A name-led approach to profile urban places based upon geotagged Twitter data

Juntao Lai, Guy Lansely, James Haworth, Tao Cheng

Abstract

Place is a concept that is fundamental to how we orientate and communicate space in our everyday lives. Crowd sourced social media data present a valuable opportunity to develop bottom-up inferences of places that are integral to social activities and settings. Conventional location-led approaches use a pre-defined spatial unit to associate data and space with places, which cannot capture the richness of urban places, i.e. spatial extents and their dynamic functions. This paper develops a name-led framework to overcome these limitations in using social media data to study urban places. The framework first derives place names from georeferenced Twitter data combining text mining and spatial point pattern analysis, then estimates the spatial extents by spatial clustering, and further extracts their dynamic functions with time, which makes up a complete place profile. The framework is tested on a case study in Camden borough of London and the results are evaluated through comparisons to the Foursquare Point of Interest (POI) data. This name-lead approach enables the shift from space-based analysis to place-based analysis of urban space.

Keywords: platial representation, place-based GIS, social media data

1. Introduction

Place is a concept that is fundamental to how people make sense of geography (Tuan, 1977). Conceptually, a place can be described as a specific named location where specific activities take place at specific time (Roche, 2015; Tuan, 1977). Urban places are particularly complex and dynamic due to the high concentration of activities and people that may change throughout the day (Batty et al., 2012; Chan, Vasardani, & Winter, 2014). A good understanding and representation of urban places would be beneficial for many applications, particularly within industries that rely on information about people and their activities, such as urban planning, retail, marketing and transportation (Arribas-Bel, 2014; Batty et al., 2012; Cronin, 2008; Davoudi, 2003).

Unfortunately, it is inherently difficult to efficiently harvest place-related datasets as places are dynamic social constructs and different members of the public may identify them differently (Jenkins, Croitoru, Crooks, & Stefanidis, 2016). Thus a place dataset should attempt to reach a consensus through the consideration of numerous human actors' perceptions. In recent years, new technologies, including Web 2.0, have greatly increased the supply, velocity and availability of Volunteered Geographic Information (VGI) (Elwood, Goodchild, & Sui, 2012). Among the sources of VGI, georeferenced social media data has both semantic information (text) and high-resolution spatial and temporal information, making it a promising resource for place-related studies (Elwood et al., 2012). Given that a large share of georeferenced social media data tend to describe places and activities in near-real time, they offer valuable opportunities to gain more insights into complex urban environments (Jenkins et al., 2016).

This paper presents a methodological framework for identifying and profiling urban local places from large sets of georeferenced social media data, with an innovative name-led approach of associating social media data to places. The framework firstly identifies place names from a large sample of georeferenced social media posts using a combination of text-mining and spatial point pattern analysis. Secondly, it identifies probable spatial extents of the identified place-names using spatial clustering. Thirdly, semantic and temporal characteristics of the places are described by analysing the associated data. The framework is demonstrated using a set of geotagged Tweets collected in Camden, Greater London.

The remainder of the paper is structured as follows: In section 2, we present the background to this study and related research. The methodological framework is described in section 3 and the case study and results are presented in section 4. Section 5 presents the conclusions and directions for further research.

2. Background

2.1 Space and place

A place is more than just a simple physical space that is solely represented by its spatial location and geometric form; it is also an experiential construction of people (Relph, 1976; Tuan, 1977). Historically, in geographic information science (GIScience), most place-related research has been simplified to spatial analysis for the ease of computation, with places represented spatially as points or polygons(Longley *et al.*, 2005; Goodchild, 2010). In contrast to space-based analysis (spatial analysis), the concept of place-based analysis (otherwise known as platial analysis), which treats place as a notion that has vague and dynamic spatial and semantic attributes, has attracted increasing attention over the last decade (Agnew, 2011; Goodchild, 2015; Goodchild & Li, 2011; Purves & Derungs, 2015; Roche, 2015). Platial analysis offers new insights beyond traditional spatial approaches when determining the interaction between people and their environment because human cognition and activity are more aligned with places rather than geometric space (Goodchild, 2015). However, to apply platial analysis, place information needs to be better extracted, organised and formalised, which remains a challenge.

2.2 Social media and place

Many efforts have been made by geographers and urban planners to construct the profiles of places. Previously, they were mostly created with data sources produced by urban institutions, such as remote sensing (RS) satellite images, census data and gazetteers {Formatting Citation}. Unlike official data sources, crowd sourced data represent a bottom-up means of generating information about the world and may provide additional insights that could have been neglected by practitioners (Goodchild, 2007). As a well-known product of VGI, OpenStreetMap (OSM) is a platform where volunteers can create and edit geographic features and the related descriptions of maps, and its overall objective is to create a set of open, free, digital maps through crowd contribution (Haklay and Weber, 2008). Although it is a crowd-sourced means of describing places, when contributing one's local knowledge in OSM, there are still many restrictions and rules, and it is mostly based on a static perspective.

Social media platforms enable users to upload information about real-world phenomena in real-time, which can be assigned to predefined check-in points or geolocated precisely using a mobile phone's positioning technology. These data can take different forms such as text, photos, videos and GPS tracks. For example, a Twitter user might describe an activity occurring at the place they are visiting, and such information can be used to reflect aspects of the place, such as its function, popularity and meaning that user attach to it. Given that places are social constructs, and many are informal and invisible to officially produced datasets, social media data present an invaluable opportunity to acquire relatable information about places. Despite their ambiguities, the integration of places into everyday life has meant that there is still well-founded interest in place data to understand people and their activities (Goodchild and Hill, 2008).

Many discussions on social media platforms tend to be place focused. Lansley and Longley (2016) segmented a large sample of geotagged Twitter posts using topic modelling techniques. They found 10% of posts to be solely attributed to locations, usually through check-ins or photography. In addition, a larger share of messages described activities and events and many of those also referenced place names. The study revealed that there is a correspondence between what people describe on Twitter and the places that they inhabit or visit. For instance, retail centres all experienced high concentrations of messages about fashion and shopping.

There have also been efforts to harness social media data to generate crowd sourced information on the geography of places. Keßler et al. (2009) for example experimented with generating bottom-up gazetteers using geotags from photo sharing web services, whilst both Hollenstein and Purves (2010) and Goodchild and Li (2011) explored the spatial extent of place names identified from georeferenced Flickr posts using kernel spatial smoothing techniques. For each of these studies a prior understanding of place names or local naming heuristics were required. Such approaches cannot detect informally named places that may have important functions in everyday life. One possible alternative is to use a crowd sourced dictionary of places. For instance, Adams and Janowicz (2015) used Wikipedia data to identify and understand places. However, no such alternative is available for spatial data and we must also be cognisant that some places might have multiple names.

2.3 Associating spatial data to place: location-led approach

There are fundamental barriers to accurately detecting and representing places from georeferenced social media data. Primarily, it is difficult to link all posts to the places they describe where pre-determined points of interest (POI) services are not used. Data typically need to be aggregated and bound to established frameworks in order to generate generalisations on places. Usually, urban place studies that harness georeferenced social media follow a location-led approach which associates point-like data to the spatial units that represent places using spatial joins. In such approach, the study area needs to be partitioned into smaller units initially, or a layer of spatial units needs to be defined. Pre-defined spatial units can be grids, administrative units, roads or user-defined catchment areas (Cranshaw, Hong, & Sadeh, 2012; Jenkins et al., 2016; Lai, Cheng, & Lansley, 2017;

Quercia, Schifanella, Aiello, & McLean, 2015). This approach is suitable for statistical analysis from a general perspective because the boundaries can be clearly defined and can be linked to alternative datasets.

However, analysing urban places with a location-led approach has many limitations. The spatial units are often designed for specific purposes, such as census surveying, but urban places in our everyday communications may not match these pre-defined units. The boundaries of urban places are vague and not fixed, and are likely to vary between people and communities, possibly also changing throughout time (Gao et al., 2017; Goodchild & Li, 2011). In addition, many places may overlap, especially in city centres, so it may not be appropriate to associate all data falling within the same spatial unit to a single location. For example, there may be shops, offices and restaurants in one area (unit) and people may have different reasons for their visit. Therefore, the generated knowledge for a single unit may be a mixture of the results of all the places within it. This diversity of information is lost by considering each unit as a single place. For all these reasons, the location-based approach has many drawbacks for studying places using social media data.

3. Generate Place Profiles With A Name-led Approach

3.1 The concept of place profile

A place can be conceptually described as a named location with a certain spatial extent, where specific activities occur at particular times (Tuan, 1977; Roche, 2015). However, a formalised definition of place, which allows these important features of a place to be digitally described and integrated has not yet been given in the existing literature. The commonly used place information formalisation schemes, such as gazetteers, present a place as a static object with attributes of place name, footprint and category (Hill, 2000). Similarly, Roche (2015) attempted to formalise the place information with three elements: name, event and location. However, places in the urban context are more complex and dynamic than such formalisations can describe. Based on the previous place studies, some essential elements of a good profile of a place can be derived.

Firstly, the name of a place is an essential element that allows it to be identified and referred to. The shared knowledge of place names allows geographic locations and other features of the place to be communicated in everyday interactions and activities, and recorded in text documents with the geographical context (Vasardani, Winter and Richter, 2013). Normally, an official set naming system (toponym) exists in every country, regulated by governing authorities for standardisation purposes. However, only certain place-names are officially authorised, and many more are unofficially recognised and adopted locally, termed as vernacular names (Vasardani, Winter and Richter, 2013; Purves and Derungs, 2015). Furthermore, in our daily communications, the name of a place can be temporally substituted by the name of the event or activities occurring there (Chan, Vasardani and Winter, 2014).

