Does better accessibility to exercise facilities mean people use them more?

by

Fangda Lu

A directed research report submitted to the Graduate College, the Geography Department of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Applied Geography
with a specialization in Geography Information Science

Summer 2018

Committee Members:

Yongmei Lu

Edwin T. Chow

Table of Contents

Acknowledgment

I would first like to thank my advisor Dr. Yongmei Lu. She consistently pushed me to produce my work, critiqued me but steered me in the right direction whenever I need.

I would also like to thank my committee member Dr. Edwin Chow who contributed to this research. Without his passionate participation, especially valuable advice, the study could not have been successfully conducted.

Finally, I must express my very profound gratitude to my parents and my closest friends who encouraged and inspired me for a lifelong love and during the research project. Their kindness and motivation are what let me overcomes any difficulties.

Abstract

GIS provides an excellent means for visualizing and analyzing health spatial data, geographically integrating large amounts of information from different sources and serving as a common platform for health data, recreation facilities, as well as indoor/outdoor activities. At the same time, social media provides an effective and accurate way for us to learn about the lifestyles of representative individuals within a region. Twitter is one of the most popular social media in the United States. Some users reveal their position with posts, which can let us know what and when are they doing, as well as where are they. The combination of GIS and social media helps to identify the characteristics of users' social media behavior that tend to be associated with their activities, both indoor and outdoor considered. In this research, keyword searching was used to filter tweets related to physical activities: parks, trails, and open spaces refer to outside activities, geocoded gyms and facilities refer to indoor activities. Service areas were defined for different levels of facilities, and then every tweet location was evaluated for its access to physical activity facilities. The same analysis was applied at the zip code level. Last, whether the accessibility affects activity was tested by statistical hypothesis testing method. As a result, the average accessibility of indoor activity-related tweets' location is not significantly different from the average of unrelated tweets' location. However, the average accessibility of outdoor activity-related tweets' location was significantly lower than the average of unrelated tweets' location. That's the

reason why we cannot say there is any expected quantity relationship was figured out.

Key Words: GIS, health, activity, social media, Twitter

Introduction

Exercise can keep our bodies healthy, mitigate obesity and bring happiness. Physical activities are very important to maintain and enhance health. According to *The State of Obesity* annual report, 80 percent of American adults do not meet the government's national physical activity recommendations for aerobic activity and muscle strengthening. In 2016, around 45 percent of adults are not sufficiently active to achieve health benefits, and physical activity rates decreased in 28 states. Texas had the 17th highest reported percentage of inactivity among adults at 25.2 percent. Around $117 billion in healthcare costs are associated with this inadequate physical activity.

Factors that affect individuals' participation in sports include time availability, affordability, skills and related knowledge, interests, physical fitness, experience, economic ability and health status; social factors include family, friends, information, convenience of physical activities resources, administrative measures and guidance, project and quantity, safety, transportation, cost, and quality; demographic factors include gender, age, race, local area, etc. This study focuses on the most basic conditions for physical activities, the influence of facilities' number, location and density to residents.

Social network service (SNS, also called "Social media") is a web-based service that allows individuals to establish public or semi-public personal documents within a closed system and clearly lists associated friends for

individuals. On this basis, individuals can browse various links between themselves and associated users (Boyd & Ellison, 2008). With the advent of the Web 2.0 era of “user participation, user-driven, and user-building,” social media has become one of the basic applications of the Internet and has been a focus of researchers from information science, communication, management, psychology, and of course, geography.

Twitter is a typical social networking site. It was recorded as one of the 10 most visited websites in 2016 and more than 200 million users posted 500 million tweets per day. Twitter not only provides a new platform for social interaction in daily life but also records users’ behavior. The most important aspect of social media for the purposes of this paper is a convenient way to get users’ locations, which provides an unprecedented opportunity for geographers who care about the relevance of social media and activity.

In this study, we believe that an area that has good accessibility to exercise facilities tends to be better served and therefore is conducive to physically active lifestyles. Traditionally, we never know whether residents use them or not, frequently or rarely. Fortunately, Twitter gave us a way to collect the truth. People often post a tweet to record their feeling after exercise, to record they’re camping or diving, to record they reached a new level on hills and so on. The most important thing is that Twitter has attached GPS position information to each tweet, so there is no need to judge where the user stayed and where that event happened. It reveals the population distribution, as well

as the difference between day and night, the difference between weekdays and weekend and many other patterns.

The purpose of this study is to explore the relationship between the accessibility to physical facilities and people's using, measured by physical activity-related tweets. The hypothesis is that residents who live in an area more accessible to physical facilities are willing to do more exercise; in other words, that accessibility affects activity.

Literature Review

Coombes et al. (2010) examined the association between objectively measured access to green space, the frequency of green space use, and physical activity. Results of the study showed that the reported frequency of green space use declined with increasing distance. The study also found that respondents living closest to a formal park were more likely to achieve the physical activity recommendation. The association with physical activity remained after adjustment for respondent characteristics, area deprivation, and a range of characteristics of the neighborhood environment.

There is a certain link between the built environment and human behavior. Handy et al. (2002) directly assessed the relationship as it influences personal health. The authors created appropriate measures for the built environment and for travel behavior. The available evidence lends itself to the argument that a combination of urban design, land use patterns, and transportation systems that promote walking and bicycling will help create active, healthier, and more livable communities. And the availability of green space significantly increases the likelihood of engaging in physical activity, especially when those green spaces have more facilities or amenities (Kaczynski et al. 2008).

Fry and Langford (2012) examined the correlation between accessibility to green space and a variety of other health conditions. The authors tend to use a combination of approaches or seek to establish the implications of incorporating alternative measures of accessibility on potential relationships. A database of

green spaces (and associated attributes) and a detailed network dataset were used to examine the sensitivity of findings to the ways in which different metrics are calculated, and this is illustrated by examining the variations in association between such metrics and a census-based deprivation index widely used in health studies to measure socioeconomic conditions.

However, the relevance is not always the same. A conflict results derived by Hillsdon, Panter, Foster, and Jones' (2006) study. They examined the association between access to quality urban green space and levels of physical activity. This study used GIS and three measures of access to open green space that were calculated based on distance only, distance and size of green space, and distance, size and quality of green space. Multiple regression models were used to determine the relationship between the three indicators of access to open green space and the level of recreational physical activity. However, different from most of the other articles, there was no evidence of clear relationships between recreational activity and access to green spaces in this case. Nielsen and Hansen (2007) came to the similar conclusion. Their statistical results indicate that access to a garden or short distance to green areas from a person's dwelling is associated with less stress and a lower likelihood of obesity. The number of visits cannot explain the effects of green areas on the health indicators.

A study conducted by Hoehner et al. (2005) analyzed perceptions and objectively measured environmental factors and their relative association with transportation or recreational physical activity. Transportation activity is not

our focus point, so we only care about the recreational activity results. After adjusting for age, gender, and education, recreational activity was positively associated with perceived access to recreational facilities and objective measures of attractive features.

Kaczynski et al. (2008) studied whether park size, number of features in the park, and distance to a park from participants' homes were related to a park being used for physical activity. They found that parks with more features were more likely to be used for physical activity; size and distance were not significant predictors. Park facilities were more important than were park amenities. Specific park features may have significant implications for park-based physical activity.

As for the measurement of accessibility, Luo and Qi's (2009) paper presented an enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility. They addressed the problem of uniform access within the catchment by applying weights to different travel time zones to account for distance decay. It reveals a spatial accessibility pattern that is more consistent with intuition and delineates more spatially explicit health professional shortage areas.

In terms of the individual and environmental predictors of residents' sports participation, Prins et al. (2010) examined whether the availability of sports facilities moderated intention-behavior relations. The authors conducted multiple logistic regression analyses to test associations between availability

of sports facilities and intention to participate in sports. The analyses showed that the intention for sports participation was stronger when sports facilities are more readily available. The results of this study indicate that the intention-sports participation association appears to be stronger when more facilities are available.

