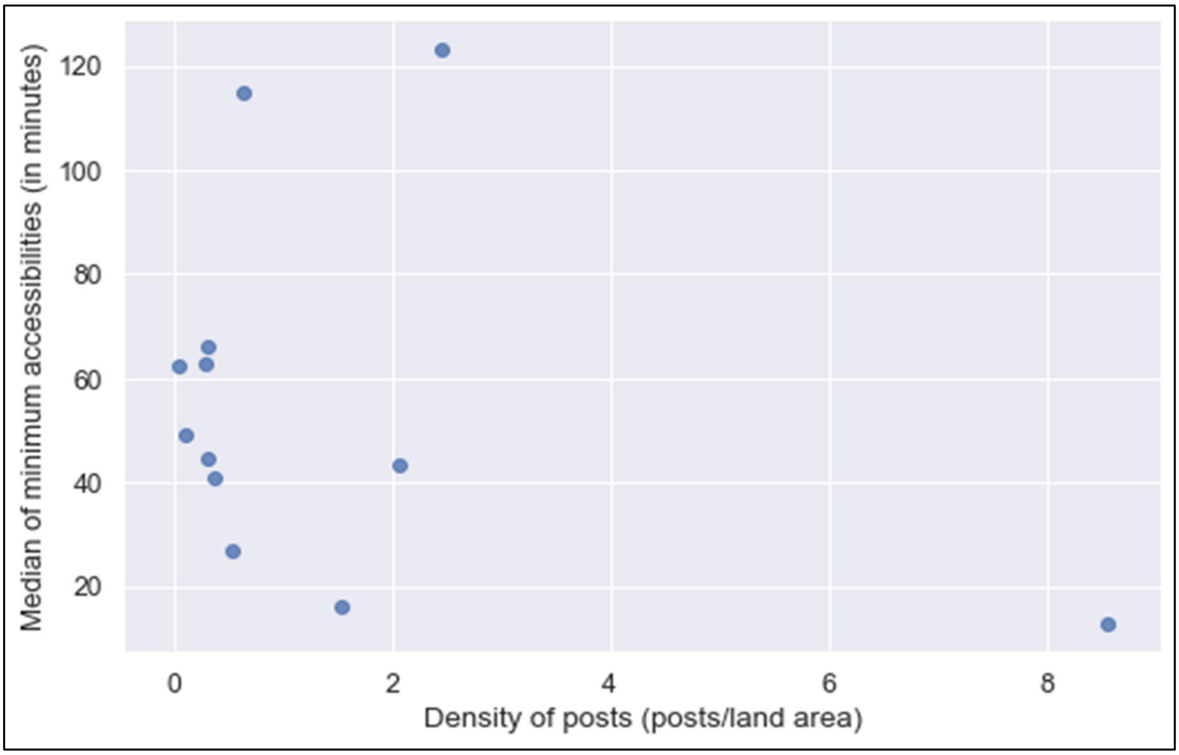


Figure 4. The accessibility and the post densities aggregated to the regional level. The Y-axis is the median of all the minimum accessibility values in the region. The X-axis is the total amount of the posts made in the region divided by the total amount of land area.

Post dispersion in parks

We first analyse the post dispersion within the parks in Sub-Saharan Africa using the quadrat count method. The assumption of the random dispersion could be rejected with a high confidence for all the parks analysed. The analysis of the Ripley's K results suggest clustering patterns for nearly all the parks at multiple distances (see Appendix B e.g. from Serengeti). Using this knowledge and visual information, we select two candidate parks for further inspection. The two resulting heatmaps from the kernel density estimation are illustrated in Figure 5 and Figure 6. Serengeti National Park (Figure 5) is an example of a park with a high concentration of Flickr posts. The concentration is high especially along the primary road which intersects the park in the northwest-southeast direction. The highest concentration of posts is in the middle of the park where the Serengeti Visitors Centre is located. The kernel density map of Nairobi national park presents a more dispersed example of Flickr posts within a park. It is much smaller than Serengeti and in the Kenyan capital city of Nairobi. The only part of the park with a slightly higher concentration of Flickr posts is located near the west-side border of the park where the Sheldrick Elephant & Rhino Orphanage is located. The park is filled with small roads which might explain the low concentration of the Flickr posts.

