Mining the Relationship between Crimes, Weather and Tweets

Joseph Alamo[1], Claudia Fortes[1], Nicole Occhiogrosso[1], Ching-yu Huang[2]
[1]NJ Center for Science, Technology and Mathematics
[2]School of Computer Science
Kean University
1000 Morris Ave., Union, NJ 07083, USA
[2]+1-908-737-6157
{alamoj, fortesc, occhiogn, chuang}@kean.edu

ABSTRACT
This research project attempts to correlate crime rates in Orlando, Florida to Orlando's weather and Twitter presence. The central dataset of interest details the crime incidents in Orlando, Florida as reported daily by the Orlando Police Department. This dataset gives the dates, categories (e.g. theft, aggravated assault, etc.), and latitude and longitude of each reported crime incident. Using a Twitter developer account, Tweets pertaining to crime are downloaded from the greater Orlando area. Tweets are filtered by the following indexed keywords: "crime", "drugs", "narcotics", "weapons", "assault", "theft", "robbery", "murder", and "larceny." Additionally, Orlando's daily weather data is collected from the National Oceanic and Atmospheric Administration. Using measures of similarity, it is discovered that crime in Orlando is concentrated most closely near Orlando's downtown center. Using regression, moderate correlations are drawn between the rates of crime and the posting of crime-related Tweets. Lastly, chi-square tests are used to show the effect of weather on crime. High crime rates are associated with average daily temperatures above 60°F. Low crime rates are associated with days with precipitation.

CCS Concepts
- **Information systems → Information systems applications → Collaborative and social computing systems and tools → Social networking sites**

Keywords
Big data; Correlation; Crime; Weather; Twitter

1. INTRODUCTION
Since its founding in 2006, Twitter has experienced rapid growth. By a 2018 estimate, there are 500 million Tweets posted per day. 100 million distinct users actively use the social media platform daily. Within the United States alone, 69 million users alone use Twitter. [1] Therefore, it stands within reason that Tweets are often representative of behavior in its users' populations. Realtime Tweets on the social media platform have served as an early warning system for developing crises, like during Germany's 2011 *E. coli* outbreak. [2] Researchers at the University of Southern California used sentiment analysis and regression to determine that Twitter users with positive tobacco-related Tweets were more likely to have smoked cigarettes. [3] Elsewhere, researchers have analyzed the efficacy of Twitter campaigns on determining voting preferences in the 2016 U.S. presidential election. [4]

In a 2014 study at the University of Virginia, Twitter data was determined as a predictor of crime in Chicago, Illinois. The Chicago Police Department categorized crime into 25 categories. This particular research study identified that their predictive model based on three months of Twitter data was successful in predicting crime for 19 of the 25 crime types. [5] While this study is substantive, little work has been done to correlate crime rates with Tweets in other cities, such as Orlando, Florida.

This research project attempts to perform correlation analysis on crime specifically in Orlando. Data, dating back to January 2017, about crime incidents in Orlando are available publicly online from the Orlando Police Department. [6] The "parent incident type" classifies each crime into one of 9 general categories: Assault, Breaking & Entering, Drugs, Homicide, Property Crime, Robbery, Theft, Theft of Vehicle, and Weapons Offense. For this study, "Theft" and "Theft of Vehicle" are condensed into a single "Theft" classification. Longitude and latitude of each crime are also provided. Twitter data is collected from a Twitter developer API over the course of 4 weeks.

Lastly, crime can be correlated to other factors besides Twitter usage, such as the weather. Analysts at the Chicago Tribune identified the effect of temperature on crime in

PRAI '19, August 26–28, 2019, Wenzhou, China

ACM ISBN 978-1-4503-7231-2/19/08 …$15.00
DOI: https://doi.org/10.1145/3357777.3357787

Chicago. They found that crime rates in the categories of battery, theft, and criminal damage increased with higher temperatures. They, however, only looked for correlations between temperature and crime in Chicago. This research project attempts to find associations between crime in Orlando and other weather-influencing factors such as temperature, wind speed, and precipitation. Orlando weather data is compiled between January 2018 and April 2019 from the National Oceanic and Atmospheric Administration (NOAA) website's Daily Summaries dataset [7].

