Using Twitter as a Means of Understanding the Impact of Distance and Park Size on Park Visiting Behavior (Case Study: London)

Seyed Taher Khalilnezhad[1]

[1]Technical University of Kaiserslautern, Kaiserslautern/Germany · Khalilne@rhrk.uni-kl.de

Abstract: This study aims to investigate how different Twitter features can be exploited for understanding people's behavior in using parks of London. We use Twitter as a source of information for collecting tweets from those who live in London. Unlike many studies that use tweets tagged with geographical coordinates, we use tweets' textual features to find the parks' tweets and the users who visit the parks. We also created a dataset of parks' names using the OpenStreetMap query. With matching tweets text with our park's names' dataset, we ended up with 271 parks mentioned in tweets more than ten times for the whole year of 2020. To evaluate the validity of using the textual features of tweets for spatial and environmental assessment, we assessed the number of parks' visitors and the mean distance that users of each park travel to visit the park. The results indicate that people see a park 2.5-3km away from their neighborhood. In addition, larger parks play more city-scale roles and attract people from more diverse communities. This finding is in accord with the studies that applied tweets coordination, indicating textual application for studying urban affairs is valid.

Keywords: London's parks, Twitter, natural language processing, textual analysis

1 Introduction

Over the last number of years, using social media as an urban survey has increased. Different studies have used different methods for exploiting social media in urban research. Twitter is one of the most familiar platforms for studying urban phenomena. Given that most Twitter active users (roughly 80%) post tweets via mobile phone (LANSLEY & LONGLEY 2016), this provides an excellent opportunity to study the dynamism of cities in a citizen-centric approach. Geotagged information attached to the tweets is beneficial when looking at the spatial behavior of citizens. “Coordinates” and “place name” are two prominent tweet’s feature that provide information about tweet dissemination locations. Over the last decade, many studies applied geotagged tweets to study emotion and sentiment analysis (KOVACS-GYÖRI et al. 2018, PLUNZ et al. 2019, RESCH et al. 2016), quality of life perception (ZIVANOVIC et al. 2020) city dynamic and land use (FRIAS-MARTINEZ & FRIAS-MARTINEZ 2014, GARCÍA-PALOMARES et al. 2018) geography of topics (LANSLEY & LONGLEY 2016), crowds’ movement (FERNÁNDEZ VILAS et al. 2019) and human mobility (FERNÁNDEZ VILAS et al. 2019, LUO et al. 2016, OSORIO-ARJONA & GARCÍA-PALOMARES 2019, YIN et al. 2016). Using tweets tagged according to their geographical coordinates is a common aspect of these studies. Although applying tweets with coordinates for some subjects of urban research like urban mobility has excellent potential, it also has some shortcomings for other sorts of research. One of the difficulties that arise when dealing with tweets with geographical coordinates is their rarity among acquired tweets. For the sake of privacy, the default setting of the Global Position System (GPS) is off on the mobile phone operating system (OIS/Android), and only 1% of the tweets are tagged with the longitude and latitude of the place they are posted (in our case study 0.5%). These problems can be alleviated by increasing the study period to

Journal of Digital Landscape Architecture, 7-2022, pp. 146-154. © Wichmann Verlag, VDE VERLAG GMBH · Berlin · Offenbach. ISBN 978-3-87907-724-3, ISSN 2367-4253, e-ISSN 2511-624X, doi:10.14627/537724015.

