City-wide mobility mapping using social media communications.

Technical Report · May 2016

DOI: 10.13140/RG.2.2.22992.61447

Antonio Remiro-Azócar
University College London

City-wide Mobility Mapping Using Social Media Communications

Research · July 2016

DOI: 10.13140/RG.2.1.3499.4161

Antonio Remiro-Azócar
University College London

City-Wide Mobility Mapping Using Social Media Communications

Philip Adams
Antonio Remiro-Azócar

Department of Physics, University of Bath, BA2 7AY
Industrial Supervisor: Becky Hayward, Buro Happold
Academic Supervisor: Frances Laughton

Report submitted to Buro Happold

Acknowledgements

We would like to express our gratitude to Buro Happold Engineering for their training and guidance. We are particularly indebted to our industrial supervisor Becky Hayward and Buro Happold research associate David Greenwood; both provided invaluable insight and co-operation. Our thanks and appreciations also go to our academic supervisor, Frances Laughton, for her attention, time and encouragement.

Abstract

In recent years, social media websites have become ubiquitous, generating huge amounts of data every day. This report presents a review of the current state-of-the-art in human mobility modelling in cities, and the use of social media data in analysing human movement in urban areas. A reproducible approach to collecting geolocated Twitter data is developed and analysed. Furthermore, an application is presented to apply clustering algorithm DBSCAN to the "tweets" collected. Topic Discovery algorithm Latent Dirichlet Allocation is utilised to infer meaning from "tweets", and Index of Dissimilarity calculations are performed successfully on the collected messages. In addition, Uber and New York taxi journey data has been sourced and analysed. The future potential of all the methods applied is examined, as is their value to Buro Happold Engineering.

Contents

1 Introduction

The four main objectives of this project are defined as follows:

1. To provide a comprehensive literature review of the current state-of-the-art in human mobility in cities, and research into how such models and publically available data sources are utilised.

2. To develop a reproducible approach for data gathering from useful data sources identified in the literature review to yield insights into human mobility and utilisation of urban space.

3. Analysis and visualisation of collected data with respect to identified performance indicators, parameters of human mobility and utilisation of urban space.

4. Validation exercise to compare and evaluate various models of human mobility and utilisation of urban space using the datasets collected.

Buro Happold Engineering is currently leading the "Make Way for Lower Manhattan" initiative, to alleviate congestion and mobility issues in the Lower Manhattan area of New York City. These issues are caused by the significant growth in the area: since the year 2000, tourism and residential numbers in the area have doubled, the number of hotel rooms has increased fourfold, and many retailers are moving into the area [1]. This has put an increased stress on the physically constrained grid system, and has led to congestion in many parts of Lower Manhattan, which the initiative aims to address. Because of this ongoing project, much of the analysis performed in this report is centred on New York City and Lower Manhattan. Nonetheless, a key aspect of the analytical techniques used is that they are reproducible and applicable to any major city.

2 Literature Review

An initial literature review presented to Buro Happold on October 2015 included the identification of mobility modelling methodologies and the evaluation of publically available data sources for their potential to improve understanding of human mobility in urban spaces. This section will give an extended review on current-century mobility modelling methodologies which have been found of interest.

2.1 Modern Modelling Methodologies

Several theories approximating human mobility with random walk or diffusion models have been introduced in the current century[2,3]. This is understandable given the vast amount of unknown factors influencing our movement. In fact, previous research on albatrosses[4], and most recently on marine predators[5] and monkeys[6], indicates that animal mobility can resemble a random walk[7]. In particular that it can be approximated by a Lévy flight[8,9]. A Lévy flight is a Markov (memory-less) process; a random walk in which steps are defined in terms of the step-lengths Δr, these giving a power-law probability distribution

$$P(\Delta r) \propto \Delta r^{-\beta}. \tag{1}$$

where $1 < \beta < 3$ is a displacement exponent. Brockmann et al. (2006)[2] creatively extended this approximation to humans using the dispersal of marked bank notes as a proxy for human mobility. The trajectories of 464000 U.S. dollar notes were tracked over a million individual displacements, obtained through a purpose built website. Brockmann at al. found that the distribution of travelling distances (in this case $\Delta r = |x_2 - x_1|$, where x_1 and x_2 are the successive report locations of a bank note) also followed a power-law distribution and was consistent with (1). The displacement exponent was measured as $\beta = 1.59 \pm 0.02$ (mean ± standard deviation). It is worth noting that particles following a Lévy flight have a significant probability of travelling very long steps. This is in agreement with observed human behaviour. Nowadays, humans travel on many different spatial scales. We tend to travel shorter distances most of the time, e.g. to the office for work, but ocassionally move much longer distances in a relatively short period of time e.g. a flight to South America.

In recent times, cellular technology, the web and social media have allowed for better empirical validation of mobility models. It is worth noticing that in 2006, mobile phone data was not yet available for scientific research due to privacy concerns. Brockmann et al. impressively manage to collect a dataset of high spatio-temporal granularity despite this. However, their bank notes cannot truly reflect the motion of individual users. Each "appearance" of a bill signifies an interaction between multiple individuals and acts as some unknown convolution of multiple trajectories[10].

In the current century, people carry mobile phones throughout most of their daily routine. Consequently, the motion of such phone should provide an optimum proxy for human mobility and obey similar statistical laws. González et al. make use of this assumption when publishing the first large scale study of mobile phone users in 2008[10]. In their groundbreaking paper, the trajectories of 100,000 anonymised individuals are tracked for a six-month period. The study suggests a similar statistical relationship than that described by (1); this time approximated by a truncated power-law with an exponential cutoff:

$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} e^{\frac{-\Delta r}{\kappa}}, \tag{2}$$

Here Δr represents the distance between an individual's position at consecutive calls; Δr_0 and κ are dataset specific parameters. In this case, the displacement exponent is measured as $\beta = 1.75 \pm 0.15$, $\Delta r_0 = 1.5$km and $\kappa = 400$km. González et al. also explore the temporal and spatial regularity of users' trajectories. They do so by ranking each visited location on the basis of the number of times an individual has been recorded in its vicinity. For example, a site with rank $L = 2$ corresponds to the second most visited locale for a specific user. It is established that the the probability of finding a selected user at a given locale scales with the reciprocal of its rank ($P(L) \propto 1/L$), and is independent of the total number of locations frequented by the user. González et al. argue that individuals display significant regularity in their mobility; a regularity not shown by the bank notes in [2].

A significant amount of current research[10-13] embraces the spatial and temporal regularity of human mobility and characterises it with the three following measures:

- The **trip distance distribution** $P(r)$, explained previously.
- An individual's **radius of gyration** r_g: the characteristic distance an individual travels during a given time period. It is expected to reveal the diversity in different people's

mobility[10,11]. Some individuals travel within a short r_g but others cover regularly longer trajectories. Each person will follow a trip distance distribution with a characteristic radius of gyration.

- The **number of visited locations** S(t) over time. Song et al. (2010)[14] measure S(t) to vary sublinearly with time ($S(t) \propto t^{\mu}$ with $\mu \approx 0.6$), illustrating individuals' tendency to revisit locations.

Previous methods used by scientists to study mobility involved collecting data through population surveys. These methods were costly, time-consuming and provided limited temporal granularity. The advent of advanced sensor technology and mobile phones in the 1990s allowed researchers to track movement with per second accuracy and relatively high geographic precision. The geographic coordinates of an individual could now be recorded at the nearest Base Transciever Station when he made a call or sent an SMS. Mobile phones opened the door for mobility mapping at massive population scales. However, privacy concerns from policy makers and economic concerns from communications providers were raised regarding the use of Call Detail Records (CDRs) for research. Additionally, sensors methods could only be utilised on a limited number of people and for relatively short periods of time. In 2009, mobile phone data became more readily available for research[10,11]. Song et al. (2010)[15] find a 93% potential predictability in user mobility across their whole user base CDRs. However, they highlight that the spatial granularity of CDRs is improvable, being accurate only up to a few hundred metres. Much more accurate and ubiquitous datasets emerged with the advent of social media.

Since the rise of social networking websites in the mid-2000s, websites such as Facebook, Twitter, Instagram and Tumblr have become increasingly ubiquitous, and their growth shows no signs of slowing down. The increased popularity of social networks was coupled with the rise of the smart-phone, which changed the way users could access the Internet. For the first time, users could access the Internet remotely from anywhere in the world. This has led very quickly to a wealth of geographically-located data. Publicly available social media data can deliver great insight into how cities are structured, populated and used, and can inform many decisions for businesses such as Buro Happold.

3 Data Sources and Collection

3.1 Twitter

Twitter is a social networking site which enables users to send and read short 140-character messages called "tweets". It is an interesting source of data for a variety of reasons:

1. **Popularity**: Within ten years, Twitter has accrued more than 500 million users, out of which more than 320 million are active[16]. Additionally, New York City has a comparatively high median household income[17] (\$58,003) and a high percentage of New Yorkers use smartphones[18] (63%). Since the spatial distribution of Twitter users is uneven, biased toward users with higher incomes[18], New York is an ideal location for exploring its data.

2. **Fast communication**: Twitter has been described in mainstream media as the "SMS of the Internet"[18]. The site puts emphasis on speed and ease of publication and its users "tweet" frequently, on average 2.02 times a day[19] as of 13 March 2016.

3. **Geo-location**: The site gives its users the option of adding their exact location information to publications. Reports have shown that about six percent of users opt-in to "geo-locate" their tweets[20]. A sample of geo-tagged tweets from "high-frequency" users can provide datasets of high temporal and spatial granularity[21-23]; a continuously-updated stream of information about human mobility.

4. **Openness in sharing data**: In the scientific community, the social network is known for openness in sharing its data[21,23]. Most competing networking sites restrict access to their information; Twitter's policy is the opposite. It provides a free Streaming API (Application Programming Interface) that allows access to at most 1% of all the tweets produced at a given time. Throughout this report, it is assumed that this sample is a is a valid representation of overall activity. The Twitter API has been used in research topics as diverse as earthquake occurrence prediction[24], and tracking dental pain[25] and post-traumatic stress disorder[26].