Thompson's (2013) paper reviews research on the relationship between attributes of outdoor environments and people's levels of activity and exercise using those environments. Gelormino (2015) carried out a review of evidence on the built environment and its health equity impact. The study found that the key features of a built environment (identified as density, functional mix and public spaces and services) may influence individual health through their impact on natural environment, social context, and population behavior and that these effects may be unequally distributed according to the social position of individuals.

Social media was also examined by health studies. Dredze (2013) employed machine learning and natural language processing to study the health content of tweets and demonstrated the potential for extracting useful public health information from aggregated social media data. He used supervised learning to filter tweets and find health-related messages, yielding 1.6 million English-language health tweets from March 2009 to October 2010. He further developed a model that discovered health topics from raw tweets for guided exploration, rather than relying on predefined illnesses.

The dramatic growth of Web 2.0 technologies and online social networks offers immense potential for the delivery of health behavior change campaigns. Maher et al. (2014) systematically reviewed evidence on the effectiveness of online social networks on health behaviors. Michael and Dredze (2012) presented the Ailment Topic Aspect Model for Twitter that associates symptoms, treatments and general words with diseases (ailments). Using the Ailment Topic Aspect Model (ATAM), Prul and Dredze (2014) aggregated self-reported health statuses across millions of Twitter users to try to characterize the variety of health information; they developed an approach to filter the general Twitter data based on health keywords and supervised classification. In 2013, they also posted a geolocation system, Carmen, that can identify the structured location information from the messages provided by the Twitter API.

More and more researchers are using social media for health-related studies. Korda and Itani’s (2011) article summarized the existing researches using social media for health-related analysis and discussed the effectiveness of various forms of social media. Chou et al. (2009) also identified the sociodemographic and health-related factors associated with current adult social media users in the United States.

Furthermore, Thackeray et al. (2013) studied the patterns of health-related social media use among adults. Heaivilin et al. (2011) focused on Twitter users who extensively share health information relating to specific disease, including actions taken and the impact. Moorhead et al. (2013) explained how social media

brings a new dimension to health care, as it offers a medium to be used by the public, patients, and health professionals to communicate about health-related issues with the possibility of potentially improving health outcomes, such as activities.

Study Area and Data

This study encompasses the Tri-County Area in Central Texas: Hays, Travis, and Williamson, which cover a total of 2,837 square miles. The Travis county seat is the city of Austin, the capital of Texas. It is also the core of the Austin-Round Rock Metropolitan Statistical Area, which had a population of 1,362,416 in 2013. In 2008-2010, the percentage of obese adults in Travis County (24.0%) was both better than that of the state (29.6%), and the Healthy People 2020 target (30.6%). In Travis County, approximately one in five adults (20.5%) indicated that they get no physical activity, which is lower than what is seen statewide (26.7%).

This study groups physical activities into indoor activities and outdoor activities based on the facilities used. Outdoor facilities refer to parks and open spaces, including natural or man-made facilities such as parks of all sizes, greenbelts, and nature preserves. All cemeteries and private green gardens were excluded from this study because they are not always available to the public for outdoor activities. The GIS data for outdoor facilities was downloaded from the Capital Area Planning Council of Governments (CapCOG) website. In total, 1,294 parks and open spaces were included in the analyses of this study.

Indoor facilities in this study are mainly fitness centers. The related activities include yoga, kickboxing, training, pools, retreats, spas, and so on. This dataset within the Tri-County area was downloaded from ReferenceUSA dataset (resource.referenceusa.com) using North American Industry Classification System

(NAICS) with Code 713940, which included industry comprised establishments primarily engaged in operating fitness and recreational sports facilities featuring exercise and other active physical fitness conditioning or recreational sports activities, such as swimming, skating, or racquet sports. These facilities were geocoded using Microsoft Bing map API based on facility address. In total, 460 facilities were used in this study.

The street network data was downloaded from the TxDOT website. It included 89,360 roads segments, with street names, locations, and measurable length information. A zip code map was downloaded from the Capital Area Planning Council of Governments (CapCOG) website and checked with the United States Postal Service (USPS). Figure 1 shows the research area and zip code map. Two highlighted zip code area will be discussed later.

Tweets data were streamed from Twitter using streaming API. The collection period was Dec. 1, 2017, to Feb. 28, 2018. It included 32,354 tweets within my total research area, including precise position attached or not.

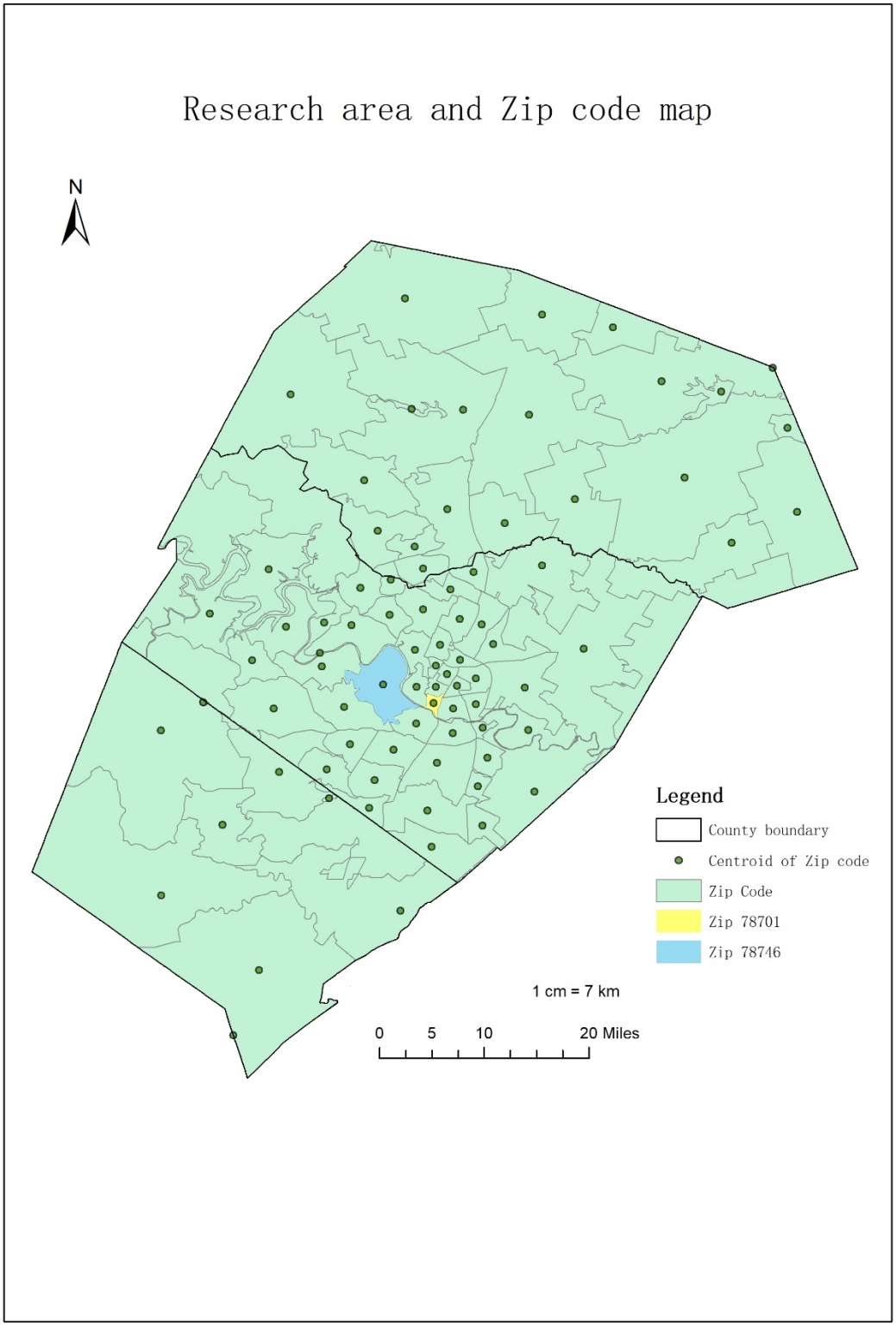

Figure 1 Research area and Zip code map

Methodology

First, the author used Twitter streaming API to collect real-time data from Twitter. The Twitter streaming API allows users to interact with its data, such as tweets and several of their attributes about tweets with a scripting language (in this research, Python). Python provides a standard library with a large number of well-designed frameworks and libraries. A Python program can request the Twitter streaming API to query a series of newline-delimited messages.