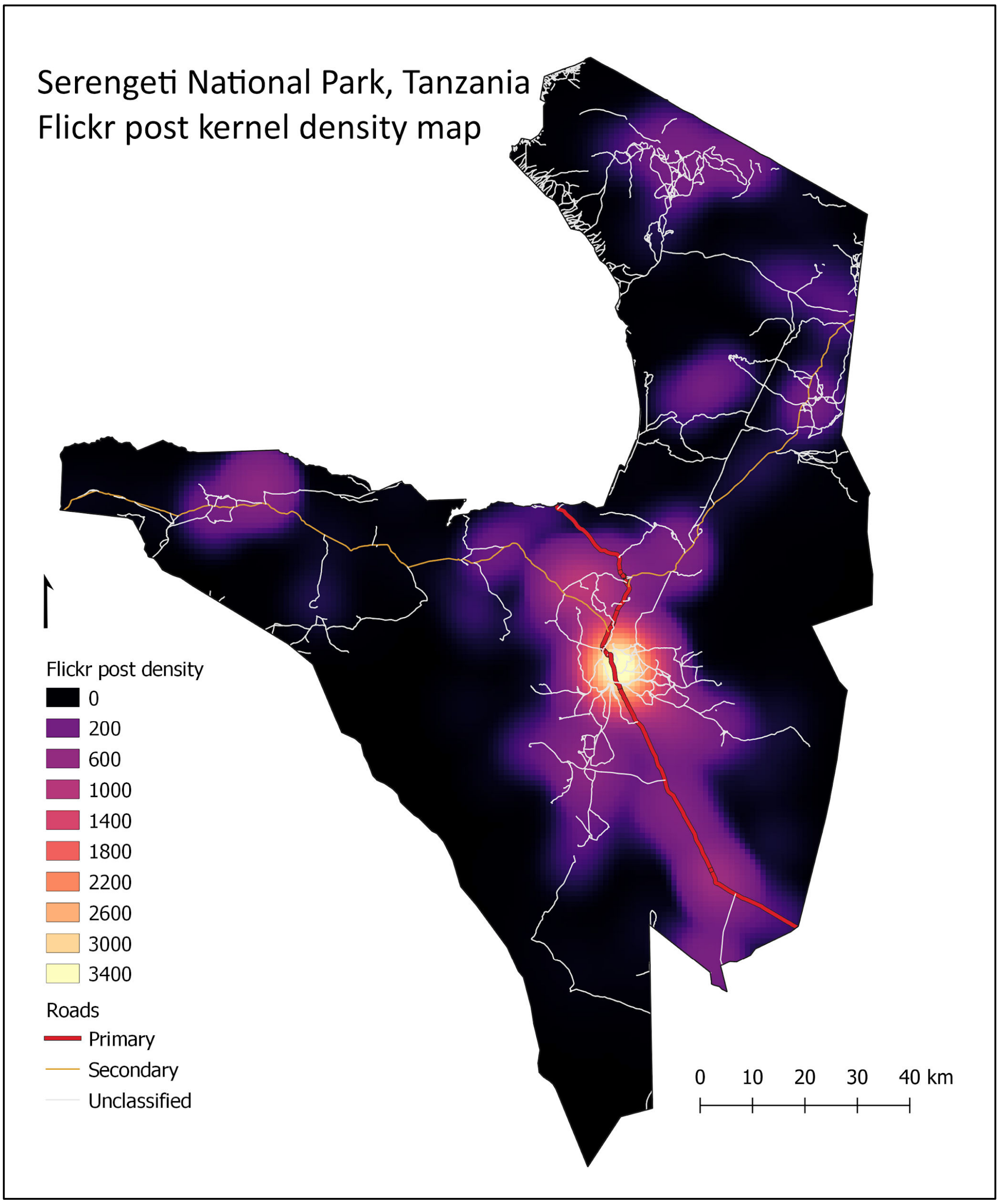


Figure 5. The Flickr post kernel density map of Serengeti National Park, Tanzania. The overall Flickr post density of the park is 1.30 posts/km². Road network © OpenStreetMap contributors.