Secondly, location, as one fundamental element of a place, defines its physical position in the space in contrast to everywhere else (Agnew, 1987, 2011). The description regarding location of a place is not just about its position, but also its geometric form and spatial extent. The spatial extent of an area-type place may be vague unless it is officially defined, hence most places in daily communication do not have a crisp boundary in the mind of people who refer to them. Moreover, the boundary of a place is likely to vary between people and communities, and may also change through time (Montello *et al.*, 2003; Goodchild and Li, 2011).

Thirdly, the locations of the lived-world that we perceive as meaningful places are differentiated because they involve a concentration of our intentions, attitudes, purposes and experience (Lynch, 1960; Agnew, 2011). Although one place may have different meanings for various individuals and groups, there is nevertheless some common ground of agreement about the meaning of that place for all the citizens interacting with the place (Relph, 1976). The meaning of place is essential to understanding the place, but is difficult for it be represented in a concise and formalised description. Activities, on the other hand, can be recorded, measured and formalised. Therefore, it has been suggested by many researchers that the meaning of place can be inferred from activities (Cheshmehzangi and Heath, 2012; Zakariya and Harun, 2013).

Lastly, as Wagner (1972) indicates, "place, person, act and time form an indivisible unity. To be oneself, one has to be somewhere definite, do certain things at appropriate times". From the perspective of describing place, a place should be described as the context where certain people do certain things at certain times. The name, location, activities and meanings of a place can change over time, which makes the static description of place incomplete and inaccurate in many cases where up-to-date information is required (Batty *et al.*, 2012; Goodchild, 2013; An *et al.*, 2015). Time, therefore, provides a continuity to the experience of place, and place should be described in a dynamic manner.

Therefore, we put forth a new concept of describing place, namely the place profile, formalised according to equation 1, where P is the place, N the name, L the location, A the activity and T the time.

$$\boldsymbol{P} = \boldsymbol{f}(\boldsymbol{N}, \boldsymbol{L}, \boldsymbol{A}, \boldsymbol{T}) \tag{1}$$

The place profile is a collection of information surrounding "what is the place called", "where it is", "what activities are occurring there" and "how these change over time". Although the information related to a place is far more than what these four elements can describe, the basic information and characteristics of a place can be described if these four elements are addressed. Describing places according to such a structural definition integrates the spatial, temporal and semantic information, which can provide relatively comprehensive insight into each place and form a standardised basis of analysing, comparing and relating places.

3.2 A name-led approach to profile places

A place is a "named domain that can occur in human discourse (by contrast, references to latitude and longitude in human discourse are of course extremely rare)" (Goodchild and Li, 2011). Individuals normally use a name to refer to a place given that people are more familiar with communicating places through names rather than coordinates (Goodchild, 2015), so a name-led approach would be better in relating social media data to urban places than a location-led approach that has been discussed in section 2.3. Some researchers may use a pre-existing database of place names (e.g. gazetteers) as a means to find place names in social media data (Hollenstein and Purves, 2010; Vasardani, Winter and Richter, 2013). However, many place names used in social media data are vernacular names and abbreviations, which are very different compared to those official recognized names listed in gazetteers. Furthermore, social media

A name-led framework is developed here to profile urban places, consisting of four major steps, by taking the geotagged Twitter data as the initial input. It first identifies probable terms as place names from the Twitter text, then estimate the spatial extents of each term, representing the geographic boundary of each place; and next extract dynamic functions and generate profile of each place, and finally evaluates the outcome with a pre-existing point of interest (POI) database. The overall framework is presented in Figure 1 and the methods used in each step are introduced in the following subsections.

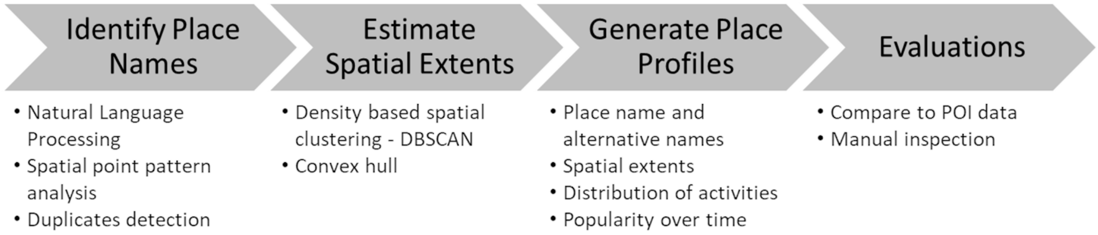


Figure 1 Framework of the name-led approach

3.3 Identifying place names

Place names should be derived from social media data itself in order to capture the full array of places discussed on social media that may or may not be captured by official datasets. Within the social media data, if a term has an unusually high concentration within a small geographic area then it might be a place name, or place-related. To extract the name of places, we combined natural language processing (NLP) and spatial point pattern analysis (SPPA) to identify candidate place names from Twitter text. An advantage of this method is that it can be repeated at any location world-wide without a prior list of place names. This includes the following three procedures: natural language processing, spatial pattern analysis and duplicates removal.

3.3.1 Natural Language Processing

A series of NLP techniques were applied to clean and format the Twitter text and reduce the size of the term vocabulary. First, numbers, punctuation and URLs were removed from the text, and then stopwords (i.e., the common words that do not have specific meaning, such as 'I', 'and', 'the') were removed according to the English stop word list from the SMART information retrieval system (Lewis *et al.*, 2004).

After text cleaning, text strings were split into terms through tokenization. Tokenization is a method that splits a string into separate terms based on the space between them. We consider that place names are not always unigrams. For example, "British Museum" is the name of a place but dividing the term into "British" and "Museum" removes the reference that is unique to that particular place. Therefore, an *n*-gram based tokenizer was adopted to allow individual tokens to represent a term of *n* words.

3.3.2 Spatial point pattern analysis with platial-score

After tokenization, the terms that are rarely used are further removed. Each remaining term in the list can be spatially represented by the distribution of georeferenced tweets that contain them. We have assumed that place-related terms will be concentrated in and around said places, therefore, we use spatial point pattern analysis (SPPA) to identify spatially concentration of terms. In this study, we use Ripley's K-function, to identify place related terms. Ripley's K-function is typically used to compare distribution patterns of a given point set with a random distribution (Ripley, 1977; Kiskowski, Hancock and Kenworthy, 2009). The point distribution is tested against the null hypothesis that the points are independent and identically distributed in space. For a given radius r, Ripley's K can be defined according to equation 2:

$$K(r) = \frac{A}{n(n-1)} \sum_i \sum_j I(d_{ij} \leq r)\, k_{ij} \tag{2}$$

Where A is the total area of a spatial point set X, n is the total number of points. $I(d_{ij} \leq r)$ is an indicator variable that equals 1 if the distance d_{ij} of a pair of points i and j is no larger than distance r, zero otherwise. If an edge correction method is defined, k_{ij} is the edge correction weight. In this research, border correction is selected. The expected value $K(r)$ for a random Poisson distribution is πr^2. The results of Ripley's K-function can be difficult to interpret and a number of variations have been proposed, such as the K-function normalized by area (L function), or by area and radius (H function). The L-function $L(r)$ (Besag, 1977) is a transformed $K(r)$, so that its expected value is a linear value r, instead of πr^2 (equation 3).

$$L(r) = \sqrt{K(r)/\pi} \tag{3}$$

$L(r)$ can be further normalized to give the H-function $H(r)$ (Kiskowski, Hancock and Kenworthy, 2009), which has an expected value of zero (equation 4)

$$H(r) = L(r) - r \tag{4}$$

Interpreting the results of $H(r)$ is much more straightforward; a positive $H(r)$ indicates clustering over that spatial scale, whereas a negative value indicates dispersion. Figure 2 shows the K-function, L-function and H-function for a sample of 1,000 geo-located Tweets, drawn randomly from the dataset described in section 4.1. On the figure, black lines represent the observed values and red dashed lines represent the expected value of a random distribution. The plots demonstrate the differences of these three functions and demonstrate that randomly selected Tweets display a clustering pattern that reflects the spatial distribution of population in the city.