5. **Information network**: Twitter's founder, Jack Dorsey, views the site more as an information network than a social network[27]. In fact, Twitter has played an important role in recent socio-political events[28]. The site's emphasis on information makes message content a massive data source for text mining[22,23] and sentiment analysis[21] applications.

3.2 Twitter Data Collection

Twitter allows public access to their Streaming Application Programming Interface (API)[29], which allows a user's application to access the real-time stream of all publically available Tweets, filtered by user-defined criteria, such as language, usernames, keywords or geographical location. The only limitation on the streaming API is that the number of Tweets available at any one time cannot exceed 1% of the total Twitter feed. If the user-defined filters return more than 1% of the total Twitter feed, the API will return a sample of these Tweets.

3.2.1 The Streaming API

The Streaming API retrieves Twitter data continuously once provided with an authenticated request. The authentication of requests is executed using the open standard OAuth, employed by Twitter to give third-party access to user information. API access is only permitted to "applications" (also called "consumers") which have been registered with Twitter. The process of registering an application at `http://dev.twitter.com` is straightforward; one is issued a "consumer key", a "consumer secret", an "access token" and an "access secret". All four user credentials never change for an application. They are used to authenticate the request and issue API calls on behalf of the application owner. The API provides the following data for every geo-located publication:

- Tweet location in latitude and longitude coordinates.

- Date, time and time zone of the tweet.

- User information: name, Twitter-specific username and ID, number of followers and tweets, place of residence, account-specified language, profile picture URL...
- Tweet content and language.

A program was developed in programming language Python to fetch tweets using the Streaming API. Twitter data is returned in JSON (JavaScript Object Notation) format. Another program was later developed to delimit relevant parameters (specifically username, ID, tweet date and time, location coordinates and tweet content) and parse the JSON data into a spreadsheet for analysis.

To allow for continuous data collection over a period of several months, the Python application was run on a Raspberry Pi: a small, Linux-based single-board computer. This influenced the choice to use Python as the primary programming language for this project, as the Raspberry Pi's Linux-based operating system, "Raspbian", is set up primarily to run Python as its main programming language. To account for some inconsistencies in the collected data (due to loss of connectivity, authentication time-outs etc.), and to extend the period for which data was available, the collected data was combined with a dataset previously collected by Buro Happold. This meant that the total New York City dataset consisted of around 2.1 million Tweets collected between the 9th of September 2015 and the 1st of March 2016. The shaded section in Figure 1 shows the area for which geo-located Tweets were collected by Buro Happold research associate David Greenwood. Such area spans over most of the borough of Manhattan and some areas of Brooklyn, Queens and Hudson County, New Jersey. The Tweets collected using the Raspberry Pi span a larger area including The Bronx and Staten Island.

Figure 2 shows two 3000 Tweet samples, from both a typical mid-week (Wednesday) afternoon and a Friday night in December 2015. These are plotted on a map of New York, to show how the distribution of people shifts throughout the day. As well as the spatial distribution, the density of Tweets also varies temporally as shown in Figure 3, where the average number of Tweets during each 15-minute interval of the day is plotted.

3.2.2 Data Filtering

The first step in analysing the collected data was to filter out unwanted Tweets. The majority of unwanted Tweets were Tweets sent by "bots"; where an account is set up to automatically post updates. Examples of "bots" are news websites, travel news accounts or job websites which post regular updates. To remove these accounts, the data was filtered by the "Tweet Source" variable. This variable provides information on how the Tweet has been sent; whether through a Twitter app on a mobile device, from the Twitter website, or through a third-party website or application. Almost all of the "bot" accounts were being updated via third-party websites, so by reducing the dataset exclusively to Tweets that had been sent via mobile devices, these "bots" were filtered out. This step was also carried out because geolocated tweets posted through websites are located by the user selecting their location. On the other hand, geolocated tweets posted from mobile devices use the device's GPS settings to find the location, which is generally accurate to a scale of around 10 metres.

The remaining "bots" were filtered out using a Python application which sent a "GET" request to the Twitter API for each user in the dataset, returning each user's 200 most recent Tweets. In cases where the location of all of these Tweets was the same, the user was flagged as a "bot" and

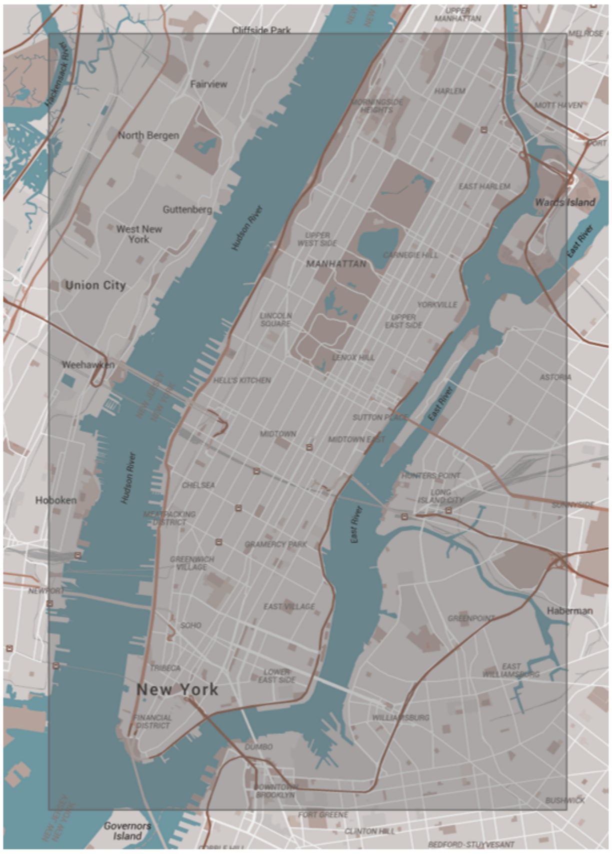


Figure 1: Map of New York City, with the shaded section showing the area for which geo-located tweets were collected by Buro Happold research associate David Greenwood. David Greenwood's dataset contains approximately 1.8 million tweets made during the period 11 September 2015 to 1 March 2016 by 204,297 individual users. All maps in this report have been produced in JavaScript using the Google Maps API.

removed from the dataset. An example of the type of accounts this method removed is weather station accounts, which post regular updates via a mobile device. These filtering techniques removed approximately 19% of Tweets in the collected data for New York City.

3.3 Uber and Taxi Data

Uber is a smartphone application which allows users to submit a trip request, then directed to Uber drivers who use their own vehicles. The application has championed the so-called "sharing-economy"[30] and upended the global taxi-cab industry[30,31]. As of 15 April 2016, Uber is available in over 400 major cities and its presence in New York is ubiquitous[31]. The nature of the platform as a mobility-providing service makes it particularly useful for the purposes of this project. The evolution of Uber usage in NYC can be compared with that of the city's official cabs.

Data collection. Uber does not tend to release its data publicly. Opportunely, two datasets

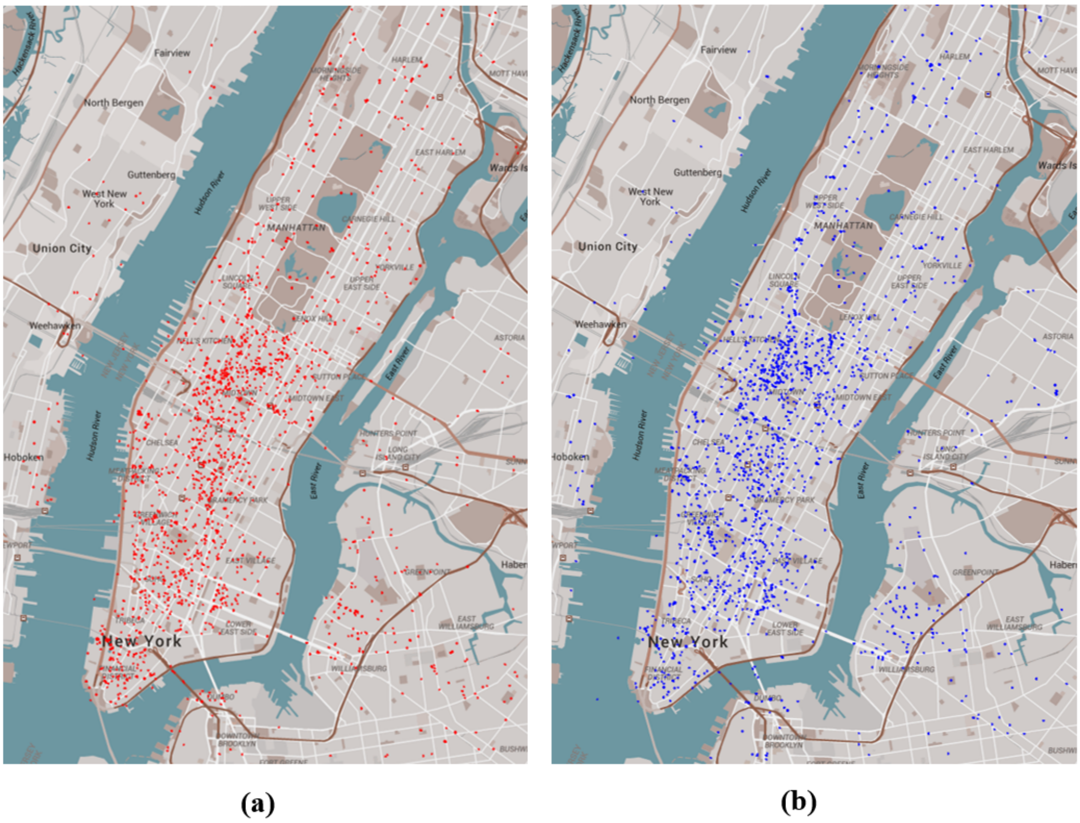


Figure 2: Shows the difference in spatial distribution of 3000 Tweets across New York City for (a) a typical mid-week afternoon, and (b) a typical Friday night.