2. METHODS

Data is collected from three different sources: Twitter, the Orlando police department, and the NAOO. Data cleaning is performed on the streamed tweets. Various data mining techniques are analyze the Orlando crime dataset, as well as correlate Tweets and weather data to Orlando crimes. These include Euclidean similarity measures, linear regression analysis, and chi-squared statistics.

2.1 Data Collection

To gather Twitter data, a developer account was requested from Twitter. A Python code that imported the "tweepy" Python package [8] were used to communicate with Twitter's data retrieval API. Twitter's data is natively stored as a json file [9]. The Python script collected Tweets in real-time that had the following indexed keywords: "crime", "drugs", "narcotics", "weapons", "assault", "theft", "robbery", "murder", and "larceny." The "json" package was used to retrieve the following attributes: user, industry, text, and created at. Additionally, geolocation coordinates were derived from the bounding box attribute. Most Tweets do not list exact longitude and latitude coordinates but instead bound the tweet location to within a set of 4 coordinates forming the endpoints of a rectangle. These 4 coordinates were parsed using a "filter" function call, and a singular latitude and longitude coordinate was approximated as the midpoints of the 4 coordinates. The Python program used the "csv" package to write each attribute to a comma-separated values file. Using a cron jab in a UNIX server, the Python script was called hourly from March 2nd to April 7th, 2019. After data collection was complete, all csv files were combined into one csv file using the following Unix command*: awk 'FNR > 1' *.csv > merged.csv.*

What was of principle concern was to be able to utilize the longitude and latitude coordinates to find the distance between each Tweet's location and Orlando. The center of downtown Orlando was pinpointed as: 28.5383 °N, 81.3792 ° W, as illustrated in Figure 1.

The Orlando crime data and NOAA weather data, dating back to January 1, 2018, were downloaded as csv files online. The Orlando crime data was also assigned a distance from the center of downtown Orlando using the Haversine formula. In total, over 24000 crimes were recorded, and over 500 days of weather data was collected. With all three datasets as csv files, they were imported into a MySQL database.

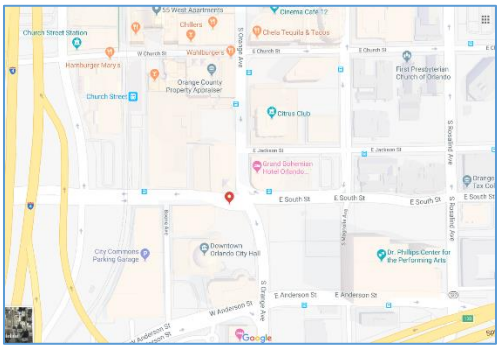


Figure 1. Map of Downtown Orlando. The coordinates for the center of downtown Orlando correspond to the address of Downtown Orlando City Hall. (Source: GoogleMaps)

In Excel VBA, a program was constructed to calculate the distance each tweet and Orlando using the Haversine formula. The VBA code is shown in Figure 2. The Haversine formula determines the shortest distance between two points on a spherical body, and is commonly used to compute the distance between two longitude and latitude coordinates [10]. In total, 116354 data points were compiled. However, after taking only a subset of the Tweets that were within a 50 km radius of Orlando, the dataset reduced to 281 Tweets.

```
Public Function getDistance(latitude1,
longitude1, latitude2, longitude2)
    earth_radius = 6371
    Pi = 3.14159265
    deg2rad = Pi/180

    dLat = deg2rad*(latitude2-latitude1)
    dLon = deg2rad*(longitude2-longitude1)
    a = Sin(dLat/2)*Sin(dLat/2) +
    Cos(deg2rad * latitude1)*Cos(deg2rad *
    latitude2) * Sin(dLon/2)*Sin(dLon/2)

    c = 2*WorksheetFunction.Asin(Sqr(a))
    d = earth_radius*c
    getDistance = d
End Function
```

Figure 2. Implementation of Haversine Formula in Excel VBA in order to calculate the distance in kilometers between two pairs of latitude/longitude coordinates.