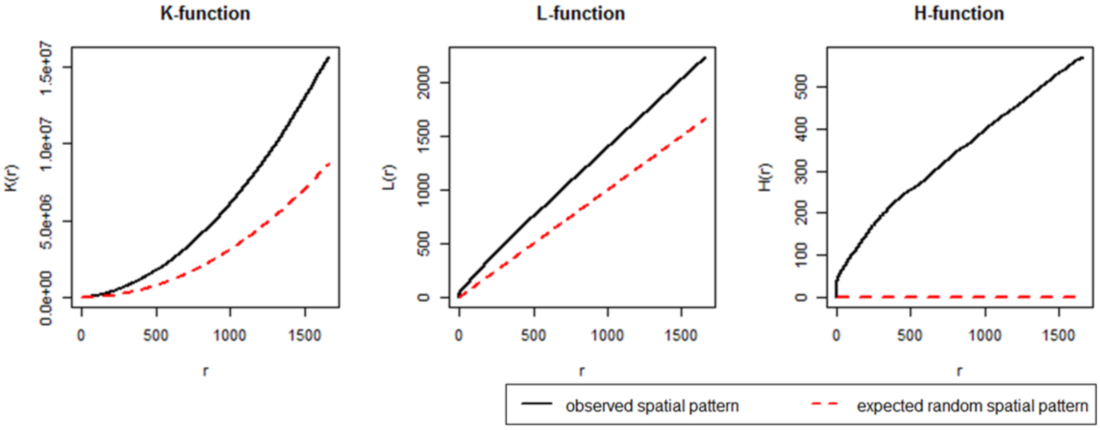


Figure 2 Ripley's K function, L function, and H function

Although Tweets in general are spatially clustered because of the underlying spatial distribution of the users, the level of clustering still varies for different terms. To demonstrate this, Tweets of three terms are selected and examined in Figure 3: 3016 Tweets containing “nice” (a term with a low *H(r)* chosen at random), 3520 Tweets containing “Euston” and 1707 Tweets containing “Euston railway station”. The map on the left shows the spatial distributions of the three terms and the plot on the right displays the H-function values of the three groups of Tweets. It can be observed from Figure 3 that Tweets of the term “nice” have a relatively dispersed pattern, spreading over the study area, while Tweets of “Euston” are more clustered. Many tweets of “Euston railway station” are concentrated in a few specific points because many of them were generated by check-in behaviours in Twitter and are geolocated to specific locations. It is worth noting that even though the number of Tweets containing these three terms is different, the results of $H(r)$ are comparable because the measure is based on point density rather than counts. The H-function values of a sample of 1000 randomly selected Tweets are also displayed as a grey line in the plot for comparison. As the plot shows, a point set which is spatially clustered has a higher $H(r)$ value. The maximum score of $H(r)$ can be extracted as a simple index of the level of clustering of a spatial point pattern. It can also be observed that the value of r that maximizes $H(r)$ indicates the radius of maximal aggregation: the radius of an area in which a centred test point on average contains the most points per area. In other words, the smaller the value of r when $H(r)$ reaches its maximum, the smaller the spatial extent of the cluster.

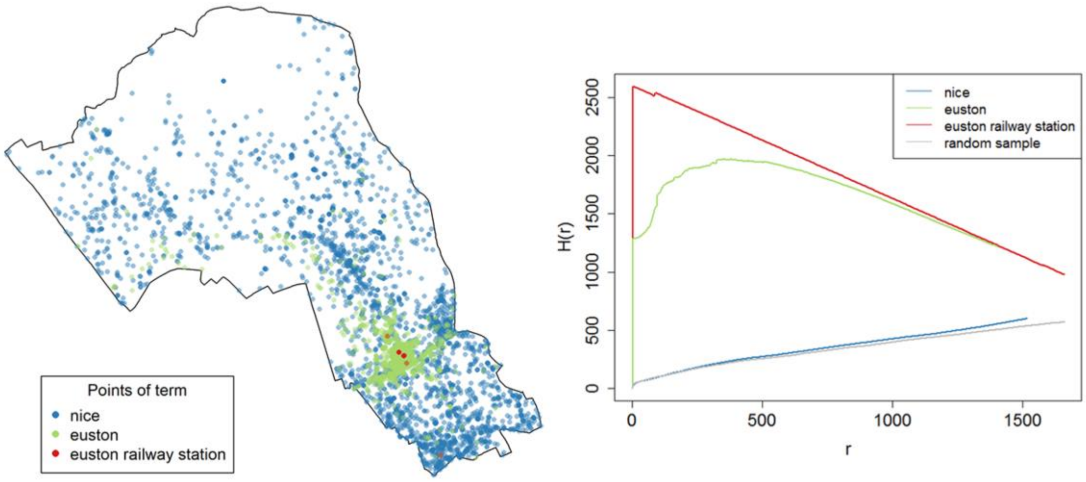


Figure 3 Spatial point distribution patterns and H-function values of different terms

Although some places are meaningful to individuals (e.g., someone's home), from a general perspective, they are not as important to identify as places that are visited and used by larger groups of people. Therefore, we only investigate places from a collective perspective. Therefore, in addition to spatial clustering patterns, it is also important to take the number of unique users that mention each term into account to prevent extremely active users from skewing the results. We introduce an index score to rank terms according to their spatial point distribution and popularity among users, which is termed the platial-score $P(t)$ and is defined in Equation 5.

$$P(t) = H_{max}(t) * \log(N_u(t)) \qquad (5)$$

Where, for each term t in the generated term list, $P(t)$ is calculated by multiplying the maximum H-function value $H_{max}(t)$ and the log of the number of users $N_u(t)$. The platial-scores of all terms are then normalised using range standardisation to facilitate comparison.

Terms with a high palatial-score are more likely to be place names. In order to separate platial terms from non-platial terms, a threshold is defined on $P(t)$. The threshold is defined quantitatively using a test statistic as follows: First, a random sample of terms (e.g. 10%) is selected and manually annotated by three volunteers. Terms that are related to place are marked as "true", while the others are marked as "false". After each volunteer marking the sample terms independently, terms are annotated by the mark that agreed by majority. For a given threshold of $P(t)$, the terms that are above the threshold are tagged as "positive", while the remainder are tagged as "negative". Terms that are marked both "true" and "positive" are the "correct" terms. The performance of this threshold can be evaluated with three indices; precision, recall and F-score. Precision measures the percentage of "correct" terms in all "positive" terms. Recall measures the percentage of "correct" terms in all "true" terms. F-score is the harmonic mean of precision and recall; a high F-core indicates both high precision and recall. The performance of different thresholds can be evaluated, and the one achieving the highest F-score is set as the final threshold.

3.3.3 Duplicate place terms detection

It is common for places to be distinguished by multiple different place names, which might be especially common in social media. For example, “eus” is the abbreviation of “Euston” and both terms are commonly used. This situation is difficult to deal with automatically because there is a risk of grouping similarly distributed place names that actually represent unique places. In addition, different tokens could be generated from one string during *n*-gram tokenization, for example, “euston railway”, “railway station eus”, “station eus” are tokens split from the string “euston railway station eus”.

Practically, each place term has a list of Twitter IDs associated with it. The Jaccard index $J(X,Y)$ was calculated for each pair of Twitter ID lists, according to equation 6. A higher $J(X,Y)$ indicates greater overlap between two lists X and Y, and a higher likelihood that these two terms originate from the same string. A Jaccard distance matrix $D(X,Y)$, was created from the pairwise $J(X,Y)$ of all terms (equation 7).

$$J(X,Y) = \frac{X \cap Y}{X \cup Y} \tag{6}$$

$$D(X,Y) = 1 - J(X,Y) \tag{7}$$

To efficiently identify the groups of similar place terms, hierarchical clustering (Ward, 1963) was applied based on the $D(X,Y)$. This approach merges place terms into groups in a bottom-up way; place terms close to each other based on their co-occurrences in Tweets are clustered. Each cluster can be labelled by the term with the highest platial-score to generate a cleaned list of place names.

3.4 Estimating spatial extents of the identified places

Once a list of place names has been derived, the Tweets in which the name within the list being mentioned can be extracted by matching the place names with Twitter text. This association process is purely based on string matching. To supply a more complete data set, the alternate names of the place (terms in the same cluster in previous step) are also used in matching. For example, “euston railway station”, “railway station eus” and “station eus” are alternate names for “London Euston railway station”, and Tweets that include any of these terms will be extracted and associated with the place. In addition, one Tweet may be associated to multiple places if it has multiple place terms simultaneously.

Tweets that associated to a place are used to refer the spatial locations of the place, and the highly concentrated area could be considered as the core spatial extent of the place. This is necessary given people communicate not only place name but would like to know where the place is spatially, which normally not a single geometric location but an area with boundary. In conventional gazetteer and POI data, the location of a place is normally represented by a point with a pair of coordinates. The point-based representation is convenient in spatial analysis, such as calculating distances and spatial joins. However, places in real life are not points but areas, and different places may have different levels of spatial influences on their neighbourhood areas, leading to varying spatial extents. The area-based representation may be more appropriate to represent the spatial extent of places.

One issue with social media data is that users might mention the names of places that they are not proximal to. For instance, Keßler et al (2009) identified that many geotags were located miles away from the places they were describing. While there is interest in harvesting these messages to estimate the social influence of places, they will confuse models that are trying to identify the geographic extent of places.

A spatial clustering approach, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester, Kriegel, Sander, & Xu, 1996), is used here to estimate the spatial extent of the place due to the following reasons: (1) clusters of arbitrary shape are detectable, such as linear, concave, oval, etc. (Gomide et al., 2011; Zhou, Frankowski, Ludford, Shekhar, & Terveen, 2007); (2) in contrast to some clustering algorithms, such as K-means, the number of clusters sought does not need to be specified; (3) the algorithm naturally handles noise by allowing isolated points to be unassigned (Sacco, Motta, You, Bertolazzo, & Chen, 2013).

DBSCAN organizes data points to obtain dense groups (clusters) that are separate from sparse data points (Zhou et al., 2007). The algorithm requires two parameters as input: (1) the radius of the search circle around a data point, usually termed ε; (2) the minimum number of points that must be in the circle of radius ε to be considered as a group of related points, usually termed $minPts$. The number of Tweets associated with a place ranges from dozens to thousands, meaning a fixed value of $minPts$ is not suitable. Therefore, the $minPts$ is expressed as a percentage of the total points that a cluster should contain. To avoid places being defined by the activity of an individual user, a filter of minimum number of users can be applied to qualify a cluster.

