Analysis of Twitter Usage in London, Paris, and New York City

Muhammad Adnan
Department of Geography,
University College London,
Pearson Building, Gower Street,
London, United Kingdom.
m.adnan@ucl.ac.uk

Paul Longley
Department of Geography,
University College London,
Pearson Building, Gower Street,
London, United Kingdom.
p.longley@ucl.ac.uk

Abstract

The penetration and use of social media services differs from city to city. This paper analyses the use of Twitter in London, Paris and New York. We develop a geographical analysis of Tweeting activity, and of the names, probable ethnicities and genders of Twitter users in these three cities. The results reveal that the majority of Twitter users are male and are of Anglo-Saxon (English speaking) extraction. The analyses also provide an insight into the ethnic diversity of populations in the three cities.

Keywords: Twitter, User generated data, Name and Ethnicity Analysis

1 Introduction

Social media services generate large quantities of data every day. The use of these data is increasing in the research and commercial sectors as recent years have seen an increase in the use of social media in service delivery. Users of the likes of Twitter, Facebook, Flickr, LinkedIn, Bebo, Sino Weibo, and Orkut are frequently mobile users of the developing range of smartphones and tablet devices. These users come from different social and ethnic backgrounds, though it is unclear how representative they are of the population at large.

This paper focuses upon Twitter, a very popular social media service. This social-networking and micro-blogging service was launched in 2006 and within six years had accrued some 200 million active users, who typically send 340 million Tweets every day [9]. This is a huge data source and the analysis of these data can provide useful insights into the places and situations in which users avail themselves of this social media service.

Previous research on Twitter data has emphasised different themes, such as the investigation of Tweeting behaviour [2], the dissemination of information across social networks [1], ethical issues in the analysis of Twitter data [4], tracking community happiness from Tweets [3], and geographic dissection of the Twitter network [6]. However, name, ethnicity, and gender analysis of Twitter usage remains largely unexplored. Previous research has shown that names provide indicators of the cultural, ethnic and linguistic characteristics of individuals ([8]). Naming and ethnicity analysis of georeferenced social media data can provide insight into the geographical structure of different ethnic groups who use a particular social media service. Gender analysis can ascertain whether male users use Twitter more than females or vice versa. Hence, this paper is aimed at analysing the name, ethnicity and gender of Twitter users.

In 2012, London, Paris, and New York City were amongst the top 10 cities with regard to the amount of Tweeting [7]. Hence, this paper undertakes an analysis of name, ethnicity and gender of Twitter users in these three cities. The structure of this paper is as follows.

Section 2 of this paper describes the data used in the analysis. The paper begins by analysing the areas of high Tweeting activity (Section 3), and goes on to analyse hourly and weekly Twitter usage (Section 4). Section 5 introduces a name and ethnicity analysis of Twitter users in the three cities. Section 6 presents a gender analysis of Twitter usage. Finally, conclusions and directions for future research are presented in Section 7.

2 Data

Twitter users produce a huge amount of data every day. Twitter provides the facility for programmers and developers to download live Twitter data using its public streaming API ([10]). The API provides a 1% sample of live geo-tagged Tweets within a user specified bounding rectangle, at any point in time.

For this paper, the Twitter Streaming API was used to download geo-tagged Tweets for London, Paris, and New York City over the period 10 September – 20 December 2012 (102 days). The numbers of geo-tagged Tweets for each city were:

London: 2, 412, 252
Paris: 740, 188
New York City: 646, 053

The data downloaded from the API included the 'User Name', 'Default Profile Language', 'Latitude of the Tweet', 'Longitude of the Tweet', and 'Tweet Message'. For the analysis reported in this paper, we used 'User Name', 'Latitude of the Tweet', and 'Longitude of the Tweet' fields.

In every city, some users send more Tweets than others. The number of unique users who sent Tweets is:

London: 140, 919 users
Paris: 42, 729 users
New York City: 59, 272 users

3   Areas of high Tweeting Activity

Areas of high relative incidence of Tweeting activity shown in Figures 1 – 3 (given at the end of this paper).

The maps were created by using the following procedure:

a)   In the first step, a grid map of 150 rows and 150 columns was created for every city. This resulted in 24, 025 individual grid squares for each grid map.

b)   Using the Twitter data ('latitude of the Tweet' and 'longitude of the Tweet') for every city, a point in polygon operation was performed to count the number of Tweets sent from the area of each grid cell.

c)   Finally, the individual grid cells were given a colour based on the number of Tweets sent from them.