gather more geotagged tweets. Still, when we aim to study in a short period, there will be much fewer tweets with coordination. This limitation causes more deficiencies when we look for the tweets posted within a specific place like parks and green spaces; therefore, the number of tweets posted from parks with geographical coordinates tags would decrease sharply. Another issue concerning the study of place through geotagged tweets is that many subjects have little relevance with their location. Also, a pilot study of the tweets tagged with coordinates reveals that many tweets talk about places posted from another site. This means that people sometimes compose a tweet after or before visiting a place. Therefore, studies that evaluate the content of tweets for emotion, sentiment, and semantic analysis to figure out the perception of the place from individuals' perspectives would have a lot of noisy data in their dataset, which will cause some deviation in the results. Although due to the massive number of tweets, we can collect millions of tweets and make decisions based on them, it still constructs a tiny portion (1%) of all tweets. As the purpose of using social media is to retrieve data as much as possible and to benefit from the advantages of big data to represent the community better, losing data due to the lack of geotagged features is in contrast to the essence of applying these means and methods. By removing non-geotagged tweets, we will lose a lot of valuable and essential information, which will always imply a probability of deviation in results. To overcome the problems mentioned above, in this research, instead of the longitude and latitude of the tweet, we use other features of tweets to extract the location that the user posts a tweet about it. In urban studies, we usually deal with large areas, and depending on the scale of our research; we do not need the exact location of the information to make decisions. For most of the study, having an approximation of the site would be fine-grained enough to understand the status of the place. In this research, we assume that textual features have enough information to approximate the location of each tweet. Tweets' text can contain information about the Area, especially when talking about the site of interest to the research. In addition to tweets' text, we use "place name," which is embedded in the tweets' metadata, to extract the center of activities of the users.

2 Method and Data

For this study, we chose Twitter as our source of information and data. Twitter APIs provide different rules and queries that allow thoroughly desired tweets retrieval. To access the tweets of the year 2020, we used full archive historical tweet endpoints, which enabled us to search for all tweets dating back to 2006. The Twitter API provides a query that can be created with operators and rules that lookup for the tweets or users matching predefined attributes, such as message keywords, hashtags, URLs, location, etc. For our study, we set a query composed of NE and SW latitude and longitude coordinates of London to define a bounding box to limit retrieving tweets area. In addition to defining our desired place for collecting tweets, English is also set as a keyword to provide us with those tweets only posted in this language. We collected 8681353 tweets between the first of February 2020 to the thirty-first of January 2021. These tweets are only the tweets that had been written in English and posted by those who were in London when they posted these tweets. Besides the "place name" other features that we selected of tweets are the text, username, and timestamp. Since we are using Natural language processing (NLP) as an approach in our study, we have to work with text features, and we only can consider those tweets that contain each of the park names in our dataset. We name these tweets as park tweets. Also, the writer of each tweet is considered a park user.

We assume that when people tweet about a park, they are in the garden at the moment of writing the tweet, or they are intended to visit the park shortly or have visited it in the near past. To build a parks dataset, we used the OpenStreetMap query in QGIS to extract all parks' and green spaces' names in London. For this purpose, also to include green spaces with a different generic name than the park, the query is composed of additional terms and phrases indicating green areas like “park,” “garden,” “greenery,” “green spaces,” “leisure.” The retrieved result was 2715 different green spaces across London. For the sake of avoiding confounding effects and ambiguity in our analysis, we excluded the playground and sports greens from our study green areas data set. We created a list of parks' names to discern parks' tweets and looked through all tweets containing each park's characters. We end up with 30044 tweets that contain the words of the gardens.

2.1 Preprocessing

Preprocessing is necessary to remove irrelevant tweets and reduce the semantic dimensions of noisy data. Therefore, in the first preprocessing step, these kinds of tweets were deleted from our dataset. In the second step, we also applied refinement methods to ensure the users could provide reliable data for our study. In the third and last step, we restrict our parks dataset to those mentioned in the tweets more often. In the prepossessing step, it turned out that a large proportion of tweets (nearly 17% of tweets) are the tweets that have the #food-waste hashtag, which is a hashtag for sharing surplus food of neighbors or the products nearing the sell-by date of local business with others in London. Since these tweets are irrelevant to our study, we eliminated them from our data set. We also found many tweets originating from the Instagram account of users, shared on Twitter with the default text “@username Just posted a photo” or “@username just posted a video” for posting a photo and video on Instagram, respectively. Regarding our textual analysis, these kinds of tweets have low value, so we did not consider these tweets in our study.

2.2 User Refinement

This study focused on the behaviors that stem from living permanently in place. To include only those who are citizens of London or at least have been living a considerable amount of time in London, we also contrived a set of rules.

1) Every user must have at least ten tweets for the time period of study.
2) Every user must have at least 90 days' time distance between the first and last of their tweets.
3) The mean time distance between users' tweets must be one day or above of one day. With this, we make sure that any user does not post all their tweets in a short period.
4) The frequency of a place where maximum tweets have been disseminated must contribute more than fifty percent of all tweets' areas of a user.