After the clustering, a convex hull is generated around each cluster to represent the spatial boundary of the place (Zhang, Noulas, Scellato, & Mascolo, 2013). A convex hull approach is a computationally efficient method that used to represent the minimum bounding shape for a set of points.

3.5 Generating place profiles

After identifying places and associating data to them, the four elements of each place (i.e. name, location, activities and time) can be revealed and its profile can be generated.

3.5.1 Name element

A list of term clusters was generated after duplicate detections in section 3.1, each cluster can be labelled by the term with the highest platial-score. The label term is suggested as the name of the place, while other terms in the same cluster can be presented as its alternative names.

3.5.2 Location element

Given a set of spatial points that are identified as related to a place, the location and core area of that place can be approximated according to spatial analysis techniques, such as kernel density estimation (KDE) (Cheng and Shen, 2018) and spatial clustering. To simplify the visualisation, in this paper, the spatial extent of each place is estimated through DBSCAN clustering as explained in section 3.2, and spatially represented by the convex hull of the clustered points. It is noted that the spatial influence of a place is not fixed and may vary through time and, although not shown here for brevity, the boundary of a place can be displayed differently according to the clustering results of data in different time periods.

3.5.3 Activities element

To infer relevant activities and semantic information of a place, a topic modelling approach, Latent Dirichlet Allocation (LDA) (Blei, Andrew and Jordan, 2003) is applied to analyse the topics discussed in Tweets that are associated to the place. As an unsupervised generative model, LDA classifies words into topics and represents documents (e.g., Tweets) as mixtures of topics with various probabilities. Detailed explanation of LDA is beyond the scope of this paper and can be found in the following references (Blei, Andrew and Jordan, 2003; Griffiths and Steyvers, 2004; Lai, Cheng and Lansley, 2017). After topic modelling, each Tweet has a probability distribution indicative of belonging to multiple topics. However, as most Tweets are very short, it is assumed that each Tweet has a single topic and they are labelled accordingly. The activity element of a place, or its function and meaning, can be derived through analysing the distributions of the topics in Tweets associated to the place.

As per the nature of user-generated social media data, there is much noise in Twitter data, where some users produce a plethora of Tweets featuring similar content. This might include, as an example, the Tweets of weather forecasting and news produced by public service accounts. Since the noise is not always identical (e.g., weather forecasting posts), to reduce the influence of such noise, instead of removing all the identical Tweets produced by each user, we limit the contribution of each individual user to the influence of each activity to the place as one. If this user visits another place, posts about another activity or does so at another time, their data will be considered as new contributions. With this concept in mind, the bias from such noise is prevented while, at the same time, other useful information is not lost because there is no need to remove Tweets in advance. For each place ($\boldsymbol{p}$), the number of unique users for each activity ($\boldsymbol{a}$), can be counted, represented as $\boldsymbol{N_{p,a}}$. The importance $S_{p,a}$ of the activity to the place can be calculated as illustrated in Equation 8.

$$\boldsymbol{s_{p,a}} = \frac{\boldsymbol{N_{p,a}}}{\sum_a \boldsymbol{N_{p,a}}} \tag{8}$$

3.5.4 Time element

Regarding the temporal information, the relative distribution of Tweets over time are analysed to reveal the dynamic popularity of the place. In this framework we present the average number of messages by hour of the week, although the data can be aggregated according to other time schemes as Twitter has detailed time stamp information. It should be noted that the spatial extent and activity distribution of one place may change in different time periods.

At this stage, the spatial, temporal and semantic information of a place can be extracted and organised as its profile, which describes its basic information and reveals its characteristics.

3.6 Evaluation

There are two steps in the evaluation. We first compare the identified place names with one of the most popular and widely used social media place databases, Foursquare POIs. Information from the Foursquare POIs are verified by the company as well as their user communities and thus may be treated as a reliable reference for a ground truth exercise (Hu, Mao and McKenzie, 2018). If a place name can be matched with a Foursquare POI, it is marked as a correctly identified result. The rest of the terms in the place names list are then manually inspected according to local knowledge and online searching, the terms will be annotated and grouped depends on their types and actual meanings, for example, whether they are real places.

4. Case study

4.1 Data and case study area

To demonstrate the methodological framework, 361,388 Tweets with geotags in the London Borough of Camden covering the entire year of 2013 are used. While Tweet data are used for the case study, the framework proposed can be applied to any spatially referenced, timestamped data with text descriptions, such as travel blogs and geotagged photos (with text). The research pipeline can also be applied in data in other language context since the terms were treated as tokens in data analysis, only the text mining and topic modelling steps will need to be changed before the spatial and statistical analysis.

Camden is located in the centre of London (see map on the left in Figure 4). Figure 4 also displays the spatial distribution of Tweets within the borough. The Tweets are spatially concentrated in the southern part of Camden, which is a major commercial sector, while the Northern half is primarily residential and has a far lower density of Tweets.

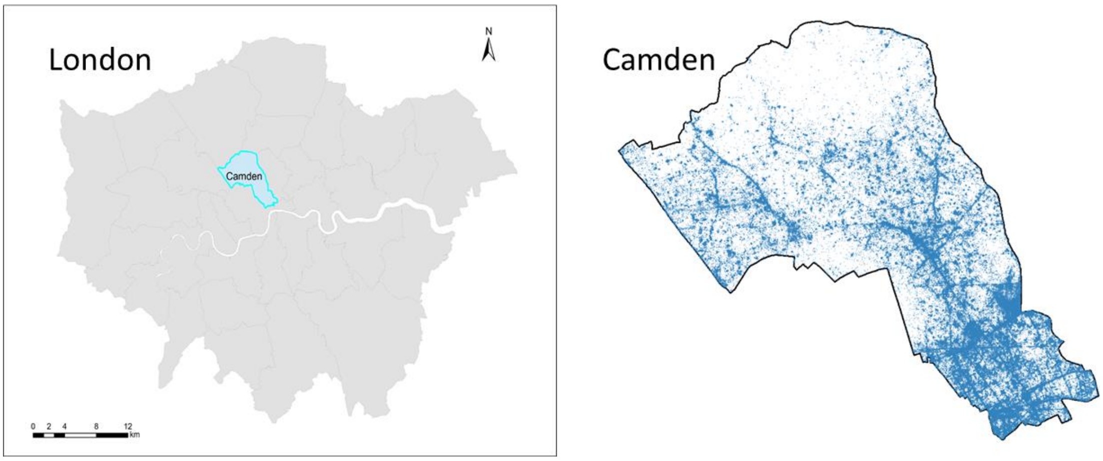


Figure 4 Case study area and Twitter data

4.2 Data preprocessing

The raw Twitter posts were cleaned and formatted to achieve a better accuracy of topic modelling (Hong and Davison, 2010; Lansley and Longley, 2016; Wang, Ye and Tsou, 2016; Lai, Cheng and Lansley, 2017). The texts of the Tweets were first imported as a “corpus”, which is a data structure to manage a collection of documents (Wallach, Mimno and Mccallum, 2009). The texts were then passed through text-mining steps, which include removing whitespaces, numbers, punctuation and URLs. Stop words were removed according to the English stop word list from the SMART information retrieval system (Lewis *et al.*, 2004). By doing these, such common characters and stop words are removed and the remaining words are more likely to be meaningful, which increases the chance of generating good quality and distinctive topics in the following topic modelling process. Furthermore, the process of “stemming” was also applied, which seeks to diminish inflected words to their stem form by removing suffixes of the words (e.g., “ing”, “ed”, “er”, etc.). These two steps are conducted to ensure different forms of words that have same meaning will be treated as one input.

A collapsed Gibbs sampler (Resnik *et al.*, 2009) was used to fit the LDA model and point estimates of the latent parameters were returned using the state of the last iteration. As suggested by Griffiths and Steyvers (2004), to result in a fine-grained decomposition of the corpus into topics that address specific activities, we chose topic number as 30, the hyper parameters α as 0.1 and β as 0.1 after empirical results of different parameter settings of LDA model and manual inspecting the results. To ensure the convergence of the model, 1000 iterations were applied. Among the 30 generated topics, many were not relevant to activities and places, like, for example, online slang and profanity. In addition, several topics were very similar, and therefore should be merged to yield more distinctive topic groups. As we were expecting to analyse activity or place-relevant topics, we extracted topics from the 30 generated topics and 10 activity-relevant topic groups were the result through referring to the topic classification schemes of POI categories[1]. The selected 10 topic groups, represented by their top 20 words ranked by their probabilities of belonging to the topic, are found in **Error! Reference source not found.**, and a label was assigned to each topic for ease of interpretation in later analysis. Tweets assigned to topics of the selected 10 groups were given the corresponding labels, while the rest of the Tweets were labelled as belonging to other topics.