In London (Figure 1: given at the end of this paper), the users in the central part of the city sent more Tweets than the users in the rest of the city. This may be because of the high number of Twitter users travelling to the centre of the city for work, shopping, and visiting purposes. The outskirts of London have low Twitter usage. However, in London, there are few areas where users do not send any Tweets.

In New York City (Figure 2), the Twitter users in certain areas are more active than others. For example, there is high Twitter usage in lower and mid-town Manhattan, West Bronx, and in certain areas of Brooklyn. Tweeting activity in some areas (Queens and Staten Island) is very low.

In Paris (Figure 3), 25 or fewer Tweets were sent from many grid squares. The central area of Paris accounts for more Tweeting activity than rest of the areas. In Paris, like London, there are few areas from where users do not sent any Tweets.

4   Hourly and Daily Twitter Activity

The previous section has discussed the areas of high Tweeting activity in London, Paris, and New York City. In this section we break down this over-all picture by time of day and time in the week.

Figure 4 (given at the end of this paper) shows the hourly Twitter activity in London, Paris, and New York City. This figure was created by aggregating the number of Tweets, by hours of the day, sent during 10 September to 20 December, 2012.
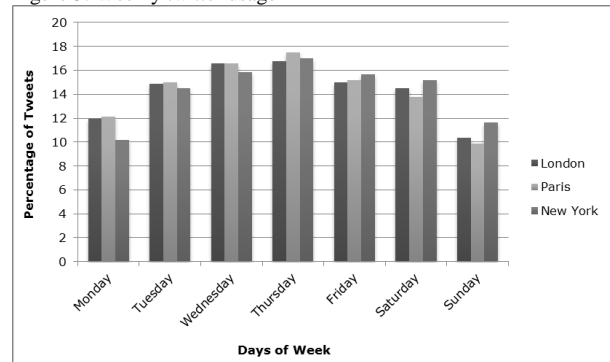
Tweeting activity in London remains low between 2AM to 5AM. It starts to increase from 5AM which is earlier than in the other two cities. Tweeting activity in London reaches its peak between 10AM – 11AM. In the night time, Twitter usage is very high between 7PM – 11PM.

Tweeting activity in Paris remains low between 2AM to 6AM. It starts increasing from 6AM. There is high Tweeting activity between 10AM – 1PM, which is different from London. At night, Twitter usage is very high between 7PM – 11PM.

Similar to Paris, Tweeting activity in New York City remains low between 2AM to 6AM. It starts increasing from 6AM. There is high Tweeting activity during 10AM – 2PM. At night, Twitter usage is very high between 7PM – 11PM.

Figure 5 shows the daily Twitter activity patterns of London, Paris, and New York City. All three cities have higher numbers of Tweets on Wednesday and Thursday. The number of Tweets sent on Sunday is less than the number of Tweets sent on any other day of the week.

Figure 5: Weekly twitter usage

5   Name and Ethnicity Analysis of Twitter Users

A name is a statement of a bearer's cultural, ethnic, and linguistic identity [8]. Knowing the name of the person can thus reveal other useful information about that person. Name analysis of Twitter users can give us an insight into the ethnic identity of individuals, which could be useful in identifying the different types of users who use this social media service.

When registering to use Twitter, users are required to enter their name or other identifying data in the 'User Name' field. In many cases, tokens other than given and family names are entered, as in 'Justins_Home', 'What is Love', etc. However, many registrants fill this field with their real forename-surname pairs.

For the three cities, we developed our own software to divide the 'User Name' field into separate 'forename' and 'surname'

fields. The following table (1) shows the result of our text analytics work, where recognisable forename-surname pairs were found in the 'User Name' field.

