Geo-Temporal Twitter Demographics

Abstract

This paper seeks uses highly disaggregate social media sources to characterize Greater London in terms of flows of people with modeled individual characteristics, as well as conventional measures of land use morphology and nighttime residence. We conduct three analyses. First, we use the Shannon Entropy measure to characterize the geography of information creation across the city. Second, we create a geotemporal demographic classification of Twitter users in London. Third, we begin to use Twitter data to characterize the links between different locations across the city. We see all three elements as data rich, highly disaggregate geo-temporal analysis of urban form and function, albeit one that pertains to no clearly defined population. Our conclusions reflect upon this severe shortcoming in analysis using social media data, and its implications for progressing our understanding of socio-spatial dstributions within cities.

Keywords

Twitter; geotemporal demographics; urban geography

1. Overview

Over the last half century or more, our understanding of the form and functioning of urban systems has been founded upon quantitative analysis of infrequently collected census data, with the identities of members of local residential populations represented by a widely used but restrictive range of socioeconomic variables. Fundamental to the analysis presented here is the premise that an integral part of identity in urban areas is the locus of daily and other activity patterns. Moreover, the advent of social media data makes it possible to monitor such patterns, making it possible for urban geography to develop beyond data that are infrequently collected and that pertain principally to night time residence. Yet in using social media data t

developing geo-temporal demographics of towns and cities it is important to remember that previous work in this tradition is much better grounded in representations of populations that are complete, or at least clearly defined. In this paper we describe how it is now possible to refocus upon urban form at ever finer granularities, along with the development of richer representation of the agents that make things happen (function) in urban systems. Yet the caveat to this is that our richer geo-temporal representations pertain to sets of individuals that are likely to provide much less complete representations of complete populations.

This has brought focus to a number of issues that we still know rather little about: the spatial dynamics of individual behaviours and the constraints upon them; the interactions between different groups in society, embedded in built environments; and the interlocution of behaviour in observable and virtual spaces. The focus of this paper is upon developing and analysing detailed representation and analysis of the ebbs and flows of human movements, or the virtual interactions that provide a complementary system wide picture of the rhythms of activity that define cities.

Geodemographic classifications are small area classifications that provide summary indicators of the social, economic and demographic characteristics of neighbourhoods (Harris et al 2005). The staple data source for most conventional geodemographic classifications is census data, augmented in some commercial classifications by 'lifestyles' data from consumer data sources. Open Data sources from government and other administrative sources are also available to augment classifications, but in most all cases the emphasis remains exclusively upon the nighttime characteristics of just residential areas, rather than the more inclusive activity patterns of citizens. Workplace classifications, derived from census data, may provide a different polarity for classifications, albeit only for the subset of the population that is in employment.

In this paper, we extend this classification framework using geotagged social media data, which are a source of volunteered (Goodchild 2007) indicators of spatial behaviour (Cox and Gollege

1981). Microblogging services such as Twitter have come to document important aspects of the daily activities of millions of users in most countries around the world, in addition to social attitudes and opinions. These services are used not only for communicating between friends, family, and colleagues, but also for real-time news feeds and to share content about venues (Pennacchiotti and Popescu, 2011). According to recent figures, the Twitter service has more than 200 million active users around the world (Twitter, 2012a). Its major user base is in European countries, with usage in London the third highest in the world (Bennett, 2012). As part of the quest to relate the functioning of urban areas to their form, the principal focus of this paper is to develop geo-temporal demographic analytics of Twitter users and the locations in which they are active. Greater London was chosen because of its high numbers of users and the wide range of activities that take place there.