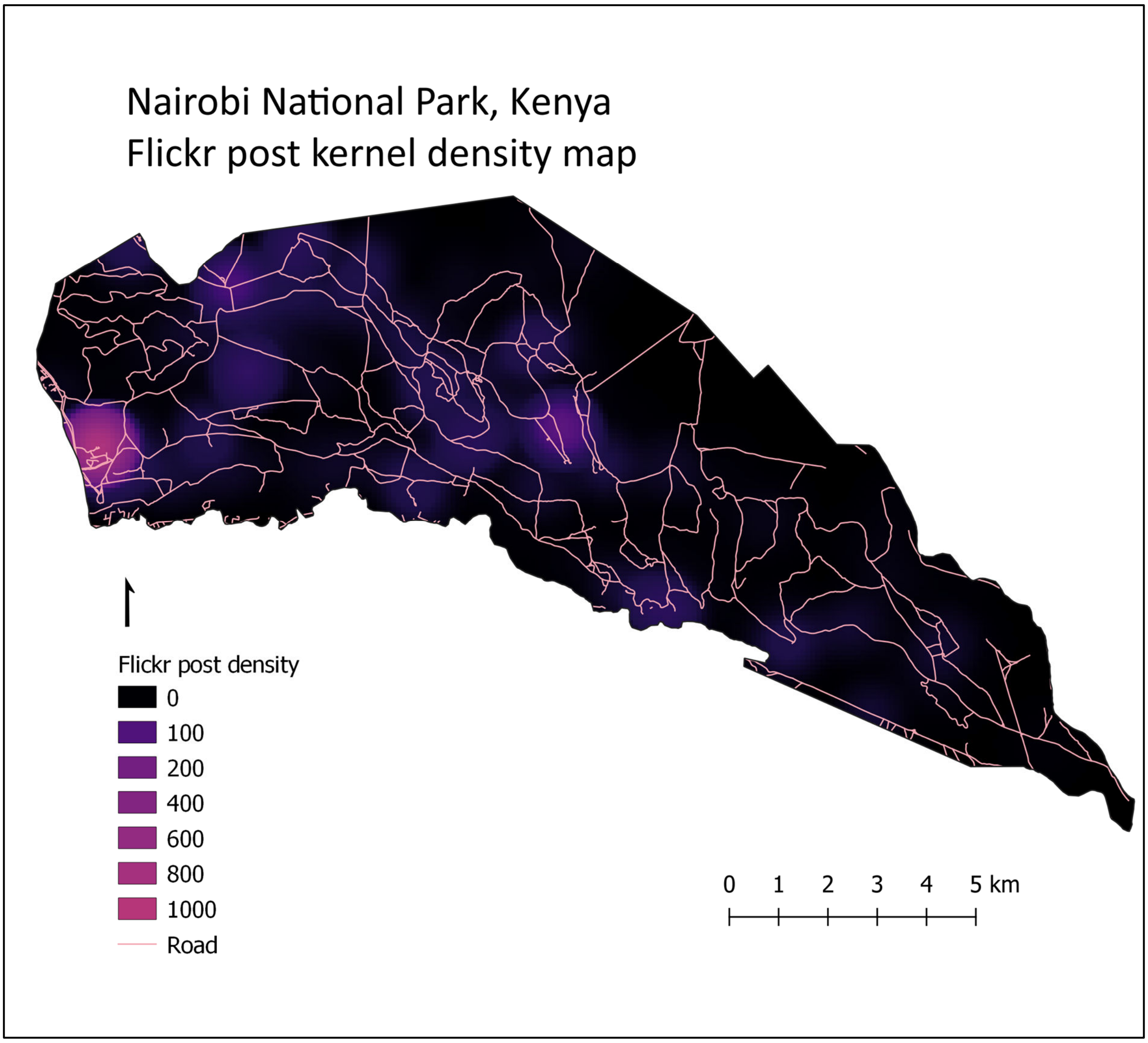


Figure 6. The Flickr post kernel density map of Nairobi National Park, Kenya. The overall Flickr post density of the park is 17.41 posts/km². Road network © OpenStreetMap contributors.

Linguistic analysis

The linguistic analysis of the posts that meet the criteria defined in the methods section shows that an overwhelming majority, about 80 % of the posts, were made in English (Figure 7). The second largest group is too uncertain to be identified, perhaps due to multilingual posts. Rest of the languages, including major world languages such as Spanish, make up only about 10 % of the posts. To learn more about the largest subsection of the posts, we look at the most common and important words in the posts identified as

English (Table 2). Simple word count highlights descriptive words of the surroundings, like *lake*, *mountain,* and *valley* as well as different US national parks (*Grand Canyon, Yellowstone*). Note that the count is case-insensitive and a word like *lake* could be a part of a place name. A tf-idf column does not share any entries with the word count, but similarly highlights the nature aspect of these parks – the words include plant and animal species (*pochard*, *acutifolia*).

