Evaluating Twitter for Nature-Related Public Health Research

December 12, 2017

Andrew Larkin

**Abstract**

Purpose: Green space is associated with myriad health outcomes, including increased longevity, cognitive function, and social cohesion. Social media, including Twitter, is used by 85-90% of young adults in the US, and can potentially facilitate public health efforts to research and promote young adult green space attitudes, perceptions, and use.

Methods: Tweets with keywords related to green space from the Portland, OR metropolitan region were collected from January 1st to December 1st, 2017 using the Twitter stream API. Ancillary tweets and major social events related to greenspace were identified using term-frequency-inverse document frequency analysis. Latent constructs for physical and emotional interactions with green space were derived using part of speech tagging coupled with topic-related word frequencies. Associations between physical, emotional, and social interactions with greenspace were evaluated using sentiment and time series analyses.

Results: Sentiment, social, visual, and physical latent construct scores were all temporally variant and positively associated. Significant social events included women’s right marches in public parks, bodies found in popular forests, selling of public forest land, and large forest fires. Scores for visual and exercise-related constructs both peaked in early summer, with several additional local maxima, suggesting both seasonal and short-term variation in greenspace use and perceptions.

Conclusions: Social, emotional, and physical interactions with greenspace are correlated, and all need to be considered when evaluating the impacts of greenspace on public health. Twitter is a viable data source for identifying self-reported trends in greenspace use and estimating the influence of greenspace on major social events.

**Background**

Natural components of urban environments, also called green spaces, are associated with multiple positive health outcomes. These outcomes span across multiple dimensions of human health. Physical health outcomes include increased birth weight (Ebisu, Holford, & Bell, 2016), decreased blood pressure (Shanahan et al., 2016), decreased risk of type 2 diabetes (Astell-Burt, Feng, & Kolt, 2014), rates of preschool childhood obesity (Lovasi et al., 2013), and mortality (Takano, Nakamura, & Watanabe, 2002). Mental health outcomes include increased cognitive function (Bratman, Hamilton, & Daily, 2012), ability to cope with stress (van den Berg, Maas, Verheij, & Groenewegen, 2010), increased positive attitude (Takano et al., 2002), and lower levels of symptoms associated with depression and anxiety (Beyer et al., 2014). Community health outcomes include increased social cohesion (Shanahan et al., 2016), sense of community (Maas, Van Dillen, Verheij, & Groenewegen, 2009), and, although dependent on neighborhood context, increased sense of safety (Maas, Spreeuwenberg, et al., 2009) and accessibility (Makris, 2015).

To leverage green space environments for improving health we must first understand how people value, perceive, and use green space, using methods that are conducive and likely to capture contextual characteristics of hard to reach groups including young adults. Social media is a data source that has potential to capture characteristics and activity patterns. In the US, between 85% (Rainie, Fox, & Duggan, 2014) to 90% (Perrin, 2015) of young adults actively participate in social media and approximately 1/3 of young adults (Mitchell & Page, 2014) completely rely on social media as their source for news. One of the most active social media sites is Twitter, with 330 million active monthly users (<https://about.twitter.com/company>).

In this paper, we collected 11 months of Tweets related to green space from the Portland, OR metropolitan region and analyzed tweets for social, emotional, and physical events related to health.

**Methods**

*Twitter Search*

Green-space related Tweets were collected from January 1st to December 1st, 2017 using Python v. 2.7 (van Rossum & Drake Jr, 2010) and the Python module Tweepy (Roesslein, 2015). Tweets were downloaded if they contained one or more keywords shown below in Table 2. In addition to Tweet content, the time of tweet and self-identified user location were downloaded. Tweets were then screened for self-identified locations within the Portland, OR area and English language. Tweets were further screened by generating and applying a list of exclusion unigram, bigram, and trigram phrases commonplace in Tweets unrelated to greenspace but not present in greenspace related Tweets (e.g. “ I **parked my car** next to yours”).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Twitter search keywords | | | | |
| backyard(s) | forest(s) | lawn(s) | pasture(s) | stream(s) |
| bush(es) | garden(s) | leaves | plant(s) | trail(s) |
| crop(s) | grass | mountain(s) | prairie(s) | tree(s) |
| field(s) | hay | nature | river(s) | woods |
| flower(s) | lake(s) | park(s) | riverside | yard(s) |

Table 2. Twitter search keywords.