2.2 Analysis of Crime Dataset

This research attempts to answer the question of whether crime is more concentrated closer to Orlando's downtown center. To decide the appropriate target features in the dataset, a "CrimeFrequency" table was created. Table 1 displays the "Incident Type" and the "Proportion" of occurrence, meaning the percentage of each type of crime out of the total crime incidence.

Table 1. Table listing the nine different crime type categories from Orlando crime database and the percentage of the total crimes occurring into those categories

IncidentType	Proportion
Theft	0.5411
Drugs	0.1046
Property Crime	0.0921
Breaking & Entering	0.0821
Assault	0.0711
Theft of Vehicle	0.0642
Robbery	0.0299
Weapons Offense	0.014
Homicide	0.0009

Base on the results it was determined that the crimes of theft (including theft of vehicle) and drugs are the two categories that display the highest percentages of crime. Then, the dataset was grouped by how far away crimes occurred from Orlando's downtown center, using an interval size of 2 kilometers. Any distance above 26 km from the center or Orlando presents very small values of crime, if any, so they were excluded. Then, the data was normalized between 0 and 1. After normalizing the data, the Pearson correlation coefficients between the distance from Orlando's center to the number of thefts and number of drugs were calculated. In addition, the Euclidean distance was calculated between each interval of distance based on the attributes of the number of thefts and drugs. The minimum Euclidean distance corresponds to the two intervals with the most similar amount of crimes.

To concentrate on just the region closest to Orlando's downtown center, this process was then repeated for a smaller radius. The same binning technique was utilized, however this time 5 bins were created of 1 kilometer each.

2.3 Correlation of Tweets and Crime

Because of errors in downloading the data, only Twitter data from March 7th to April 4th was used in analysis. The corresponding Orlando crime data for those dates was also isolated. The number of Tweets and crimes per day were queried. In addition, the number of crimes categorized as thefts and as drug-related crimes were also tabulated. In Excel, the number of total crimes, thefts, and drug-related crimes were normalized between 0 and 1 and then plotted with the number of Tweets versus time. The Pearson correlation coefficient was calculated between each crime metric and the number of Tweets. To confirm the accuracy of the correlation coefficient calculations, each crime metric was plotted individually against the number of Tweets in a scatter plot.

2.4 Association between Weather and Crime

The number of crimes per day was calculated by counting entries with the same date this was then processed against three different features of the weather dataset, rain, wind speed, and average temperature. Everything was compared based on the date for 451 days. Chi-Squared Tables were created along with calculating the Chi-Squared and p values. This was then used to determine if the absence or presence of rain cause the crime rate to be high or low. Afterward, a correlation calculation was performed on the precipitation amounts versus the crime counts for the day. Everything was repeated for both the wind speed above and below 8 and average temperature above and below 60. Wind speed of 8 was chosen because that was the median of the data, wind speed of 60 was chosen based on being the approximate median. Chi-squared calculations were all completed assuming 60 crimes or less as a low crime and anything higher as high crime rate. Initially a crime rate of 53 was chosen based on the median of the data but this rate yielded no significant p-values so 60 was chosen as approximately one standard deviation step above the median.

3. RESULTS

Results are tabulated to depict the concentration of crimes in Orlando, the strength of the linear correlation between Tweets and the number of daily crimes, and the association between weather factors and crime incidence.

3.1 Analysis of Crime Dataset

Table 2 tabulates the number of thefts and drug-related crimes for each region outside of Orlando's center. Table 3 depicts the same data but the count of theft and drug crimes are normalized between 0 and 1. The Pearson correlation coefficients were calculated between Figure 1, as shown in Table 3.

