Twitter-Based Activity Patterns - A Case Study From Post-Circuit Breaker Singapore

GE5228 Group Project Report

Group 5

Lim Zhu An
Phang Yong Xin
Sherie Loh Wei
Xu Yuting [1]

[1] Corresponding author: email address: xu_yuting@hotmail.com

1.0 Abstract

Since 19 July 2020, Singapore entered Phase 2 of re-opening after one and half month's "Circuit Breaker" measures to curb the spread of COVID 19. Although most businesses and public places have resumed operation at a reduced capacity, individuals are strongly advised to practice social distancing and avoid crowds. Both implicit and explicit measures to prevent overcrowding may have changed how people visit places in Singapore. This study used geotagged Twitter data from September to October 2020 to examine the spatial and temporal patterns of residents' locations in Singapore and explore if service amenities remain "attractive" to residents. While distinct temporal patterns of tweets generally aligned with an office-hour behaviour, spatial analysis revealed no statistically significant clusters. The regression analysis showed that distances to service amenities do not provide strong explanations for tweeting patterns. This study then used a Random Forest Supervised Machine Learning Model to train and predict spatial distribution of activities during off-work recreational hours using service amenities POIs and land use mix. The top five explanatory variables are Parks, Parks and malls public links, Taxi stands, Residential areas, and Shopping malls. These variables have the strongest influence in driving the model prediction of spatial distribution of activities in off-work recreational hours.

2.0 Introduction

The ongoing COVID-19 pandemic has severely impacted the Singapore economy and society. On 7 April 2020, the Circuit-Breaker (CB) measures were implemented to restrict the usage of public spaces as well as social gathering (gov.sg, 3 Apr 2020). The Circuit Breaker was subsequently lifted on 2 June (gov.sg, 28 May 2020) and the country cautiously entered more relaxed phases of re-opening. At the time of writing in late October 2020, small group gatherings are allowed and most businesses reopened, while high-risk settings such as bars, pubs and karaoke services remain closed (gov.sg, 20 Oct 2020), reflecting a careful balance of the economic interest against the risk of social congregation to curb the spreading of virus. Individuals are advised to practice social distancing, while businesses are advised to reduce service capacity to avoid forming crowds. Government agencies rolled out apps such as URA's Space Out and NParks' Safe Distancing @ Parks to inform residents on the crowd level in respective public places, allowing people to make informed decisions on when and where to go for services. Besides the implicit consideration of risk in place visiting, more explicit movement control measures, such as zoning and no-visitor rules, have also been implemented by some organisations such as NUS (Davie, 2020). At the same time, more Singaporeans have opted for having services delivered to them, instead of them visiting the amenities (Toh, 2020). The unique context post-CB generates interest to understand what activity patterns are like currently, and whether services and amenities still attract people to visit them.

Location-based social media data, contributed by individual users voluntarily, offers a wealth of information, including temporal, spatial and semantic attributes, and has been used in a growing body of literature (Yan et al., 2020). In domains of urban science and GIS, social media data has been used to understand individual activity-space and spatiotemporal public space visiting patterns (Kovacs-Gyori et al., 2018), sentiment of the public near places (Kovacs-Gyori et al., 2018, and Wei and Lan, 2015), or crowd behaviours during a major event (Lee and Sumiya, 2010). Soliman, Soltani, Yin, Padmanabhan and Wang (2017) explored urban land use classification using movement patterns of Twitter users' in Chicago, achieving an accuracy of 0.78 with key locations derived from temporal patterns. In a similar vein, other social media platforms like Foursquare that have check-in features have been employed in the study of land use in New York City, in addition to data extracted from Twitter (Zhan, Ukkusuri and Zhu, 2014). They then adopted a clustering inference approach in tandem with a supervised learning approach (random forest classifier) for land use inference, where the latter method returned a higher overall accuracy of approximately 79%. Tweets have been analysed to assess public sentiments and mobility patterns during the novel coronavirus pandemic (Kabir and Madria, 2020) in the USA. The findings established a positive correlation between the volume of movement and infection rates.

In Singapore, however, most studies using Twitter data focused on analysing semantic information, and very few used Tweets to understand urban places. Prasetyo, Achanauparp and Lim (2016) found an association between the location of schools, type and competitiveness and a connection between these educational institutions and shopping malls. There is also a gap in knowledge on how the unique context of post-Circuit Breaker re-opening has changed place-visiting patterns, and whether Twitter data is suitable for studying Singapore's urban spaces. In this study, we used Twitter data to explore the spatial-temporal patterns of activities in post-CB Singapore. Based on the place-visiting patterns established, we then modeled and predicted visitors at places during non-office recreational hours, and examined how such patterns relate to service amenities. As an exploratory exercise to acquire an overview of spatio-temporal tweet patterns, the findings will also enable a discussion on the suitability of Twitter data to understand geospatial questions in the context of Singapore.

3.0 Methodology

3.1 Data Collection

Tweets

Twitter public standard API (v 1.1) (https://developer.twitter.com/en/docs/twitter-api/v1/) was used to search for all tweets posted from 18 September 2020 to 10 October 2020. A spatial extent of 25km radius from the Central Catchment was defined for tweet collection via the API, an area that effectively covers most of the Singapore main island (Figure 3.1).

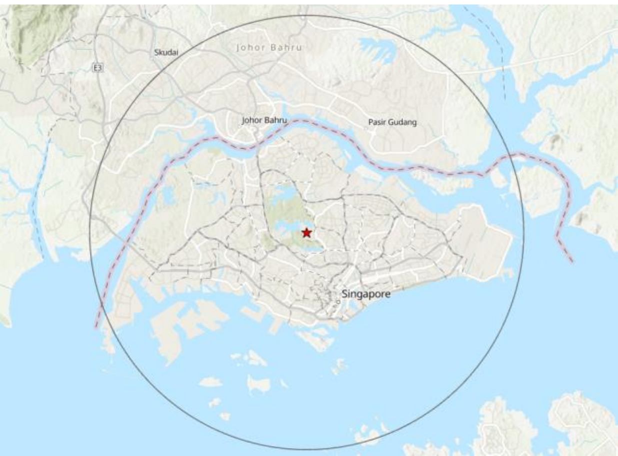


Figure 3.1. Spatial extent of tweet collection

Due to unstable network connection and search rate limit imposed by Twitter, tweet collection was unsuccessful for the time periods specified in Table 3.1.

Date	**Day of Week**	**Status**
23 Sep 2020	Wednesday	Missing 9 hours (8 AM - 5 PM)
24 Sep 2020	Thursday	Missing 8 hours (8 AM - 4 PM)
27 Sep 2020	Sunday	Missing 6 hours (9 AM - 3 PM)
28 Sep 2020	Monday	Missing 11 hours (8 AM - 7 PM)
29 Sep 2020	Tuesday	Missing 12 hours (8 AM - 8 PM)
30 Sep 2020	Wednesday	Missing 13 hours (8 AM - 9 PM)

1 Oct 2020	Thursday	Missing 11 hours (8 AM - 7 PM)
2 Oct 2020	Friday	Missing 10 hours (8 AM - 6 PM)
3 Oct 2020	Saturday	Missing 10 hours (8 AM - 6 PM)
4 Oct 2020	Sunday	Missing 12 hours (8 AM - 8 PM)
5 Oct 2020	Monday	Missing 14 hours (8 AM - 10 PM)
6 Oct 2020	Tuesday	Missing 16 hours (8 AM - 12 AM)
8 Oct 2020	Thursday	Missing 15 hours (9 AM - 12 AM)
9 Oct 2020	Friday	Missing 22 hours (12 AM - 10 PM)
10 Oct 2020	Saturday	Missing 16 hours (9 AM - 12 AM)

Table 3.1. Time periods with unsuccessful tweet collection.

Twitter API returns each tweet object with relevant attributes, including user details, full text of tweets, timestamp and tweet location (full list of attributes in Appendix 8.1). Varied levels of spatial precision were observed with the tweets collected (Table 3.2) as a result of user account setting. According to Twitter product policy, Twitter users who chose to enable precise location on their account will have all their tweets geotagged automatically to specific coordinates, while users who chose not to enable precise location may still optionally tag individual tweets to a location by keying in a place name. Tweets with exact X/Y coordinates were digitised into point features directly, while tweets with a specific landmark as place name are geocoded using OneMap search API (https://docs.onemap.sg/#search) by providing the Place_Name attribute as keywords to obtain the coordinates. Because some place names in the tweets are erroneous or not recognised in the OneMap API, only 306 tweets with specific landmark place names were successfully geocoded. Tweets with spatial information at the scale of a neighbourhood and above were recognised to give a poor indication of the user device location, and were not used in spatial analysis.