Table 1: Result of the text analytics work

City	Total number of users	Number of users where forename-surname pairs found
London	140, 919	106, 917
Paris	42, 729	28, 775
New York City	59, 272	42, 674

The next step describes the Ethnicity Analysis performed on the users where forename-surname pairs were found.
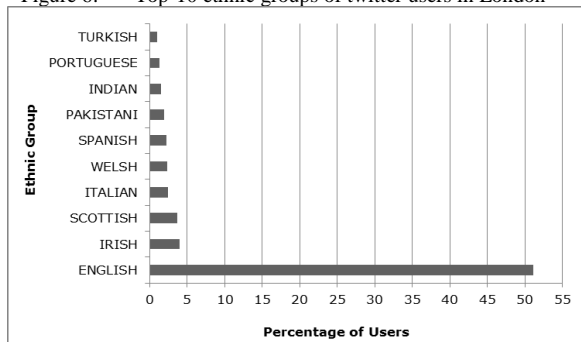
5.1 Ethnicity Analysis

Onomap [8] is software which may be used to assign a predicted ethnic group to a forename-surname pairing. Onomap was run on the forename-surname pairs of Twitter users. The following table shows the number of users for each city, where an ethnic group was successfully assigned to the forename-surname pairs.

Table 2: Ethnicity analysis of twitter users in the three cities

City	Number of Users	Number of users where an ethnic group was found
London	106,917	99, 974
Paris	28, 775	23, 450
New York City	42, 674	39, 110

For London, 69 ethnic groups were found. The following figure (6) shows the top-10 ethnic groups of Twitter users in London.
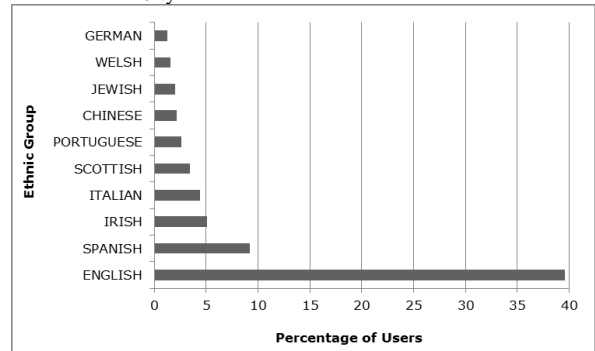
Figure 6: Top-10 ethnic groups of twitter users in London

In London, 'ENGLISH' Twitter users are in very high numbers (51%). The users of second most common group ('IRISH') account for 4% of Tweets. The graph gives us an indication of long established and recent immigrants living in London, as indicated by the considerable number of Twitter users from the 'SPANISH' (2.25%), 'PAKISTANI' (1.9%), 'INDIAN' (1.56%), 'PORTUGUESE' (1.34%), and 'TURKISH' (1.024%) ethnic groups.

For New York City, 67 ethnic groups were found. The following figure (7) shows the top-10 ethnic groups of Twitter users in New York City.
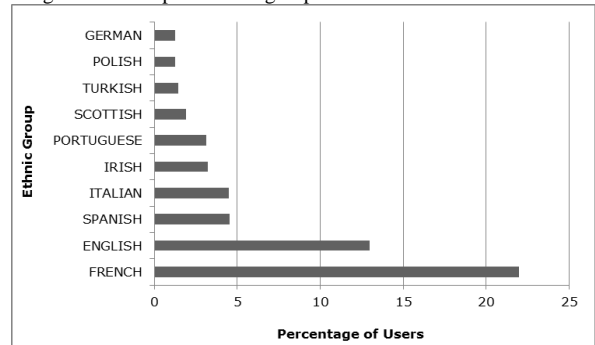
Figure 7: Top-10 ethnic groups of twitter users in New York City

In New York City, the highest number of Twitter users are described as 'ENGLISH' (39%). The above graph provides an indication of prevalence of Tweeting amongst different European ethnic groups in New York. Another interesting ethnic group is 'JEWISH' which is 8th in terms of Twitter usage.

For Paris, 69 ethnic groups were found. The following figure (8) shows the top-10 ethnic groups of Twitter users in Paris.

Figure 8: Top-10 ethnic groups of twitter users in Paris

In Paris, the high number of Twitter users are 'FRENCH' (22%). However, there are also high number of 'ENGLISH' users as well (13%). This could be becuase of a high number of 'ENGLISH' visitors Tweeting in Paris (during the period of 10th September - 20th December, 2012). However, it also shows the dominance of 'ENGLISH' users on Twitter. Other large groups of Twitter users in Paris are 'SPANISH' (4.5%), 'ITALIAN' (4.5%), 'IRISH' (3.2%), and 'PORTUGUESE' (3.1%).