Table 1. The selected 10 topics with labels

Topic ID	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10
Top 20 Words	eat	watch	park	great	day	year	hair	train	night	game
	food	love	photo	event	work	day	wear	station	tonight	play
	drink	song	post	job	sleep	time	girl	bus	uk	win
	coffe	film	street	work	back	school	black	run	show	arsenal
	breakfast	listen	garden	today	night	today	love	walk	parti	team
	dinner	music	st	busi	time	work	dress	home	great	footbal
	tea	time	pic	meet	bed	week	white	car	amaz	fan
	lunch	play	hotel	interest	home	start	red	railway	hous	good
	chocol	show	squar	detect	feel	ago	today	stop	pm	chelsea
	chicken	good	bridg	talk	week	tomorrow	nice	cross	club	goal
	wine	video	hous	social	tomorrow	havent	cut	underground	museum	player
	cake	tv	road	market	wait	long	shoe	drive	theatr	season
	cook	amaz	uk	good	hour	gonna	colour	road	love	mate
	hot	sing	hill	manag	today	exam	short	time	art	today
	pizza	movi	town	day	tire	life	nail	heathrow	excit	man
	bar	make	tower	googl	weekend	miss	blue	lhr	live	hes
	burger	live	market	uk	morn	back	top	wait	ticket	tonight
	egg	night	camden	team	earli	month	boy	ben	hall	england
	beer	album	palac	check	start	uni	eye	tube	music	great
	make	tune	view	read	tonight	finish	put	airport	royal	score
Label	**Food**	**Entertainments**	**Outdoor & Sightseeings**	**Social & Business**	**Work & Life**	**Education**	**Fashion & Style**	**Travel & Transport**	**Arts & Show**	**Sports & Games**

[1] https://developer.foursquare.com/docs/resources/categories

4.3 Identified place terms

After text cleaning, over 2.7 million tokens were generated using an *n*-gram tokenizer for values of *n*=1,2,3. To reduce the computational intensity of the subsequent steps, the vocabulary is pruned by removing the terms that occur in less than 100 documents, resulting in 2,847 tokens. The Tweets that correspond with each token were then extracted. The user number was summed by counting the unique user IDs within the Tweet list for each token. Ripley's L-function was then applied to the Tweet coordinates to estimate the spatial point distribution pattern.

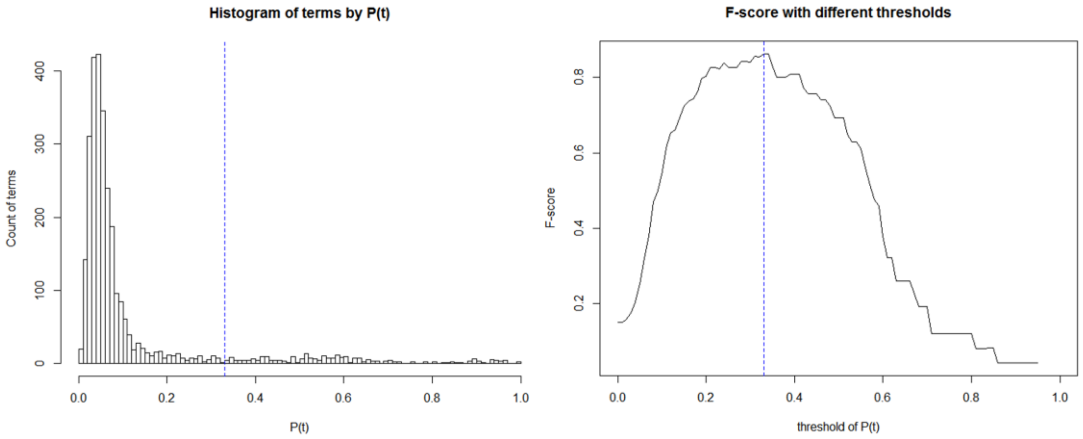


Figure 5 Distribution of standardized platial score and the identified threshold

The platial-score $P(t)$ of each token in the pruned vocabulary was calculated. A histogram of the normalised platial-scores of all tokens is displayed in Figure 5 (left). To identify the threshold of $P(t)$ we conducted an evaluation. A random 20% sample of the tokens (568 in total) were selected and manually annotated. We iterate the threshold from 0 to 1 with a step of 0.01, calculating the F-score at each threshold. The result is displayed in Figure 5 (right). The F-score reaches its maximum (0.86) at threshold P(t) = 0.33. Using this threshold, which is indicated by the blue dashed line in the charts, we identified a total of 252 terms that are likely to be place names.

To identify overlapping terms, the pairwise Jaccard index of the selected terms was calculated. Next, hierarchical clustering was applied to the terms based on the Jaccard distance matrix. The resulting dendrogram is shown in Figure 6. The height of hierarchical tree corresponds to the Jaccard distance, which ranges from 0 to 1. Users can cut the tree to produce any number of clusters from 1 to 252. After inspecting the clustering tree at different heights, we specify the height at 0.5 to generate 138 clusters, which successfully merges many overlapping terms while avoiding the merging of different places. A Jaccard distance of 0.5 means 50% of Tweets in the subsample contain overlapping terms, which suggests these two terms are very likely to originated from longer terms that refer to the same place. The term that has the highest platial-score within each cluster is chosen as the place name label of this cluster. The remaining terms in a cluster are labelled as its alternative place names.

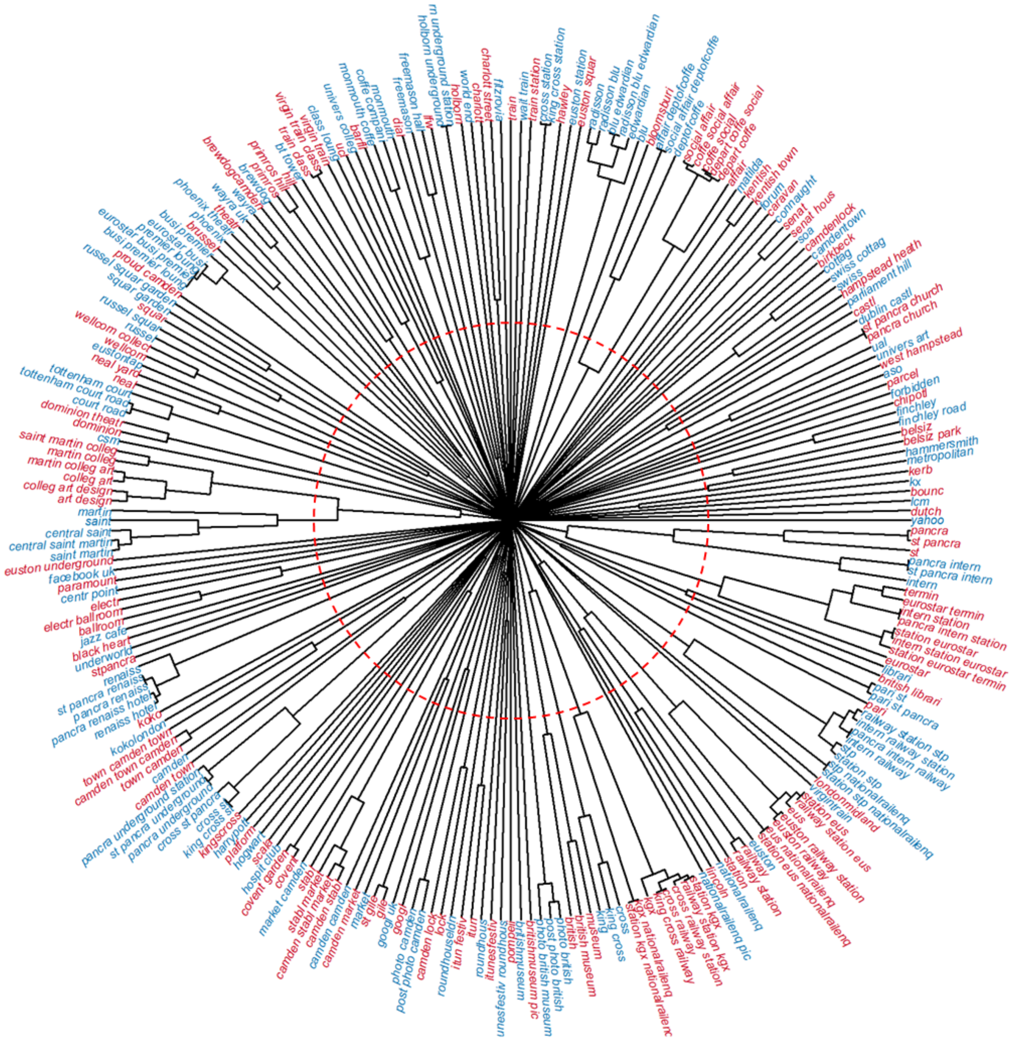


Figure 6 Hierarchical clustering dendrogram of the place terms

After these steps, the majority of the noise terms which are not relevant to places are removed and only the terms that are most likely to be place names remain. The parameters in this filtering process can be tuned by the user according to their needs and preferences. At this stage, it is feasible to go through the filtered term list and identify terms that refer to the same place which cannot be detected in the previous step, based on local knowledge. A summary of the steps in identifying place names is illustrated in Table 2.