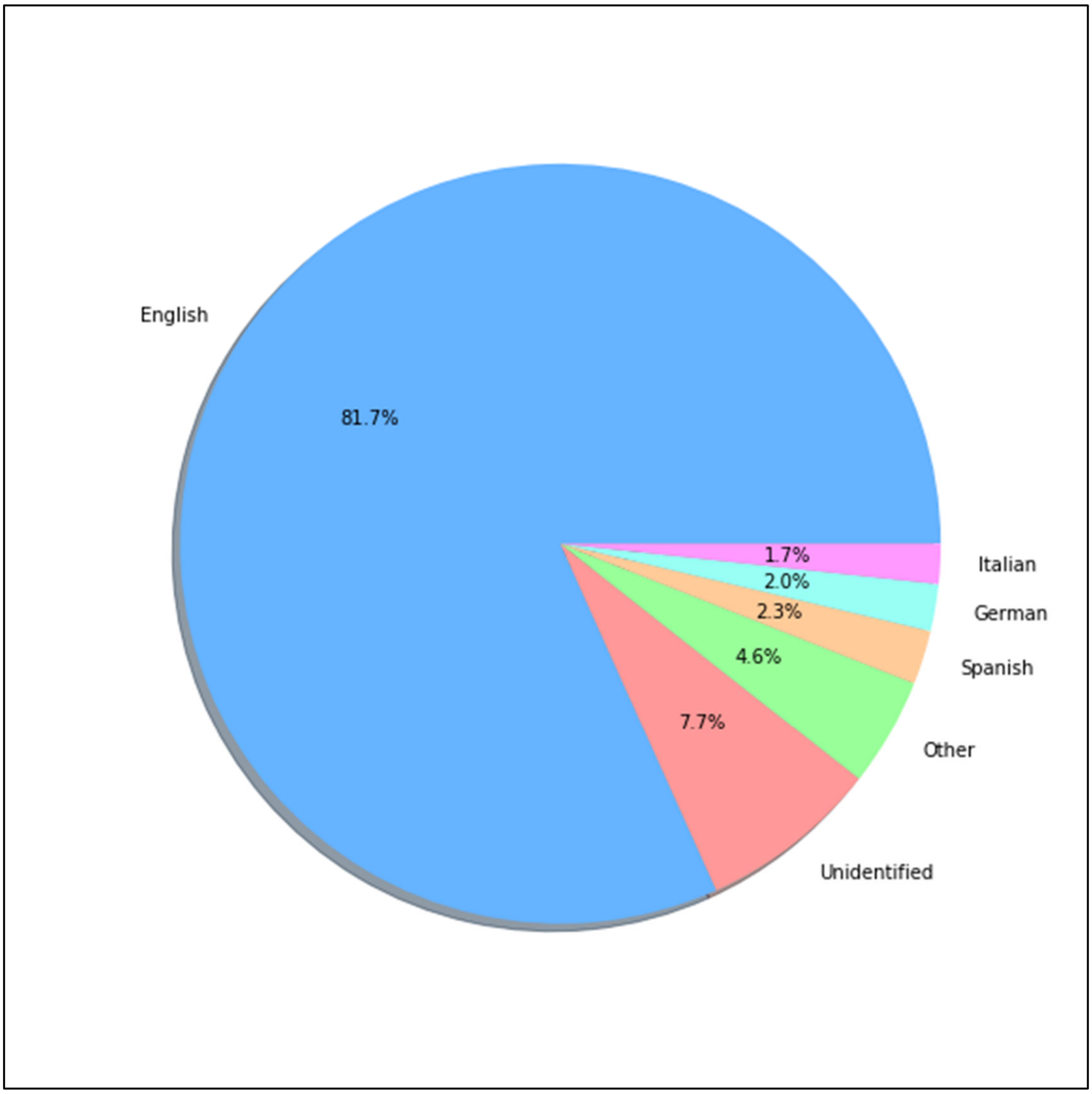


Figure 7. The proportion of the languages identified in the posts.

Table 2. The top 15 words of interest in the English language Flickr posts by the word count and the tf-idf methods.

Word count	***Tf-idf***
park	*pochard*
national	*therates*
canyon	*corunastylis*
lake	*troglodytes*
trail	*sochi*
view	*albiventris*
one	*arabian*
river	*jabba*
south	*queulat*
valley	*culprit*
grand	*acutifolia*
mountain	*tartally*
yellowstone	*flowerpecker*
yosemite	*whitehorn*
day	*abys*

Discussion and Conclusions

The Flickr post density seems to depend on the popularity of the application. The three regions with the highest number of Flickr posts (Northern America, Eastern Asia, and Europe) also have the highest post density within their national parks. However, when examining the national parks inside a country or a region, the results can be comparable. For example, the differences in the Flickr post densities between Canada’s remote and large national parks and Yosemite national park in the United States can be considered representative of the differences in the visitation rates. However, additional information, for example, from other sources of digital data or information about the population and the tourism in the areas, should be included for more in depth conclusion of the actual rates.

No notable correlations between the accessibility and the post density of the parks are found either at the global or regionally aggregated scale. Multiple reasons can explain this. First, the national park dataset is heterogeneous: the parks are of widely different scales, some consist mostly of water, some have next to none posts, and some are on remote islands. The term ‘national park’ thus encompasses so many different types of areas that studying them together at the global scale proved difficult. The WDPA dataset does not seem to be exhaustive either, as, for example, there are no parks marked in e.g. France, Spain and China. Further research might benefit from studying only a clearly defined subset of the parks, such as parks of predefined scale, in a single country, or with a minimum number of posts in them.

Second, the accessibility dataset by Weiss et. al (2018) has some shortcomings when used in this manner. It measures the accessibility to major population centres which might not represent how accessible the park is to its visitors. For example, the population in the surrounding areas might be dispersed enough not to qualify as a population centre. The parks that are popular with tourists might also be more accessible due to regular transportation connections. It is also possible that the accessibility in general is not that important of a pull factor for national park visitors. Further research is needed to determine the relation between the accessibility and the visitations rates as indicated by the post density. Perhaps different accessibility datasets that have a higher resolution or are calculated specifically for the national parks at hand could prove useful.

The results on the point pattern and the Kernel density analyses in Sub-Saharan Africa show that the Flickr posts are fairly concentrated in the national parks of the region. Inspection of Serengeti and Nairobi national parks show that the most concentrated areas are located near the tourist attractions like the visitor centre in Serengeti or the elephant and rhino orphanage in Nairobi. Also, the roads and the paths seem to dictate the concentration of the Flickr posts. The comparatively lower clustering of posts in Nairobi national park might be caused by the extensive road network within the park. Further research could be done into the importance of roads as a driving factor of the point patterns by, for example, measuring how far the points are on average from the nearest road. Another plausible explanation for the concentration of the posts in parks like Serengeti may be that tourists visit the parks with dangerous wildlife as a part of a guided tour. Topography of the parks, which makes some sections unreachable, is also an important factor to consider in further research. Including accessibility data of different regions inside the park to the post concentration analysis could provide more answers for the clustering. Methods other that the ones used here (Ripley's K, KDE) for studying the patterns should be considered as well.