The individual messages streamed by this API are JSON encoded, which can be easily parsed and read. This API can collect data by keyword or bounding box. In this research, the author collected tweets in a rectangle area covering Tri-County, then use ArcGIS to delete out-of-boundary tweets. The API based on an HTTP connection is not as reliable as a math formula. The collected data might have some missing parts and duplicate information. All problems above can be dealt with through a phrasing step. A Python script can parse JSON-formatted tweets plus their metadata (i.e. data produced by the Twitter status tracking API). It logs and interprets specific data from each tweet, then written in CSV (Comma-Separated Values) files. It consists of creating time, tweet ID, tweet text, user ID, follower count, retweet count, time zone, and most importantly, position.

Both place and exact location geographic metadata provided by Twitter API. However, usually, only 1.1% of all tweets provide exact location coordinates. The author got 23,922 tweets that have an exact location, filtered from 32,354

coarse geo-referenced tweets within the study area. Next, the author must identify the physical activity-related tweets from geo-referenced tweets. The semantic analysis was conducted to identify physical activity-related information from a tweet's text. Particularly, in order to simplify the model, keyword searching was used.

For the indoor activity keywords selection, the author used seed words: health, weight loss, wellness, exercise, fitness, workout, diet, nutrition, gym, weight control, fat loss, lose weight, training, muscle, bodybuilding, etc. Then the author used a hashtag search from https://hashtagify.me/hashtag/ to identify the related indoor activity keywords by seed words. A hashtag is a type of metadata tag used on social networks such as Twitter and other microblogging services, allowing users to apply dynamic, user-generated tagging, which makes it possible for others to easily find messages with a specific theme or content; it allows for easy and informal markup of folk taxonomy without need of any formal taxonomy or markup language. In total, the final keywords list for indoor activities has 100 words, and 759 tweets were identified from 23,922 tweets. As for outdoor activities, the author used every Olympic sport which is classified as an outdoor activity. The final keywords list contains 75 words, and 640 tweets were identified from 23,922 tweets. The full words list is attached to the appendix in this report.

All spatial analyses were conducted using ESRI's ArcMap 10.5.1. The next step of the project was evaluating the accessibility to indoor and outdoor

facilities for the population in the Tri-County study area. First of all, the author created the service area for every facility. In this step, calculating distance is normally done in one of two ways, Euclidean distance or network distance. Using Euclidean distance is easy to execute but is not accurate. The street network dataset allows the author to use a more accurate measurement of travel cost from one place to another. Because we lack information on the speed limits of the roads, a three-zone distance estimation was used, as shown in the following table (Table 1). The mileage was decided by the common sense of the extents to which people feel about distance in the study area.

Table 1 Facilities service area table

Facilities		Class interval (Miles)	Benefit ratio
Indoor		3 5 8	1 0.6 0.129
Outdoor	<650 Acres	3 5 7	1 0.6 0.129
	650-3000 Acres	5 8 12	1 0.6 0.129

	>3000 Acres	10 25 50	1 0.6 0.129

The indoor facilities are all treated the same. The outdoor facilities are classified into three categories (See Table 1). The author used the K-means clustering algorithm to assess and classify the parks and open spaces into three categories based on areal size. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters, in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. After classification 1,252 were classified as small parks, 35 as medium parks, and 7 as large parks, which is compatible with our general knowledge.

Indoor activity facilities are point features, but parks and open spaces are polygon features, one problem remaining so far is we don't know where the parks' entrances are located. They may have more than one entrance, some might be motor-only/pedestrian-only entrances, or they might be open completely. To simplify this problem, the author calculated the centroid of every park and open space, so they can participate in the follow-up process as with indoor facilities.

After we calculated the service covered area as in Table 1, we needed to give every site a score to depict its overall accessibility to the facilities. There is a two-step floating catchment area (E2SFCA) method proposed by Radke

and Mu (2000) and later improved by Luo and Wang (2003). It's suitable and has been used in a number of studies measuring the potential accessibility to health-related facilities. The 2SFCA method delineates travel cost zones to a facility and assigns weights to these zones, thereby accounting for distance decay instead of assuming a uniform access within a catchment. The function looks like this:

$$A = e^{-\frac{t^2}{\alpha}}$$

where A is the relative accessibility, e is the base of the natural exponent, t is the travel time equivalent cost, and α is the parameter. The A-value, known as cost path score of the zones, falls into a Gaussian function. And for each zone, the probability of accessing decreases.

According to the break distance shown above, the author divided the three class intervals of different facilities into three regions, each of which was deemed to have the same travel cost. Through the three equations for Zone1, Zone2, and Zone3, α can be solved. In this study, the author used α as 1.303 in all zoomed out the equivalent cost. Then the function can be written in this way:

$$\text{Accessibility} = e^{-\frac{(equivalent\ cost\ mileage\ (From\ a\ tweet\ location\ to\ a\ facility)\ unit)^2}{1.303}}$$

The decided values by this function were shown in Table 1.

Next, the sum of all accessibility scores for the various facilities for each tweet location represents the overall accessibility score for a tweet to all physical activity facilities. The score may be larger than 1 because a score does not equate possibility but a quantitative index reflecting how convenient that a tweet place can access physical activity facilities.

Taking a similar approach, the accessibility score can be derived for all zip code areas. The author wanted to explore whether the average accessibility score of each region has any correlation with physical activity-related tweets within the regions. Every zip code area centroid was calculated as a residents' point. It was used to compare the convenience of sports facilities in one area with the enthusiasm of residents in using these facilities. The author then joined the physical activity-related tweets, indoor and outdoor separately, to the zip code regions by location. We can get a correlation table between two variables, the number of tweets and the sum accessibility score. The correlation coefficient can be applied to measure which is a better fitting function and how is the correlation between the two variables.

Finally, a hypothesis test was conducted to examine if the accessibility scores were statistically different between the locations for the physical activity-related tweets and those for unrelated tweets. In general, the physical activity-related tweets were a specific sample set from 23,922 geo-referenced tweets. This will be compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets. The comparison was deemed statistically significant if the relationship between the data sets would be an unlikely realization of the null hypothesis according to a threshold probability—the significance level.

Results

Figure 2 shows the distribution of green spaces in the Tri-County area. There are 706 parks located in Travis County, more specifically within the Austin metropolitan area, covering 327.624 acres in total. There are 451 parks located in Williamson County, covering 71.229 acres in total. And there are 137 parks located in Hays County, covering 99.227 acres in total. Further details will be explained later.

Figure 3 and Figure 4 show the service area for each facility, which were produced by different service zones depending on the service level. In terms of indoor facility service, downtown Austin should have a high accessibility to the many facilities. Other hotspots should be around some popular shopping plazas, such as the Domain and Barton Creek Square. Other service facilities are also found in middle-sized cities, such as Buda, Kyle, and San Marcos, but they rarely cover vast rural areas.

Parks and open areas are in a different story. Because some parks have very large service areas, many of the service areas overlap (See Figure 4). Although Travis County has the greatest number of parks, the average area is the lowest out of the three counties. Hays County has fewer parks, however, some of the biggest parks and open areas are in Hays County. Their service area covered up to a 50 miles distance, reaching Williamson County. Except for State parks and other high-level open spaces, numerous green spaces are in cities within the study area. For instance, San Marcos residents have more green space than Hays

on average; Austin also has more green space available than Travis county it located.