Type of Spatial Attributes (in decreasing order of precision)	**Percentage of Tweets**
With exact coordinates (X/Y)	0.25%
With place name (a specific landmark), but without X/Y coordinates	0.31%
With place name (a neighbourhood/town), but without X/Y coordinates	0.01%
With place name (a region), but without X/Y coordinates	2.43%
With place name (a city), but without X/Y coordinates	0.03%
With place name (a state/province), but without X/Y coordinates	0.00%
With place name (a country), but without X/Y coordinates	0.05%
No explicit spatial information	96.91%

Table 3.2. Spatial precision of Tweets

Points of Interest (POIs) and Other Geospatial Data

Geospatial locations of service amenity POIs in GIS format were collected from various official data sources and open data sources.

Data	**Source**

Hawker Centres Residential with 1st Storey Commercial Community Clubs Parks Park Connector Network Dual Use Scheme (DUS) Sports Facilities SCDP Park Mall Master Plan 2019	data.gov.sg
Shopping Malls	List of shopping malls in Singapore from Wikipedia (https://en.wikipedia.org/wiki/List_of_shopping_malls_in_Singapore) and geocoded using OneMap search API
MRT Stations Bus Stops Taxi Stands	mytransport.sg
Medical Facilities	KML file extracted from Google Maps

Table 3.3. Summary of data sources for POIs

3.2 Spatiotemporal Analysis and Linear Regression Analysis

Digitised tweet point features were analysed for their spatial and temporal patterns. The spatiotemporal analysis framework used in Rao et al. (2012) was adapted for this study.

	Temporal Analysis	**Spatial Analysis**	**Dynamic Spatiotemporal Analysis**	**Static Spatiotemporal Analysis**
Temporal Attribute	Independent	Fixed	Independent	Fixed
Spatial Attribute	Fixed	Independent	Dependent	Dependent
Thematic Attribute	Dependent	Dependent	Fixed	Fixed
Examples of questions for this study	In the same location, how do activity patterns vary with time?	In the same time period, how do activity patterns vary in places?	How do spatial patterns of activities vary with time?	How do spatial patterns vary in different locations at a fixed point of time?

Table 3.4. Categories of spatiotemporal analysis, adapted from Rao et al. (2012)

Building on the assumption that presence in space, proxied by posting of tweets, is driven by the purpose to receive services, we then performed Ordinary Least Squares (OLS) regression to estimate the relationship between aggregated tweet counts against the distance to the nearest service amenity POIs to explain the spatial patterns of tweets. The OLS regression model was first run with distances to all 12 types of service amenity POIs as explanatory variables. Subsequently, we assessed model accuracy, individual variables' Variance Inflation Factors (VIFs) and variable collinearity, and removed explanatory variables that give high VIFs (>7.5) and high collinearity (R2 > 0.6) from the model. The final OLS regression model was then analysed for its accuracy and implication.

3.3 Random Forest Model

A random forest regression model offered on ArcGIS was chosen to model the number of tweets at different locations across the study area. ArcGIS forest-based classification and regression tool provides a ready-to-use solution and is well suited for the purposes of this study. This tool uses an adaption of Leo Breiman's random forest algorithm, which is a supervised machine learning model. With the input of service amenity POIs as point/polygon feature layers, the tool is able to generate distances from each training feature to be used for the training of the regression model. Presence of land uses and distance to amenities were conceptualised as potential explanatory variables to model and predict spatial distribution of people in the area.

Model Inputs

To create the point feature layer with attributes that describe the area, the Singapore main island extent was split into square fishnet grids of 1 km by 1 km. This spatial resolution was chosen based on computation power and time required to run the model. It also potentially increases the number of land use types per fishnet grid. The presence of land use (based on Master Plan 2019) in each grid was denoted with categorical attributes (1 for presence, 0 for absence). Tweets posted on Weekends as well as Weekdays from 6pm to midnight were aggregated in each grid to produce a total count to represent spatial distribution of people during off-work recreational hours. A total of 3982 tweets were used in this model input. The land use and tweet count attributes were then joined to the fishnet centroid point features for model input.

Besides land uses, service amenity POIs (Table 3.3) were used as explanatory training distance features. The tool generates the distance between each training feature point and all the distance variable features, which will then be used as continuous numerical inputs for the training of the model. Attributes from the different fields within the training feature layer are used as variables for the model as well.

Model Parameter Settings

The model was initially trained with 100% of the training data to first determine the key explanatory variables that were driving the results of the model. The model parameter "Number of trees" was modified in order to determine a stable model in terms of the ranked importance variables. After a few iterations, the number of trees chosen for this model was 500 trees. The other model parameters such as minimum leaf size, maximum tree depth, data available per tree were all based on ArcGIS default settings.

Model Validation

After a stable model with regards to variable importance is trained, the model was then validated by setting the percentage of training data to be used for validation to determine the accuracy of the model. 10% of the training data was used to validate the accuracy of the model. The model was trained without this random subset of data, and the observed values for those features were compared to the predicted values to validate model performance. Because the choosing of the subset data for validation is random, a total of 15 validation runs was performed to get a maximum, minimum, and median R-Squared value for the validation runs.

4.0 Result and Analysis

4.1 Twitter Users in Singapore

As illustrated in Figure 4.1, a total of 748,533 users contributed to 2,776,651 tweets collected in the full circular buffer area from 18 September 2020 to 10 October 2020, with an average of 3.71 tweets per user. The count of tweets posted by each user varies remarkably, with one top user posting over 9000 tweets, while a majority of others posting only 1 tweet within the study period. A similar pattern was found among users who have posted geotagged tweets within Singapore. A total of 1396 users had posted geotagged tweets on Singapore main island, among which, more than half of the users posted only 1 tweet within the study period. The figures suggest that only a small handful of users were contributing to most number of tweets. This finding is consistent with several studies using Twitter data in Singapore such as Prasetyo et al. (2016) and is consistent with studies in other countries such as Bruns and Stieglitz (2012).

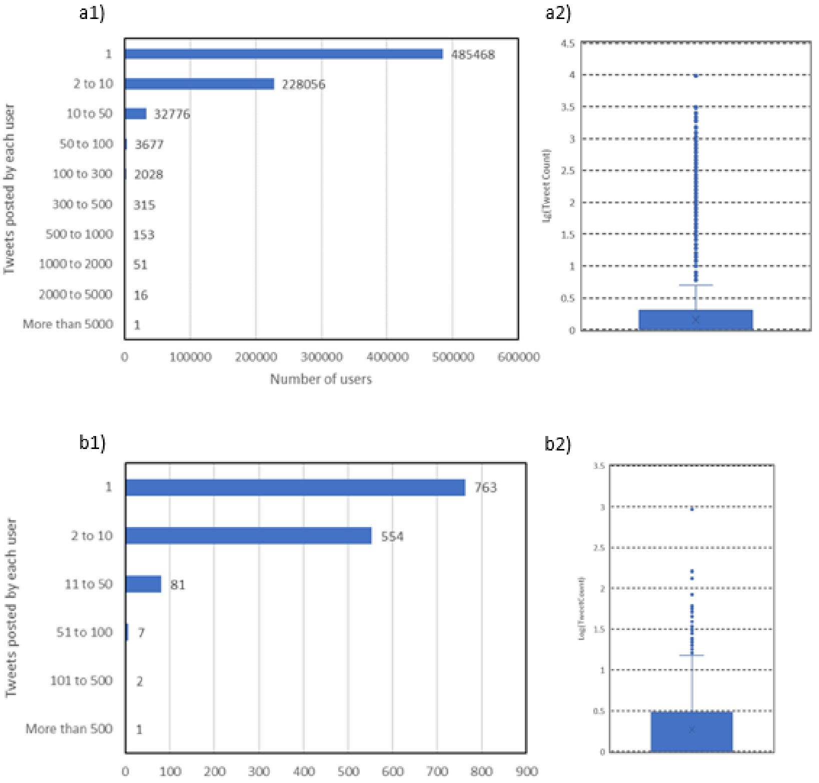


Figure 4.1. a) Distribution of all tweets (total tweets = 2,776,651, total user = 748,533, mean tweets/user = 3.71, median tweets/user = 1, sd = 25.46), b) Distribution of geotagged tweets in Singapore (total tweet count = 5,754, total user = 1,396, mean tweets/user = 4.12, median tweets/user = 1, sd = 25.87)

4.2 Spatio-temporal Analysis

Daily total count of tweets and geotagged tweets in Singapore shows a highly variable pattern over the data collection period (Figure 4.2). Days with lower than average tweets and geotagged tweets generally correspond with days with missing periods of tweet collection (Table 3.1).