*Identifying Social Themes*

Tweets were then partitioned into eight day intervals referred to this paper as eight day weeks and major themes within the tweet content for each week were identified using term-frequency-inverse document frequency (TF-IDF) with the NLTK Python module (insert reference). The author intended to partition Tweets into seven day weeks, but made a coding error that was caught in time to update the paper but not reanalyze study results. TF-IDF analysis identified several themes unrelated to greenspace. The responsible ancillary tweets were removed from the dataset before further analysis.

*Developing Latent Constructs for Exercise and Aesthetic Stimulation*

Visual stimulation and physical exercise are two hypothesized pathways of action for greenspace effects on mental and physical health. Latent constructs for these pathways were derived by developing topic word lists for each construct (Table 3) and counting the weekly frequency of topic words in collected Tweets. To ensure topic words were directly related to greenspace rather than other sentence nouns, Tweets were lemmatized and tagged using the Stanford NLP Part of Speech Tagger (Manning et al., 2014). Only topic words with a syntactic relationship to one of the Twitter search keywords listed in Table 2 were included in frequency counts.

|  |  |
| --- | --- |
| Exercise | Aesthetic |
| Hike | View |
| Walk | Beautiful |
| play | Sunset |
| run | green |

Table 2. Topic words for Exercise and Aesthetic latent constructs.