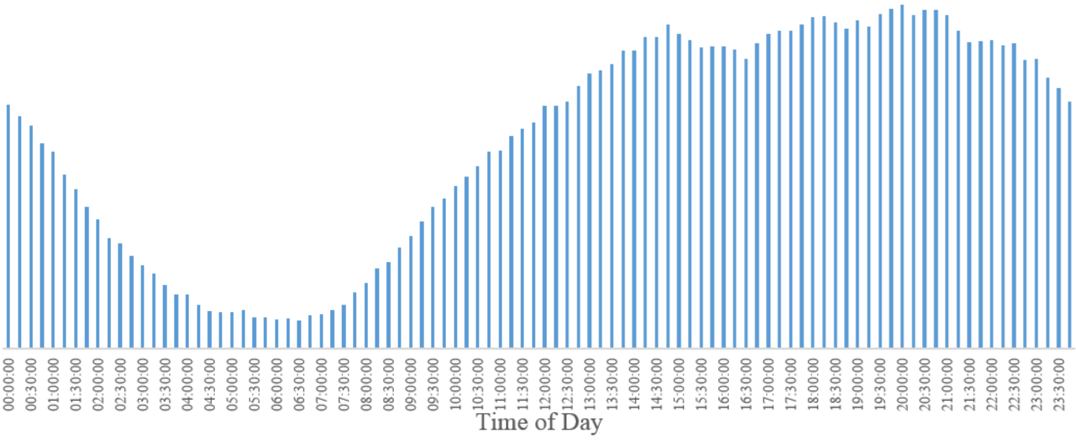


Figure 3: Shows the average temporal distribution of Tweets in New York City, in 15 minute intervals.

were downloaded from blog FiveThirtyEight[a]; these were made available to the site via two Freedom of Information Law requests to the NYC Taxi & Limousine Commission (TLC)[b] . The datasets span the periods April-September 2014 and January-June 2015. The data provided is fairly limited and only specifies the date, time and location of a "pick-up". The datasets collected contain pick-up information for 4,534,327 rides for April-September 2014 and 14,270,489 rides for January-June 2015. The increase in pick-ups between both six-month periods should come as no surprise; Uber expanded aggressively towards the end of 2014[32].

The TLC releases publicly data for official taxis; New York taxis can either be green or yellow. The green "boro" programme was introduced in August 2013 to service the outer boroughs of the city (Brooklyn, Queens, The Bronx, Staten Island) and Upper Manhattan, which had significantly lower access to taxi rides[32]. Green cabs can only pick up passengers in the aforementioned areas. On the contrary, yellow "medallion" cabs have traditionally serviced New Yorkers and are authorised to pick up passengers anywhere in the city. The TLC datasets for official taxis are extremely detailed. They are made up of precise location coordinates and timestamps for both pick-ups and drop-offs for over 402 million yellow taxi trips from April 2013 to December 2015 and 33.4 million green taxi journeys from August 2013 to December 2015. Other information, like distance travelled, number of passengers, fare amount, and even payment method and tip amount is included.

3.4 Location-Based Social Networks

In addition to social media websites such as Twitter, where a user can choose whether or not to share their location, there has also been an increase in popularity of location-based social networks (LBSNs), such as Foursquare, Gowalla and BrightKite. LBSNs are distinguished from other social networking websites because their primary function is location-based. The most popular of these networks is Foursquare, a search-and-discovery mobile application where users can "check-in" at venues (and share this with their friends), search for new locations, and receive recommendations for new places to visit.

The most readily available source of LBSN data is Foursquare and there have been many studies of Foursquare data[33,34]. LBSN data is more easily analysed, in the context of mobility, than Twitter data for several reasons. Firstly, once users check-in at a specific place, they can definitively be grouped together as being at the same place. This is much harder with geo-located data from Twitter. For example, it is hard to tell the difference between someone tweeting from a certain location and someone who happened to post a Tweet as they were walking past. With LBSN data it is clear that the user was at the specified location. Furthermore, with a LBSN such as Foursquare, it is also very easy to classify what a user was actually doing at their location; Foursquare venues are categorised by activity. Foursquare venues' activites could potentially be used as an input in models such as the PWO model, developed by Yan et al. (2014)[13] .

The most common areas of research in this field fall under two categories: POI (point of interest) recommendation and "next check-in" prediction. Recent work towards POI recommendation[33] has developed a framework for recommending POIs based on a user's previous behaviour, which is independent of location, i.e. the recommendation framework will work for a user travelling to a

[a] https://fivethirtyeight.com

[b] www.nyc.gov/tlcresearch

new city. Due to the nature of POI recommendation, it is very difficult to quantify how accurate or reliable these recommendations actually are. The "next check-in" problem has also been addressed in recent studies, and several predictive methodologies have been tested for their accuracy in predicting a user's next check-in. It is found that currently, the most reliable technique is to use M5 Model trees based on factors such as previous visits, physical distance and time of day[34].

4 Data Analysis

Analysis of the collected Twitter data is focused on two different aspects of it. The first is location information, commonly explored through clustering algorithms like DBSCAN (Density-Based Spatial Clustering of Applications with Noise). The second aspect of interest is the message content - analysed in this report using the topic discovery algorithm LDA (Latent Dirichlet Allocation) and performing IoD (Index of Dissimilarity) calculations.

4.1 The DBSCAN Algorithm

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a spatial clustering algorithm which will group points with many nearby neighbours and discard lone points in low-density regions. Given the obtained Twitter data, the algorithm is used to cluster an individual's tweets by location and determine the likely origin of his/her trips. Previous literature has assumed that the cluster with the highest number of tweets corresponds to a residential locale[35-37]. Even though this is not always the case, such cluster will be referred to as an user's "home". The DBSCAN algorithm intakes two user-specified parameters:

- epsilon (ϵ),
- minimum number of points (*minpts*).

Epsilon represents how close points should be to each other to be considered part of a cluster. One can think of it as the minimum cluster size. If ϵ is too small, a substantial number of points will not be clustered and potential clusters which are sparser will be identified as noise. If ϵ is too large, a considerable number of points will belong to the same cluster and denser clusters may merge together. The parameter *minpts* characterises the number of neighbours a point should have to be contained by a cluster. It defines the minimum number of neighbours within a cluster of "radius" epsilon. Figure 4 illustrates a typical DBSCAN cluster.

4.1.1 Implementation

A program was written in Python to perform DBSCAN. The program inputs a .csv file with a Twitter user's location coordinates. An arbitrary location is chosen as a starting point for the algorithm. Subsequently, its "neighbourhood" (defined by parameter epsilon) is computed. If such "neighbourhood" contains at least minpts points, a cluster is started. Otherwise, the point is labelled as noise. This procedure is repeated for all points in the file.

User selection. An important step in implementing the algorithm is finding a suitable user base with which to execute it. An analysis of the collected Tweets shows that some users are much

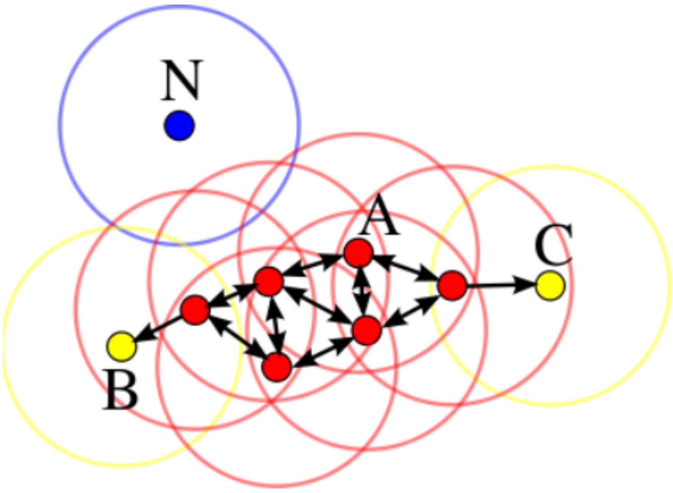


Figure 4: A DBSCAN cluster, with *minpts* set to 3, and ϵ represented as the radius of each circle. Red points are "core" cluster points, and yellow points B and C are "border" points, still considered part of the cluster. N is a "noise" point, outside of the cluster. Reproduced from [38].

more active than others. Figure 5 shows the distribution of a subset of 321,934 Tweets collected between 27 January 2016 and 1 March 2016. Over 44% of users only sent one tweet during the sample period. A small percentage of users (just over 15%) account for over 65% of all the tweets sent. Additionally, the top 1% of Twitter users contribute around 20 % of all tweets, while the median number of geo-located tweets per account is only 2. This is not at all surprising given the nature of Twitter; many people use Twitter as a stream of news and information but rarely post content[37].

It is clearly more complicated to use DBSCAN for users with a small number of tweets. On the contrary, the coordinates of "high frequency" users provide location data of high temporal granularity and inferences on their mobility can be drawn. It was decided to avoid the tweets published by the top 1% of users. Despite the attempts made to filter bots, some of these accounts appeared reporting news, incidents or advertising. Users between the 2nd (roughly 33 tweets per user for this data subset) and 1st percentile (75 tweets per user) were selected to perform DBSCAN.

Coordinate transformation. Twitter data presents users' location in a decimalised latitude-longitude coordinate system. Latitude-longitude coordinates follow the curvature of the Earth and produce inconsistent results when input to the DBSCAN algorithm[36]. The problem arises because the distance in metres between degrees of longitude depends on latitude; it tends to zero when approaching the poles. Furthermore, epsilon represents an Euclidean distance in metres. Therefore, a necessary data pre-processing step is required to convert Tweet latitude-longitudes to a Cartesian grid value. For purposes of comparison to residential land-use data for New York, the "New York-Long Island State Plane Coordinate System" was used. The "State Plane Coordinate System" is applicable to many cities in the USA, and was converted in this case using the "stateplane" Python package[39].