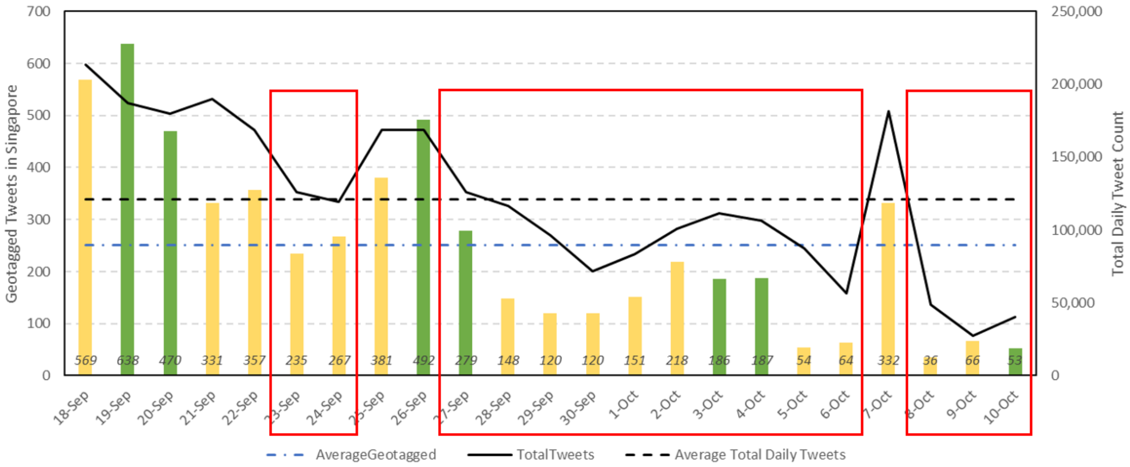
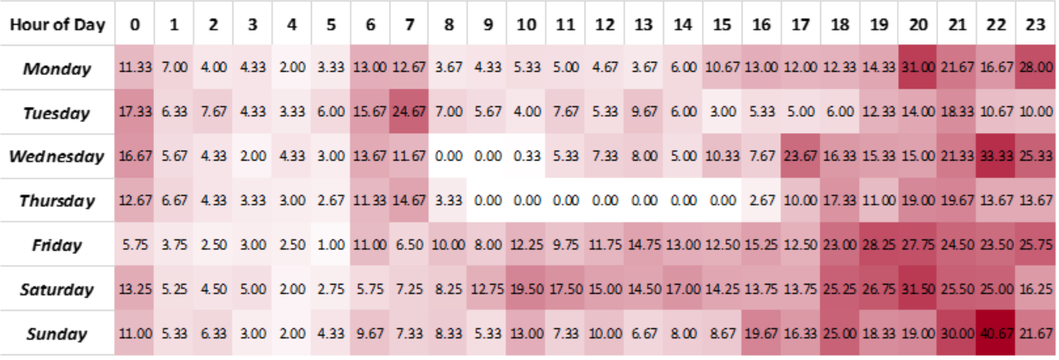


Figure 4.2. Total daily count of tweets and geotagged tweets in Singapore collected over the period from 18 September 2020 to 10 October 2020. Values in the red boxes indicate days with time periods of unsuccessful tweet collection.

Two peak periods for tweet activities on weekdays from Figure 4.3: during morning commuting hours (6 - 8 AM) and evening after-office hours (after 6 PM). This could be due to availability of users to engage in online activities. On weekends, the hourly plot shows a smoothly increasing trend for average tweet counts, suggesting that on weekends, online activities grow in intensity gradually over the course of a day, and finally reaching the peak in the late evening. Over the course of a week, fewer tweets are posted during office hours (weekday 9AM - 6PM) than other time periods.



Hour of Day	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Monday	11.33	7.00	4.00	4.33	2.00	3.33	13.00	12.67	3.67	4.33	5.33	5.00	4.67	3.67	6.00	10.67	13.00	12.00	12.33	14.33	31.00	21.67	16.67	28.00
Tuesday	17.33	6.33	7.67	4.33	3.33	6.00	15.67	24.67	7.00	5.67	4.00	7.67	5.33	9.67	6.00	3.00	5.33	5.00	6.00	12.33	14.00	18.33	10.67	10.00
Wednesday	16.67	5.67	4.33	2.00	4.33	3.00	13.67	11.67	0.00	0.00	0.33	5.33	7.33	8.00	5.00	10.33	7.67	23.67	16.33	15.33	15.00	21.33	33.33	25.33
Thursday	12.67	6.67	4.33	3.33	3.00	2.67	11.33	14.67	3.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.67	10.00	17.33	11.00	19.00	19.67	13.67	13.67
Friday	5.75	3.75	2.50	3.00	2.50	1.00	11.00	6.50	10.00	8.00	12.25	9.75	11.75	14.75	13.00	12.50	15.25	12.50	23.00	28.25	27.75	24.50	23.50	25.75
Saturday	13.25	5.25	4.50	5.00	2.00	2.75	5.75	7.25	8.25	12.75	19.50	17.50	15.00	14.50	17.00	14.25	13.75	13.75	25.25	26.75	31.50	25.50	25.00	16.25
Sunday	11.00	5.33	6.33	3.00	2.00	4.33	9.67	7.33	8.33	5.33	13.00	7.33	10.00	6.67	8.00	8.67	19.67	16.33	25.00	18.33	19.00	30.00	40.67	21.67

Figure 4.3. Average hourly count of tweets posted within Singapore main island spatial extent. Time periods when tweet collection was not successful were excluded from computation.

Results from Incremental Spatial Autocorrelation in ArcGIS on the entire Singapore geotagged tweets dataset show peak distance at around 1km, a spatial scale at which clustering could be observed, but a low z-score suggests that such clustering is not statistically significant. Tweets were further segmented into 2-hour intervals and Incremental Spatial Autocorrelation was conducted on each time block. For all time blocks, the result shows no statistically significant spatial clusters. This suggests that the processes that promote global spatial clustering could be random.

Kernel density of tweet point features (Figure 4.4) show some areas with visually higher density of tweets, such as the Central Business District (CBD), and several large residential estates and regional hubs e.g. Choa Chu Kang, Bukit Batok, Woodlands, Hougang-Sengkang-Punggol stretch, Serangoon, Bishan-Ang Mo Kio, Changi Airport. We then summed up tweet count based on the land use types from Singapore's Master Plan they fall on, and found that over 27% of tweets were posted on a Residential land plots, followed by White[2] and Commercial land use types (Figure 4.5), possibly because work-

[2] A white site allows for multiple uses within a single plot of land e.g. a combination of residential, hotel, commercial uses etc.

from-home was still the default mode of working for most businesses and organisations during the data collection period.

The central area presents higher density of tweets than other areas of Singapore in all time periods, suggesting a consistently higher intensity of tweet activities than other areas in Singapore. The observation was also made by Prasetyo et al. (2016) on tweets in 2012. This can be possibly explained by a greater diversity of land use mix and more important services in the central area, that may attract both work and leisure purpose-driven trips.

In Figure 4.4, over the course of a day, the 04:00 - 06:00 time period represents the time block with the lowest spatial coverage of tweet activities, forming distinct islands of tweet density at the CBD and big residential estates such as Bukit Batok, Serangoon, Bedok and Seng Kang. Night time tweets cover a much bigger and contiguous geographical area, suggesting more distributed patterns of tweeting activities in off-work hours.

From the spatiotemporal analysis above we conclude that tweeting patterns exhibit distinct diurnal and weekly trends. Although visually we can identify distinct areas with higher density, these areas do not form statistically significant spatial clusters, suggesting that random processes could have contributed to the formation of such high density areas in post-CB Singapore.