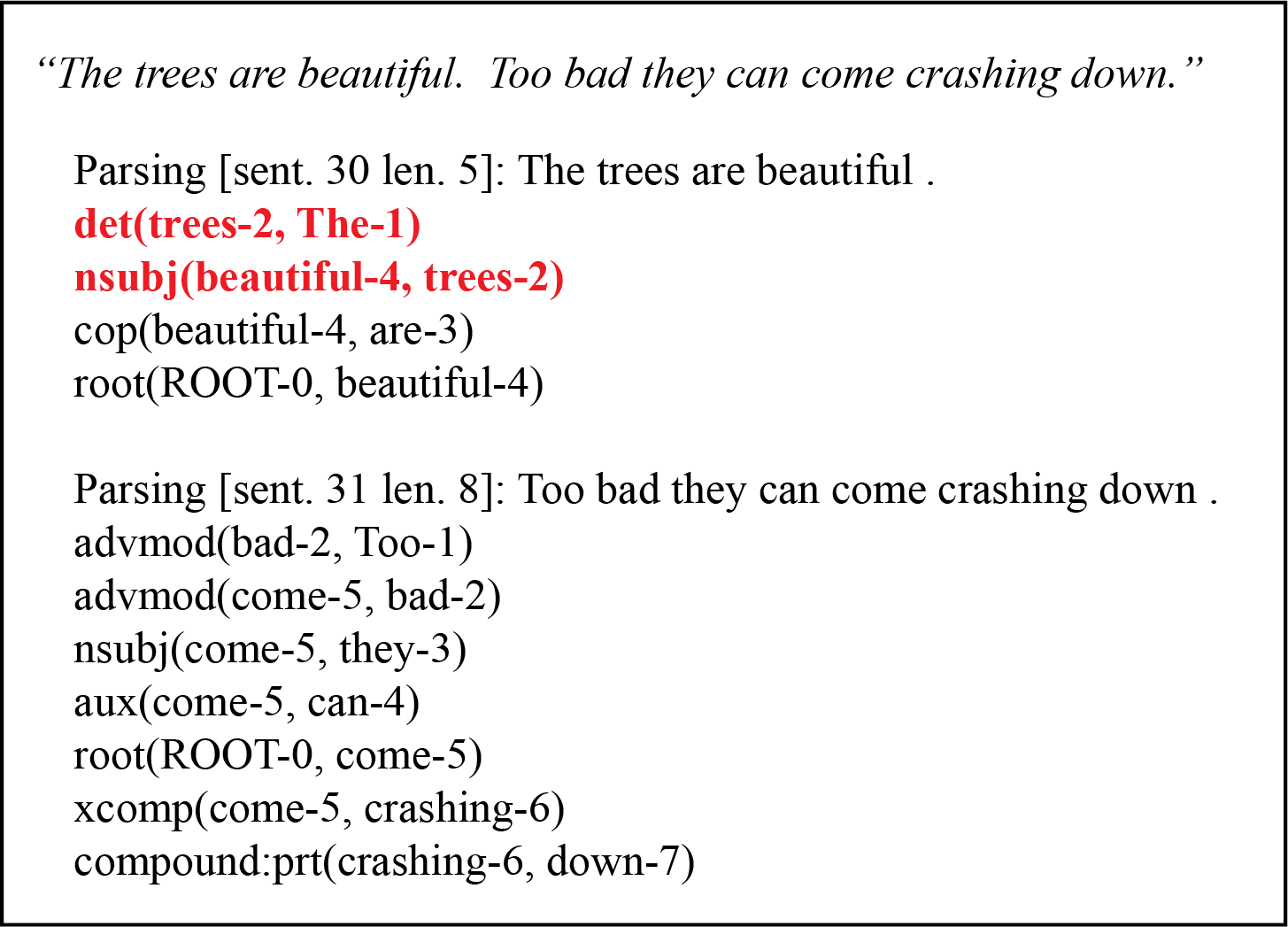


Figure 1. Example tweet and output produced by the Stanford NLP. Keyword syntactic relationships are highlighted in red. In this example sentence, only the words ‘the’ and ‘beautiful’ have a syntactic relationship between the word ‘trees’ and would be considered when deriving word frequencies.

*Sentiment Analysis*

Estimated sentiment for each tweet was derived using the VADER sentiment intensity analyzer within the Python NLTK library (Hutto and Gilbert, 2014). Sentiments were averaged for each week, excluding sentiments with positive and negative partial likelihood scores of 0 (i.e. no positive or negative content in tweet).

*Time series analysis*

Weekly word frequency, latent construct scores, and sentiment scores were plotted as a function of time. Correlations scores and sentiment were derived and compared to identify interactions between hypothesized social, psychological, and physical pathways of action.

*Results*

We collected 72,007 Tweets that met inclusion criteria. Among 100 tweets randomly sampled from included tweets, 84 were directly relevant to greenspace, 81 of which directly relevant to greenspace conditions in the Portland, OR region. Out of 300 days of coverage, 12 days (4%) have partial temporal coverage due to logistical difficulties. Inspection of partial tweet content and historical records suggest no major events occurred on the days of partial coverage.

Trigrams with TF-IDF scores greater than 0.1 are shown below in Table 3. Each trigram corresponds to a significant local or national event related to greenspace. For example, during week 2, the Tom McCall Waterfront Park served as the site for a large women’s march. Collected tweets during the event describe greenspace as a public area for gathering, a source of visual stimulation, and a safe place for physical exercise (walking and marching) (Table 3).

|  |  |  |  |
| --- | --- | --- | --- |
| Week | Keywords | Score | Example Tweet |
| 2 | mccall waterfront park | 0.112 | A picture I took earlier of the snow at waterfront :) @ Tom McCall Waterfront Park https://t.co/EALSy8SUpB |
| 3 | national park service | 0.156 | Rogue National Park Tweets Climate Change Facts in Defiance of Trump! How dare they smear our beloved oil barons!? https://t.co/QnxLalgMp9 |
| 6 | elliott state forest | 0.118 | Stop Sale of Elliot State Forest CALL Treasurer Read &amp; DEMAND that he switch his vote:PHONE 503-378-4329EMAIL : CONTACT@TOBIASREAD.COM |
| 11 | happy earth day | 0.143 | Happy Earth Day! We started our day on the trails, so grateful for Forest Park and these runners? https://t.co/IXWghJXmr0 |
| 31 | eagle creek trail | 0.140 | Still-burning Indian Creek Fire extends trail closures to much of Eagle Creek trail https://t.co/YKFWdeuPi6 |
| 39 | near hideaway lake | 0.114 | Officials say missing man near Hideaway Lake is from Molalla and is a lost hunter. #LiveonK2? https://t.co/4fKxarQFmh |

Table 3. Trigrams from weekly tweets with TF-IDF scores greater than 0.1

Weekly word frequency and latent variable scores are shown below in Figures 2 and 3. To adjust for low sampling rates because of technical errors during week 0, scores were proportionally adjusted based on the ratio of total tweets in the partial coverage week to total number of tweets in the following week (4.48). Due to large inter-week variability, 4 week moving averages are shown in Figures 2 and 3 to improve visual interpretation. The terms ‘beautiful’ and ‘walk’ both peaked globally at 15-16 weeks, with local peaks at weeks 30-31. The terms ‘run’ and ‘hike’ globally peaked at weeks 22-23 (also a local peak for the term ‘beautiful’). The term ‘view’ peaked at weeks 35-37. Weeks of score peaks appear to be independent of major social events, except for week 31 where tweets are a mix of positive and negative comment regarding wildlife beauty, danger, and inability to access nature due to wild files.