5.2 Top Surnames

This sub-section is aimed at identifying the top-10 surnames of Twitter users and their ethnic groups in London, New York City and Paris. The following tables (3-5) list the top-10 surnames on Twitter in the three cities.

Table 3: Top Surnames of Twitter users in London

Surname	Ethnic Group	# of Users
Smith	English	681
Brown	English	368
Jones	Welsh	362
Taylor	English	336
Williams	Welsh	249
Lee	English	199
Patel	Indian Hindi	190
Johnson	English	189
Harris	English	188

'Smith' is the top surname of Twitter users in London, with 681 users bearing the surname. The top 10 surnames are of 'English' or 'Welsh' ethnic groups, except for the surame 'Patel' which is of 'Indian Hindi' ethnic group.

Table 4: Top Surnames of Twitter users in New York City

Surname	Ethnic Group	# of Users
Rodriguez	Spanish	168
Smith	English	143
Lopez	Spanish	127
Lee	English	113
Johnson	English	113
Garcia	Spanish	111
Perez	Spanish	100
Gonzalez	Spanish	94
Jones	Welsh	92

'Rodriguez' is the top surname of Twitter users in New York City, with 168 users bearing the surname. The top 10 surnames are of 'English' and 'Spanish' ethnic groups, and shows a high number immigrants having Spanish roots.

Table 5: Top Surnames of Twitter users in Paris

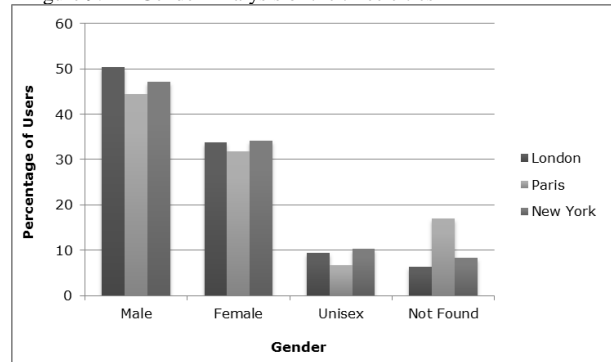| Surname	Ethnic Group	# of Users
Lee	English	42
Smith	English	36
Garcia	Spanish	31
Rodriguez	Spanish	28
Sanchez	Spanish	24
Nicolas	French	23
Martinez	Spanish	22
Lopez	Spanish	20
Gonzalez	Spanish	20

In Paris, the top surnames 'Lee' and 'Smith' are drawn from the 'English' ethnic group. The top Twitter names in Paris have a different ethnic composition in Paris to the population at large, with only one name identified as 'French' and all the others belonging to 'English' and 'Spanish' ethnic groups. This may be because of the high number of 'English' and 'Spanish' visitors visting Paris during the study period, or the core usage groups of the Twitter medium.

6   Gender Analysis

Forename can also be used to determine the gender of a person. GenderChecker [5] is a database which allows checking whether a forename is male, female, or unisex. Forenames of Twitter users, retrieved in section 5, were assigned a gender by using the GenderChecker database. The users were assigned 'Male', 'Female', or 'Unisex' genders. 'Not Found' was assigned when a forename was not found in the database. The result of the analysis is shown in the following figure (9).

Figure 9:   Gender Analysis of the three cities

Following table (6) lists the absolute numbers of different genders used in the analysis.

Table 6: Gender Analysis of the three cities

Gender	London	Paris	New York
Male	53,872	12,779	20,116
Female	36,180	9,175	14,541
Unisex	10,101	1,939	4,435
Not Found	6,764	4,882	3,582
Total Users	**106,917**	**28,775**	**42,674**

The graph, in figure 9, shows that almost 50% of the users are Male, and 33% of the users are Female in the three cities. This shows a dominance of Male users on Twitter. In all the three cities, less than 10% of users were assigned to the 'Unisex' category.

7   Conclusion and Future work

Social media services generate large quantities of user generated content, and the exploitation of this resource is increasing in the research and commercial sectors. The use of the Twitter social media services varies between different cities and countries. This paper has presented a preliminary descriptive analysis of Tweeting activity, specifically according to location, name, ethnicity and gender of Twitter users in the cities of London, Paris, and New York City. The analysis suggests that the majority of Twitter users are male users and have Anglo-Saxon roots (i.e. English speaking). The analysis also provides an insight into the ethnic diversity of population in the three cities. However, it must be remembered that the sources and operation of bias in our (very large) dataset are unknown: by no means everybody Tweets, and there is no a priori reason to assume that

Tweeters who disclose their locations are representative of those that do not.