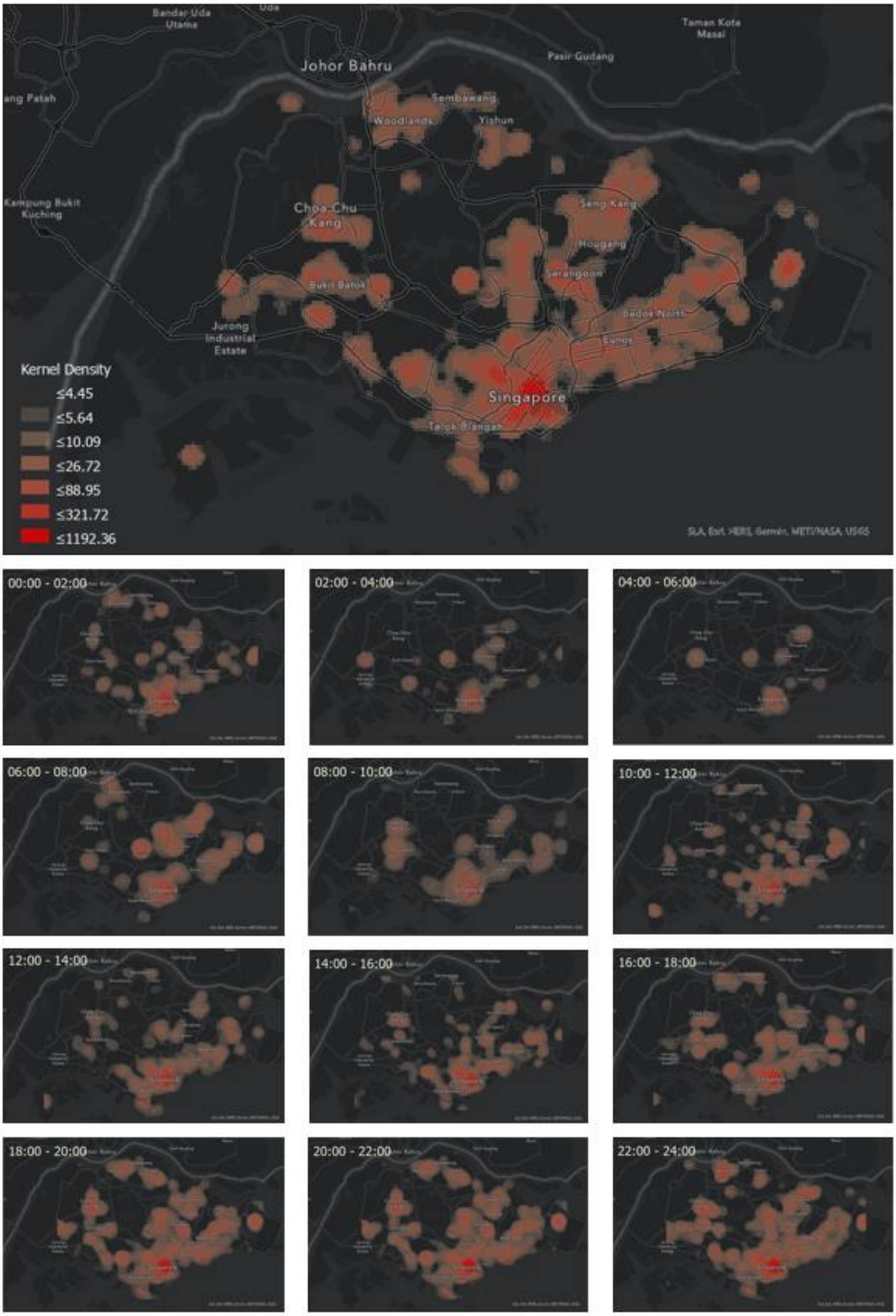


Figure 4.4. Kernel density of tweets during the entire study period (output raster cell size = 200m, using default search radius to ensure at least one neighbour within search distance)

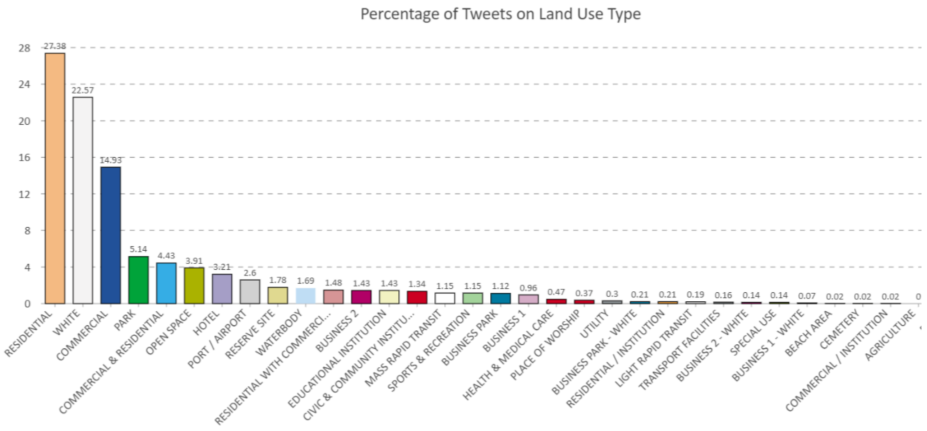


Figure 4.5. Breakdown of land use types with tweet counts over the study period. Tweets posted on Road land use type were tagged to the nearest adjacent land use

4.3 Ordinary Least Squares (OLS) Regression Analysis

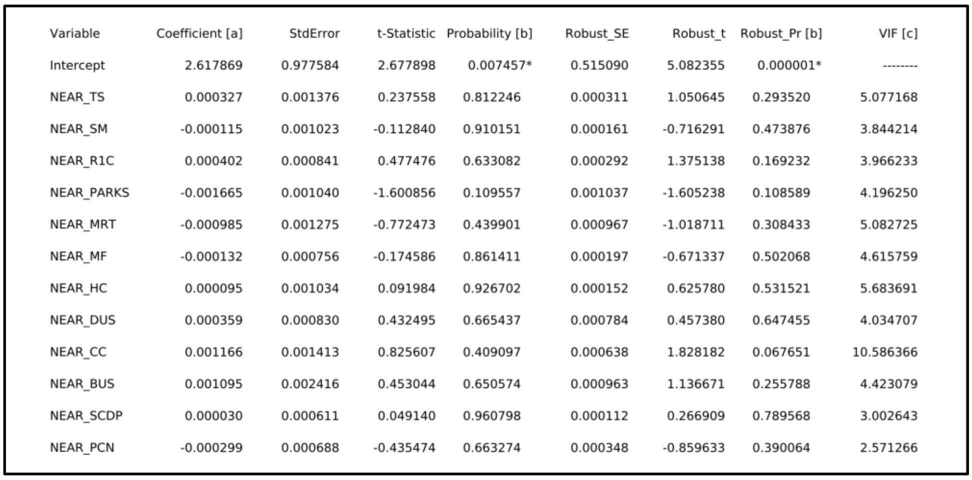
Among the 12 explanatory variables in the OLS regression model, distance to Community Centres was removed for reflecting a high VIF value, and MRT, Hawker Centres, Bus Stops and Medical Facilities were removed for high correlation R-Squared values. The final OLS regression model used the remaining seven explanatory variables: Taxi Stands, Shopping Malls, Residential with 1st Storey Commercial, Parks, Dual Use Scheme (DUS) sports facilities, SCDP Park Malls, and Park Connector Network.

Results from the final model (Figures 4.8 and 4.9) show that it has improved substantially from the first model (Figures 4.6 and 4.7). VIF values for the remaining variables are lower, suggesting that the remaining variables exhibit lower redundancy and multicollinearity, and are suitable explanatory variables. The significance of the Koenker (BP) Statistic has increased, as shown in a reduced probability from 0.95 to 0.71. Hence, we can conclude that the 7 variables surfaced in Figure 4.11 have the greatest influence on the number of tweets recorded. To illustrate further (with reference to Figure 4.7), the relationship between "NEAR_DUS" and the dependent variable is a positive one, where for every 1 metre increase in distance from this facility, the tweet count increases by 0.000877. On the contrary, there is an inverse relationship between "NEAR_PARKS" and tweet count - with a 1 metre increase in distance from a park, the number of tweets fall by 0.001257. Taking the absolute number from the coefficients, the influence of the 7 remaining variables are ranked as follows (Table 4.1):

Ranking	**Variable**	**Coefficient (Absolute)**
1	Park (NEAR_PARKS)	0.001257
2	Dual Use Scheme (NEAR_DUS)	0.000877
3	Residential with 1st Storey Commercial (NEAR_R1C)	0.000613
4	Shopping Mall (NEAR_SM)	0.000301
5	Park Connector Network (NEAR_PCN)	0.000271
6	Taxi Stand (NEAR_TS)	0.000199
7	SCDP Park Malls Public Link (NEAR_SCDP)	0.000045

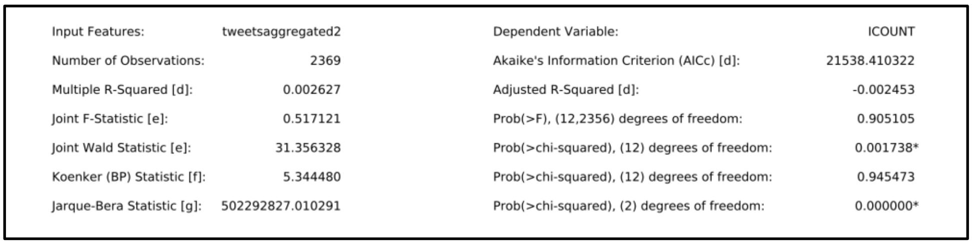
Table 4.1. Ranking of Variables

A similar pattern is exhibited in the following section (Section 4.4) where most of the aforementioned variables are ranked highly in terms of importance, but with slight variation.