Since we are studying the behavior and the attitudes of London citizens, the first two conditions exclude those that had tweeted from London, but they are a tourist and only have traveled to London for a short trip. The third condition is to make sure that users do not send all their tweets in a short period (in one day or part of a day). Also, they have a consistent behavior of posting tweets. This is important because, with this condition, there would be more probability for the user to tweet from different places and, therefore, more chances for extracting their centers of activity. With the last condition, we only choose those whose frequen-

cy of their top spot is more than 50 percent of all places that the user has tweeted from. All these four conditions have to meet by the user to be included in our study. Also, with these conditions, we can be confident enough that the extracted place for the user could be represented as their center of activities. The whole number of Twitter users that we retrieved is 11014 which after applying the refinement rules, we end up with 5405 users as the presumable residents of London.

2.3 Park Refinement

To find the parks that have a reasonable number of tweets, therefore, it is more likely that they have the potential for extracting insights. We used the N-gram language classification model to limit this research to the most mentioned parks in tweets. N-gram classifier, given a document, is defined as a set of all unique contiguous subsequence of words or tokens of size n in the document (BRODER et al. 1997). For instance, the bigram (N=2) of the sentences of "I am in Hyde Park" will be "I am," "am in," "in Hyde," "Hyde Park." Since each park name is composed of two or three words, we applied the bigram(n=2) and trigram(n=3) machine learning approach to classify park names that have been mentioned in the text of the tweets based on their frequency. Interesting, the most and the top mentioned terms were the parks' names, which means these tweets are talking about gardens. It can be interpreted as a correct refinement method for extracting park tweets. From 2715 parks' names in our dataset, 843 parks were mentioned in our tweet's dataset. Since the less a park is mentioned in the tweet, the fewer are the chances for getting correct information from that park, and we removed those parks that were mentioned less than ten times in the tweets. Finally, we got 272 parks.

3 Defining Neighborhoods

As discussed early in the introduction, we decided to use place names as an attribute for defining and finding users' homes in our study. As we investigated the precision of variety representation of the place name of all tweets, it turned out that this precision for the city of Great London is in the level neighborhood or even smaller areas. In other words, each place name is dedicated to an area in size of neighborhood or smaller. Considering that two main centers of activities of people are home and workplace, we can identify each center of activities based on the pattern of people's life. For example, much previous research identified the home as the most frequently visited place for a user during a nigh time (HUANG et al. 2014, LUO et al. 2016). Therefore, we can identify these centers of activity by dividing tweets into day tweets and night tweets. In this research, we assume that due to the Covid-19 pandemic and lockdown in 2020, people were restricted to work from their homes which caused their two centers of activities to be geographically matched together. We interpret the most frequent place name of each user's tweets as their center of activities and classify this center of activity in the corresponding neighborhood. In the user refinement section, we end up with 5405 final users, which we used all their tweets in the entire year 2020 to extract their center of activity, namely their neighborhood. Notice that the most frequent place of posting tweets for these users constitutes more than 50% of all their posting places. For 5405 users, we ended up with 119 neighborhood names. Given 5405 final users out of all 11014 users, to make sure that all parks have lost some of their users proportionally equal, we executed a regular distri-

bution test (D'Agostino's K^2 test) of the user ratio. Here, the user ratio is defined as the ratio between the number of users of each park before and after the users' refinement process. The P-value for D'Agostino's K^2 test is 0.08, which means the distribution is normal. Therefore, comparison among parks is possible.

4 Geocoding

To use these locations in our case study, next, we will extract their longitude/latitude coordinates and geocode them in GIS. For converting a place name to a coordinate of that place, we use Nominatim.