Table 2. Table listing the distance ranges from Orlando's city center along with the total number of thefts and drug-related crimes in those respective areas

ID	Distance in Km from Orlando	Theft_Count	Drugs_Count
1	0-2	2711	879
2	2-4	2316	499
3	4-6	1940	355
4	6-8	2422	303
5	8-10	2372	305
6	10-12	1571	134
7	12-14	915	39
8	14-16	121	3
9	16-18	66	6
10	18-20	79	6
11	20-22	131	5
12	22-24	36	1
13	24-26	2	2

The calculations revealed that the theft and drug crimes had a positive correlation of 0.86319, meaning that they both increase or decrease together. The correlation between distance from the Orlando center and drug crimes is -0.86254, meaning the two attributes are inversely correlated, when one increases the other one decreases. The same type

of results was obtained when calculating the correlation between distance from Orlando center and theft crime, the correlation value is -0.92646, so they are also negatively correlated.

Table 3. Normalized data (between 0 and 1) for theft and drug-related crimes based on Table 2

ID	Theft_Count	Drugs_Count
1	1	1
2	0.8542	0.5672
3	0.7154	0.4032
4	0.8933	0.344
5	0.8749	0.3462
6	0.5792	0.1515
7	0.337	0.0433
8	0.0439	0.0023
9	0.0236	0.0057
10	0.0284	0.0057
11	0.0476	0.0046
12	0.0126	0
13	0	0.0011

Table 4. Pearson Correlation Coefficients

The correlation between the crimes of theft and drugs is: 0.863187872
The correlation between distance from Orlando and drug crime is: -0.862546438
The correlation between distance from Orlando and theft crime is: -0.926459958

Euclidean distance was calculated to determine which pairs of distances from the Orlando center are more similar. The result was the pair (11, 8) which have the minimum Euclidean distance of 0.004356. The bin 11 corresponds to a distance of 20 to 22 kilometers away from the Orlando center, and bin 8 accounts for crimes 14 to 16 kilometers from Orlando. Because the results obtained were too far away from the area of highest crime incidence, the same procedure was repeated for a smaller radius. The elected distance was 5 kilometers. The same binning technique was utilized, however this time 5 bins were created of 1 kilometer each. The count of thefts and drug crimes in these regions are summarized in Table 5.

Table 5. Table listing the total crime counts for thefts and drugs within 5 km of Orlando's city center

ID	Distance in Km from Orlando	Theft_Count	Drugs_Count
1	0-1	1438	385
2	1-2	1273	494
3	2-3	1144	325
4	3-4	1172	174
5	4-5	870	149

The data was normalized between 0 and 1, and the Euclidean distance was calculated. Table 5 represents the corresponding Euclidean distance matrix. The results revealed that the smallest Euclidean distance is between the pair 1 and 2. Therefore, it can be concluded that the highest similarity between the crimes of drugs and theft occur closer to the center of Orlando.

Table 6. Euclidean Distance Matrix for each coordinate in Table 5. The most similar amounts of thefts and drugs occur in 0-1 and 1-2 km distances away from Orlando's city center.

Euclidean Distance					
Pair (t2, d2)	**1(1.0000, 0.6841)**	**2(0.7095, 1.0000)**	**3(0.4824, 0.5101)**	**4(0.5317, 0.0725)**	**5(0.0000, 0.0000)**
1(1.0000, 0.6841)	0				
2(0.7095, 1.0000)	0.429	0			
3(0.4824, 0.5101)	0.546	0.540	0		
4(0.5317, 0.0725)	0.770	0.944	0.440	0	
5(0.0000, 0.0000)	1.21	1.226	0.702	0.537	0

3.2 Correlation of Tweets and Crime

After normalizing the Twitter and crime data to be between 0 and 1, they were plotted versus time in Figure 3. The Pearson correlation coefficient (r) was calculated between Tweets and all crimes, thefts, and drug-related crimes. Results are shown in Table 7. All values of *r* are close to zero indicating no significant correlation between the datasets.

Table 7. Correlations between Tweets and Crime.