The linguistic analysis of the posts gives further insight on both the user base and what the visitors do at the parks. First, an overwhelming majority of the analysed posts (82 %) are made in English, which might indicate an overrepresentation of people from English speaking countries and in part explain the high post densities in North America and Europe. Though the analysis of Instagram posts from Finland reveals that the users post in English in significant amounts regardless of their home country (Hiippala et al., 2019). The country of origin of the posters and thus the number of tourists from each country could be estimated using a similar algorithm as Hiippala et al. use (2019, pp. 301). The simple word count analysis could be expanded in future research with a more complex topic modelling (see, e.g. Hiippala et al., 2019, pp. 303). This analysis could be used to learn more about the interest of the visitors even at a park-level.

In this article, we examined Flickr data in national parks worldwide. We first described the patterns the posts make at the regional and national scales. We learned that the link between the accessibility and the social media post density is not clear at the global or regional level. We then looked at the dispersion of the posts in the parks in Sub-Saharan Africa and found that clustering, as opposed to dispersing or forming random patterns, is extremely common. Finally, we described some features found in the linguistic content of the posts. We attempted to explain each result and offer ideas for future studies. In summary, we conclude that we achieved the study aims of both exploring the datasets and their possibilities in conservation science and offering methods for future research.

References

Africa Sun News. (2003). Africa national parks list. *Africa Sun News.* Retrieved 26.3.2020, available: http://www.africasunnews.com/national_parks.html

Alsaedi, N., Burnap, P. & Rana, O. (2016). *Temporal TF-IDF: A high performance approach for event summarization in Twitter*. IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, 2016. https://doi.org/10.1109/WI.2016.0087

Anselin, L. (2015). *Point pattern analysis: quadrat counts* [video file]. Retrieved 11.5.2020, available: https://www.youtube.com/watch?v=Ww95WKxUoZk

Balmford, A., Green, J. M. H., Anderson, M., Beresford, J., Huang, C., Naidoo, R., Walpole, M. & Manica, A. (2015). Walk on the wild side: estimating the global magnitude of visits to protected areas. *PLoS Biology 13*. https://doi.org/10.1371/journal.pbio.1002074

Bouton, S. N., Frederick, P. C., Dosualdo Rocha, C., Barbosa Dos Santos, A. T. & Bouton, T. C. (2009). Effects of tourist disturbance on wood stork nesting success and breeding behaviour in the Brazilian pantanal. *Waterbirds 28*, 487-497. https://doi.org/10.1675/1524-4695(2005)28[487:EOTDOW]2.0.CO;2

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

Buckley, R. (2011). Tourism and environment. *Annual Review of Environment and Resources 36*, 397-416. https://doi.org/10.1146/annurev-environ-041210-132637

Buckley, R. C., Morrison, C. & Castley, J. G. (2016). Net effects of ecotourism on threatened species survival. *PLoS ONE 11*, 23-25. https://doi.org/10.1371/journal.pone.0147988

Cessford, G. & Muhar, A. (2003). Monitoring options for visitor numbers in national parks and natural areas. *Journal for Nature Conservation 11(4)*, 240-250. https://doi.org/10.1078/1617-1381-00055

Crush, J. S. (1980). National parks in Africa: a note on a problem of indigenization. *African Studies Review 23(3)*, 21-32. https://doi.org/10.2307/523669

Di Minin, E., MacMillan, D. C., Goodman, P. S., Slotow, R. & Moilanen, A. (2013). Conservation businesses and conservation planning in a biological diversity hotspot. *Conservation Biology 27*, 808-820. https://doi.org/10.1111/cobi.12048

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

Eagles, P. F. J. & Wade, D. (2006). Tourism in Tanzania: Serengeti National Park. *Bois et forêts des tropiques 290(4)*, 73-80.

Encyclopaedia Britannica. (2020). National park. *Encyclopaedia Britannica.* Retrieved 13.3.2020, available: https://www.britannica.com/science/national-park

Elith, J., Graham, C. H., Anderson, R. P. et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography 2,* 129-151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Frank, L., Bradley, M., Kavage, S., Chapman, J. & Lawton, T.K. (2008). Urban form, travel time, and cost relationships with tour complexity and mode choice. *Transportation 35 (1)*, 37–54. https://doi.org/10.1007/s11116-007-9136-6

García, R., Lopez, M., Pérez, S., Juan, C. & Raúl., P. (2015). *The kernel density estimation for the visualization of spatial patterns in urban studies*. 15th International Multidisciplinary Scientific GeoConference SGEM, Albena, Bulgaria, 2015 https://doi.org/10.5593/SGEM2015/B21/S8.111