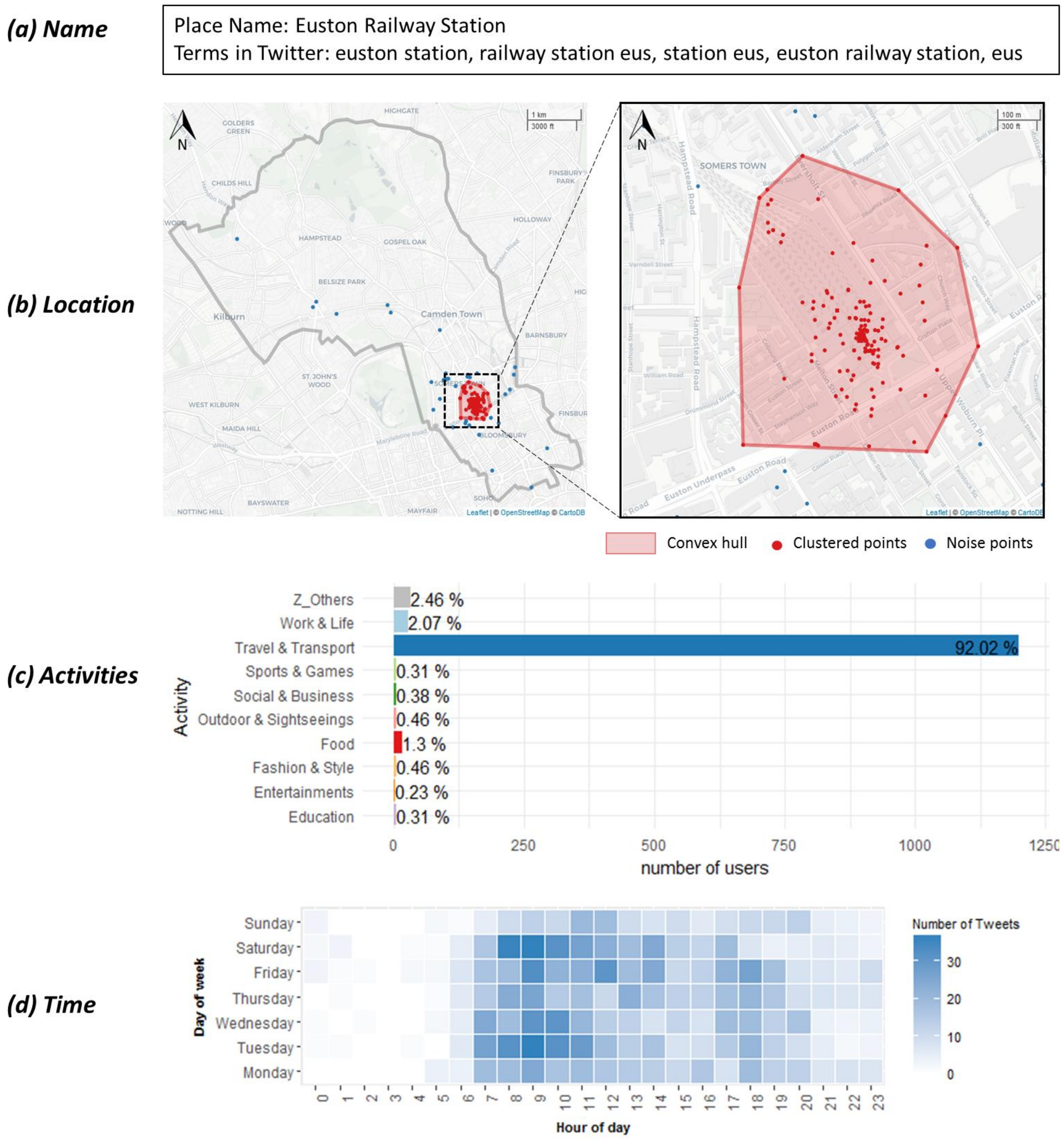
Table 2. Procedures for extracting place-terms

Process	Description	Count of terms
Tokenization	Extract n-gram tokens from all the Tweets.	2,737,897
Filtering by doc frequency	Remove tokens which occur in less than 100 Tweets.	2,859
Filtering by platial score	Filter terms according to the platial score, threshold is set as 0.33.	252

Clustering duplicated terms	Cluster terms that are referring to the same place, pick the term with highest platial score in each cluster.	138
Manual inspection	Manually inspect the place term list. Merge terms of the same place.	111

4.4 Spatial boundaries of the identified places

The distribution of the number of Twitter points associated with each of the 111 identified places follows an approximate power law, ranging from 100 to 8,274 with a mean value of 685. For each Twitter point set, we apply DBSCAN clustering to identify the dense clusters, which are the core locations of that place. Based on the result of empirical tests with different sets of parameters, to have a reasonable constraint on the size of the major cluster without splitting the data into too many small clusters, the search radius is set as 200 metres and the minimum points threshold is set to 10% of the investigating point set in this case. Most of the places only have one cluster, while a few places may have more than one cluster. The extracted clusters indicate the core locations of a place. A convex hull is used to approximate the spatial extent of the place. An example is shown in Figure 7(b).



Place Name: Euston Railway Station
Terms in Twitter: euston station, railway station eus, station eus, euston railway station, eus

Figure 7 Place profile of "Euston Railway Station"

4.5 Place profiles

The occurrence of place names on social media throughout the day and week may reflect its characteristics and the related activities. An example of the generated place profile of Euston railway station is observed in **Error! Reference source not found.**. The profile consists of four elements: 1) place name: “Euston railway station”. Other terms extracted from GSM data that are identified as associated to the place are also listed here as one place may have multiple terms (alternate names), such as “euston station”, “station eus”; 2) location: all the associated Tweets are plotted on the map, where the major cluster detected by DBSCAN is highlighted with a convex hull. The location and a rough spatial extent can be observed from the map; 3) activities: the distribution of the 11 activity topics are displayed as a bar chart here, the x-axis also indicates the number of users; 4) time: the Tweet counts of 24-hours over 7 days of the week are visualised here, each cell represents one hourly slot. The morning peaks and evening peaks at the Euston railway station can be observed from the temporal pattern shown in Figure 7(d). From the profiles presented, we can obtain basic information and some insights about this place. Based on our local knowledge of the place (i.e., the Euston railway station), the place names, location, prominent activity and temporal variance are correctly identified.

4.6 Evaluation of the results

Through matching, 75 of our 111 identified places can be directly linked to the POIs recorded in Foursquare database. Grouped according to the top-level category of POI, the counts and cases of the matched places are presented in Table 3. Most of the identified places are in “Art & Entertainment”, “Outdoor & Recreation” and “Travel & Transport” categories. In contrast, no places in the “residences” category are identified in this case study. It is noted that there are 2,085 Points of Interest (POIs) in total in Camden recorded in Foursquare (as collected in 2017), only a small portion of them have been identified through our approach. It is possibly because of popularities of these POIs vary a lot, a large proportion of which refer to singular outlets or small places that are rarely discussed by the users, especially in Twitter. However, according to our local knowledge of the study area, most popular places were successfully identified. More places may be detected if we decrease the threshold of term frequency or platial score, while which may increase the chance of getting noise terms at the same time.

Table 3. Summary of the identified places matched with POIs

Place Category	Count	Examples
Arts & Entertainment	15	British Museum, RoundHouse, KOKO
Colleges & Universities	4	University College London, Central Saint Martins College, Senate House, Birkbeck
Food	3	Caravan, Chipotle Mexican Grill, Monmouth Coffee Company
Nightlife Spots	4	Electric Ballroom, The World's End, Hawley Arms, The Parcel Yard
Outdoors & Recreation	20	Camden Town, Primrose Hill, Russell Square
Professional & Other Places	10	Google UK, Facebook London, British Library
Shops & Services	4	Camden Market, Camden Lock, Camden Stables Market, Forbidden Planet
Travel & Transport	13	Euston Railway Station, King's Cross Railway Station, Eurostar Business Premier Lounge
Others	2	BrewDog Camden (beer bar), KERB KX (street food)
Residences	0	

The rest of the places in the list that do not match the POI data were subsequently manually investigated and labelled. The results are shown in Table 4. The analysis confirmed that the Twitter data successfully identified some places that were not in the POI data. The Tweets also located events as places, since the terms are frequently mentioned in specific locations, which were not considered in the POI database. However, there were some discrepancies. For instance, the Twitter data also identified some peripheral places where the POI fell just outside of the study area. In addition, the terms “Paris” and “Brussels” were found because of the discussions of the international trips in the St Pancras international railway station. The methodology also misidentified train brands as places because consumers tend to complain about train delays to train companies when they are at the major stations.

Table 4. Summary of places that do not match with POIs

Type	Count	Examples (Descriptions)
Events	4	itunesfestiv (itunes Festival) lfw (London Fashion Week) lcm (London Fashion Week Men's) pompeii (Pompeii and Herculaneum exhibition in British Museum)
Places not recorded in POI	7	soa (SOAS university of London) barfly (Barfly, a pub) paramount (Paramount, a closed restaurant) …
POI outside of the area	3	covent garden (place in the boarder of study area) univers art (University of the Arts London, at the boarder of study area) camden (name of the general study area)
Place outside of the area	2	pari (Paris, destination of international train journey) brussel (Brussels, destination of international train journey)
Place-type terms	9	train station market …
Others	11	londonmidland (London Midland, train company) nationalrailenq (national rail enquiry, hashtag) harrypott (Harry potter 9 3/4 platform inside of King's Cross station) …

4.7 Advantages of the place profile over POI

Although the number of places we identified are smaller than existing POIs, when comparing such place profiles with conventional place databases such as gazetteers, or the POI data, many advantages can be observed: 1) Some irregular expressions of place names are identified using the name-led approach. This is helpful in detecting more associated data in research using online documents. If the query is made using the exact place name recorded in conventional data bases, for example “Euston railway station”, only part of the relevant data can be retrieved and the data that use other alternate terms such as “station eus” will be missed. 2) The location information stored in either gazetteers or POI databases are mostly a pair of coordinates. As well as the location, the areas influenced by the place can be identified using place profiles. This is important because it reveals where people view themselves to be, which may be different from where gazatteers, POIs or other geographies place them. 3) Normally only one category-tag is attached to the place in conventional databases, while the activity information in the profile reflects the values relating to multiple activities. 4) A detailed temporal variation can be observed from the profile, which helps to better understand the dynamic nature of the place. It should be noted that, while the example shown in Figure 7 (place profile for Euston station) is intuitive due to its clearly defined function, this may not be the case for other places whose function may be more diverse.