This is preliminary research in a promising research area and could be extended in different ways. One possible extension is a detailed geo-spatial analysis of different ethnic groups in the three cities by the identification of their residential and work place areas. The second possible extension is to conduct a country wide analysis of Twitter users in United Kingdom, United States of America, and France. This kind of analysis can give us a detailed insight into the identity of Twitter users and how they use this particular social media service.

8   Acknowledgements

References

[1] Antonio Lima and Mirco Musolesi. Spatial Dissemination Metrics for Location-Based Social Networks. In *Proceedings of the UbiComp'12*. 2012.

[2] Bernardo A. Huberman and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, Volume 14, Number 1. 2009.

[3] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. Tracking "Goss Community Happiness" from Tweets. In *Proceedings of the ACM CSCW*, 2012. 2012.

[4] Fabian Neuhaus and Timothy Webmoor. Agile Ethics for Massified Research and Visualization. *Information, Communication & Society*, DOI: 10. 1080/1369118X.2011.616519. 2011.

[5] GenderChecker. GenderChecker. Retrieved 22nd January, 2012, from http://genderchecker.com. 2012.

[6] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P. Gummadi. Geographic Dissection of the Twitter Network. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. 2012.

[7] Mediabistro. Revealed: The Top 20 Countries and Cities of Twitter [STATS]. Retrieved 31st December, 2012, from http://www.mediabistro.com/alltwitter/twitter-top-countries_b26726. 2012.

[8] Pablo Mateos, Paul A. Longley, and David O'Sullivan. Ethnicity and population structure in personal naming networks. *PLoS (Public Library of Science)*, 6(9) e22943, 1-12. 2011.

[9] Twitter. What is Twitter?. Retrieved 31st December, 2012, from https://business.twitter.com/basics/what-is-twitter/. 2012.

[10] Twitter. The Streaming APIs. Retrieved 22nd January, 2012, from https://dev.twitter.com/docs/streaming-apis. 2012.