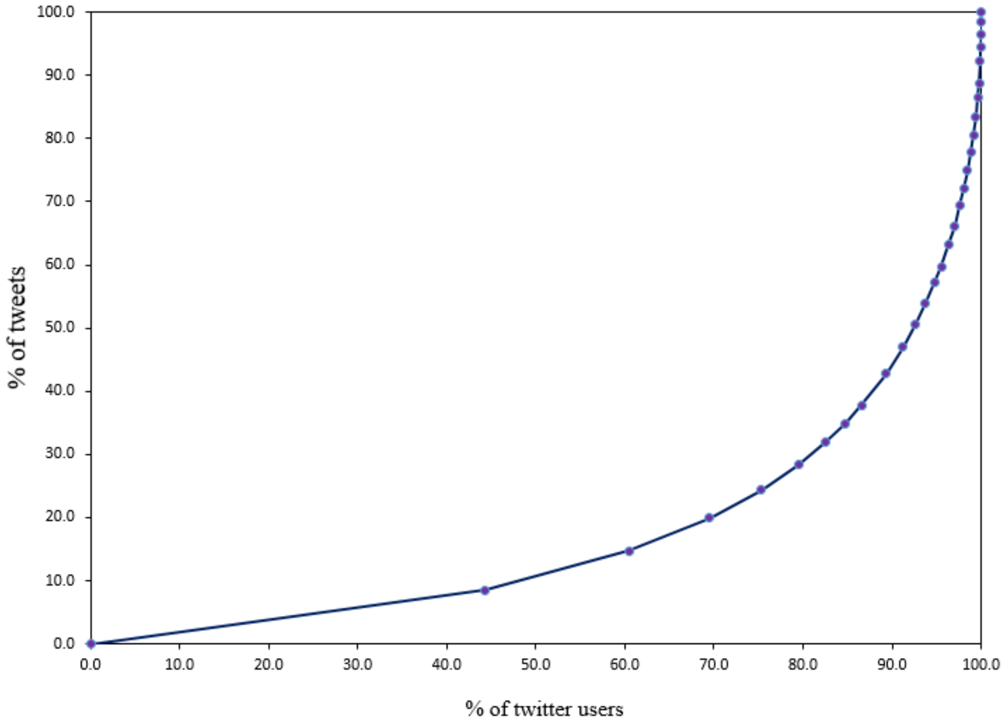


Figure 5: Plot for the proportion of tweets published against the proportion of users making them. This dataset contains 321,934 tweets made by 61,910 individual users.

4.1.2 Results and Discussion

The process of obtaining DBSCAN results was perhaps the most time-consuming and least rewarding part of the project; it took the team several weeks to fine-tune the model. This process was highlighted by an inability to find parameter values (for ϵ and and *minpts*) giving valid results on a consistent basis.

Unfortunately, the wrong premise was used selecting a dataset for this exercise. At the time, the dataset provided by David Greenwood was utilised for analysis because it had a much greater number of tweets that those collected from the Raspberry Pi. However, the fact that it is mainly limited to the borough of Manhattan is an important inconvenience. Roughly 2 million New Yorkers commute to Manhattan for work on a daily basis[40]. The borough has a very dynamic population; it serves approximately 4 million people on a typical weekday with a night population of 2.05 million[40]. Evidently, a significant number of the users tracked will not live in Manhattan. It is therefore unlikely that measured most visited locations are the users' origin locations ("homes"). Perhaps they constitute "work" locales (a cluster with the second highest number of Tweets [35-37]). These hypotheses were taken into consideration when choosing a dataset. However, at the time, data was still being collected from the Raspberry Pi. There were concerns on if sufficient data would be gathered and on the project's time constraints. Ultimately, the boundaries placed on the dataset provided by Buro Happold will significantly underestimate any mobility measurements.

Despite the difficulties encountered, the DBSCAN model shows some promise for future applications. Principally because it can calculate many of the measures utilised in current mobility research. Provided sufficient data is collected over a large enough geographic range, the radius of gyration

$$r_g = \sqrt{\sum_{i=1}^{n}(r_i - r_h)^2} \tag{3}$$

of an user can be calculated. Here n would represent the number of recorded tweets from the user, r_i the location of tweet i and r_h the "home" location. Some radii of gyration could be calculated for certain users using parameters $\epsilon = 4000$ and $minpts = 3$. Not surprisingly, the values obtained were considerably smaller than typical results recorded by Kurkcu et al. (2015)[35] in New York City. Another measure which could potentially be inferred is each user's trip distance distibution $P(r)$, commonly represented by a power law:

$$P(\Delta r) = \Delta r^{-\beta}. \tag{4}$$

In this case, Δr would be the distance between consecutive tweets and β an estimated displacement exponent. It is worth noting that when using the free Twitter Streaming API, user tweets are rarely consecutive (the free API allows access to at most 1% of the tweets produced at a given time). A possible solution would be to use the very expensive but comprehensive Firehose API - a feed provided by Twitter that allows access to 100% of tweets[41].

4.2 Latent Dirichlet Allocation

First proposed in 2003[42], Latent Dirichlet Allocation (LDA) is a generative probabilistic model to identify various topics within text documents. The model first presumes a generative procedure of documents and words, and uses the observed data to infer the key parameters of the probabilistic distributions in the generation process. LDA assumes the following generative process for each text document of N words within the set M of all documents:

1. Choose $\theta \sim \text{Dir}(\alpha)$. Here, θ is a topic distribution, which is chosen from $\text{Dir}(\alpha)$: a Dirichlet distribution with parameter α. α is a weight vector representing the topic importance in the set of documents.

2. For each of the N words w_N:

   a.) Choose a topic $z_n \sim$Multinomial(θ). Generate a topic z_n from a multinomial distribution with parameter θ.

   b.) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n. Here, w_n is drawn from a keyword vocabulary distribution of topic z_n by using another multinomial distribution with parameter β. That is, each topic z_n has a keyword distribution $\beta_(z_n)$ over the total vocabulary, following the multinomial distribution with parameter θ [42,43].

There are three distributions in this model. The topic weight distribution θ is the first, describing how topics are distributed within the set of documents. The second is the document-specific topic distribution determining which topic is to be used for a word. Such is used to generate z_n for the nth word in the document. Finally, the keyword distribution of a given topic z_n is used to generate a word w_n in the topic z_n from the total vocabulary[43]. The relationship between these distributions is shown in Figure 6.

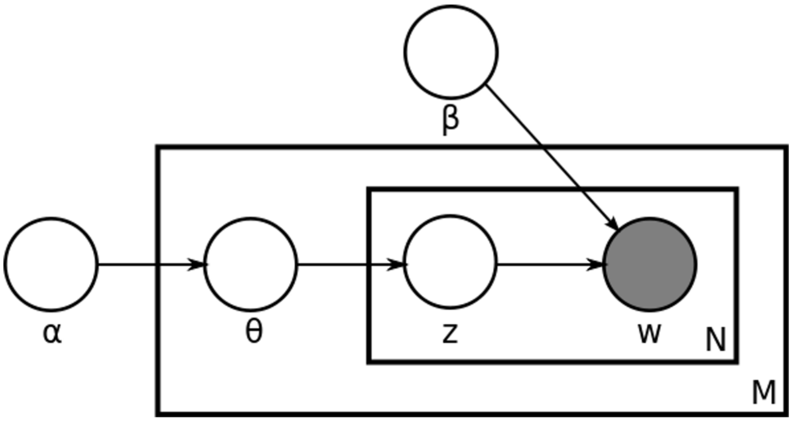


Figure 6: Graphical model representation of Latent Dirichlet Allocation in "Plate" notation, where the boxes are "plates" representing replicates. The outer plate represents documents, and the inner plate represents the repeated choice of topics and words within a document. Reproduced from [42].

Topic Discovery models such as LDA are usually applied to large corpora of text documents, such as collections of news articles or scientific papers. Each of these documents will typically contain several hundred or thousand words, in contrast to the 140 character limit of Twitter messages. Therefore, the assumption that LDA will be applicable to Twitter messages is not trivial, and needs to be investigated. One study[44] runs an LDA model on a body of Twitter messages with known topics (eg. Tweets collected from a sports news account, all assigned to the topic "sports"). This study and several others[45-47] have demonstrated that LDA is effective in identifying topics in Twitter messages. Consequently, it was determined that LDA would be the most suitable method of Topic Discovery to implement in this project. LDA is one of the most commonly used Topic Discovery models and so, it can be implemented with a number of existing programming libraries.

4.2.1 Implementation

This section describes the implementation of an LDA model to identify topics in the Tweet content of the New York City dataset. This model was implemented as a Python application using several pre-existing Python libraries, as shown below. As discussed in the previous section, LDA is suitable for Topic Discovery in Twitter messages. However, Twitter messages must go through several text pre-processing steps before they can be suitable as an input to the model. This is because there are significant differences between Twitter messages and the text corpora LDA would normally be applied to, such as news articles. Twitter messages contain much more "noise" in their content which must be filtered out. This "noise" can include Tweets in different languages, words spelled incorrectly, Tweets which contain only a link to another website, or messages which contain "hashtags". An example Tweet "HOLA! HELLO! Let me show you how text processing works! #Physics http://www.bath.ac.uk/", will be used to illustrate the pre-processing steps in their respective order:

1. **Language filtering**: Performed using the Python library "langid", which is a stand-alone language identification tool[48]. For this analysis, only Tweets in English were considered. This step filtered out approximately 30% of all Tweets. This figure initially seems very high, but is largely due to the amount of Tweets which contain only a URL, spelling mistakes or shortened words; having therefore have no words which can be identified as English. These Tweets would not be relevant for the model, and would be filtered out in the following step, so the amount of Tweets filtered out in this step is entirely appropriate. The example Tweet becomes "HELLO! Let me show you how text processing works! #physics http://www.bath.ac.uk/".