Gillan, J. & Gonzalez, L. (2012). Ripley's K function and pair correlation function. *The Landscape Toolbox*. Retrieved 4.5.2020, available: https://wiki.landscapetoolbox.org/doku.php/spatial_analysis_methods:ripley_s_k_and_pair_correlation_function

Goodwin, H. (1996). In pursuit of ecotourism. *Biodiversity and Conservation 5*, 277-291. https://doi.org/10.1007/BF00051774

Gössling, S. (1999). Ecotourism: a means to safeguard biodiversity and ecosystem functions? *Ecological Economics 29(2)*, 303-320. https://doi.org/10.1016/S0921-8009(99)00012-9

Gössling, S. (2002). Global environmental consequences of tourism. *Global Environmental Change 12*, 283-302. https://doi.org/10.1016/S0959-3780(02)00044-4

Hart, T. & Zandbergen, P. (2014). Kernel density estimation and hotspot mapping: Examining the influence of interpolation method, grid cell size, and bandwidth on crime forecasting. *Policing: An International Journal 37(2)*, 305-323. https://doi.org/10.1108/PIJPSM-04-2013-0039

Hausmann, A., Slotow, R., Fraser, I. & Di Minin, E. (2016). Ecotourism marketing alternative to charismatic megafauna can also support biodiversity conservation. *Animal Conservation 1*, 208-211. https://doi.org/10.1111/acv.12292

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

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., Di Minin, E. (2017a). Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conservation Letters 11(1)*. https://doi.org/10.1111/conl.12343

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

Levin, N., Kark, S. & Crandall, D. (2015). Where have all the people gone? Enhancing global conservation using night lights and social media. *Ecological Applications 25*, 2153-2167. https://doi.org/10.1890/15-0113.1

Margules, C. R. & Pressey, R. L. (2000). Systematic conservation planning. *Nature 405,* 243-253. https://doi.org/10.1038/35012251

Mavoa, S., Witten, K., McCreanor, T. & O'Sullivan, D. (2012). GIS based destination accessibility via public transit and walking in Auckland, New Zealand. *Journal of Transport Geography 20 (1),* 15–22. https://doi.org/10.1016/j.jtrangeo.2011.10.001

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: a revolution that will transform how we live, work, and think*. Mifflin Harcourt Publishing Company, New York, USA. Pp. 1-242.

Kaplan, A. & Haenlein. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons 53(1)*, 59-68. https://doi.org/10.1016/j.bushor.2009.09.003

Knight, A. T., Cowling, R. M. & Campbell, B. M. (2006). An operational model for implementing conservation action. *Conservation Biology 20*, 408-419. https://doi.org/10.1111/j.1523-1739.2006.00305.x

Krüger, O. (2005). The role of ecotourism in conservation: panacea or Pandora's box? *Biodiversity & Conservation 14*, 579-600. https://doi.org/10.1007/s10531-004-3917-4

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

Pickering, C. M. & Hill, W. (2007). Impacts of recreation and tourism on plant biodiversity and vegetation in protected areas in Australia, *Journal of Environmental Management 85*, 791-800. https://doi.org/10.1016/j.jenvman.2006.11.021

Ranaweerage, E., Ranjeewa, A. D. G. & Sugimoto, K. (2015). Tourism-induced disturbance of wildlife in protected areas: a case study of free ranging elephants in Sri Lanka. *Global Ecological Conservation 4*, 625-631. https://doi.org/10.1016/j.gecco.2015.10.013

Richards, D. R. & Friess, D. A. (2015). A rapid indicator of cultural ecosystem service usage at a fine spatial scale: content analysis of social media photographs. *Ecological Indicators 53*, 187-195. https://doi.org/10.1016/j.ecolind.2015.01.034

Serengeti National Park. Serengeti National Park. *Serengeti National Park.* Retrieved 26.3.2020, available: https://www.serengetinationalpark.com/

Shuyo, N. (2010). *Language Detection Library for Java*. Retrieved 11.5.2020, available: https://github.com/shuyo/language-detection

Siegfried, W. R., Benn, G. A. & Gelderblom, C. M. (1998). Regional assessment and conservation implications of landscape characteristics of African national parks. *Biological Conservation 84(2)*, 131-140. https://doi.org/10.1016/S0006-3207(97)00110-9

Smith, R. J., Verissimo, D. & Macmillan, D. C. (2010). Marketing and conservation: how to lose friends and influence people. *In* Leader Williams, N., Adams, W. & Smith, R. (Eds.): *Trade-offs in conservation: deciding what to save,* pp. 215-232. Blackwells, Oxford, UK.