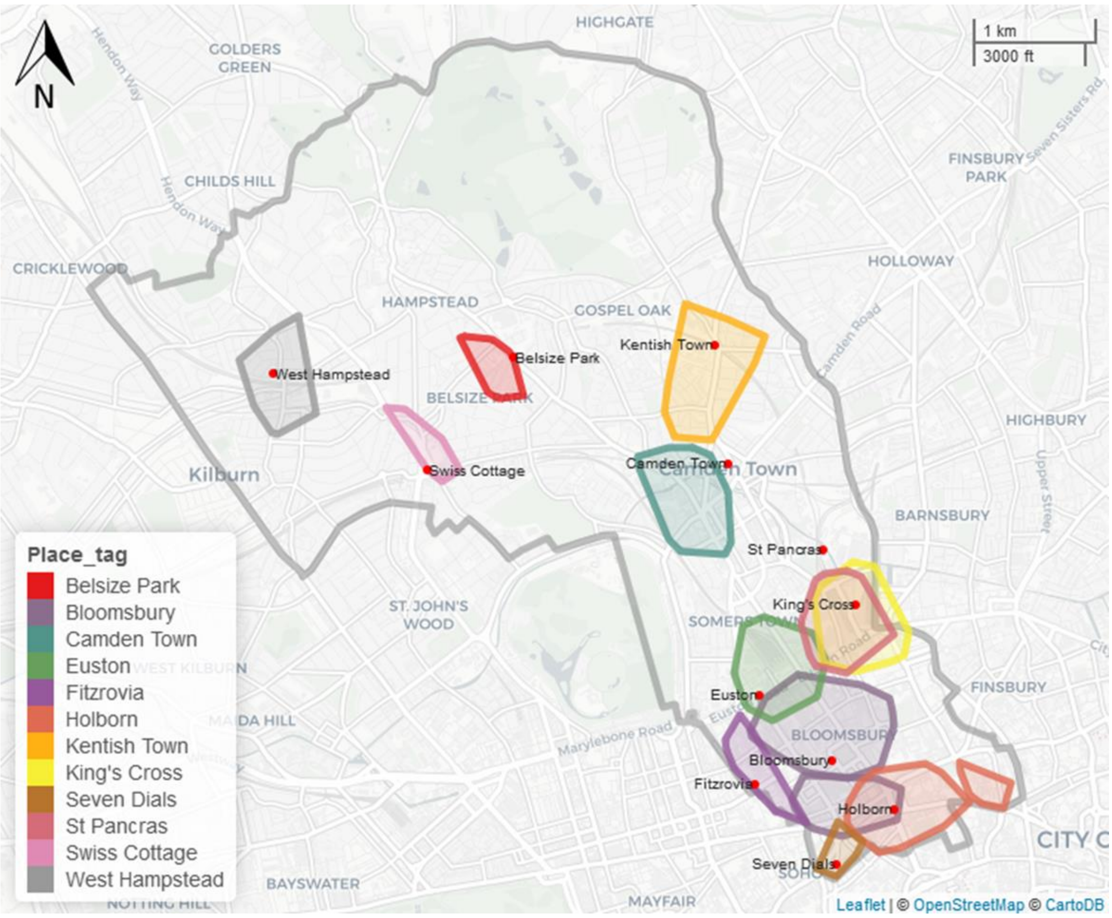


Figure 8 The identified neighborhood-level places and corresponding POIs

Another advantage of the framework presented here over standard POI type data is that it identifies the spatial extent of larger places, instead of simply representing them as points. For example, the method successfully identifies polygons for neighbourhoods in the study area. The neighbourhoods are mapped in Figure 8 and the POIs are shown for comparison. Interestingly, the location of the "St Pancras" POI falls outside of the extent identified by the Twitter data. This is largely because "St Pancras" is a neighbourhood-level administrative unit, and the location of it recorded in the POI data is the centroid of this unit. However, people mostly refer to "St Pancras international railway station" when such term (i.e. "St Pancras") is mentioned in daily communications or online posts. This highlights the mismatch when relating places in human cognition to officially defined place locations. Name-led place definition is able to identify these differences. In addition, by associating Tweets to places via place names instead of rigid pre-defined spatial units, the boundary of each individual place can be better represented, and the overlaps between places can also be clearly shown. This suggests that the name-led approach proposed in this research provides a new way of associating and representing platial data, from a place-based thinking.

5. Conclusions and Future work

This paper has presented a name-led framework for harnessing information on places from geotagged social media data. Demonstrating this through a sample of Tweets in Camden, London, we have presented a means of firstly, harvesting probable place names directly from Tweet text and then secondly, estimating the spatial extents of the identified places. Then profiles of the places are generated by analysing semantic and temporal information of the associated Tweets.

We should note that bias exists since only Twitter data is used to conduct this research, however, the research framework and methods proposed in this paper can be applied on many other data sources, especially geotagged textural documents. Each big data source is likely to over or under-represent certain groups, activities and places when they are repurposed to represent real-world phenomena (Lansley et al, 2018). For instance, social media may be more useful for understanding urban tourist attractions than they are for understanding sleepy rural villages. However, social media remains unique in that they generate large volumes of georeferenced information at a high velocity and from large numbers of people. It is fundamentally distinctive from official datasets which are infrequently updated and do not reflect the full spectrum of public perspectives.

Indeed, the quality of the list of identified place names could also be improved by exploring more advanced NLP techniques to filter out place names mentioned in such informal online documents. The overlap between the Twitter inferred places and the pre-existing POI database from Foursquare could also suggest that there is merit in using both types of place data in conjunction with each other to improve our understandings about social media activity and the real-world. We also have not explicitly taken into account varying accuracies of the geolocations associated with the Tweets. Depending on the device and location settings, the geolocations may be determined by GPS, mobile phone masts or both. Therefore, the accuracy may vary from a few metres to hundreds of metres, especially in dense urban canyons where satellite visibility and multipath have an effect. The design of the DBSCAN algorithm accounts for this to a certain extent, but future work could examine alternatives to the convex hull for determining the boundaries of places that take into account uncertainty.

Despite the limitations mentioned above, the general research framework and the innovative name-led approach proposed in this research breaks the conventional limitations of using pre-defined rigid spatial units to analyse places and create a more flexible and people-centred way of perceiving places. Instead of simply joining data to the rigid pre-defined spatial units, data are joined to places via place names, which corresponds to how people perceive and communicate about place. More importantly, this research is unique in that little to no prior knowledge of the bounding locations are required to harvest the place information, thus this could be repeated in other locations around the world to profile urban places without the need of a well-established place database. If dealing with data in another language, only the text mining and topic modelling steps will need to be modified by adopting the tokenization tool corresponding to that language. The rest of the research pipeline, which are mainly spatial and statistical analysis, can be easily transferred. In addition, a concept of place profile was proposed in this research, which helps to structurally organise the basic information of urban places, (including place name, location and boundary, activities and temporal information) that can be extracted and inferred from social media data. Thus, it is hoped that this research can assist the general shift from space-based analysis to place-based analysis in the context of geographical information science.

After profiling urban places, their connections can be further explored in future studies, which will benefit a wide range of urban place-related applications. For example, with such detailed and dynamic understanding of urban places achieved, site-selections for retail business can be more convenient. In fields such as urban planning, such knowledge can aid more efficient distribution of resources and planning of new infrastructure to maximize their usage. In marketing and tourism, more targeted planning can be made using the information gathered about places, specific to location, target audience and even time of day. In general, it is hoped that this research can assist in advancing the development of smart cities, where information about places can be better gathered, processed and utilised.

References

Adams, B., & Janowicz, K. (2015). Thematic signatures for cleansing and enriching place-related linked data. *International Journal of Geographical Information Science*, *8816*(October), 1–24. https://doi.org/10.1080/13658816.2014.989855

Agnew, J. (2011). Space and place. In *Handbook of geographical knowledge* (Vol. 2011, pp. 316–330). https://doi.org/http://dx.doi.org/10.4135/9781446201091.n24

Agnew, J. A. (1987). *Place and politics : the geographical mediation of state and society*. Retrieved from https://books.google.co.uk/books/about/Place_and_Politics.html?id=9EUVAAAAIAAJ

An, L., Tsou, M.-H., Crook, S. E. S., Chun, Y., Spitzberg, B., Gawron, J. M., & Gupta, D. K. (2015). Space–Time Analysis: Concepts, Quantitative Methods, and Future Directions. *Annals of the Association of American Geographers*, *105*(5), 891–914. https://doi.org/10.1080/00045608.2015.1064510

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, *49*, 45–53. https://doi.org/10.1016/j.apgeog.2013.09.012

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., … Portugali, Y. (2012). Smart cities of the future. *European Physical Journal: Special Topics*, *214*(1), 481–518. https://doi.org/10.1140/epjst/e2012-01703-3

Besag, J. (1977). Contribution to the discussion on Dr Ripley’s paper. *JR Stat. Soc.*, 193–195.