	Tweets and Crime	Tweets and Thefts	Tweets and Drug-Crimes
R :	0.009535096	0.079812899	-0.13594667

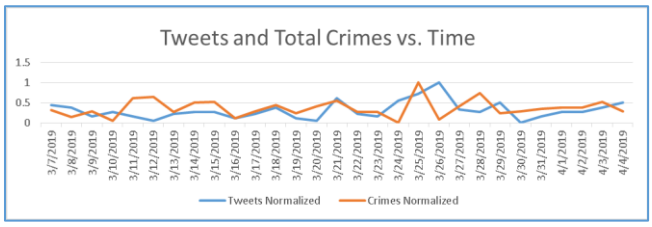
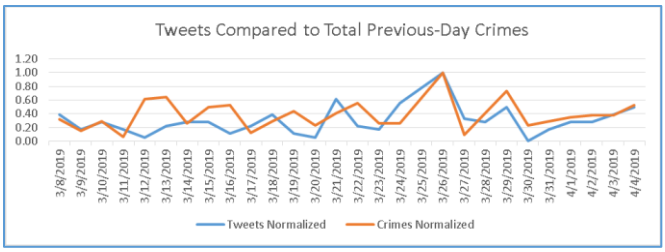
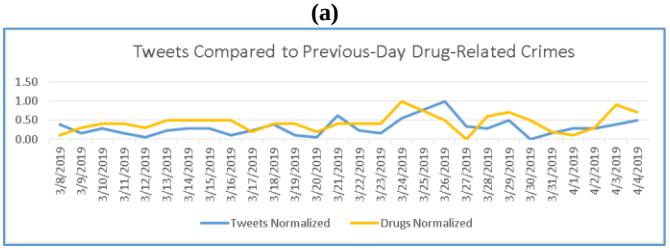


Figure 3. The Pearson correlation coefficients (r) are shown for the number of Tweets versus the number of crimes, thefts, and drug-related crimes. All values of r are close to zero indicating no significant correlation between the datasets. The plots of Tweets and total crimes versus time is shown for reference.

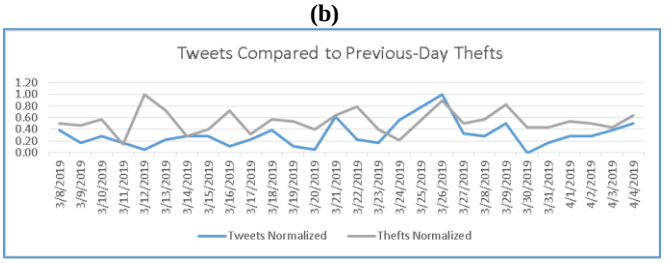
This finding was not expected. An outlier of very low crime occurs on 3/24, so that data point was removed from the dataset. It was also suspected that the timing of Tweets and crime were not offset properly. The number of Tweets in a given day were then plotted against the number of crimes that occurred the day before (Figure 4). More peaks between the two curves seem to line up, specifically for the second half of the data. This process is repeated for the number of thefts and drug-related crimes as well.



(a)



(b)



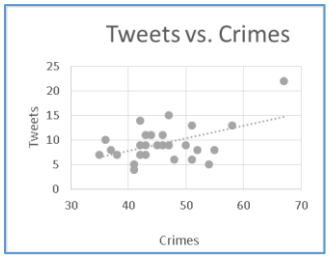
(c)

Figure 4. Line graphs depicting the progression of number of daily tweets versus the number of reported crimes occurring the day before. (a) Tweets about crimes compared to all crimes occurring in Orlando the day before. (b) Tweets about crimes compared to the number of drug-related incidents occurring in Orlando the day before. (c) Tweets about crimes compared to the number of reported thefts occurring in Orlando the day before .

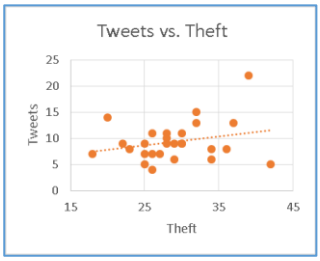
The Pearson correlation coefficients were recalculated and are presented in Table 8. These findings then showed moderate correlation between the number of Tweets and previous-day crimes. The correlation between Tweets and previous-day thefts, and Tweets and previous-day drug crimes were also weakly correlated. This indicates that more people post about crimes on Twitter the day after crimes occur. Figure 5(a-c) shows the associated scatterplots for these trends with regression lines plotted as well.