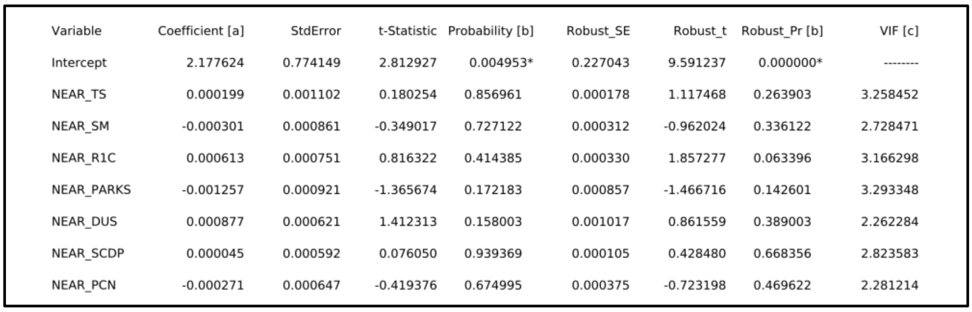
Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	2.617869	0.977584	2.677898	0.007457*	0.515090	5.082355	0.000001*	--------
NEAR_TS	0.000327	0.001376	0.237558	0.812246	0.000311	1.050645	0.293520	5.077168
NEAR_SM	-0.000115	0.001023	-0.112840	0.910151	0.000161	-0.716291	0.473876	3.844214
NEAR_R1C	0.000402	0.000841	0.477476	0.633082	0.000292	1.375138	0.169232	3.966233
NEAR_PARKS	-0.001665	0.001040	-1.600856	0.109557	0.001037	-1.605238	0.108589	4.196250
NEAR_MRT	-0.000985	0.001275	-0.772473	0.439901	0.000967	-1.018711	0.308433	5.082725
NEAR_MF	-0.000132	0.000756	-0.174586	0.861411	0.000197	-0.671337	0.502068	4.615759
NEAR_HC	0.000095	0.001034	0.091984	0.926702	0.000152	0.625780	0.531521	5.683691
NEAR_DUS	0.000359	0.000830	0.432495	0.665437	0.000784	0.457380	0.647455	4.034707
NEAR_CC	0.001166	0.001413	0.825607	0.409097	0.000638	1.828182	0.067651	10.586366
NEAR_BUS	0.001095	0.002416	0.453044	0.650574	0.000963	1.136671	0.255788	4.423079
NEAR_SCDP	0.000030	0.000611	0.049140	0.960798	0.000112	0.266909	0.789568	3.002643
NEAR_PCN	-0.000299	0.000688	-0.435474	0.663274	0.000348	-0.859633	0.390064	2.571266

Figure 4.6. OLS results for first model with all 12 explanatory variables



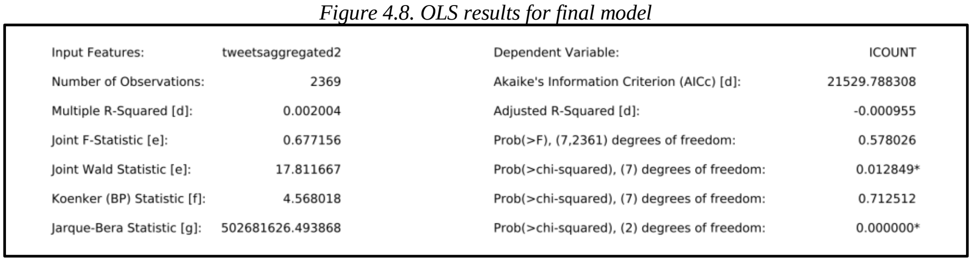
Input Features:	tweetsaggregated2	Dependent Variable:	ICOUNT
Number of Observations:	2369	Akaike's Information Criterion (AICc) [d]:	21538.410322
Multiple R-Squared [d]:	0.002627	Adjusted R-Squared [d]:	-0.002453
Joint F-Statistic [e]:	0.517121	Prob(>F), (12,2356) degrees of freedom:	0.905105
Joint Wald Statistic [e]:	31.356328	Prob(>chi-squared), (12) degrees of freedom:	0.001738*
Koenker (BP) Statistic [f]:	5.344480	Prob(>chi-squared), (12) degrees of freedom:	0.945473
Jarque-Bera Statistic [g]:	502292827.010291	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

Figure 4.7. OLS diagnostics for first OLS model with all 12 explanatory variables.



Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	2.177624	0.774149	2.812927	0.004953*	0.227043	9.591237	0.000000*	--------
NEAR_TS	0.000199	0.001102	0.180254	0.856961	0.000178	1.117468	0.263903	3.258452
NEAR_SM	-0.000301	0.000861	-0.349017	0.727122	0.000312	-0.962024	0.336122	2.728471
NEAR_R1C	0.000613	0.000751	0.816322	0.414385	0.000330	1.857277	0.063396	3.166298
NEAR_PARKS	-0.001257	0.000921	-1.365674	0.172183	0.000857	-1.466716	0.142601	3.293348
NEAR_DUS	0.000877	0.000621	1.412313	0.158003	0.001017	0.861559	0.389003	2.262284
NEAR_SCDP	0.000045	0.000592	0.076050	0.939369	0.000105	0.428480	0.668356	2.823583
NEAR_PCN	-0.000271	0.000647	-0.419376	0.674995	0.000375	-0.723198	0.469622	2.281214

Figure 4.8. OLS results for final model



Input Features:	tweetsaggregated2	Dependent Variable:	ICOUNT
Number of Observations:	2369	Akaike's Information Criterion (AICc) [d]:	21529.788308
Multiple R-Squared [d]:	0.002004	Adjusted R-Squared [d]:	-0.000955
Joint F-Statistic [e]:	0.677156	Prob(>F), (7,2361) degrees of freedom:	0.578026
Joint Wald Statistic [e]:	17.811667	Prob(>chi-squared), (7) degrees of freedom:	0.012849*
Koenker (BP) Statistic [f]:	4.568018	Prob(>chi-squared), (7) degrees of freedom:	0.712512
Jarque-Bera Statistic [g]:	502681626.493868	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

Figure 4.9. OLS diagnostics for final model

Despite showing improved variable suitability, the final model remained weak in performance, as seen in a low Multiple R-Squared and Adjusted R-Squared value (Figure 4.9), suggesting that the model is able to account for a very low percentage of dependent variables. This could be because the model is missing key explanatory variables or that a linear regression may not be a suitable model choice. The low coefficient values for all explanatory variables suggest that the relationships between all of them with the dependent variable (tweet count) are weakly positive or negative. It should also be noted that none of the explanatory variables are associated with a statistically significant coefficient, and a significant probability of Jarque-Bera Statistic suggests that the model is highly biased. From Figure 4.10, we observe that the residues deviate from the line of the best fit and are thus not normally distributed. This goes against one of the classical assumptions of OLS, giving us estimates that are biased with high variances. Additionally, the skewness (normally distributed = 0 vs. dataset = 47) and kurtosis (normally distributed = 3 vs. dataset = 2266) values are far from that of a typical normally distributed dataset.

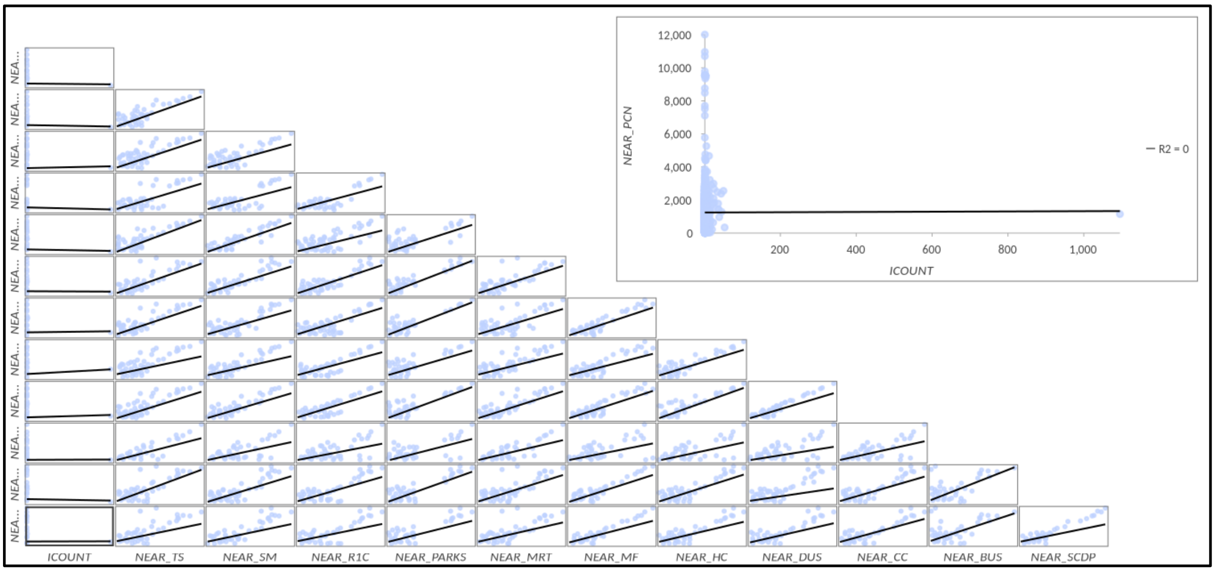


Figure 4.10. Variable Residuals

In this view, non-parametric tests that are free from any underlying assumptions of the dataset distribution could be more suitable for this particular or adjacent studies (e.g. Mann Whitney U Test, Kruskal Wallis H Test). Adopting a log-linear scale in the model may improve the estimates and their significance as well.