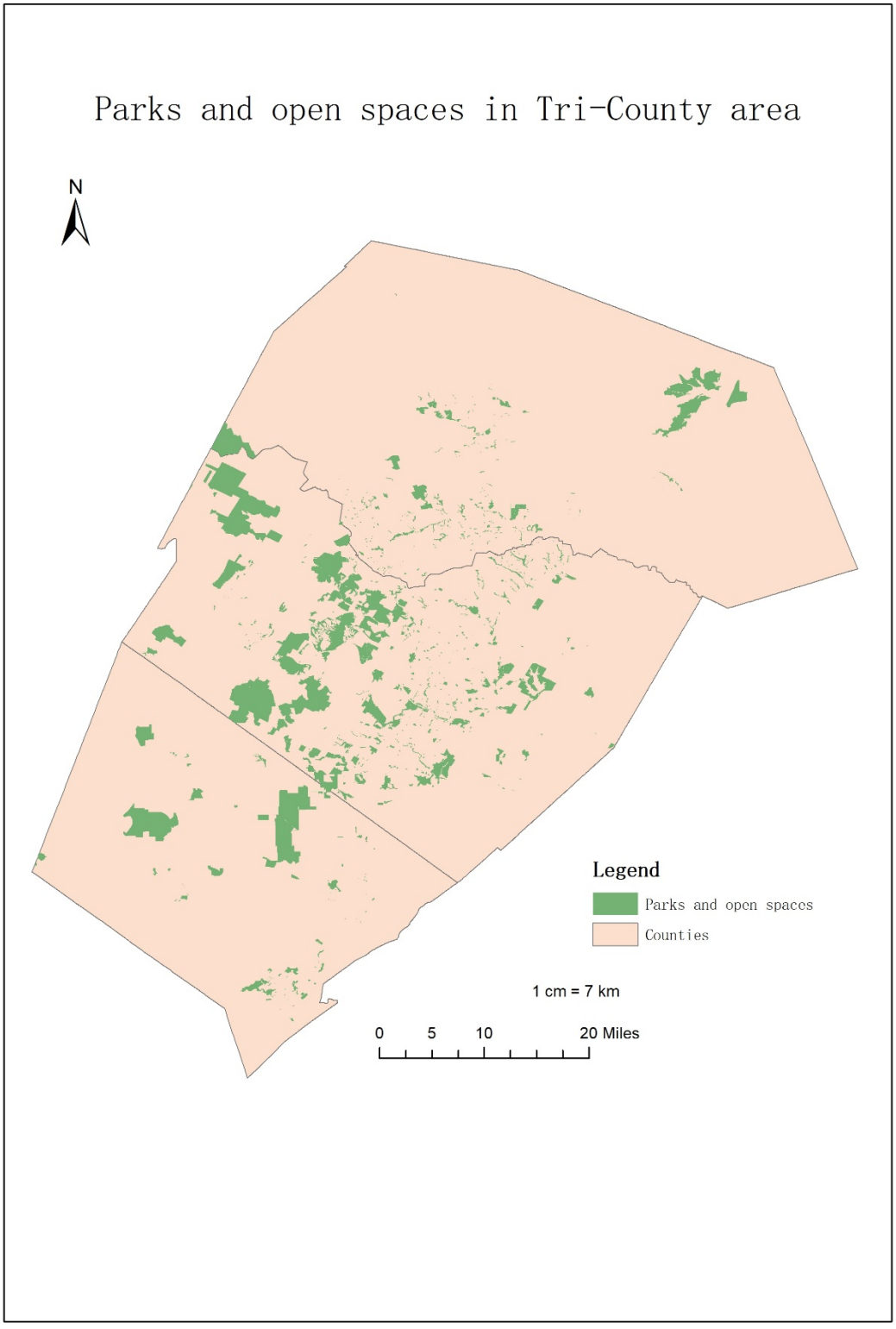

Figure 2 Parks and open areas distribution map

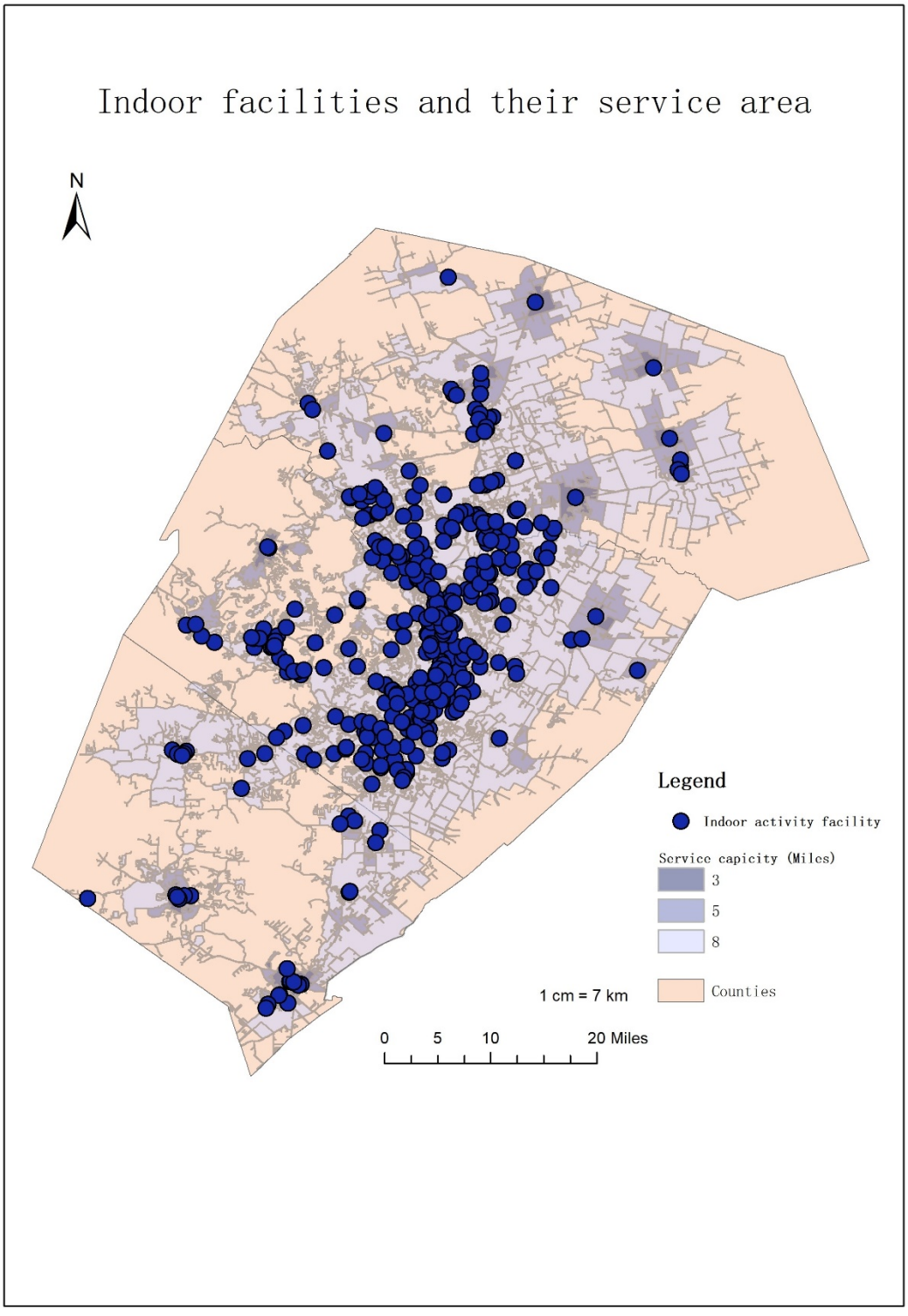

Figure 3 Indoor facilities and their service area

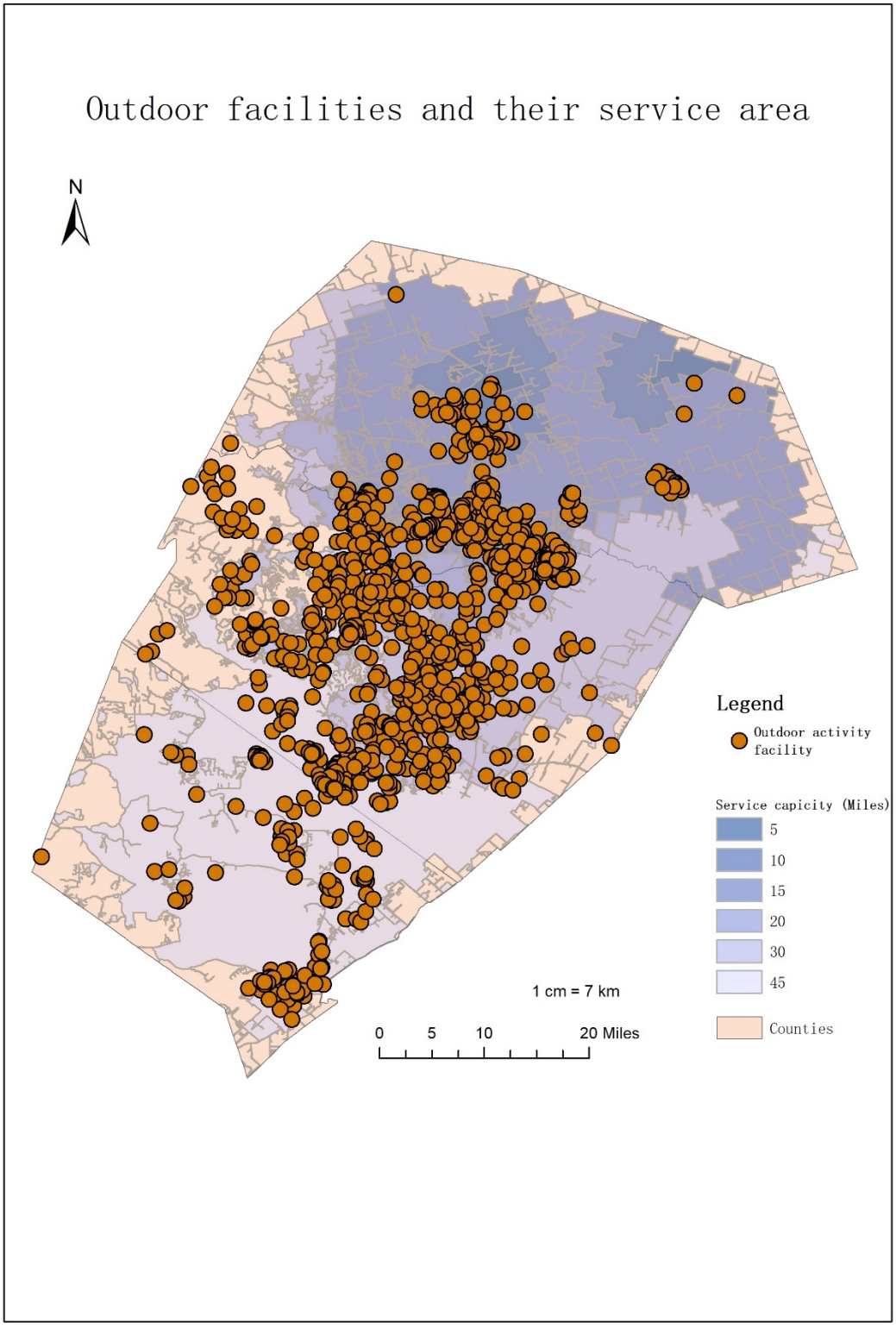

Figure 4 Outdoor facilities and their service areas

After integrating this with the Twitter data, we expected that the locations

affiliated with the physical activity-related tweets to have significantly different accessibility scores than the rest of the tweets, both for indoor and outdoor activities. If the hypothesis is accepted, the accessibility of physical activity-related facilities would appear to be related to physical activities, that are reflected by the related tweets.

Overall, our 23,922 geo-referenced tweets were a subset of all tweets in the research area. In other words, the tweets we randomly collected for this study can be seen as a sample from all tweets posted from the study area and during the study period. (Twitter users are also a sample from the population, of course.) The physical activity-related tweets are a sample of the geo-referenced tweets, and the other tweets are the other sample of geo-referenced tweets. Therefore, we can calculate the mean and standard deviation for the samples and estimate the mean and standard deviation for the population. In other words, we were doing a two-sample t-test of the null hypothesis that the sample is an unbiased representation of the whole population.

$$Hypothesis: \begin{cases} H_0: The\ mean\ accessibility\ score\ for\ the\ physical\ activity-related\ tweets\ is\ NOT\ significant\ differnet\ from\ the\ mean\ score\ for\ the\ unrelated\ tweets \\ H_1: The\ mean\ accessibility\ score\ for\ the\ physical\ activity-related\ tweets\ is\ significant\ different\ from\ the\ mean\ score\ for\ the\ unrelated\ tweets \end{cases}$$

Because we were not sure the sample means will higher or lower than the population, a two-tailed hypothesis should be applied.

First of all, based on the central limit theorem, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally

distributed. Our sample size is large enough in study scale, so we don't need to consider the Mann-Whitney U test that does not require the assumption of normal distributions. In order to decide whether to use the Student's t-test or Welch's t-test, we have to strictly check if the variances of the two populations are not significantly unequal. The Student t-test assumes that the two data sets came from distributions with the same variances, as known as a homoscedastic t-test, but Welch's does not. This procedure was a two-sample analysis of variance, and so we should use a homogeneity of variance test. The homogeneity of variance test uses an F-test to test the null hypothesis that the variance is equal across groups. The result is the following:

Table 2 F-test result (Indoor)

Indoor	unrelated tweets	related tweets
mean	52.31803	50.74246
variance	925.3185	881.8967
observations	23164	758
df	23163	757
F Stat value	1.049237	
P(F<=f) one-tail	0.184884	
F one-tail threshold	1.091955	

Table 3 F-test result (Outdoor)

Outdoor	unrelated tweets	related tweets
mean	220.1836	208.7032
variance	5544	5453.689
observations	23282	640
df	23281	639
F Stat value	1.01656	

P(F<=f) one-tail	0.393215		F one-tail threshold	1.100458	

In the above tables, the p-value, as known as probability value, is the minimum significant level to reject null hypothesis. Fortunately, if we use significant level as α=0.01, now P>2α=0.02, so the null hypothesis cannot be rejected, both indoor and outdoor situations satisfied the homoscedastic hypothesis. In other words, although the large different sample sizes between related and unrelated tweets, they have an insignificant enough difference in variances. So, we can move on to the t-test below:

Table 4 T-test result (Indoor)

Indoor	unrelated tweets	related tweets
mean	52.31803	50.74246
variance	925.3185	881.8967
observations	23164	758
covariance	923.9443	
df	23920	
t Stat value	1.404291	
P(T<=t) two-tail	0.160245	
t two-tail threshold	1.960063	

Table 5 T-test result (Outdoor)

Outdoor	unrelated tweets	related tweets
mean	220.1836	208.7032
variance	5544	5453.689
observations	23282	640
covariance	5541.587	
df	23920	
t Stat value	3.848956	

P(T<=t) two-tail	0.000119		t two-tail threshold	1.960063	

In this case, we selected the significant level α=0.01, which means we have 99% confident to accept or reject the hypothesis. We can compare the p-value with the significant level interpret the results as we did in F-test. According to Table 4, the two-tail p-value=0.160245, larger than α=0.01, so we cannot reject the null hypothesis H_0, that is, the mean accessibility for indoor activity-related tweets' location was not significantly different from the mean accessibility for unrelated tweets' location on a 99% confidence level. On the other hand, the difference was significant between the mean accessibility for outdoor activity-related tweets' location and the mean accessibility for unrelated tweets location, because the two-tail p-value is much smaller than the significant level α=0.01, as well as confidence level 99%. Obviously, the mean accessibility score of related tweets is lower than the mean score of unrelated tweets. So, we can say that the mean accessibility score for the outdoor activity-related tweets is significantly lower than the mean score for unrelated tweets.