Table 8. Correlations between Tweets and Crime.

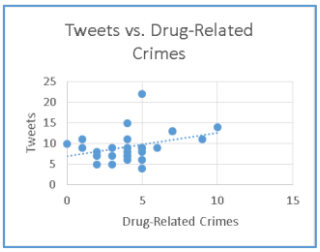
	Tweets and Crime	Tweets and Thefts	Tweets and Drug-Crimes
R:	0.492815917	0.252833586	0.339814376



(a)



(b)



(c)

Figure 5. The Pearson correlation coefficients (r) are computed for the number of Tweets versus the number of previous-day crimes (a), thefts (b), and drug-related crimes (c). The associated scatter plots are shown with regression lines drawn in. Tweets and total previous-day crimes are moderately correlated.

3.3 Association between Weather and Crime

The results of chi-squared tests are summarized in Tables 9-11. Table 9 depicts moderate association between crime and precipitation. Little to no association was found between wind speed and crime, as shown in Table 10. Table 11 shows the association between crime and temperatures above 60°F.

Crime Rate vs Precipitation

Correlation= -0.052949987153331
Chi-Square Value= 5.2718361950643
p-value= 0.021673

Table 9. Chi-square test of Rain vs Crime

	Rain	Not Rain	Row Total
High Crime	40	51	91
Low Crime	114	251	365
Column Total	154	302	456

Crime Rate vs Wind Speed

Correlation= -0.015588924352478
Chi-Square Value= 0.60318099846717
p-value= 0.437367

Table 10. Chi-square test of Wind vs Crime

	High Wind (>8)	Low Wind (<=8)	Row Total
High Crime	43	48	91
Low Crime	156	209	365
Column Total	199	257	456

Crime Rate vs Average Temperature

Correlation= 0.17954568953367

Chi-Square Value= 6.4537603538061
p-value= 0.011072

Table 11. Chi-square test of Temp vs Crime

	High Temp (>60)	Low Temp (<=60)	Row Total
High Crime	85	6	91
Low Crime	302	63	365
Column Total	387	69	456

4. CONCLUSIONS

This research found that crime incidents (namely thefts and drug offenses) occur disproportionately closer to Orlando's downtown center. The results of this study are partially skewed by the irregular, asymmetric shape of Orlando, as illustrated in Figure 6. To get a better sense of the localized regions where crime is prevalent, a clustering technique could be implemented in future work.

It was found that there is little correlation between the number of Tweets in the greater Orlando area on a given day and the number of crimes. However, there is a moderate positive correlation between the number of Tweets on a given day and the number of crimes occurring the previous day. This indicates that people in Orlando are most likely to tweet about crime the following day after crimes occurred. This research does not indicate anything about the predictive power of Tweets in predicting future crimes. The biggest hindrance in this analysis came from the limited sample size of Tweets available. While over 100,000 Tweets in total were collected in four weeks, less than 300 came from within a 50 km radius of Orlando. This can be attributed to the fact that not all Twitter users have geolocation settings enabled and there are rate limiting restrictions in streaming real-time Twitter data. Given a larger sample size, correlation analysis could be made between crimes and Tweets containing specific individual keywords.

Crime rates and the presence of rain showed significance and at $p<0.05$ but not at $p<0.01$. The correlation calculation showed essentially no linear correlation between precipitation rates and crime rates for a given date. Crime rates and wind speed had no significant p-value along with also showing little linear correlation. Crime rates and average temperature had a significant p-value and a slight positive correlation. Based on the observed data calculations the most likely weather factor to affect crime rate is the average daily temperature. While there was no found linear correlation between weather factors and crime rate there was an apparent association between temperature, precipitation, and crime. Low crime rates are associated with days with precipitation, while high crime rates are associated with daily average temperatures above 60^{o}F.