Su, S., Wan, C., Hu, Y. & Cai, Z. (2016). Characterizing geographical preferences of international tourists and the local influential factors in China using geo-tagged photos on social media. *Applied Geography 73*, 26-37. https://doi.org/10.1016/j.apgeog.2016.06.001

Steven, R., Pickering, C. & Castley, G. J. (2011). A review of the impacts of nature based recreation on birds. *Journal of Environmental Management 92*, 2287-2294. https://doi.org/10.1016/j.jenvman.2011.05.005

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

UNEP-WCMC (2019). *User manual for the World Database on Protected Areas and world database on other effective area-based conservation measures: 1.6*. UNEP-WCMC: Cambridge, UK. Retrieved 11.5.2020, available: http://wcmc.io/WDPA_Manual

UNEP-WCMC & IUCN (2020), *Protected planet: The World Database on Protected Areas (WDPA)* [On-line], version 01/2020, Cambridge, UK: UNEP-WCMC and IUCN. Retrieved 11.5.2020, available: www.protectedplanet.net

University of Helsinki. (2015). Social media data could contribute to conservation science. *ScienceDaily*. Retrieved 11.3.2020, available: https://www.sciencedaily.com/releases/2015/09/150915105007.htm

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

Weiss, D. J. et al. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, 553(7688), 333-336. https://doi.org/10.1038/nature25181

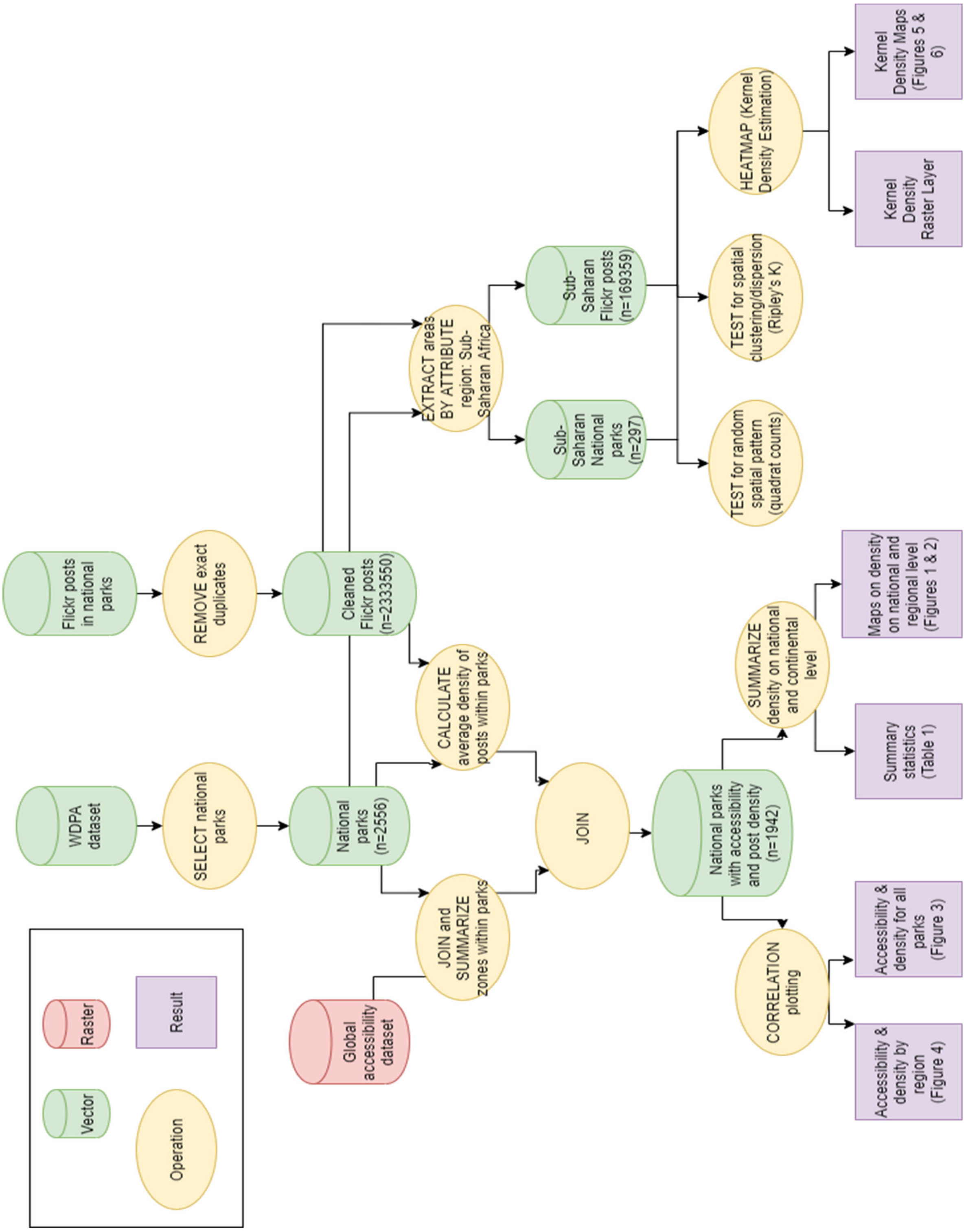
Willemen, L., Cottam, A. J., Drakou, E. G. & Burgess, N. D. (2015). Using social media to measure the contribution of red list species to the nature-based tourism potential of African protected areas. *PLoS ONE 10*. https://doi.org/10.1371/journal.pone.0129785