Figure 2. Four week moving average for aesthetic categories.

Figure 3. Four week moving average for exercise categories

Sentiment scores are shown in Figure 4. Sentiment scores peak during weeks 14-15, during which the terms ‘beauty’ and ‘walk’ reach their global maxima. Sentiment was lowest during week 31, which coincides with the observed wildfire social event. Sentiment scores reached another local minimum during weeks 39-40, which coincided with multiple bodies found in a popular recreational park. Correlations between sentiment, and 4 week moving averages of word frequencies, and latent variables are shown in Figure 4. Sentiment was most strongly associated with the latent construct ‘exercise’, and was also correlated with ‘walk’ and ‘hike’ activities and visual attributes ‘green’, ‘beautiful’, and the ‘aesthetic’ latent construct. Interestingly, the ‘aesthetic’ and ‘exercise’ latent constructs are strongly correlated (r = 0.77), suggesting that visual and exercise pathways to greenspace health benefits are interdependent.

Figure 4. Weekly sentimental scores. Sentiment was lowest during forest fire season (week 31) and discovery of multiple bodies in a public park (week 39)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | sentiment | view | sunset | green | beautiful | aesthetic | run | hike | play | walk | exercises |
| sentiment | 1.00 |  |  |  |  |  |  |  |  |  |  |
| view | -0.13 | 1.00 |  |  |  |  |  |  |  |  |  |
| sunset | -0.01 | 0.19 | 1.00 |  |  |  |  |  |  |  |  |
| green | 0.39 | -0.49 | -0.04 | 1.00 |  |  |  |  |  |  |  |
| beautiful | 0.35 | -0.17 | -0.23 | 0.57 | 1.00 |  |  |  |  |  |  |
| aesthetic | 0.32 | 0.33 | 0.23 | 0.49 | 0.77 | 1.00 |  |  |  |  |  |
| run | 0.15 | 0.03 | -0.18 | 0.42 | 0.40 | 0.41 | 1.00 |  |  |  |  |
| hike | 0.41 | -0.58 | -0.08 | 0.69 | 0.64 | 0.36 | 0.21 | 1.00 |  |  |  |
| play | 0.25 | -0.13 | 0.14 | 0.52 | 0.38 | 0.43 | -0.02 | 0.55 | 1.00 |  |  |
| walk | 0.41 | -0.10 | -0.04 | 0.55 | 0.71 | 0.65 | 0.18 | 0.35 | 0.27 | 1.00 |  |
| exercises | 0.48 | -0.32 | -0.06 | 0.81 | 0.82 | 0.70 | 0.43 | 0.80 | 0.65 | 0.75 | 1.00 |

Table 4. Correlation matrix between aesthetic, exercise, and sentiment scores.

**Discussion**

Previous research suggests green space impacts human health through social, psychological, and physical pathways of action. While experimental and observational studies tend to isolate pathways to the greatest extent possible, in real world applications these pathways are strongly interconnected and should be concomitantly considered. In our analysis of tweets from January through November 2017, self-reported visual characteristics of greenspace were strongly associated with self-reported exercise and sentiment. Significant social events, most notably adverse events were also associated with greenspace sentiment. These preliminary results support the use of Twitter as a platform for evaluating complex physical-psychological-sociological interactions between greenspace and human health.

While the derived methodology provides a useful starting point, there are several limitations that need to be addressed. First, 16% of tweets from a random sample of screened tweets were not directly relevant to green space research. Example non-relevant tweets include bible scriptures, and professional athletics (most often using bi- tri-grams with the keyword ‘field’, ‘yard’, and/or ‘park’). While a small degree of noise is permissible in a large dataset, developing more additional more refined latent constructs than those shown here will require reducing the number of ancillary tweets in the twitter dataset. Machine learning, such as developing a greenspace classifier, coupled with error rate classifications can potentially reduce the percentage of false positive tweets. Second, in this analysis we’ve removed emojis and hyperlinks from our twitter text and subsequent analysis. However, these are rich data sources that should be considered in future analyses. Finally, the current analysis is ecological in nature and doesn’t follow individual authors over time. We chose in this initial analysis not to include individual information for ethical reasons. However, a carefully designed study can minimize identifiable information while tracking tweet authors over time, providing context that’s essential for moving beyond ecological and toward prospective observational study designs.