The implications of these findings are relevant to Orlando's police force. The Orlando community (especially closer to Orlando's center) could benefit from more police reinforcement on warmer days and on days when it is not raining. It is also worthwhile for law enforcement to monitor and analyze local Twitter data, as Tweets are typically reflective of local crimes.

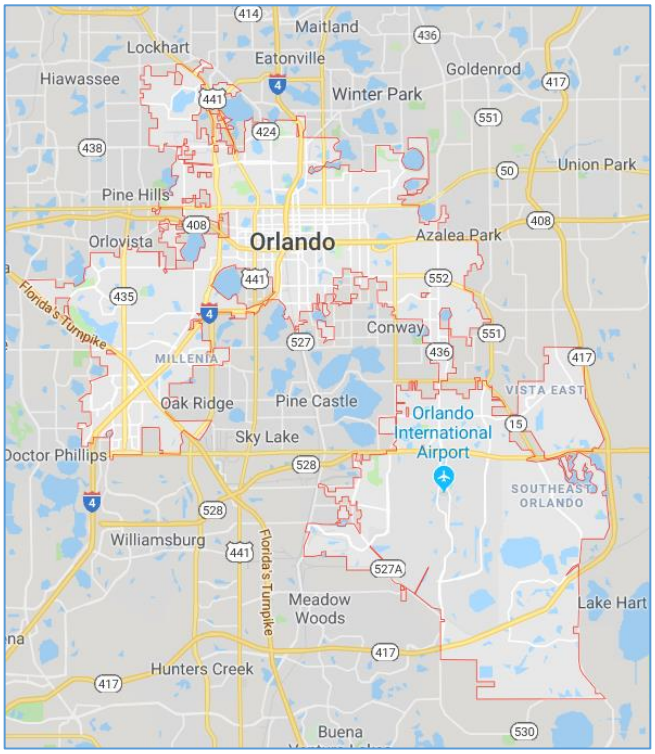


Figure 6. Irregular Boundary Line of the City of Orlando (Source: GoogleMaps).

5. REFERENCES

[1] S. Aslam, "Twitter by the Numbers: Stats, Demographics & Fun Facts," *Omnicore*, 2018. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/. [Accessed: 20-Apr-2019].

[2] C. Orellana-Rodriguez and M. T. Keane, "Attention to news and its dissemination on Twitter: A survey," *Comput. Sci. Rev.*, vol. 29, pp. 74–94, 2018.

[3] J. B. Unger *et al.*, "Talking about tobacco on Twitter is associated with tobacco product use," *Prev. Med. (Baltim).*, vol. 114, no. June, pp. 54–56, 2018.

[4] P. Grover, A. Kumar, Y. K. Dwivedi, and M. Janssen, "Technological Forecasting & Social Change Polarization and acculturation in US Election 2016 outcomes – Can twitter analytics predict changes in voting preferences," *Technol. Forecast. Soc. Chang.*, no. September, pp. 1–23, 2018.

[5] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decis. Support Syst.*, vol. 61, pp. 115–125, 2014.

[6] "Orlando Police Department Incident Dataset." [Online]. Available: https://moto.data.socrata.com/dataset/Orlando-Police-Department/hirq-qvex/data. [Accessed: 20-Apr-2019].

[7] National Oceanic and Atmospheric Administration, "Climate Data Online," 2019. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/datasets. [Accessed: 20-Mar-2019].

[8] J. Roesslein, "Tweepy Documentation," 2019. [Online]. Available: http://docs.tweepy.org/en/3.7.0/. [Accessed: 20-Apr-2019].

[9] Twitter, "Tweet Objects," 2019. [Online]. Available: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json.html.

[10] N. Chopde and M. Nichat, "Landmark Based Shortest Path Detection by Using Dijkstra Algorithm and Haversine Formula," *Int. J. Eng. Res. Appl.*, vol. 3, no. 3, pp. 162–165, 2013.