According to my hypothesis, the zip code area accessibility scores should be positively related to the number of physical activity-related tweets in that area. In this case, the author did a simple linear fitting at first. Here are the results:

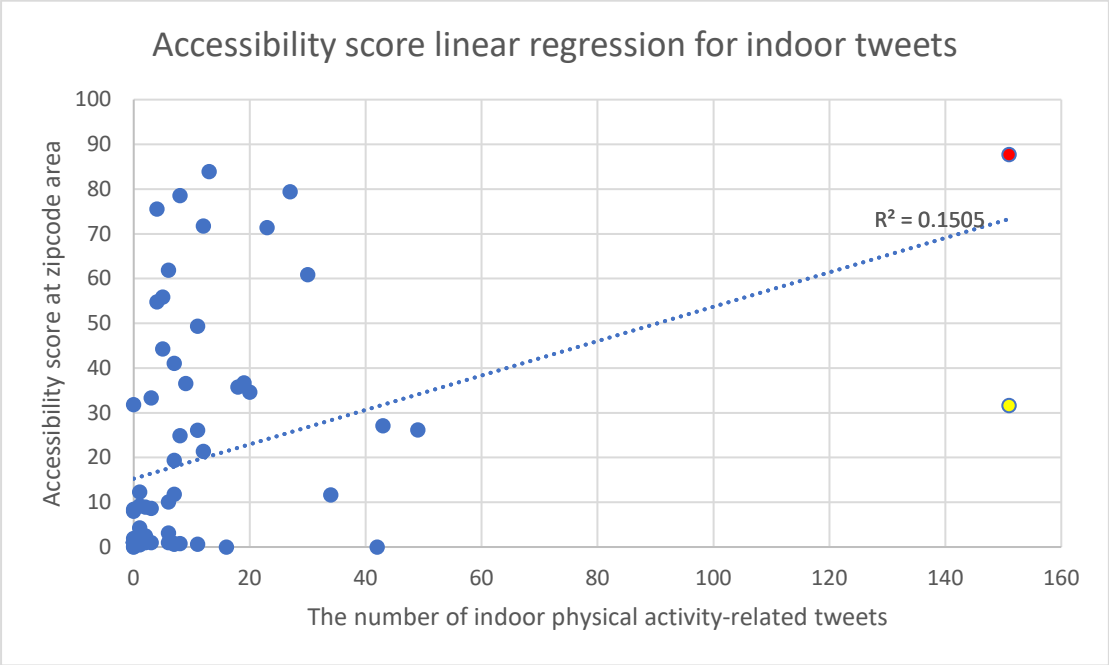

Figure 5 Accessibility score at zip code area & Indoor physical activity-related tweets

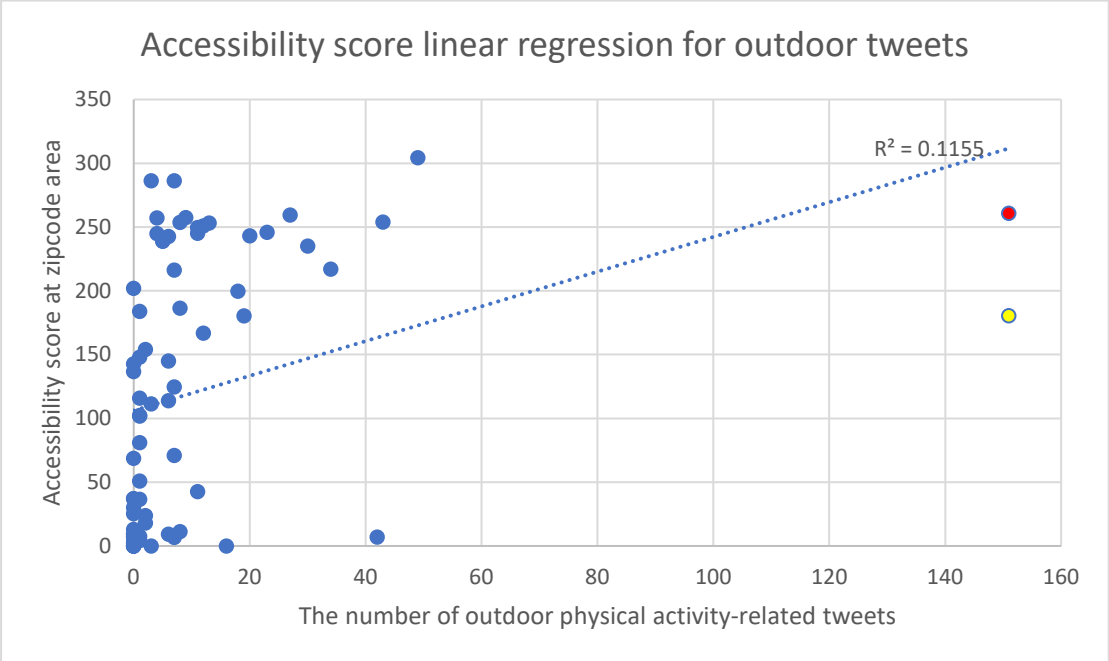

Figure 6 Accessibility score at zip code area & Outdoor physical activity-related tweets

Obviously, the linear fitting was too rough in this case: the correlation coefficient was very low, and the fitting effect was not good. We also noticed that there were two obvious outliers in both the indoor figure and outdoor figure. They were ZIP 78701 (shown red in Figure 5 - 6) and ZIP 78746 (shown yellow in

Figure 5 - 6). By observing the data points, the author tried a logarithmic fit. This requires deleting all zero points, in other words, deleting regions where there were not any physical activity-related tweets. The effect was better, but not enough. After removing outliers, the results were shown below in Figures 7 - 10.

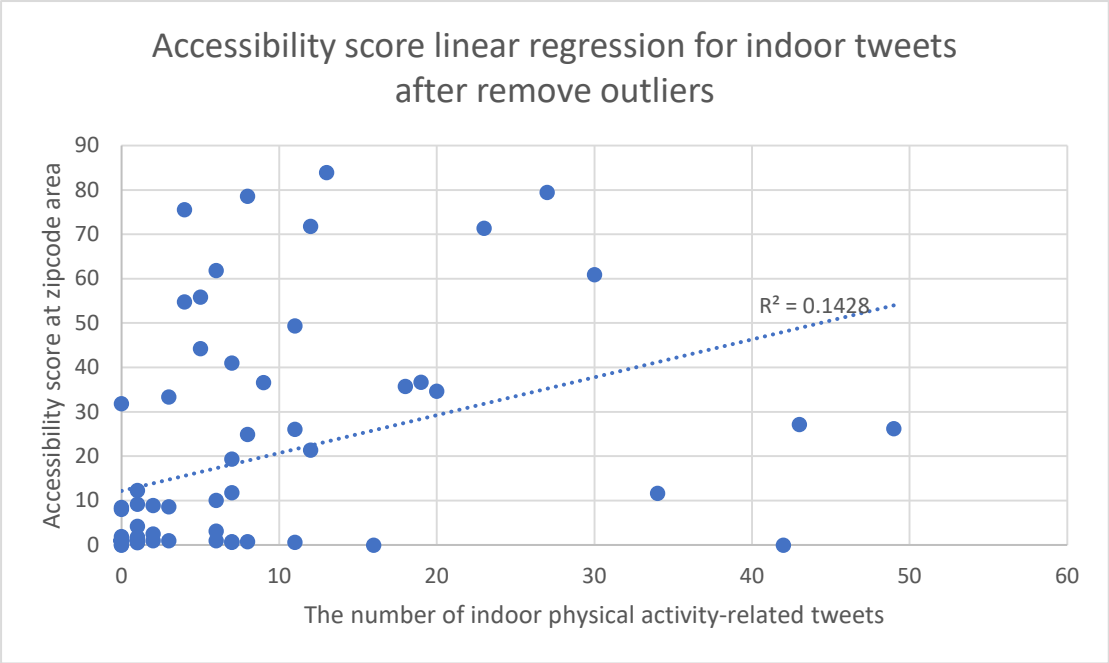

Figure 7 Accessibility score at zip code area without outliers & Indoor physical activity-related tweets