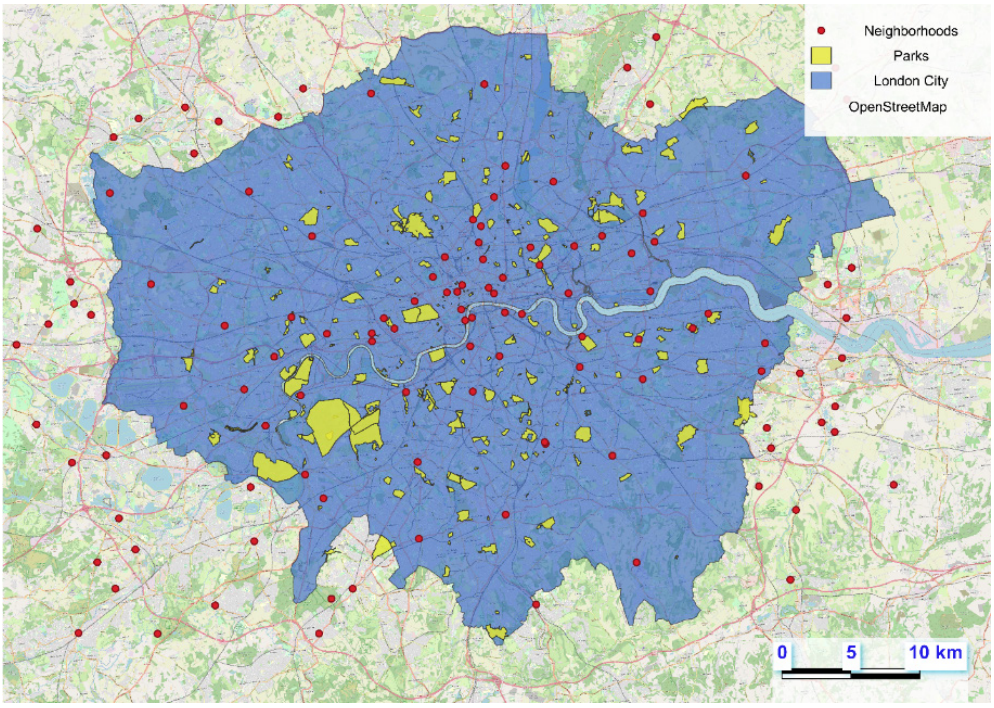

Fig. 1: Visualization of extracted parks and neighborhood from parks' tweets

This geocoder algorithm uses OpenStreetMap data, gets the place name as a string in input, and returns the longitude/latitude of that location. Finally, we imported this longitude/latitude into GIS to use for further analysis. Figure 1 represents the extracted places of the tweets (Neighborhoods).

5 Analysis and Result

To evaluate the visiting behavior of people, especially the distance that users tend to travel to visit a park, we use the Euclidean distance between the centroid of each park and the corresponding coordinate of each neighborhood in QGIS. Figure 2 shows the frequency of mean length that each park's users have traveled to visit the park. The frequency of visiting a park over long distances decreases dramatically as it is presented. The maximum frequency appeared at a distance of 2.5-3km, which means most people tweeted this far from their homes. Considering even distribution of parks in the study area, people can access a park from almost any city point within a 500 m walk (KOVACS-GYÖRI et al. 2018).