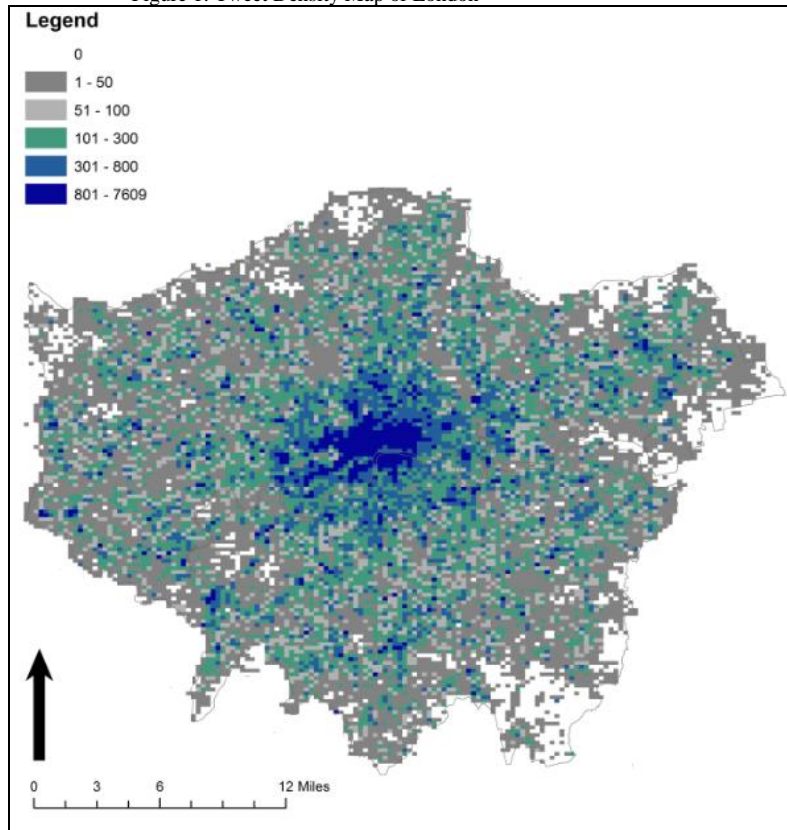Figure 1: Tweet Density Map of London

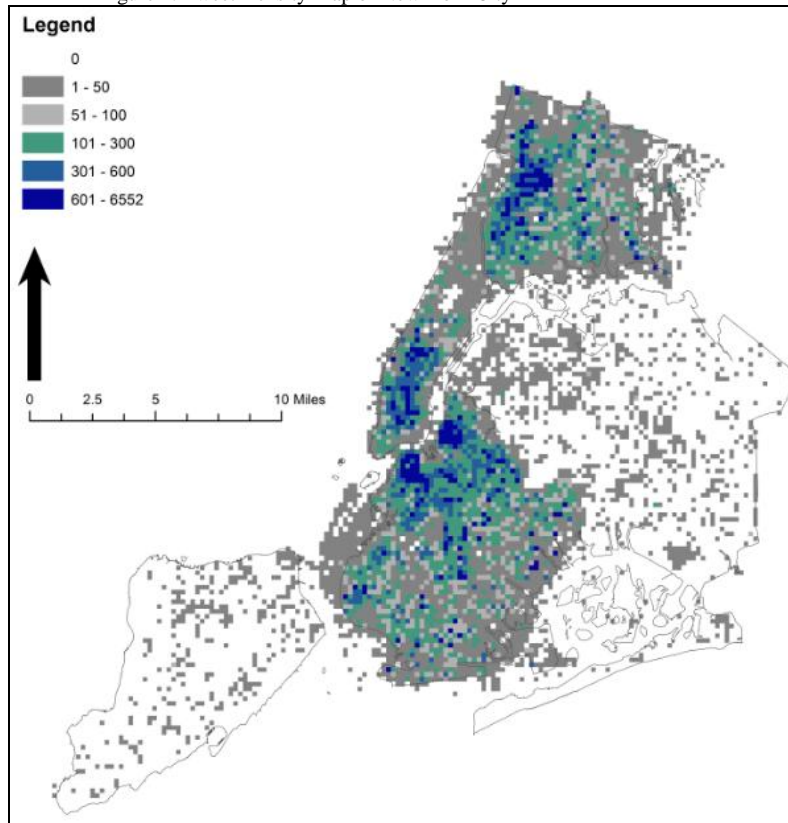Figure 2: Tweet Density Map of New York City
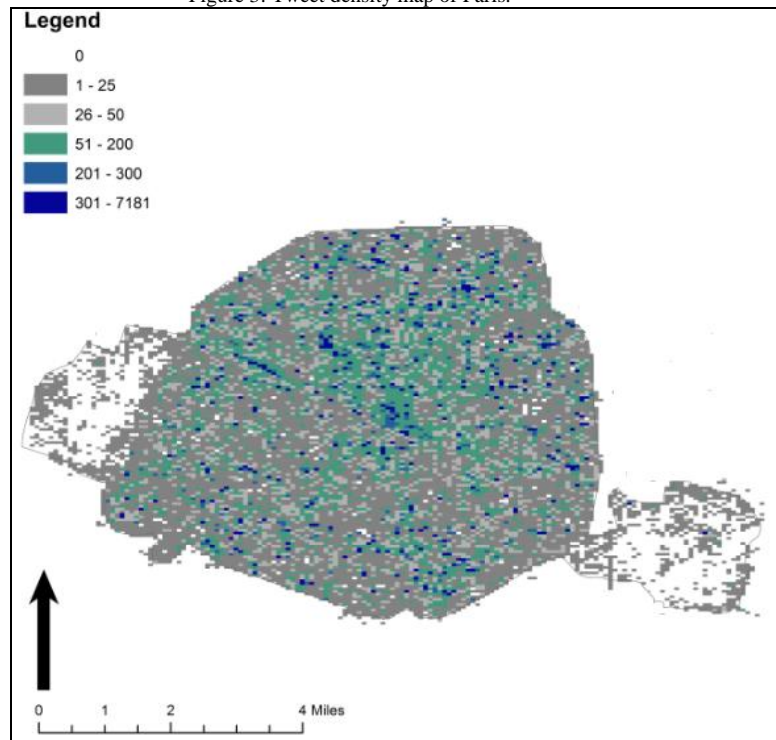
Figure 3: Tweet density map of Paris.

Figure 4: Hourly Twitter usage in London, Paris and New York City