4.4 Random Forest Classification and Modelling

The summary of the regression diagnostics results is shown in Table 4.2. R-Squared value of the validation data suggests that the model is able to explain 50.7% of the observed variation, and it is statistically significant (i.e. P-value <0.025). This indicates that the model might still lack key explanatory variables, which will improve the performance of the model. The uncertainty in the input data due to the nature of Volunteered Geographic Information (VGI) data also affects the performance and accuracy of the random forest model.

	Training Data	**Validation Data**
R-Squared	0.868	0.507
P-Value	0.000	0.000
Standard Error	0.005	0.086

Table 4.2. Summary of Random-Forest Model Validation Results

The top 20 variables are shown in Table 4.3. These variables are ranked according to their importance in driving the results of the random forest model. From the results of the top variable importance ranking, it is identified that the location of parks has the most significant influence in terms of driving the model outcome when training with the total tweet count data per 1km cell. It was also noted that the top 5 variables of Parks, SDCP Park and mall public link, Taxi stands, Residential, and Shopping malls makes up more than half of the variable importance. This indicates that these 5 variables are highly influential in training of the random forest model.

Rank	Variable	Importance	Percentage (%)
1	PARKS	157902.84	22
2	SDCP PARK MALLS PUBLIC LINK	104611.88	14
3	TAXI STANDS	75827.32	10
4	RESIDENTIAL	72528.73	10
5	SHOPPING MALLS	54777.42	8
6	RESIDENTIAL WITH COMMERCIAL AT 1ST STOREY (Continuous, Distance Feature)	48148.46	7
7	SCHOOLS	30755.03	4
8	HOTEL	30067.48	4
9	MEDICAL FACILITIES	24600.19	3
10	MRT STATIONS	21550.71	3
11	COMMUNITY CLUBS	21163.32	3
12	PARK CONNECTORS	14511.27	2
13	HAWKER	13645.71	2
14	COMMERCIAL & RESIDENTIAL	13147.71	2
15	RESIDENTIAL WITH COMMERCIAL AT 1ST STOREY (Categorical, Landuse type)	9161.57	1
16	SPORTS & RECREATION	8389.23	1
17	COMMERCIAL	6670.87	1
18	BUS STOPS	6664.18	1
19	PLACE OF WORSHIP	4576.85	1
20	EDUCATIONAL INSTITUTION	3058.23	0

Table 4.3. List of explanatory variables ranked by importance

5.0 Discussion

5.1 Activity Patterns and Contribution from Service Amenities and Land Use Mix

In this study, we used tweet activity locations to gauge physical activity patterns in urban places. We observed distinct diurnal and weekly activity patterns that closely resemble a typical office work temporal profile. Generally, more activities occur outside office hours, and weekend activity increases in count gradually over the day, rather than showing distinct peak activity periods. This also shows that the majority of Twitter users in Singapore could be engaged with typical working-hour schedules, and tend to post more tweets outside of office hours.

Another important observation from the spatiotemporal analysis is that there is no statistically significant spatial cluster, despite showing areas with visually higher tweet activity density. The processes that contribute to the observed varied spatial activity density could be a result of random processes, and further segmentation is required to uncover the mechanism of such processes and their influence on activity patterns. Further studies can be conducted to assess if such patterns could support the hypothesis that the current safe distancing measures, both implicit and explicit, are effective in preventing overcrowding.

Activity patterns in post-CB Singapore could not be adequately explained by distance to service amenities. OLS regression was performed with the assumption that service amenities could attract human activities, but the result suggests poor model accuracy and absence of significant variable coefficients. Hence, distances to service amenities are not strong explanatory factors for activity patterns in post-CB Singapore. The Random Forest Model showed improved model performance by focusing on activities in recreational hours and also accounting for land use mix. The result suggests that several amenities and land use types, such as parks, shopping malls, residential developments and HDB shop houses are among the more important explanatory variables for activity patterns post-CB. It should be noted that residential land is also the top land use type where tweet activities are found, but were not included in the OLS regression model, possibly contributing to the poor OLS regression model performance. The high percentage of tweets from residential land and its relatively strong explanatory factor might be explained by the fact that work-from-home is still the default mode of working during the data collection period. Therefore, locations of residential land might have accounted for a majority of tweet activities post-CB.

5.2 Suitability of Twitter Data for Geospatial Research on Singapore

Prior to this case study, Twitter data was rarely used to study activity patterns in Singapore and there was no evaluation of suitability of Twitter data to study urban place related issues.

Our study shows that Twitter user group is a highly biased representation of the Singapore resident population. Only a small number of users, as compared to the resident population, posted tweets that can be geotagged to a precise location in Singapore. Due to absence of other user details, such as age, race and gender, we are unable to analyse if the Twitter user group constitutes a representative sample of the resident population. However, according to Statista (n.d.), a provider of market and consumer data, Twitter is the sixth most used social media in Singapore, with estimated 1.37 million users in 2020, mostly aged between 25 and 34. This suggests that the tweets collected only represent a small group of young working adults, which might support the temporal patterns that closely resemble a typical office work patterns. Among the Twitter users, each user is unevenly represented, with the majority of users only posting 1 tweet over the whole period of data collection, and a small handful of active users contributes to the a large majority geotagged tweets. This means that the spatial patterns derived from geotagged tweets are contributed by a small number of active users who geotagged many of their tweets.

On the quality of spatial attributes, we found varied levels of spatial precision of tweet location. Only 0.28% of all tweets were geotagged to precise X/Y coordinates and another 0.31% were geotagged to a specific landmark, while over 90% tweets do not have any discernible location attributes. Among the 0.31% tweets that were geotagged to a specific landmark, a large majority could not be directly geocoded with OneMap API, due to erroneous entries. The remaining tweets, despite having place name attributes, are too generic to provide insight on exact location of users and activities.

Tweeting habits may influence the data quality as well. The study relies on voluntarily posted tweets to understand temporal and spatial activity patterns. It is well documented that not all places receive equal coverage on social media, as users may selectively report locations only when they are deemed important to be shared (Sloan and Morgan, 2015).

To prevent heavy API usage from disrupting its network, Twitter has imposed a limit on how many tweets can be queried with its free-of-charge public API. The amount of tweets collected will not exceed 1% of all tweets posted by users. Twitter does not disclose the detailed mechanism to allow 1% of tweets to be accessed via the API. Assuming that a random sampling process is deployed by the API, the tweets collected could similarly be more representative for users who have posted more, than the infrequent users. This limitation can be addressed by using the chargeable premium API from Twitter, which allows access to the full Twitter dataset. Prasetyo et al. (2016) collected one month’s full Tweet dataset with premium API and managed to gather 38,646 Singapore Twitter users who posted geocoded tweets over one month in 2012. Although Prasetyo et al. (2016) used user profile location instead of tweet location to filter the tweets and it was known that user profile location could differ from real-time tweet device location (Graham, Hale and Gaffney, 2014), data from premium API captures much more users and tweets and could present a viable solution for several limitations mentioned above.

Many of the above-mentioned weaknesses of Twitter data are characteristic to VGI big data due to absence of quality control on user-generated content. However, Twitter data still represents a valuable dataset that is accessible to the general public, and contains rich temporal, spatial and semantic attributes that can be further exploited with a suitable research question that adequately addresses data bias and quality issues.

5.3 Other Limitations

Absence of Pre-CB Activity Pattern as a Benchmark

In this study, we presented findings on activity patterns in the context of post-CB re-opening. An understanding of the pre-CB activity patterns and correlations with service amenities will allow a comparison of pre- and post-CB activity patterns, and assess the effectiveness of safe distancing and risk communication measures on physical activity patterns.

Missing Time Periods in Data Collection

The missing blocks of time when tweet collection was unsuccessful pose challenges for an accurate understanding of temporal tweet patterns. In this study, we had attempted to minimise the impact by segmenting the datasets into time periods for separate analysis. However, a full and complete period of data collection would still be beneficial to present an unbiased study on the activity patterns post-CB.

Geocoding

There are three limitations to the geocoding of the tweets data that this study had to consider. Firstly, the data collected in the Place_Name field is not always valid. It may contain invalid tokens, improper place names, spelling mistakes, typographical errors, etc. This will result in an unsuccessful conversion and indicates that these records will be excluded from the geocoding process. Secondly, it was noted that there were many place names with a general location rather than a specific location name (e.g. Central Region). These place names will return a valid coordinate, but are deemed as inaccurate. Lastly, some place names might occur in more than one location. For example, restaurants or supermarkets might have multiple branches throughout Singapore (e.g. Fairprice), therefore the coordinates returned might not be accurate. The geocoding pipeline in this study accounts for the limitations mentioned above.

In studying mobility behaviours using the tweets itself, the risk of skewed data remains with the attachment of fake locations that is not easily detected by GIS tools (Hecht, Hong, Suh and Chi, 2011). The accuracy of the geocoding process is only as accurate as the quality of the input data.