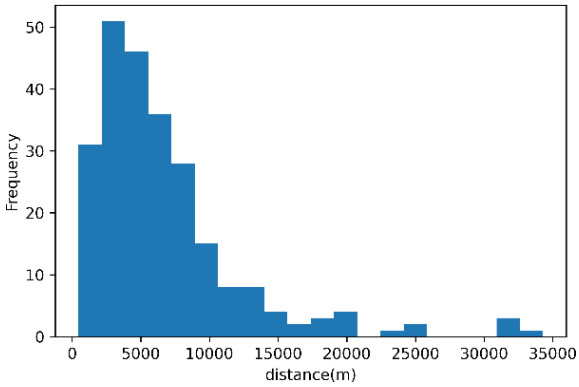

Fig. 2:
Frequency of mean distances that users' park traveled to visit each park

To understand the actual reason why people, prefer to visit a park 2.5-3km far from their home while they can access a local park in a shorter distance, we need more in-depth scrutiny. For this aim, we intersected different park characteristics and applied a correlation coefficient test (Table 1). Although it is possible to extract other features of the parks, which is interesting for visitors from the text of tweens, in this study, we only focus on the largeness (Area), number of tweets, number of users (visitor), distance from visitors' neighborhood, neighborhood diversity, the population density of each park. Population density especially was added to understand how increasing the number of people in an area will discourage people from using the park due to the Covid_19 and social distancing rules. Neighborhood diversity also is an index of the number of diverse neighborhoods that users of each park belong to. Table 3 shows how different park characteristics are correlated to each other. As shown in Table 1, there is a meaningful correlation between parks' areas and neighborhood diversity (p=0.51). Number of users variable also has a relatively high correlation with Area (size of each park) which means the larger parks attract more people. Interestingly, distance does not correlate with the number of tweets (p = -0.03), suggesting people do not necessarily tweet when they visit a new park far from their homes. This can be the result of dropping the number of people visiting a place in long-distance. If we only consider 2.5-3km distance from individuals' neighborhoods to eliminate the impact of the rarity of long-distance visiting, the correlation between distance and number of tweets increment to 0.27, which indicates a slightly positive correlation.

Table 1: Pearson's correlation coefficient test among parks' different characteristics

	Number of Tweets	Number of Users	Neighborhood Diversity	Distance	Area	Population Density	Tweets per User
Number of Tweets	1.00	0.96	0.89	-0.03	0.57	0.13	-0.02
Number of Users	0.96	1.00	0.84	-0.07	0.50	0.12	-0.12
Neighborhood Diversity	0.89	0.84	1.00	0.08	0.51	0.22	-0.10
Distance	-0.03	-0.07	0.08	1.00	0.05	0.05	0.19
Area	0.57	0.50	0.51	0.05	1.00	-0.09	-0.02
Population Density	0.13	0.12	0.22	0.05	-0.09	1.00	0.02
Tweets per User	-0.02	-0.12	-0.10	0.19	-0.02	0.02	1.00

Population density has a very slight correlation with the number of visitors (P = 0.12), which is negligible. This can be interpreted alongside the resulting correlation between the number of park users and the parks' area. Since these two variables have a meaningful correlation (p = 0.5), population density would relatively stay constant while the size and number of users in different parks change.

6 Discussion

To evaluate the accuracy of the results of this research, acquired by applying textual analysis of tweets, we compare these results with results of other studies exerted with different methods and tools. Comparing the visiting distance of parks does not show any significant differences from previous research (KOVACS-GYÖRI et al. 2018, ROSSI et al. 2015). Therefore, it indicates the reliability of the textual analysis of tweets for finding helpful insight and information of people's behavior and specifically travel distance in this study. Interestingly, since previous studies have been done before the Covid-19 Pandemic, results also implicitly indicate that pandemic has had no impact on visiting distance of park users. However, this is a distance from park visitors' neighborhoods that they mentioned a park in their tweets, which can be because of different reasons. For example, due to the acquaintance of the local park, people do not tweet when they visit a garden in their neighborhood. However, examining distance and number of tweets in the previous section does not show a meaningful correlation, so people do not necessarily tweet when they visit a park far from their homes. Therefore, this resulting distance is not a consequence of the application of Twitter and can be referred to as a mean distance that people tend to visit a park. Results also show that the largeness of the park is also a characteristic that explicitly influences the number of visitors, which is an expected result.

What is more interesting is that our results indicate the larger parks attract people from the more diverse neighborhood. This means larger parks play a city-wide scale role rather than a

local. As expressed, these results are not very surprising, but they show that using textual analysis would provide us with reliable results. The population density in parks came into account to show how much social distancing causes people to avoid gathering in the park. Results show that there is not any significant correlation between population density and the other number of park users. However, this could result from taking the whole area of the park to calculate park density. For more accurate results, it is suggested that the areas that are more attractive for people within a park be considered in the calculation of park density rather than the entire park area. This goal is beyond the method and purpose of this research and leaves for future research. Beyond the distance and park size, other parks' features can significantly attract people to visit a park. This can be due to the limited services that local parks provide for people. Other than functionality, factors like aesthetic features, presence of people, safety, wideness, degree of naturalness can influence visiting distance. Textual features of tweets can also be used to understand how these features influence people's visiting behavior, which is beyond the scope of this study and can be the main aim for further studies.