Blei, D., Andrew, Y., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1020. Retrieved from https://stat.duke.edu/~scs/Courses/Stat376/Papers/Variational/BleiJordan2003.pdf

Chan, K., Vasardani, M., & Winter, S. (2014). Leveraging Twitter to detect event names associated with a place. *Journal of Spatial Science*, *59*(1), 137–155.

https://doi.org/10.1080/14498596.2014.852073

Cheng, T., & Shen, J. (2018). *Grouping People in Cities: From Space-Time to Place-Time Based Profiling*. https://doi.org/10.1007/978-3-319-73247-3_10

Cheshmehzangi, A., & Heath, T. (2012). Effects of temporary markets on spatial inter-relations: A behavioural analysis of a public realm in the UK. *Journal of Asian Behavioural Studies*, *2*(4), 21–32. Retrieved from http://fspu.uitm.edu.my/cebs/images/stories/cebs/jabsv2n4jan2012c3.pdf

Cranshaw, J., Hong, J. I., & Sadeh, N. (2012). The Livehoods Project : Utilizing Social Media to Understand the Dynamics of a City. *Icwsm*, 58–65.

Cronin, A. M. (2008). Mobility and Market Research: Outdoor Advertising and the Commercial Ontology of the City. *Mobilities*, *3*(1), 95–115. https://doi.org/10.1080/17450100701797349

Davoudi, S. (2003). EUROPEAN BRIEFING: Polycentricity in European spatial planning: from an analytical tool to a normative agenda. *European Planning Studies*, *11*(8), 979–999. https://doi.org/10.1080/0965431032000146169

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information : Researching Volunteered Geographic Information : Spatial Data , Geographic Research , and New Social Practice. *Annals of the Association of American Geographers*, *102*(3), 571–590. https://doi.org/10.1080/00045608.2011.595657

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, *96*(34), 226–231. Retrieved from www.aaai.org

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., … Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, 1–27. https://doi.org/10.1080/13658816.2016.1273357

Gomide, J., Veloso, A., Meira, W., Almeida, V., Benevenuto, F., Ferraz, F., & Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. *Proceedings of the ACM WebSci'11, June 14-17 2011, Koblenz, Germany.*, 1–8. https://doi.org/10.1145/2527031.2527049

Goodchild, M. F. (2007). Citizens as sensors: web 2.0 and the volunteering of geographic information. In *GeoFocus* (pp. 8–10). Retrieved from http://geofocus.rediris.es/2007/Editorial3_2007.pdf

Goodchild, M. F. (2010). Twenty years of progress: GIScience in 2010. *Journal of Spatial Information Science*, *1*(1), 3–20. https://doi.org/10.5311/JOSIS.2010.1.2

Goodchild, M. F. (2013). Prospects for a Space–Time GIS. *Annals of the Association of American Geographers*, *103*(5), 1072–1077. https://doi.org/10.1080/00045608.2013.792175

Goodchild, M. F. (2015). Space, place and health. *Annals of GIS*, *21*(2), 97–100. https://doi.org/10.1080/19475683.2015.1007895

Goodchild, M. F., & Hill, L. L. (2008). Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, *22*(10), 1039–1044. https://doi.org/10.1080/13658810701850497

Goodchild, M. F., & Li, L. (2011). Formalizing space and place. *CIST2011-Fonder Les Sciences Du Territoire*, 177–183. Retrieved from https://hal.archives-ouvertes.fr/hal-01353206/

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101 Suppl*, 5228–5235. https://doi.org/10.1073/pnas.0307752101

Haklay, M., & Weber, P. (2008). OpenStreet map: User-generated street maps. *IEEE Pervasive Computing*, *7*(4), 12–18. https://doi.org/10.1109/MPRV.2008.80

Hill, L. L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. *Research and Advanced Technology for Digital Libraries*, 280–290. https://doi.org/10.1007/3-540-45268-0_26

Hollenstein, L., & Purves, R. (2010). Exploring Place through User-Generated Content: Using Flickr to Describe City Cores. *Journal of Spatial Information Science*, *1*(1), 21–48. https://doi.org/10.5311/JOSIS.2010.1.3

Hong, L., & Davison, B. (2010). Empirical study of topic modeling in twitter. *Proceedings of the First Workshop on Social …*, 80–88. https://doi.org/10.1145/1964858.1964870

Hu, Y., Mao, H., & McKenzie, G. (2018). A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science*, *00*(00), 1–25. https://doi.org/10.1080/13658816.2018.1458986

Jenkins, A., Croitoru, A., Crooks, A. T., & Stefanidis, A. (2016). Crowdsourcing a collective sense of place. *PLoS ONE*, *11*(4), e0152932. https://doi.org/10.1371/journal.pone.0152932

Keßler, C., Janowicz, K., & Bishr, M. (2009). An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval. *GIS '09*, 91–100. https://doi.org/10.1145/1653771.1653787

Kiskowski, M. A., Hancock, J. F., & Kenworthy, A. K. (2009). On the use of Ripley's K-function and its derivatives to analyze domain size. *Biophysical Journal*, *97*(4), 1095–1103. https://doi.org/10.1016/j.bpj.2009.05.039

Lai, J., Cheng, T., & Lansley, G. (2017). Improved targeted outdoor advertising based on geotagged social media data. *Annals of GIS*, 1–14. https://doi.org/10.1080/19475683.2017.1382571

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, *5*(Apr), 361–397.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., … Shi, L. (2015). Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers*, *105*(3), 512–530. https://doi.org/10.1080/00045608.2015.1018773

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographic Information Systems & Science*. Wiley.

Lynch, K. (1960). The Image of the City. In *The M.I.T Press*. https://doi.org/10.2307/427643

Montello, D. R., Goodchild, M. F., Gottsegen, J., & Fohl, P. (2003). Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition & Computation*, *3*(2–3), 185–204. https://doi.org/10.1080/13875868.2003.9683761

Purves, R. S., & Derungs, C. (2015). From Space to Place: Place-Based Explorations of Text. *International Journal of Humanities and Arts Computing*, *9*(1), 74–94. https://doi.org/10.3366/ijhac.2015.0139

Quercia, D., Schifanella, R., Aiello, L. M., & McLean, K. (2015). *Smelly Maps: The Digital Life of Urban Smellscapes*. Retrieved from https://arxiv.org/pdf/1505.06851v1.pdf

Relph, E. C. (1976). Place and Placelessness. In *London:Pion*. London:Pion.

Resnik, P., Resnik, P., Hardisty, E., & Hardisty, E. (2009). Gibbs Sampling for the Uninitiated. *Umiacs.Umd.Edu*, (June), 1–23. https://doi.org/10.1017/CBO9781107415324.004

Ripley, B. D. (1977). Modelling Spatial Patterns. In *Source: Journal of the Royal Statistical Society. Series B (Methodological)* (Vol. 39). Retrieved from https://www.jstor.org/stable/pdf/2984796.pdf?refreqid=excelsior%3A0d25bd429a8ec1a94706e5d323d1726f

Roche, S. (2015). Geographic information science II: Less space, more places in smart cities. *Progress in Human Geography*, *40*(4), 1–10. https://doi.org/10.1177/0309132515586296

Sacco, D., Motta, G., You, L., Bertolazzo, N., & Chen, C. (2013). Smart Cities, Urban Sensing and Big Data: Mining Geo-location in Social Networks. *Congresso Nazionale AICA*. Retrieved from http://camellia.unipv.it/servizi/images/publication/2013/2013-88.pdf

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS One*, *10*(3), e0115545. https://doi.org/10.1371/journal.pone.0115545

Tuan, Y.-F. (1977). *Space and place : the perspective of experience*. Minneapolis : University of Minnesota Press .

Vasardani, M., Winter, S., & Richter, K. (2013). Locating place names from place descriptions. *International Journal of Geographical Information Science*, *27*(12), 2509–2532. https://doi.org/10.1080/13658816.2013.785550

Wagner, P. L. (1972). *Environments and Peoples*. Retrieved from https://books.google.co.uk/books/about/Environments_and_Peoples.html?id=dBtjQgAACAAJ&redir_esc=y

Wallach, H. M., Mimno, D., & Mccallum, A. (2009). Rethinking LDA : Why Priors Matter. *Advances in Neural Information Processing Systems 22*. https://doi.org/10.1007/s10708-008-9161-9

Wang, Z., Ye, X., & Tsou, M.-H. (2016). Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Natural Hazards*, *83*(1), 523–540. https://doi.org/10.1007/s11069-016-2329-6

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Zakariya, K., & Harun, N. Z. (2013). The People's Dataran: Celebrating Historic Square as a

Potential Temporary Market Space. *Procedia - Social and Behavioral Sciences*, *85*, 592–601. https://doi.org/10.1016/j.sbspro.2013.08.388

Zhang, A. X., Noulas, A., Scellato, S., & Mascolo, C. (2013). Hoodsquare: Modeling and Recommending Neighborhoods in Location-Based Social Networks. *2013 International Conference on Social Computing*, 69–74. https://doi.org/10.1109/SocialCom.2013.17

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., & Terveen, L. (2007). Discovering personally meaningful places. *ACM Transactions on Information Systems*, *25*(3), 12-es. https://doi.org/10.1145/1247715.1247718