**Conclusions**

Green space-related tweets capture self-reported social, physical, and psychological interactions with green space. Real time tweet collection and analysis presents opportunities to capture unique, episodic beneficial and harmful green space interactions. Next steps for include refining and expanding exclusion criteria and latent constructs, following individual tweeters or groups over time, and inter and intracity comparisons to compare and contrast city greenspace characteristics.

**Acknowledgements**

I would like to acknowledge Professors John Piorkowski and Ian McCulloh for lending their experience in social media analytics throughout this course project.

**References**

Beyer, K. M., Kaltenbach, A., Szabo, A., Bogar, S., Nieto, F. J., & Malecki, K. M. (2014). Exposure to neighborhood green space and mental health: evidence from the survey of the health of Wisconsin. *International Journal of Environmental Research and Public Health*, *11*(3), 3453–3472.

Bratman, G. N., Hamilton, J. P., & Daily, G. C. (2012). The impacts of nature experience on human cognitive function and mental health. *Annals of the New York Academy of Sciences*, *1249*(1), 118–136.

Ebisu, K., Holford, T. R., & Bell, M. L. (2016). Association between greenness, urbanicity, and birth weight. *Science of The Total Environment*, *542*, 750–756.

Hutto, C.J., & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.  *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI, June 2014.

Lovasi, G. S., Schwartz-Soicher, O., Quinn, J. W., Berger, D. K., Neckerman, K. M., Jaslow, R., … Rundle, A. (2013). Neighborhood safety and green space as predictors of obesity among preschool children from low-income families in New York City. *Preventive Medicine*, *57*(3), 189–193.

Maas, J., Spreeuwenberg, P., Van Winsum-Westra, M., Verheij, R. A., Vries, S., & Groenewegen, P. P. (2009). Is green space in the living environment associated with people’s feelings of social safety? *Environment and Planning A*, *41*(7), 1763–1777.

Maas, J., Van Dillen, S. M., Verheij, R. A., & Groenewegen, P. P. (2009). Social contacts as a possible mechanism behind the relation between green space and health. *Health & Place*, *15*(2), 586–595.

Makris, M. V. (2015). Separate, Different, but Not Isolated: How Youth in Public Housing Relate to Their Gentrified Community. In *Public Housing and School Choice in a Gentrified City* (pp. 171–189). Springer. Retrieved from http://link.springer.com/chapter/10.1057/9781137412386\_7

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)* (pp. 55–60). Retrieved from http://www.aclweb.org/website/old\_anthology/P/P14/P14-5.pdf#page=67

Mitchell, A., & Page, D. (2014). State of the news media 2014. *Pew Research Journalism Project*. Retrieved from http://radiogood.ru/goto/http://www.journalism.org/2014/03/26/state-of-the-news-media-2014-overview/

Perrin, A. (2015). Social media usage. *Pew Research Center*. Retrieved from http://www.gbcinteractive.com/s/PI\_2015-10-08\_Social-Networking-Usage-2005-2015\_FINAL.pdf

Rainie, L., Fox, S., & Duggan, M. (2014). The web at 25 in the US. *Pew Research Center’s Internet & American Life Project*.

Roesslein, J. (2015). Tweepy. *Python Programming Language Module*.

Shanahan, D. F., Bush, R., Gaston, K. J., Lin, B. B., Dean, J., Barber, E., & Fuller, R. A. (2016). Health benefits from nature experiences depend on dose. *Scientific Reports*, *6*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4917833/

Takano, T., Nakamura, K., & Watanabe, M. (2002). Urban residential environments and senior citizens’ longevity in megacity areas: the importance of walkable green spaces. *Journal of Epidemiology and Community Health*, *56*(12), 913–918.

van den Berg, A. E., Maas, J., Verheij, R. A., & Groenewegen, P. P. (2010). Green space as a buffer between stressful life events and health. *Social Science & Medicine*, *70*(8), 1203–1210. https://doi.org/10.1016/j.socscimed.2010.01.002

van Rossum, G., & Drake Jr, F. L. (2010). Extending and embedding Python, Release 2.7. *Python Software Foundation, Wolfeboro Falls*.Visser, K., Sichling, F., & Chaskin, R. J. (2016). Hot times, hot places. Youths’ risk perceptions and risk management in Chicago and Rotterdam. *Journal of Youth Studies*, 1–17.