6.0 Conclusion

This study presents an attempt to explore spatiotemporal activity patterns and model the relationship between activity patterns with service amenities in post-CB Singapore using Twitter data. The findings offered preliminary insights into place-visiting patterns in Singapore under the explicit and implicit orders to prevent overcrowding to curb disease spread in Phase 2 of re-opening after Circuit Breaker. We showed that distances to service amenities are not strong explanatory variables for the observed activity patterns in post-CB Singapore. A Random Forest Model accounting for land use mix can predict the activity patterns in off-work recreational hours with up to 50.7% accuracy, but key explanatory variables could still be missing and other processes could be driving the observed activity patterns. At the time of writing, the Singapore government has announced that the city state will not return to pre-COVID situation in Phase 3, but more activities have resumed continually considering low community cases. Longer periods of data collection on people movement will enable an understanding of how place-visiting behaviours respond to the changing regulations. Bearing in mind the limitations of Twitter data, such as lack of quality control and inconsistent spatial precision, the research questions can be narrowed down further so that it can be adequately and accurately answered with such data.

7.0 References

Bruns, A. and Stieglitz, S. (2012). Quantitative Approaches to Comparing Communication Patterns on Twitter. Journal of Technology in Human Services, 30(3-4), 160-185.

Davie, S. (2020). NUS plans to keep students within zones on campus. *The Straits Times*, 24 May 2020. Accessed 20 October 2020 from https://www.straitstimes.com/singapore/education/nus-plans-to-keep-students-within-zones-on-campus.

Gov.sg (3 April 2020). *PM Lee: The COVID-19 Situation in Singapore (3 Apr)*. Accessed 9 Sept 2020 on https://www.gov.sg/article/pm-lee-hsien-loong-on-the-covid-19-situation-in-singapore-3-apr.

Gov.sg (28 May 2020). *Ending circuit breaker: phased approach to resuming activities safely*. Accessed 31 Oct 2020 from https://www.gov.sg/article/ending-circuit-breaker-phased-approach-to-resuming-activities-safely.

Gov.sg (20 Oct 2020). *Roadmap to Phase 3*. Accessed on 31 Oct 2020 from https://www.gov.sg/article/roadmap-to-phase-3.

Graham, M., Hale, S.A. and Gaffney, D. (2014). Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional Geographer*, 66(4), 568-578.

Hecht, B., Hong, L., Suh, B. and Chi, E.H. (2011). Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles. CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 237-246.

Kabir, M.Y. and Madria, S. (2020). CoronaVIS: A Real-time COVID-19 Tweets Data Analyzer and Data Repository.

Kang, C., Sobolevsky, S., Liu, Y. and Ratti, C. (2013). Human Movements in Singapore: A Comparative Analysis Based on Mobile Phone and Taxicab Usages. UrbComp '13: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, 1-8.

Kovacs-Gyori, A., Ristea, A., Kolcsar, R., Resch, B., Crivellari, A., and Blaschke, T. (2018). Beyond Spatial Proximity - Classifying Parks and Their Visitors in London Based on Spatiotemporal and Sentiment Analysis of Twitter Data.International Journal of Geo-Information, 2018, 7, 378.

Lai, J., Lansley, G., Haworth, J., Cheng, T. (2019). A name-led approach to profile urban places based on geotagged Twitter data. Transactions in GIS, 2020 (24), 858-879.

Lee, R. and Sumiya, K. (2010). Measuring Geographical Regularities of Crowd Behaviors for Twitter-based Geo-social Event Detection. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks (LBSN '10). Association for Computing Machinery, New York, NY, USA, 1–10. DOI:https://doi-org.libproxy1.nus.edu.sg/10.1145/1867699.1867701.

Liu, X. and Long, Y. (2016). Automated identification and characterisation of parcels with OpenStreetMaps and points of interest. Environment and Planning B: Planning and Design, 43(2), 341-360.

Marti, P., Serrano-Estrada, L, Nolasco-Cirugeda, A. (2017). Using locative social media and urban cartographies to identify and locate successful urban plazas. *Cities*, 64 (2017), 66-78.

Prasetyo, P.K., Achanauparp, P. and Lim, E. (2016). On Analysing Geotagged Tweets for Location-Based Patterns. ICDCN' 16: Proceedings of the 17th International Conference on Distributed Computing and Networking, 45, 1-6.

Rao, K. V., Govardhan, A., and Rao, K. V. C. (2012) .Spatiotemporal Data Mining: Issues, Tasks and Applications. *International Journal of Computer Science and Engineering Survey*, 3(1), 39 – 52.

Resch, B., Uslander, F., Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, 2018 (45), 4, 362-376.

Sloan, L. and Morgan, J. (2015). Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLoS ONE*, 10(11).

Soliman, A., Soltani, K., Yin, J., Padmanabhan, A. and Wang, S. (2017). Social sensing of urban land use based on analysis of Twitter users' mobility patterns. *PLoS ONE*, 12(7).

Statista (n.d.). Number of Twitter users in Singapore in 2019 and 2020 (in millions)*. Accessed 31 Oct 2020 from https://www.statista.com/statistics/490600/twitter-users-singapore/.

Toh, T. W. (26 April 2020). Food delivery sector booms in a time of coronavirus. *The Straits Times*. Accessed on 31 Oct 2020 from https://www.straitstimes.com/singapore/transport/sector-booms-in-a-time-of-coronavirus.

Wei, Y., and Lan, M. (2015). GIS analysis of depression among Twitter users. Applied Geography, 60 (2015), 217-223.

Yan, Y., Feng, C-C., Huang, W., Fan, H., Wang, Y-C and Zipf, A. (2020). Volunteered geographic information research in the first decade: a narrative review of selected journal articles in GIScience. International Journal of Geographical Information Science.

Zhan, X., Ukkusuri, S.V. and Zhu, F. (2014). Inferring Urban Land Use Using Large-Scale Social Media Check-in Data. *Networks and Spatial Economics*, 14, 647-667.