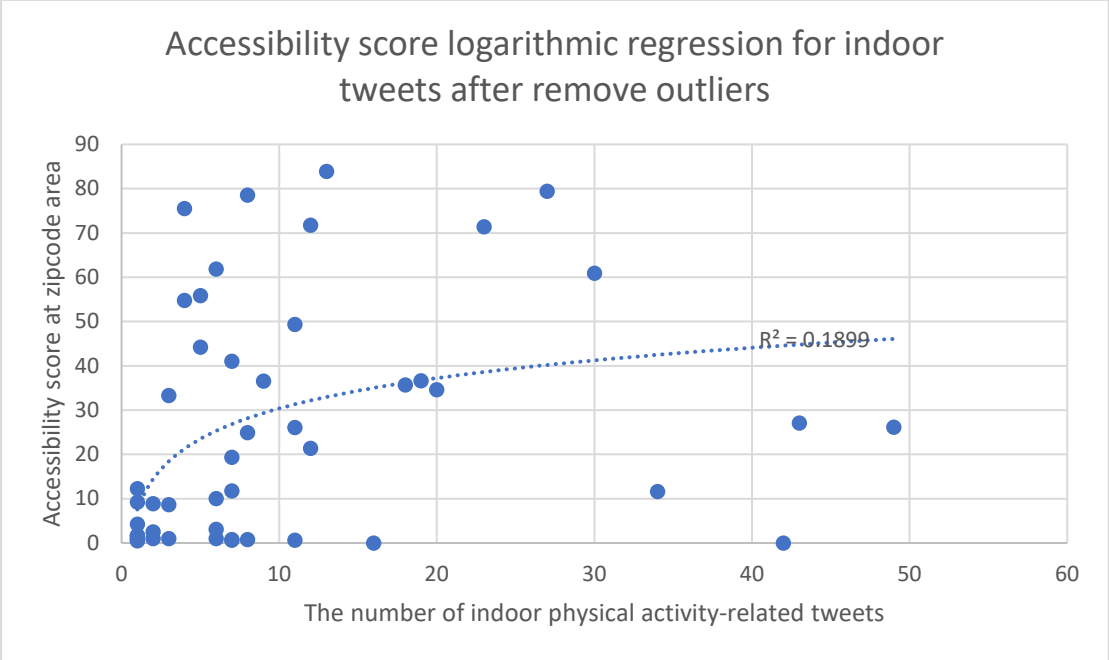

Figure 8 Accessibility score at zip code area without outliers & Indoor physical activity-related tweets (Logarithmic function)

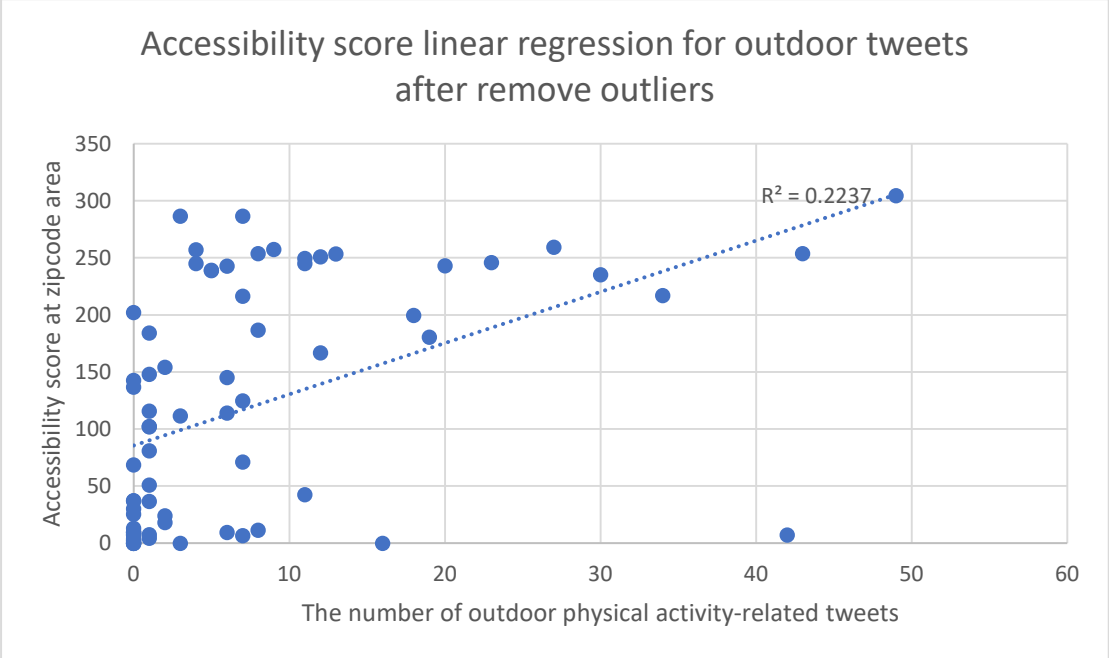

Figure 9 Accessibility score at zip code area without outliers & Outdoor physical activity-related tweets

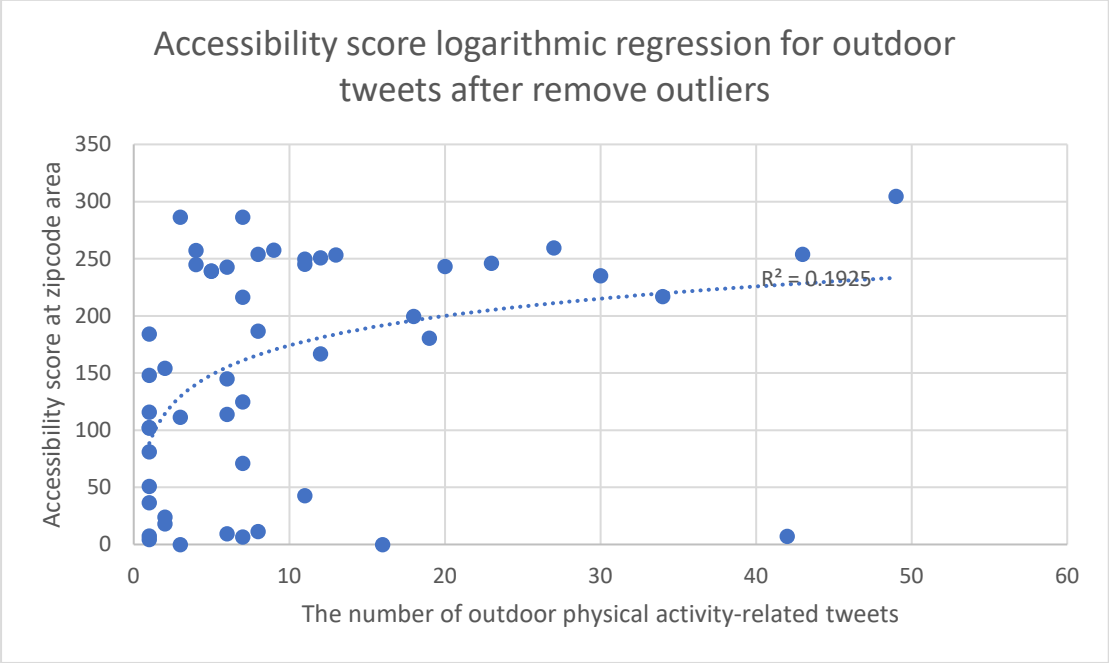

Figure 10 Accessibility score at zip code area without outliers & Outdoor physical activity-related tweets (Logarithmic function)

The logarithmic fitting helps with the indoor situation, but no improvement with the outdoor situation. Except this, although the linear regression doing better after removed outliers, however, we still cannot say it has any relationship because of the much lower correlation R^2.

Discussion

This study does have some strengths and weakness. One of the strengths was the use of street distance, which results in higher accuracy. The most important point is the combination with social media, which provides an indirect way to measure to what extent people are involved in indoor and outdoor physical activities. We don't have a direct measure of the usage of the physical activity facilities but used tweet mentioning as a surrogate. However, some weaknesses still need to be pointed out.

The time period for this study is winter. Even if the winter in Texas is not very cold, the climatic factors still significantly affect people's participation in activities, especially the enthusiasm for outdoor activities. Since some summer activities were in the keywords list, for example, kayaking and rafting, as well as people may don't want to swim in the winter, the activities amount may not accurate. In future studies, it may be worthwhile to distinguish between physical activities that are climate sensitive versus those not. The same reason, if our research area is in the north, some winter activities such as ski and ice hockey may bloom. Of course, it's best to sort by popularity at the local if possible whatever place and season.

Continuing the above considerations, a number of people do not go out because they do not have transportation, live far away from the facilities, or need to take care of their children or other family obligations. They may participate in physical activities at home, such as yoga and squash. When such activities were

monitored on Twitter, it is difficult to connect their locations with particular facilities. And the residents who live in apartments usually do not need to go to a specialty gym because most of the apartments have their own fitting center, which didn't involve in our study. In particular, a highly concentrated area of Twitter users, around University of Texas Austin campus, where students are more likely to use the university gym doesn't list either. They may be improved in further studies.