2. **Lowercase**: All message content is converted from uppercase to lowercase using regular expression functions in the standard Python library. For the example Tweet, "hello! let me show you how text processing works! #Physics http://wwtwo w.bath.ac.uk/".

3. **URL and special character removal**: Many of the Tweets in the New York City dataset have been posted through third-party applications, and therefore contain a URL link to another website. Many Tweets also contain "hashtags" and other special characters, which may not be a useful input to the model. For this specific analysis, hashtags were removed. However, as hashtags are used to tag Tweets into topics, these could potentially be used as a method of Topic Discovery. This was not investigated in this project but is a key area for further work. In this step of the code, both URLs and special characters were removed using regular expression functions in the standard Python library. The example Tweet is now "hello! let me show you how text processing works! physics".

4. **Removal of stop words, unigrams and bigrams**: Stop words are words such as "this", "that", "the" and "and", which are very commonly used, and do not impart meaning on their own. These were removed by filtering against a list of stop words provided by the "Natural Language Toolkit" Python library[49]. Unigrams and bigrams are special cases of "n-grams": a sequence of n items within a sequence of text. This can be used to refer to any aspect of language, i.e. a sequence of words, letters, syllables etc. In this case, unigrams and bigrams refer to words of 1 and 2 letters respectively. These are filtered out because they are generally likely to be stop words, and therefore not going to be useful inputs for the LDA model. The example Tweet becomes "hello let show text processing works physics".

5. **Tokenisation**: The Tweet is converted to a string of tokens delimited by the spaces between words. This step also eliminates punctuation marks from the message. In the example, [hello, let, show, text, processing, works].

6. **Vectorisation**: Finally, two arrays: a dictionary and a corpus, are produced. The dictionary is a numbered list of the vocabulary of all the Tweets, where each word used will appear once. The corpus is a list of arrays, where each Tweet is represented as an array of numbers, each number corresponding to a word in the dictionary. This is commonly referred to in natural language processing as the "bag-of-words" format. In its final pre-processing form, the example Tweet is $\{$ 0 : "hello"; 1 : "let"; 2 : "show"; 3 : "text"; 4 : "processing"; 5 : "works"; 6 : "physics"$\}$.

Two other common techniques in natural language processing are lemmatisation and stemming: these techniques are similar and both aim to reduce words to a common base. For example, stemming would reduce “plays”, “playing” and “played” to the base word “play”. Lemmatisation performs the same process as stemming, but will also reduce related words into one, such as reducing “am”, “is” and “are” to the common base “be”. This is often done in natural language processing to improve efficiency in models, and was initially implemented in this code using the “Natural Language Toolkit” Python Library[49]. However, it was found that this process did not improve the output of the LDA model, and made it more difficult to interpret the output topics. Therefore, this step was not used in the final implementation of the LDA model.

A topic modelling package for python called ‘gensim’, developed by Radim Rehurek[50,51], is used for implementing LDA. This package was chosen due to its efficiency when dealing with large corpora and its ease of implementation. Additionally, the Gensim package also has methods to perform Latent Semantic Analysis and Hierarchical Dirichlet Processes. These other methods of analysis have not been considered relevant for this project, but could be taken into account in the future by Buro Happold to perform other processes. Another advantage of the Gensim package is that the algorithm can be made to iterate several times, re-evaluating the values α and β (see Figure 6) without these having to be input again by the user. It was found that the values of these parameters converge very quickly, so 5 iterations of the algorithm were deemed to be the best compromise between efficiency and accuracy.

4.2.2 Results and Discussion

The LDA model was run for the full New York City dataset to identify topics within the Twitter messages. The output of this model is the keyword distribution $\beta_(z_n)$ for each topic, showing the probability for each word of being in such. These keyword distributions are then used to classify all of the Tweets into their most likely topics, based on the words within each Tweet. For this analysis, it was decided to only classify Tweets which had a probability of greater than 50% of belonging to a particular topic; all others were disregarded.

To first test the validity of this model, the algorithm was run on two different subsets of the New York City dataset: both of which contained only a smaller geographical area centred in Lower Manhattan. One contained only Tweets posted by tourists and the other contained Tweets posted by locals. To identify users who were tourists, the “Bot Checking” Python application described in **3.2.2** was altered to find the average location of a user’s most recent 200 Tweets. If such user’s “centre of mass” was outside a set distance of New York, he/she was classified as a tourist. By hand-checking a sample of users from the Tourist dataset, it was shown that the “Tourist Checker” application was consistently accurate. The reason for conducting this analysis was to assist Buro Happold’s “Make Way for Lower Manhattan” initiative: by identifying the topics which tourists Tweet about in Lower Manhattan, one can assess the key tourist attractors in the area. These then can inform development of the “Tourist Trail” recommended by Buro Happold.

The results of this initial analysis were as expected and confirmed the validity of the LDA model: tourist Tweets were organised into topics which were related to specific locations. For example, one topic relating to the 9/11 Memorial and Museum containing words “wtc”, “memorial”, “September” etc. Other topics focused on landmarks in Lower Manhattan such as Wall St., the Financial District and Stone Street (a popular area with a high concentration of bars/restaurants).

A more general spread of topic was seen in the non-tourist dataset, referring to topics such as work, travel, arts and entertainment. One topic common to both datasets was Stone Street: showing that the area is popular with both tourists and locals. Table 1 shows an example topic output by the program for Tweets collected during New York Fashion Week. The name of the topic is left to use interpretation; it could be defined under "Nightlife", for example.

Word	Probability
"central"	8.7%
"show"	2.3%
"nyfw"	2.1%
"fashion"	2.0%
"week"	1.7%
"party"	1.4%
"still"	1.4%
"tonight"	1.2%
"friends"	1.0%
"rock"	0.9%

Table 1: An example of a nightlife-related topic obtained through LDA.

The temporal patterns of topics are analysed, an example of which is shown in Figure 7. In 7(a), the temporal distribution of Tweets associated with a "Work & Home" theme can be observed. These appear to peak in the early afternoon, having a general spread throughout the day and a similar distribution to that of Figure 3. In contrast, Tweets associated with "Arts & Culture" or "Nightlife" topics appear to have a much more temporally concentrated distribution. Figure 7(b) shows the temporal distribution for the "Arts & Culture" topic, which has two clear peaks in the afternoon and evening, and then declines very quickly with little overnight activity. These distributions match initial expectations and are similar to results of previous analyses of this kind[22].

To further validate the effectiveness of the LDA model, a similar analysis to that performed by Lai et al.[52] (2015) was carried out on the New York City dataset. Here, Tweets within a 10 minute walking distance of New York Subway stations were clustered together, and the most common Tweet topic for each station plotted for a weekend morning and evening. The plot is shown in Figure 8. This analysis shows some spatio-temporal trends, but there are not as clear or meaningful as those established by Lai et al.. New York subway stations are less densely spaced than the London Underground stations in [52], while the population is more densely populated. Therefore, there are a much greater number of Tweets assigned to each cluster; there being significantly less granularity in the results. For this analytical technique to be applicable to New York City, or any other city, we must explore other methods of clustering groups of spatially concentrated Tweets together.

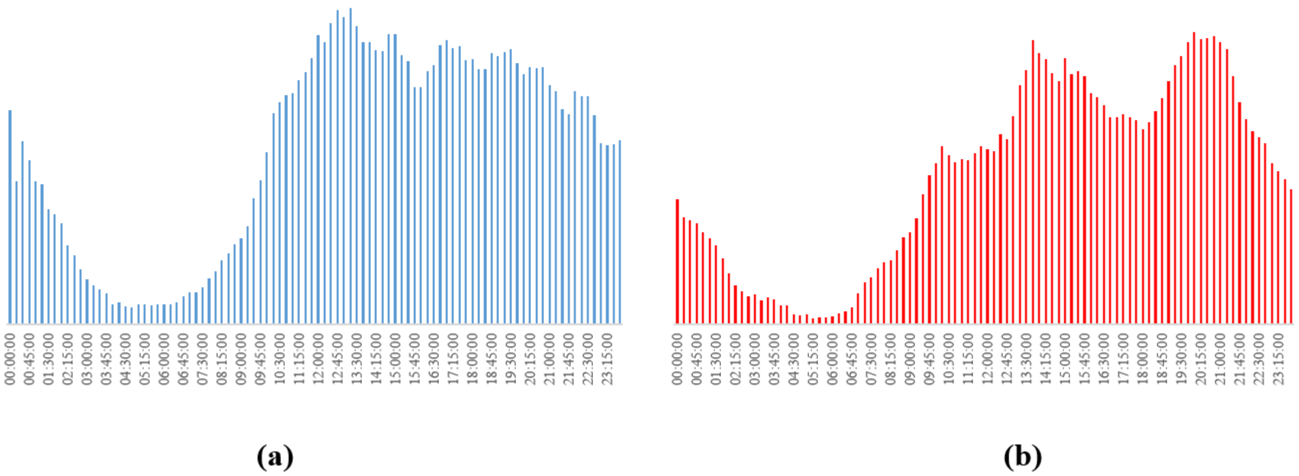


Figure 7: Shows the temporal distributions of Tweets classified with the topics (a) Work & Home, and (b) Arts & Culture.