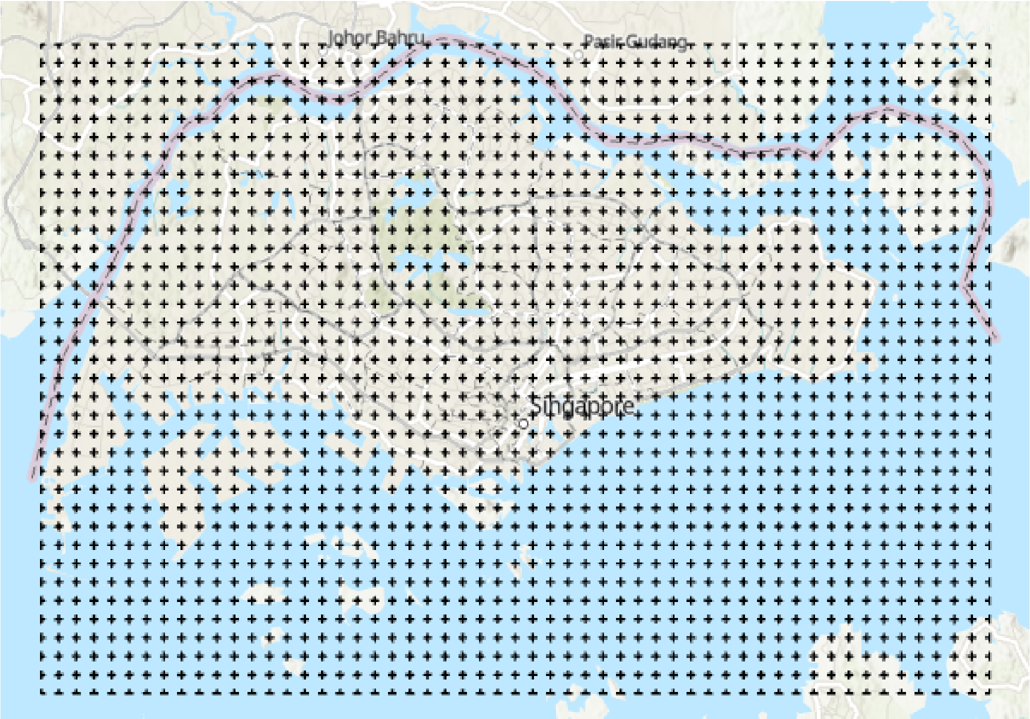
8.0 Appendix

8.1 List of Tweet Attribute Returned From Twitter Standard Search API

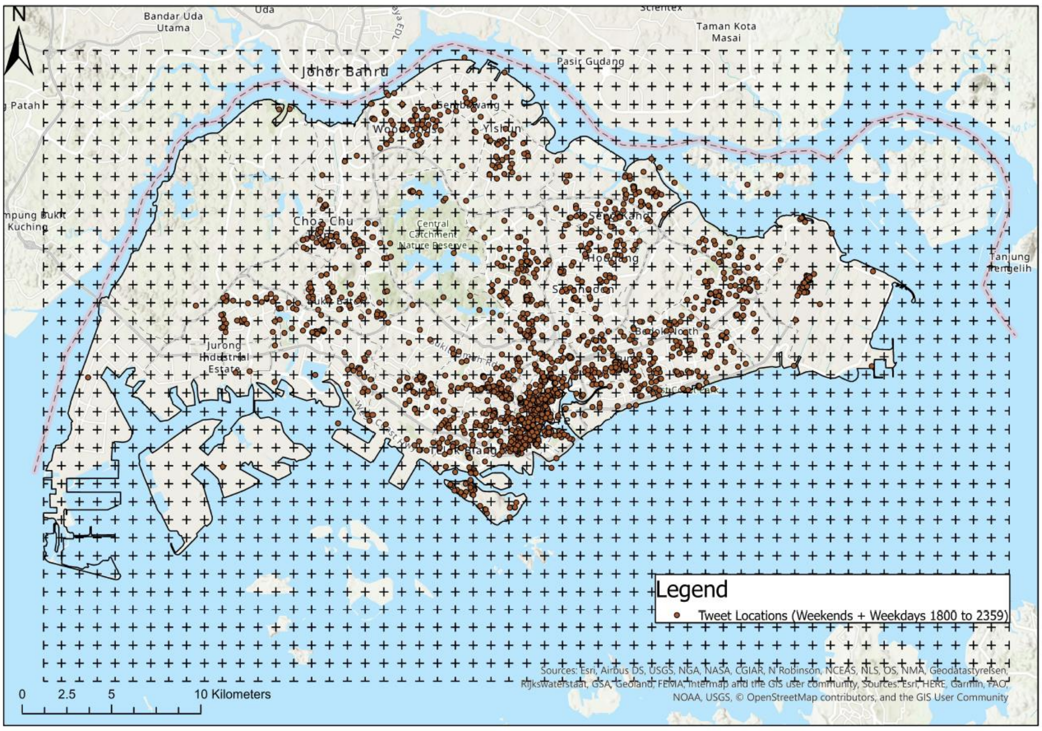
Tweet objects returned from Twitter Standard Search API contain the following attributes:

- Created_at: timing of posting, in UTC time zone
- Id_str: unique ID of tweet
- Truncated: boolean value indicating whether the tweet text has been truncated for the text limit
- Text: Tweet content
- Source: The source application where the tweet is posted
- Source_url: The source service through which the tweet is posted
- User
    - Name
    - Screen_name
    - Location: Location in the user profile
- Coordinates
    - Type: Point or None
    - Coordinates: Latitude/Longitude
- Place
    - Id: a unique ID in Twitter location database
    - Url
    - Place_type
    - Name
    - Full_name
    - Country
    - Bounding_box: if the place is a vague location (e.g. town, neighbourhood), a bounding box will be defined with 4 corner coordinates
        - Coordinates
        - Type
    - Attributes
- Lang: language in which the tweet is posted

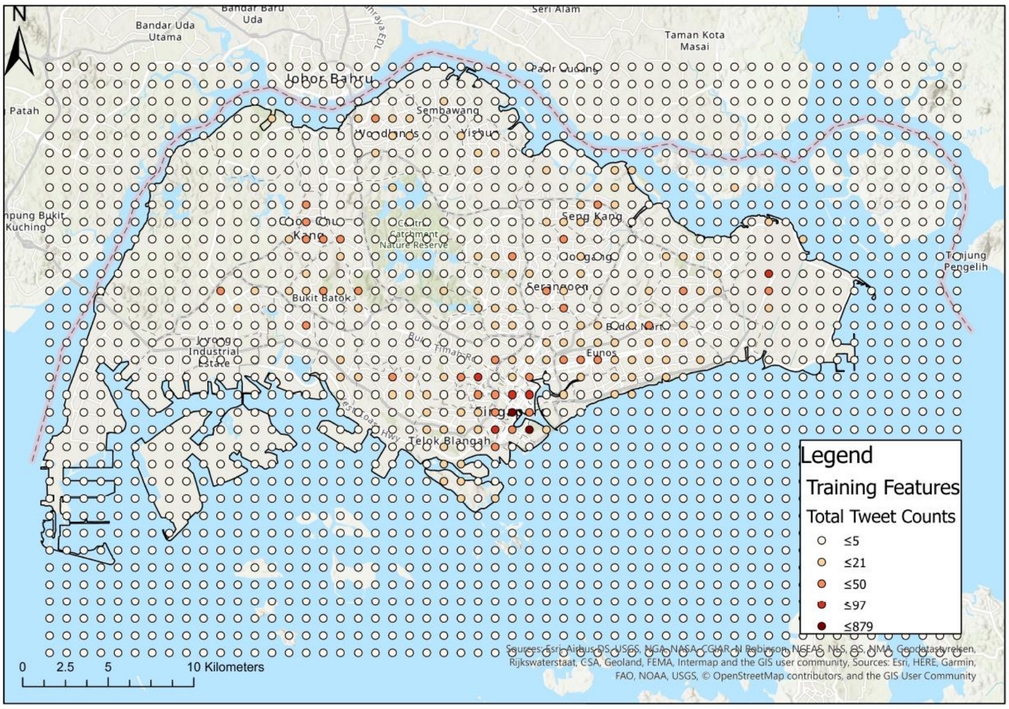
8.2 Map of fishnet grid cells generated for the study area



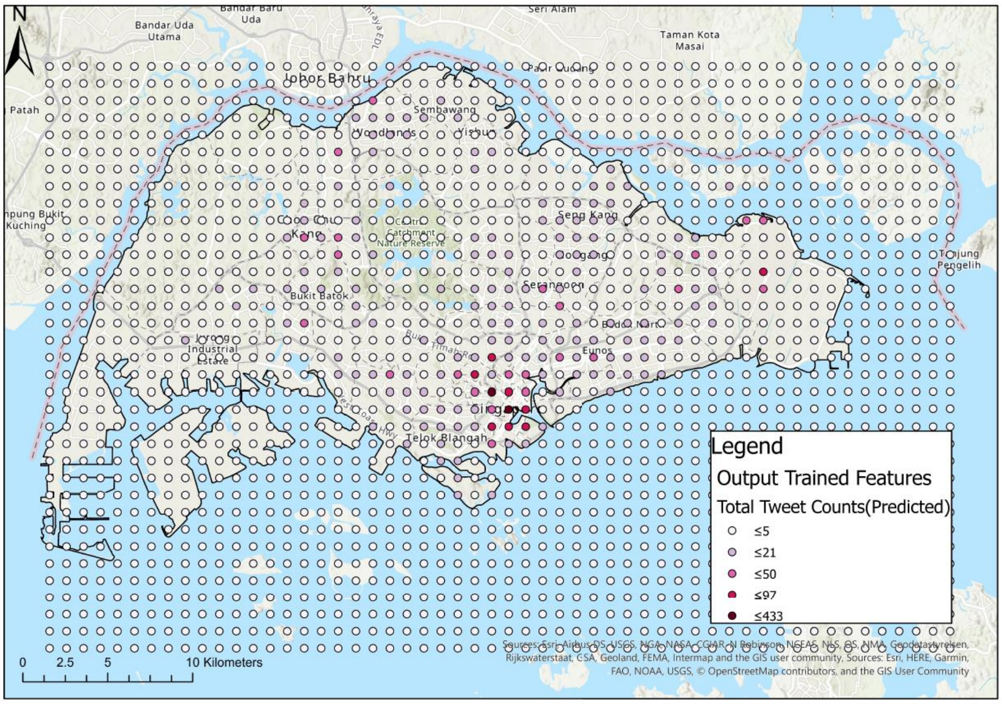
8.3 Map of tweets occurring during weekends and weekdays (1800-2359 Hrs)


N
Bandar Uda Utama
Taman Kota Masai
Johor Bahru
Pasir Gudang
Sembawang
Woodlands
Yishun
Choa Chu
Central Catchment Nature Reserve
Sengkang
Hougang
Serangoon
Jurong Industrial Estate
Tanjung Pengelih
Legend
● Tweet Locations (Weekends + Weekdays 1800 to 2359)
0
2.5
5
10 Kilometers
Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

8.4 Map of total tweet counts for input training feature



8.5 Map of predicted tweet counts based on trained features


N
Bandar Baru Uda
Bandar Uda Utama
Seri Alam
Taman Kota Masai
Johor Bahru
Pasir Gudang
Sembawang
Woodlands
Yishun
Patah
Kampung Bukit Kuching
Choa Chu Kang
Central Catchment Nature Reserve
Seng Kang
Punggol
Tanjung Pengelih
Bukit Batok
Serangoon
Bedok North
Jurong Industrial Estate
Eunos
Singapore
Telok Blangah
Legend
Output Trained Features
Total Tweet Counts(Predicted)
≤5
≤21
≤50
≤97
≤433
0
2.5
5
10 Kilometers
Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community