The most important limitation of the study is that there are factors that affect the relationship between activity facilities and the activity-related tweets that went unaccounted for in the statistical analyses. Only the existence and distance were accounted for, but other factors we mentioned before also play a role in these patterns. For instance, someone cannot afford the membership of a gym, or someone must take care of their babies at home. They may post tweets about physical activities but do not attend.

In this study, the author only considered the impact of the size of the parks and green spaces on its service capacity. In fact, the park's amenities will have a major impact. For instance, families with pets will prefer an outdoor space with a dog park. In addition, the size of the park will also affect the number of tweets was posted, and parks with rich landscapes are more likely to attract people to send tweets. As for the distance determining, the author used the centroid of parks and open spaces to take the place of the polygon. This method missed the entrance information to simplify the calculation but caused

the modifiable areal unit problem (MAUP). The imperfection by this method may hinder any significant difference one can find in spatial accessibility.

Finally, the outliers zip code areas (shown in Figure 1) we observed are worth some special attention. The high tweets amount-high accessibility score outlier, Zip 78701, has the most convenient access to both indoor and outdoor activity facilities, as well as the greatest number of tweets. However, it only has a population of 7,401 as estimated in 2016, both small in area and population. This zip code area is in the downtown district. It has more commuters than residents and is rich in physical activity facilities.

The other outlier, the high tweets amount-low accessibility score one, Zip 78746, has many tweets but there is poor accessibility to physical activity facilities. There is a significantly high level of physical activities in this area, which is disproportionate with its population of 27,971. A number of other regions, such as Zip 78704 with 46,474 people and Zip 78705 with 32,424 people have far fewer numbers of physical activity-related tweets. If integrated into socioeconomic status, we noticed that the median income for a household in Zip 78746, the city of Westlake Hills, was $128,556, which is obviously higher than the other Zip areas. Zip area 78746 is a rich community enclaved by the City of Austin, where there is a very high level of health literate and where physically-active lifestyle is widely adopted.

Conclusion

The use of activities facilities is important to many areas. Urban planners need to predict which parts of a city have the greatest activity needs, merchants need to locate their gyms wisely for maximum benefit, medical researchers need to be prepared to respond to the health consequences from physically inactive lifestyle, and geographers want to know the connection between people's activities and the environment. From the data perspective, Twitter date provides a unique opportunity for us to grab an indicator regarding what extent residents are active at what locations.

This research examined the trends between the accessibility to indoor and outdoor physical activity facilities and people' s possible physical activity level. According to the results, the outdoor activity facilities accessibility score is negatively correlated with the outdoor physical activity-related tweets, contrary to expectations, as well as the indoor situation didn' t display significantly. We don' t have a straightforward answer to this elusive relationship since the study is lacking in the control of some variables. They include social and subjective such as ethnicity, age, and socioeconomic status, among other limitations.

In summary, the study found that the physical activity facilities accessibility was related to the volume of tweets in which people talk about physical activities. Additionally, learning from Zip code level study, we found that we can identify some communities that care about physical activity

significantly more than neighbors who have a similar accessibility, by looking at the outliers in regression analysis.

Reference

[1] Paul, Michael J., and Mark Dredze. "A model for mining public health topics from Twitter." Health 11 (2012): 16-6.

[2] Thompson, Catharine Ward. "Activity, exercise and the planning and design of outdoor spaces." Journal of Environmental Psychology 34 (2013): 79-96.

[3] Maher, Carol A., Lucy K. Lewis, Katia Ferrar, Simon Marshall, Ilse De Bourdeaudhuij, and Corneel Vandelanotte. "Are health behavior change interventions that use online social networks effective? A systematic review." Journal of medical Internet research 16, no. 2 (2014).

[4] MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering". Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284–292. ISBN 0-521-64298-1. MR 2012999.

[5] Dredze, Mark, Michael J. Paul, Shane Bergsma, and Hieu Tran. "Carmen: A twitter geolocation system with applications to public health." In AAAI workshop on expanding the boundaries of health informatics using AI (HIAI), vol. 23, p. 45. 2013.

[6] Thackeray, Rosemary, Benjamin T. Crookston, and Joshua H. West. "Correlates of health-related social media use among adults." Journal of medical Internet research 15, no. 1 (2013).

[7] Gelormino, Elena, Giulia Melis, Cristina Marietta, and Giuseppe Costa. "From built environment to health inequalities: An explanatory framework based on evidence." Preventive medicine reports 2 (2015): 737-745.

[8] Heaivilin, N., B. Gerbert, J. E. Page, and J. L. Gibbs. "Public health surveillance of dental pain via Twitter." Journal of dental research 90, no. 9 (2011): 1047-1051.

[9] Centola, Damon. "Social media and the science of health behavior." Circulation 127, no. 21 (2013): 2135-2144.

[10]Chou, Wen-Ying Sylvia, Yvonne M. Hunt, Ellen Burke Beckjord, Richard P. Moser, and Bradford W. Hesse. "Social media use in the United States: implications for health communication." Journal of medical Internet research 11, no. 4 (2009).

[11]Paul, Michael J., and Mark Dredze. "You are what you Tweet: Analyzing Twitter for public health." Icwsm 20 (2011): 265-272.

[12]Austin/Travis County Health and Human Services Department (A/TCHHSD). "Community Health Assessment Austin/Travis County Texas" (2012)

[13]Sprinthall, R. C. (2011). Basic Statistical Analysis (9th ed.). Pearson Education. ISBN 978-0-205-05217-2.

[14]Zimmerman, Donald W. (1997). "A Note on Interpretation of the Paired-Samples t-Test". Journal of Educational and Behavioral Statistics. 22 (3): 349–360. doi:10.3102/10769986022003349. JSTOR 1165289.

Appendix

Keywords list

Seed words	Indoor related words	Outdoor-related words
health	crunch	jog
weight loss	sit-up	walk
wellness	trunk rotation	hike
exercise	leg pull-in	hiking
fitness	side plank	run
workout	hyperextension	5k
diet	row	10k
nutrition	back fly	pentathlon
gym	pulldown	marathon
fit fam	pull-down	soccer
weight	pullup	football
control	pull-up	baseball
healing	leg curl	fishing
mindfulness	squat	frisky
fat loss	front lunge	hockey
lose weight	hip	kayak
training	shoulder press	rafting
fitspo	shoulder	softball
sport	extension	tennis
muscle	lateral arm pull	triathlon
bodybuilding	biceps curl	frisbee
	chin-up	wakeboard
	triceps	wrestle
	extension	wrestling
	cycling	ski
	bowling	backpacking
	aerobox	camping
	aero-kickboxing	climbing
	aero-step-toning	flying gliding
	aero-toning	golf
	bootcamp	horse riding
	bosu	powerboats
	boxing	sailing
	core strength	snorkeling
	bootcamp	shooting
	hiit	scrambling
	jump rope	wilderness
	nordic walking	survival

	physical conditioning piloxing piyo pilates toning tabata total sculpt y bar essentrics gentle pilates meditation yoga pilates qi gong tchi kung stretching martial arts aikido capoeira karate tai chi self-defence aero-belly dancing aero-dance aero-dance- pilates aero-latin african dance ballet workout belly dancing contemporary dance djamboola hip-hop line dance the groove ? tango teen dance zumba myofascial massage trx	surf trekking wildlife safari snorkel angling gliding camel safari rock climbing camping diving canyoning ballooning desert jeep safari bicycling biking bird-watching bird watching birdwatching parasailing motorbike expedition elephant safari paramotoring tree climbing windsurfing skydiving adventure park canoeing paragliding snowshoeing hunting clam digging wingsuit flying ice climbing orienteering mountain climbing atv riding paintball skateboarding

	cardio-graphy aquafit aqua arthritis aqua bootcamp aqua cardio aqua core aqua dance aqua interval aqua jogging aqua parent & baby aqua zumba prenatal aqua swim taekwondo judo badminton pickleball ping pong squash basketball futsal indoor soccer volleyball racquetball table tennis	