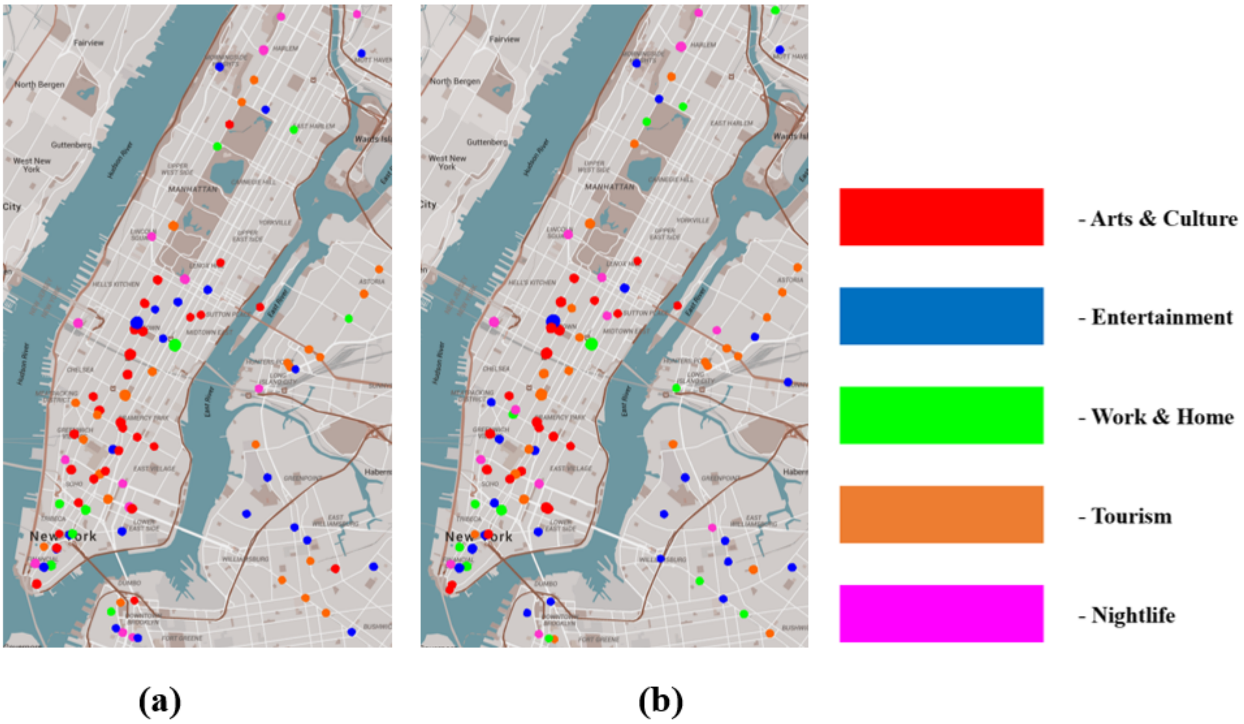


Figure 8: Clusters of Tweets attached to Subway stations in New York City for (a) a typical weekend during the day, and (b) a typical weekend night. Cluster colours represent the most common topic of Tweets for that cluster, as determined by the LDA model.

4.3 Index of Dissimilarity

Another method of analysing the content of geo-located Twitter messages is to perform an Index of Dissimilarity (IoD) calculation: this is a measure of the evenness of a spatial distribution. To calculate the IoD, the distribution must first be divided into unique geographical areas; the IoD of a word y is then defined as:

$$\theta(x,z) = \frac{1}{2}\sum_{x}\left(\frac{X_x{}^{\mathrm{y}}}{X_*{}^{\mathrm{y}}} - \frac{X_x{}^{*}}{X_*{}^{*}}\right) \tag{5}$$

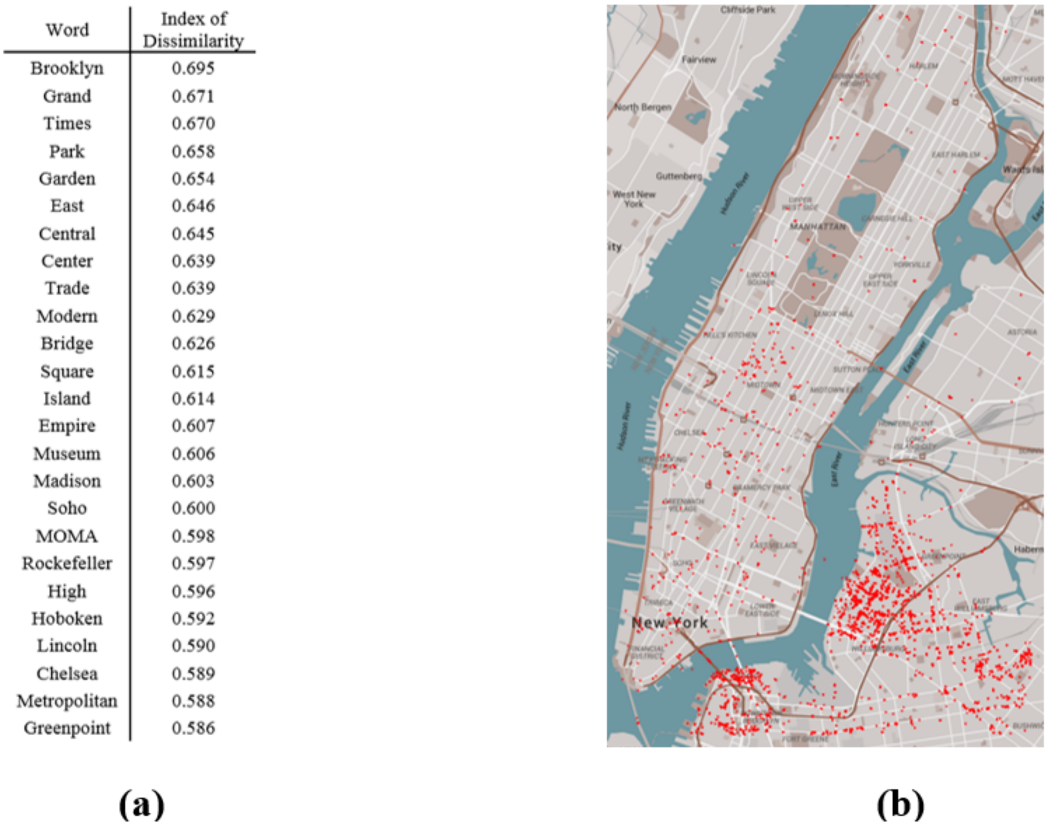
where is x the set of geographical areas, $X_x{}^{\mathrm{y}}$ represents the number of times the word y is used in area x and (*) represents a summation across the missing index. The result of this calculation is a standardised value between 0 and 1, where 0 indicates a uniform distribution and 1 indicates a spatial concentration[22].

To perform this calculation, an application was produced in Python, which performed the same text pre-processing steps outlined in **4.2**, and then iterated over every word to calculate the IoD. To reduce anomalous results, the calculation was only performed on words which appeared in the corpus more than 10 times. There is very little academic research in applying IoD to Twitter messages[22,53] and these analyses have all divided Tweets into pre-defined geographical areas (by electoral wards, for example). One supposes that not all cities will have obvious means of dividing Tweets into geographical areas. To create an approach which is reproducible for other cities, it was necessary to investigate potential methods of geographically dividing Tweets.

The first approach taken to geographically divide Tweets in the New York City dataset was very simple: the city was divided into a grid of equally sized squares, and the IoD calculation performed on these. However, this approach is flawed, since the number of Tweets in each grid can vary greatly. To attempt to improve this method, DBSCAN was utilised to cluster Tweets. Once the Tweets have been clustered, the IoD calculation can be performed using each cluster as an area x. This approach has a smaller amount of variation in the number of Tweets in each cluster. Figure 9 shows results for the 25 most spatially concentrated words of the dataset. The results of both approaches were very similar for the most spatially concentrated words. As expected, the most concentrated words throughout the whole city, refer to locations, boroughs, landmarks and places of interest.

Another method of applying the IoD calculation is to divide data temporally, rather than geographically. In equation (5), each group x would represent the Tweets from a time period rather than a geographic area. This method would identify words relating to events, holidays and news. It was determined that this approach would not be of interest to Buro Happold, but the IoD application developed would not need any changes to perform this analysis.

IoD calculations can be valuable to Buro Happold for two main reasons. Firstly, words concentrated in the Lower Manhattan area can provide good insight into identifying the key attractors in the area. Such insight could be extremely valuable when designing the Lower Manhattan "Tourist Trail". Secondly, it has been shown that spatially concentrated, or Location Indicative Words (LIW), can be used to predict the location of Tweets that are not geolocated[54].The output of the IoD application could potentially be used to develop a reliable geo-prediction framework to infer the location of a Tweet based on the content of the message. This is of interest as it would greatly increase the amount of geolocated data available. Due to the time constraints of this project, it was not possible to conduct any work on a geo-prediction model, but this is an area with great potential for further work.



Word	Index of Dissimilarity
Brooklyn	0.695
Grand	0.671
Times	0.670
Park	0.658
Garden	0.654
East	0.646
Central	0.645
Center	0.639
Trade	0.639
Modern	0.629
Bridge	0.626
Square	0.615
Island	0.614
Empire	0.607
Museum	0.606
Madison	0.603
Soho	0.600
MOMA	0.598
Rockefeller	0.597
High	0.596
Hoboken	0.592
Lincoln	0.590
Chelsea	0.589
Metropolitan	0.588
Greenpoint	0.586

Figure 9: (a) shows the 25 most spatially concentrated words in the New York City dataset, as determined by the Index of Dissimilarity calculation. (b) shows the distribution of Tweets containing the most spatially concentrated word, “Brooklyn”.

4.4 Uber and Taxi Data

4.4.1 Trends in Official Taxi Usage

As mentioned in section **3.3**, the borough at which a pick-up or drop-off takes place can be inferred from the taxi datasets’ location coordinates. A set of bounding boxes is assigned to each borough to approximate its area and filter the data accordingly. These area estimations are coarse but accurate enough to give insight on the number of official taxi (yellow and green) and Uber pick-ups in each borough. Initial observation suggests that whereas pick-ups are more concentrated in Manhattan (86.8% of all pick-ups take place here for April 2013-December 2015), drop-offs disperse into the outer boroughs.