Woodroffe, R. & Ginsberg, J. R. (1998). Edge effects and the extinction of populations inside protected areas. *Science 280*, 2126-2128. https://doi.org/10.1126/science.280.5372.2126

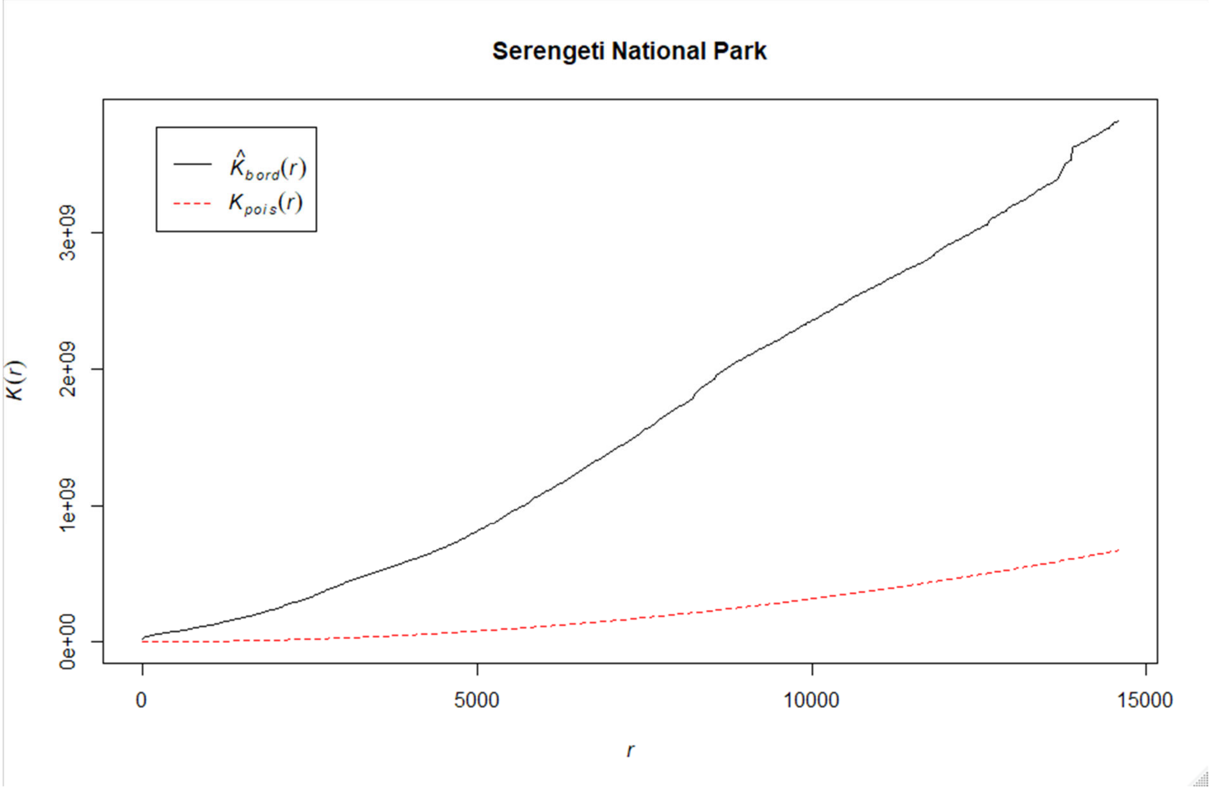
World Tourism Organization. (2015). Towards measuring the economic value of wildlife watching tourism in Africa – briefing paper. *World Tourism Organization.*

Appendices

Appendix A: The spatial analysis workflow


Flickr posts in national parks
REMOVE exact duplicates
Cleaned Flickr posts (n=2333550)
EXTRACT areas BY ATTRIBUTE region: Sub-Saharan Africa
Sub-Saharan Flickr posts (n=169359)
Sub-Saharan National parks (n=297)
HEATMAP (Kernel Density Estimation)
TEST for spatial clustering/dispersion (Ripley's K)
TEST for random spatial pattern (quadrat counts)
Kernel Density Maps (Figures 5 & 6)
Kernel Density Raster Layer
WDPA dataset
SELECT national parks
National parks (n=2556)
CALCULATE average density of posts within parks
JOIN and SUMMARIZE zones within parks
Global accessibility dataset
JOIN
National parks with accessibility and post density (n=1942)
SUMMARIZE density on national and continental level
Maps on density on national and regional level (Figures 1 & 2)
Summary statistics (Table 1)
CORRELATION plotting
Accessibility & density for all parks (Figure 3)
Accessibility & density by region (Figure 4)
Raster
Vector
Result
Operation

Appendix B: The example plot of Ripley's K Function. Interpret it like this: if the black line is above the red dotted line (the result of a Poisson point process), it is clustered at that distance (X-axis). If below, it is dispersed and if roughly the same, the points are randomly located.


Serengeti National Park
$\hat{K}_{bord}(r)$
$K_{pois}(r)$
$K(r)$
0e+00
1e+09
2e+09
3e+09
r