7 Conclusion

In this study, we applied a different method for finding the parks' tweets. Instead of using the common process of collecting tweets with precise coordination to study the tweet disseminated, we used tweets' text to find the those that talk about the desired location. To examine our method, we investigated how far people travel to visit a park. The results of our analysis indicate that people usually see a park 2.5-3km away from their neighborhood. This can be due to different factors, like people not posting a tweet when they visit a garden in their community due to familiarity with the environment. However, Pearson's correlation test does not show any relationship between distance and the number of tweets (users). Also, results indicate that large parks attract people from different parts of the city, which means they play more of a city-scale role than a green space for their immediate neighborhood. The results are in accord with the previous studies, which show how textual features of tweets can be useful for urban and spatial analyses.

References

FERNÁNDEZ VILAS, A., DÍAZ REDONDO, R. P. & BEN KHALIFA, M. (2019), Analysis of crowds' movement using Twitter. Computational Intelligence, 35 (2), 448-472. https://doi.org/https://doi.org/10.1111/coin.12205.

FRIAS-MARTINEZ, V. & FRIAS-MARTINEZ, E. (2014), Spectral clustering for sensing urban land use using Twitter activity. Engineering Applications of Artificial Intelligence, 35, 237-245. https://doi.org/https://doi.org/10.1016/j.engappai.2014.06.019.

GARCÍA-PALOMARES, J. C., SALAS-OLMEDO, M. H., MOYA-GÓMEZ, B., CONDEÇO-MELHORADO, A. & GUTIÉRREZ, J. (2018), City dynamics through Twitter: Relationships between land use and spatiotemporal demographics. Cities, 72, 310-319. https://doi.org/https://doi.org/10.1016/j.cities.2017.09.007.

HUANG, Q., CAO, G. & WANG, C. (2014). From where do tweets originate? A GIS approach for user location inference. https://doi.org/10.1145/2755492.2755494.

KOVACS-GYÖRI, A., RISTEA, A., KOLCSAR, R., RESCH, B., CRIVELLARI, A. & BLASCHKE, T. (2018), Beyond Spatial Proximity – Classifying Parks and Their Visitors in London Based on Spatiotemporal and Sentiment Analysis of Twitter Data. ISPRS International Journal of Geo-Information, 7 (9), 378. https://www.mdpi.com/2220-9964/7/9/378.

LANSLEY, G. & LONGLEY, P. A. (2016), The geography of Twitter topics in London. Computers, Environment and Urban Systems, 58, 85-96. https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2016.04.002.

LUO, F., CAO, G., MULLIGAN, K. & LI, X. (2016), Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. Applied Geography, 70, 11-25. https://doi.org/https://doi.org/10.1016/j.apgeog.2016.03.001.

OSORIO-ARJONA, J. & GARCÍA-PALOMARES, J. C. (2019), Social media and urban mobility: Using twitter to calculate home-work travel matrices. Cities, 89, 268-280. https://doi.org/https://doi.org/10.1016/j.cities.2019.03.006.

PLUNZ, R. A., ZHOU, Y., CARRASCO VINTIMILLA, M. I., MCKEOWN, K., YU, T., UGUCCIONI, L. & SUTTO, M. P. (2019), Twitter sentiment in New York City parks as measure of well-being. Landscape and Urban Planning, 189, 235-246. https://doi.org/https://doi.org/10.1016/j.landurbplan.2019.04.024.

RESCH, B., SUMMA, A., ZEILE, P. & STRUBE, M. (2016), Citizen-Centric Urban Planning through Extracting Emotion Information from Twitter in an Interdisciplinary Space-Time-Linguistics Algorithm. Urban Planning, 1, 114. https://doi.org/10.17645/up.v1i2.617.

YIN, J., GAO, Y., DU, Z. & WANG, S. (2016), Exploring Multi-Scale Spatiotemporal Twitter User Mobility Patterns with a Visual-Analytics Approach. ISPRS International Journal of Geo-Information, 5 (10), 187. https://www.mdpi.com/2220-9964/5/10/187.

ZIVANOVIC, S., MARTINEZ, J. & VERPLANKE, J. (2020), Capturing and mapping quality of life using Twitter data. GeoJournal, 85 (1), 237-255. https://doi.org/10.1007/s10708-018-9960-6.