Figure 10 presents the monthly evolution of cab pick-ups in New York’s five boroughs and in JFK and La Guardia airports for the time period April 2013-December 2015. Notice how the green taxi programme is introduced into Brooklyn in August 2013 and subsequently into Queens, The Bronx, Staten Island and Upper Manhattan. In Brooklyn and The Bronx, the programme has incremented overall taxi activity dramatically. As of December 2015, green cabs constitute 61.4% of official taxi pick-ups in Brooklyn and 92.3% of official taxi pick-ups in The Bronx. The green taxi programme appears to be less effective in Queens with 37.9% of official taxi pick-ups for the month. This figure is misleading and arises from the the location of JFK and La Guardia airports, both in Queens. Green taxis do not serve the airports, which contribute to a 52.6% of all pick-up activity in the borough from April 2013 to December 2015. When discarding airport activity, "boro" cabs make up 80.6% of official taxi pick-ups in Queens for December 2015. Not surprisingly, green taxis account for only 2.81% of official pick-ups in Manhattan for the month; these cabs can only operate in the borough above East 96th and West 110th Streets. It is worth noticing that The Bronx and Staten Island have significantly less taxi traffic throughout the studied period of time (0.665% and 0.004% of overall pick-up activity respectively).

4.4.2 The Rise of Uber

Uber's entrance into the NYC taxi market has been at least as disruptive as the deploying of green taxis. The Uber datasets obtained allow for year-over-year comparison between equal three-month time frames (April-June 2014 and April-June 2015). Table 2 shows the change in April-June official taxi pick-ups from to 2014 to 2015 for New York City, each of its boroughs and JFK and La Guardia airports.

Borough	Official Taxi	Uber	Overall Change
Airports	+40,114 (+2.55%)	+225,129 (+330%)	+ 265,243 (+16.2%)
The Bronx	-1,015 (-0.286%)	+119,003 (+8,570%)	+117,987 (+32.4%)
Brooklyn	+288,226 (+13.2%)	+1,066,938 (+576%)	+1,355,164 (+57.1%)
Manhattan	-3,562,345 (-10.6%)	+3,890,009 (+294%)	+327,664 (+0.854%)
Queens	+170,025 (+5.38%)	+610,409 (+566%)	+780,434 (+10.0%)
Staten Island	+191 (+19.1%)	+3,918 (+1270%)	+4,108 (+315%)
Net NYC	-3,104,919 (-7.26%)	+5,690,276 (+350%)	+2,585,397(+5.82%)

Table 2: Growth of official taxi and Uber pickups from April-June 2014 to April-June 2015.

It can be observed in Figure 10, that the number of green-taxi pickups roughly stabilises in April 2014. Hence, one can argue that Uber is the main responsible for the 5.2% net growth in taxi (official and Uber) activity from April-June 2014 to April-June 2015. Interestingly enough, Uber's growth has significantly effected a greater usage of taxis in the outer boroughs. However, pick-ups in Manhattan have increased barely by a 0.854%, despite the dramatic rise of Uber rides. This contradicts the discourse of politicians and taxi companies blaming Uber for increased congestion[32], at least in New York City. Congestion is not a problem in the outer boroughs[39]; the green taxi programme was introduced precisely because these boroughs were underserviced[32,55]. Uber is replacing yellow cabs in Manhattan but is probably not leading to increased congestion or gridlock there.

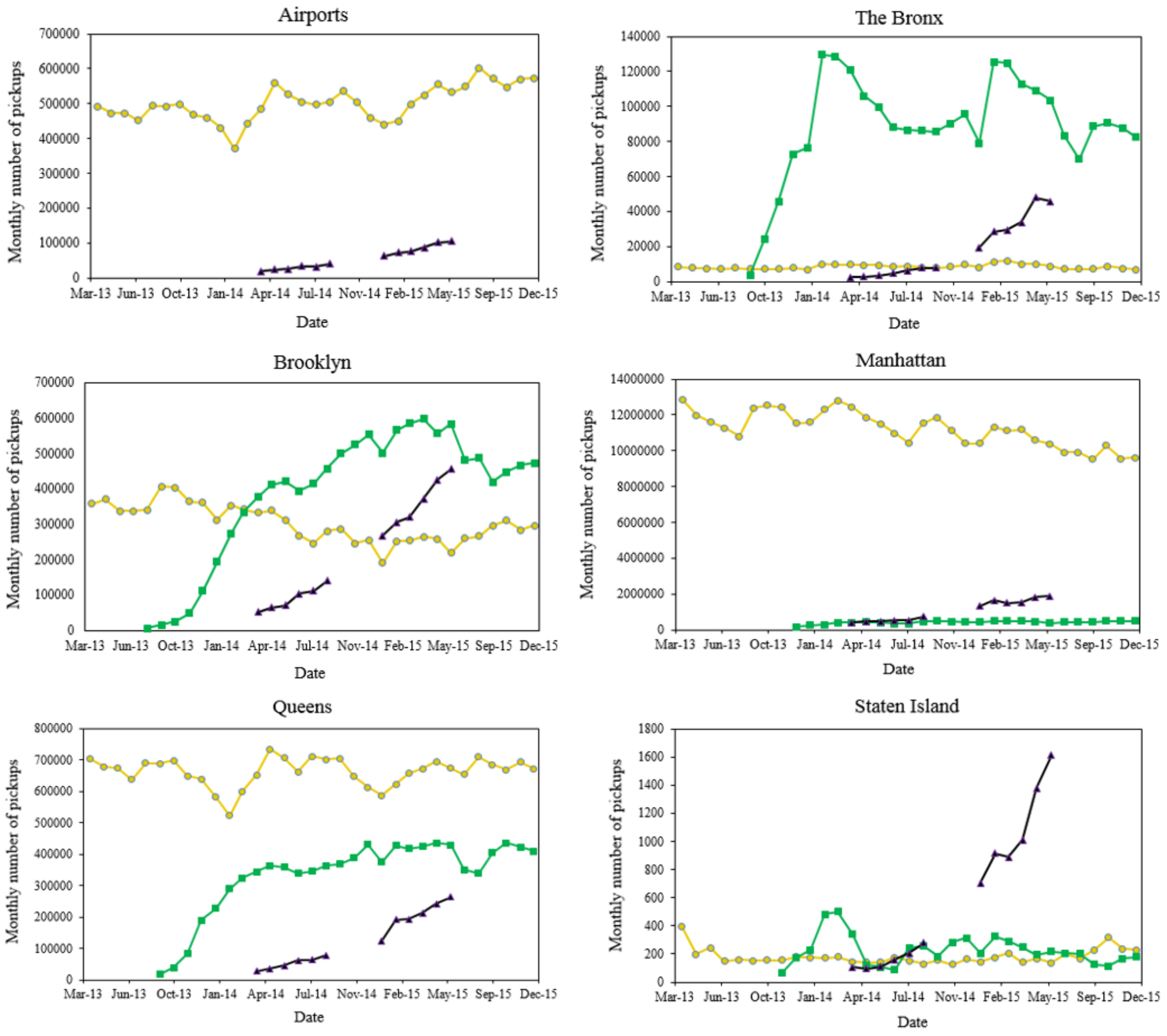


Figure 10: Set of graphs showing the evolution of the monthly number of taxi and Uber pickups from April 2013 to December 2015 for the five NYC boroughs and JFK and La Guardia Airports. The yellow line (circles) represents yellow taxis, the green line (squares) green taxis and the black line (triangles) Uber rides. Data for Uber rides has been obtained from two Freedom of Information Law requests made by blog FiveThirtyEight to the New York City TLC (Transport and Limousine Commission). Data for Uber rides is only available for the periods of time April to September 2014 and January to June 2015.

5 Conclusions and Future Work

Recalling the four objectives of the project:

1. To provide a comprehensive literature review of the current state-of-the-art in human mobility in cities, and research into how such models and publically available data sources are utilised.

2. To develop a reproducible approach for data gathering from useful data sources identified in the literature review to yield insights into human mobility and utilisation of urban space.

3. Analysis and visualisation of collected data with respect to identified performance indicators, parameters of human mobility and utilisation of urban space.

4. Validation exercise to compare and evaluate various models of human mobility and utilisation of urban space using the datasets collected.

All of these objectives have been achieved, with varying degrees of success. Section 2 gives a comprehensive summary of the current state-of-the-art in mobility modelling. Section 3 outlines an approach for gathering geolocated data from Twitter using a Python application, and details the steps necessary to filter and analyse this data. This approach is reproducible for any major city.

There were several analyses performed on the collected data, including topic discovery analysis using Latent Dirichlet Allocation, clustering Tweets with DBSCAN, and calculating the Index of Dissimilarity for words appearing in the Twitter dataset. The topic discovery analysis was successful, but a clustering method for Tweets of similar topics could not be implemented. The application of DBSCAN to find "home" and "work" locales shows potential. However, it was not entirely successful due to the geographic bounds of the collected data and the inability to find parameter values giving valid results on a consistent basis. Additionally, the method implemented requires a more systematic way of selecting users and less "hand-picking". It is not easy to quantify the success of the performed Index of Dissimilarity calculations. However, the results match expected values and other published results in the field. The Index of Dissimilarity could potentially be used in future work with Location Indicative Words in a geo-prediction framework.

This report has identified several open data sources with interesting features to analyse. These have included geo-located tweets from Twitter and pick-up coordinates from Uber. Additionally, open data released by New York City has been used to contrast the findings obtained. This data has consisted of taxi trip records from the city's Taxi and Limousine Commission. Turnstile data from the Metropolitan Transport Authority was also used to validate station populations when looking at the Twitter activity in their neighbourhoods. It is worth noting that there are more data sources available in the City; for example, Central Park daily weather data is available from the National Climatic Data Center[c] (NCDC), bike usage data is available from Citi Bike[d], New York City's bike share system, and detailed land use data can be found at Open Data NYC[e]. Data from the NCDC has in fact been used to show how daily snowfall and rainfall affect the number of

[c] https://www.ncdc.noaa.gov

[d] https://www.citibikenyc.com/

[e] https://nycopendata.socrata.com/

Uber pick-ups (not included in this report). Rainfall does not appear to have a significant effect on taxi ridership, whereas snowfall may cause a dramatic increase. Citi Bike data could be analysed more carefully since it relates directly to the topic of this project and is extremely detailed. It contains station locations and timestamps for when bike rides start and end, rider gender, rider birth year and whether the rider is an annual subscriber or a short term customer.

Analysis has also been performed on taxi rides on a smaller scale, looking at individual neighbourhoods in New York City. Once again, these have been delimited by placing sets of bounding boxes enclosing their area. The analysis in question is incomplete; this is partly due to the time consuming process of filtering the data. Knowledge of software like PostgreSQL and PostGIS would help storing the data or performing geographic calculations. When looking at individual neighbourhoods, the goal has been to find out which are more active at night (pick-up/drop-off percentages corresponding to 10pm-5am) and which are more active at day (7am-9am on weekdays). Those more active at night can be classified as areas providing nightlife and entretainment, and those more active on weekday mornings can be classified as residential areas. These results could then be used to validate the LDA results. Additionally, the detailed land use data provided by the city could be another source of validation.

References

[1] *Make Way For Lower Manhattan*, Buro Happold. Viewed 3rd April 2016. http://www.burohappold.com/knowledge-and-news/article/make-way-for-lower-manhattan-3708/.

[2] Brockmann, D.D., Hufnagel, L., Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075).

[3] Havlin, S., Ben-Avraham, D. (2002). Diffusion in disordered media. *Advances in Physics*, 51(1).

[4] Viswanathan, G.M., Afanasyev, V., Buldyrev, S.V., Murphy, E.J., Prince, P.A., Stanley, H.E. (1996). Lévy flight search patterns of wandering albatrosses. *Nature*, 381(6581).

[5] Sims, D.W., Southall, E.J., Humphries, N.E. et al. (2008). Scaling laws of marine predator search behaviour. *Nature*, 451(7182).

[6] Ramos-Fernandez, G., Mateos, J.L., Miramontes, O., Cocho, G., Larralde, H., Ayala-Orozco, B. Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi). *Behavioural Ecology and Sociobiology*, 55(3).

[7] Buchanan, M. (2008). Ecological modelling: The mathematical mirror to animal nature. *Nature*, 453(7196).

[8] Klafter, J., Shlesinger, M.F., Zumofen, G. (1996). Beyond brownian motion. *Physics Today*, 49(2).

[9] Mantegna, R.N., Stanley, H.E. (1994). Stochastic process with ultraslow convergence to a gaussian: the truncated Lévy flight. *Physical Review Letters*, 73(22).

[10] González, M.C., Hidalgo, C.A., Barabási, A. (2008). Understanding individual human mobility patterns. *Nature*, 456(7196).

[11] Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C. (2013). Unravelling daily human mobility motifs *Journal of the Royal Society Interface*, 10(84).

[12] Di Lorenzo, G., Reades, J., Calabrese, F., Ratti, C. (2012). Predicting personal mobility with individual and group travel histories *Environment and Planning B*, 39(5).

[13] Yan, X., Zhao, C., Fan, Y., Di, Z., Wang, W. (2014). Universal predictability of mobility patterns in cities. *Journal of the Royal Society Interface*, 11(100).

[14] Song, C., Koren, T., Wang, P., Barabási, A. (2010). Modelling the scaling properties of human mobility. *Nature Physics*, 6(10).

[15] Song, C., Qu, Z., Blumm, M., Barabási, A. (2010). Limits of Predictability in Human Mobility. *Science*, 327(5968)

[16] twitter.com (2016). Twitter's official website. [online] Available at: http://www.twitter.com [Accessed 1 Oct. 2015].

[17] Siena Research Institute. (2015). Cell Phones Used by 90 Percent of New Yorkers; Smartphones Used by Nearly Two-Thirds. *Siena Research Institute*, Siena, New York.

[18] Kurkcu, A. et al. (2015). Evaluating the usability of geolocated Twitter data as a tool for human mobility and mobility patterns: a case study for NYC. *Transportation Research Board's 95th Annual Meeting.*

[19] D'Monte, L. (2009). Swine Flu's Tweet Tweet Causes Online Flutter. Business Standard, [online]. Available at: http://www.business-standard.com [Accessed 1 Mar. 2015].

[20] *Twitter Statistics* (2016). Twitter Live Stats. Viewed 13th March 2016. http://www.internetlivestats.com/twitter-statistics/.

[21] Frank, M.R., Dodds,P.S., Danforth, C.M., Mitchell, L. (2013). Happiness and the Patterns of Life: A Study of Geolocated Tweets. *Scientific Reports*, 3(2625).

[22] Adnan, M., Lansley, G., Longley, P.A. (2015). Exploring the geo-temporal patterns of Twitter messages *UCL.*

[23] Adnan, M., Lansley, G., Longley, P.A. (2014). The geotemporal demographics of Twitter usage. *Environment and Planning A*, 47(2).

[24] Qu, Y. Huang, C., Zhang, P., Zhang. J. (2011). Microblogging After a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake. In: *Computer Supported Cooperative Work and Social Computing*, pp 25-34.

[25] Heaivilin, N. et al. (2011). Public health surveillance of dental pain via Twitter. *Journal of Dental Research*, 90(9).

[26] Coppersmith, G., Dredze, M., Harman, C. (2014) Quantifying Mental Health Signals in Twitter *Johns Hopkins University.*

[27] Lapowsky, I. (2013). "Ev Williams on Twitter's Early Years". Inc, [online] Available at: http://www.inc.com [Accessed 19 Mar 2016].

[28] Kassim, S. (2012). Twitter Revolution: How the Arab Spring was Helped by Social Media. Policy.Mic, [online] Available at: http://www.mic.com [Accessed 19 Mar 2016].

[29] *The Streaming APIs*, Twitter. Viewed 9th April 2016. https://dev.twitter.com/streaming/overview.

[30] Slate (2014). The Year in Uber. [online] Available at: http://www.slate.com [Accessed 4 Apr. 2016].

[31] Wallsten, S. (2015). The Competitive Effects of the Sharing Economy: How is Uber Changing Taxis? Technology Policy Institute, New York.

[32] FiveThirtyEight (2015). Uber is Taking Millions of Manhattan Rides Away From Taxis. [online] Available at: http://fivethirtyeight.com/ [Accessed 30 Mar. 2016].

[33] Zhang, C., Wang, K. (2016). POI Recommendation Through Cross-Region Collaborative Filtering. *Knowledge and Information Systems*, 46(2), pp.369-387.

[34] Noulas, A., Scellato, S., Lathia, N. et al. (2012). Mining User Mobility Features of Next Place Prediction in Location-Based Services. 12th IEEE International Conference on Data Mining.

[35] Kurkcu, A. et al. (2015). Evaluating the usability of geolocated Twitter data as a tool for human mobility and mobility patterns: a case study for NYC. *Transportation Research Board's 95th Annual Meeting.*

[36] Swier, N. et al. (2015). Using geolocated Twitter traces to infer residence and mobility. *GSS Methodology Series No. 41*, Office for National Statistics.

[37] Jurdak, R. et al. (2015). Understanding Human Mobility from Twitter *PLOS ONE*, 10(7).

[38] *DBSCAN Illustration*, Wikipedia. Viewed 16th April 2016. https://en.wikipedia.org/wiki/DBSCAN#/media/File:DBSCAN-Illustration.svg.

[39] Moss, M.L., Qing, C. (2012). The Dynamic Population of Manhattan. *Rudin Center for Transportation Policy and Management*, NYU, New York.

[40] *stateplane*, Viewed 17th April 2014. https://pypi.python.org/pypi/stateplane/0.1.1.

[41] Morsatter, F. et al. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *International Conference on Weblogs and Social Media*, pp 400-408.

[42] Blei, D., Ng, A., Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5), pp.993-1022.

[43] Zhang, L., Sun, X., Zhuge, H., (2015). Topic Discovery of Clusters from Documents with Geographical Location. *Concurrency and Computation: Practice and Experience.* 27(15), pp.4015-4038.

[44] Risch, J. (2016). Detecting Twitter Topics Using Latent Dirichlet Allocation. Thesis. Uppsala University.

[45] Fujino, I. (2014). Refining LDA Results and Ranking Topics in Order of Quantity and Quality with an Application to Twitter Streaming Data. Proceedings – 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp.209-216.

[46] Tan, S., Li, Y., Sun, H. et al. (2014). Interpreting the Public Sentiment Variations on Twitter. *IEEE Transactions on Knowledge and Data Engineering.* 26(5), pp.1158-1170.

[47] Ghosh, D., Guha, R. (2013). What are we 'Tweeting' About Obesity? Mapping Tweets with Topic Modeling and Geographic Information Systems. *Cartography and Geographic Information Science*, 40(2), pp.90-102.

[48] *langid.py readme*, langid. Viewed 13th April 2016. https://github.com/saffsd/langid.py.

[49] *NLTK 3.0 Documentation*, Natural Language Toolkit. Viewed 14th April 2016. http://www.nltk.org/.

[50] *genism*, Radim Řehůřek. Viewed 14th April 2016. http://radimrehurek.com/gensim/.

[51] Řehůřek, R., Sojka, P. (2010). Software Framework for Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp.45-50.

[52] Lai, J., Cheng, T., Lansley, G. (2015). Spatio-Temporal Patterns of Passengers' Interests at London Tube Stations. Presented at: GISRUK 2015, University of Leeds.

[53] Birkin, M., Harland, K., Malleson, N. (2013). The Classification of Space-Time Behavior Patterns in a British City from Crowd-Sourced Data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7974, pp.179-192.

[54] Han, B., Cook, P., Baldwin, T., (2014). Text-Based Twitter User Geolocation Prediction. *Journal of Artificial Intelligence Research*, 49, pp.451-500.

[55] Todd W. Schneider (2015). Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. [online] Available at: http://toddwschneider.com/ [Accessed 30 Mar. 2016].