Previous research using Twitter data has explored a number of different themes. Stephens and Poorthuis (2014) compared the social properties of Twitter users' networks with the spatial proximity of the networks by undertaking an analysis of network density and transitivity; Shelton et al (2015) developed a conceptual and methodological framework for the analysis of social media data through analysis of the socio-spatial relations embodied in such data; Zook and Poorthuis (2014) examined the spatial distribution of geocoded social media data that referenced "beer" and related terms; Twitter activity was investigated during Hurricane Sandy in order to elucidate the complex relationship between the material world and its digital representation (Shelton et el, 2015); the political orientation, gender and ethnicity of Twitter users has been inferred using a machine learning approach (Pennacchiotti and Popescu, 2011); network metrics have been used to compare the social dynamics of Twitter usage with those of physical communities (Quercia, Capra, and Crowcroft 2012); Foursquare social media data have been used to populate a clustering model to study the composition of a city (Cranshaw et el, 2012); a latent attribute inference method has been used to infer the age, gender, and political affiliation of Twitter users (Al Zamal et al, 2012); and Birkin et al (2014) have used the content of the social media messages to classify different areas of the city of Leeds, UK. Our objectives in this paper are to focus upon the tasks of developing geo-temporal analysis of the flows of

population and information in London. In so doing, we geo-locate and selectively mine social media data in order to ascribe characteristics of the information source (the source of a social media tweet) and the characteristics of the setting in which it is made.

In the first stage of our analysis, we apply the Shannon Entropy measure (Batty, 2010) of Twitter activity to measure the geography of information creation across Greater London. The resulting map is of information generating activity rather than any conventional population geography. The individual basis to Tweets makes it possible to cluster the identifiable characteristics of human individuals, meaning that unlike the conventional areal aggregations that characterize conventional geodemographics, the issue of ecological fallacy does not arise. The second stage of our analysis thus attempt to devise a Twitter geotemporal demographic classification, on the basis of the results of inferences made at the individual level. The final strand to our analysis is more speculative: we use the locations at which co-occurrences of Twitter activity by unique individuals occur as indicators of connectivity within the city. Our motivation for doing so is to identify the patterns of use of social media, albeit that this is likely to be dominated by transport corridors in the first instance.

2. Computing Occurrences, Characteristics and Connectivity

Under terms and conditions that are current at the time of writing, the Twitter Streaming API (Twitter, 2012b) can be used to download a 1% sample of tweets in any prescribed time period. The characteristics of this sample can be limited to a subset of all tweets with a user-specified characteristic, and in our case is restricted only to those that are geotagged. By no means all of Twitter users opt in to the geotagging option, and in practice the 1% sample of all tweets is more than sufficient to acquire all of the tweets that are geocoded in any given time period in Greater London during 2013. A total of 8 million (8,027,646) geo-tagged Tweets were downloaded, sent by a total of 385,050 unique users. The fields downloaded from the API included the user name, latitude and longitude from which the Tweet was sent, time and date

of the tweet message, and message content. It is clear from this that the results of this study represent a small and self-selecting sample of all Twitter users in London, who in sum are unlikely to represent any wider population defined according to any broad-based measure. Our motivation is nevertheless to identify how social media sources can be used to establish a geo-temporal basis to classifying urban areas in terms of activity and interaction.

Social media data are important in this context because they are highly disaggregate and differentiated in both space and time. This potentially frees geodemographic analysis from the night time geography of residential locations, but with the very major caveat that self-selecting individuals who supply volunteered data are most unlikely to be representative of any clearly defined population: see Longley et al (2015) for an extended discussion. Here, we will proceed regardless of this major caveat however, in order to illustrate the potential of 'Big' social media data in devising geo-temporal measures of urban form and functioning in the spirit of Batty and others.

Over the time period of our analysis (January – December, 2013), the majority of our 385,050 Twitter users sent four or fewer tweet messages with London geocodes. We confine our analysis to the 155,249 users who sent five or more tweets, since our objectives included ascertaining the probable residence of the Twitter users. Our database thus comprised 7,609,574 geo-referenced tweets sent by 155,249 users. Figure 1 shows the proportions of users who sent particular numbers of tweets. The Figure shows that majority of users sent less than 100 tweets, but that the over-all distribution has a long tail of users sending much higher numbers.

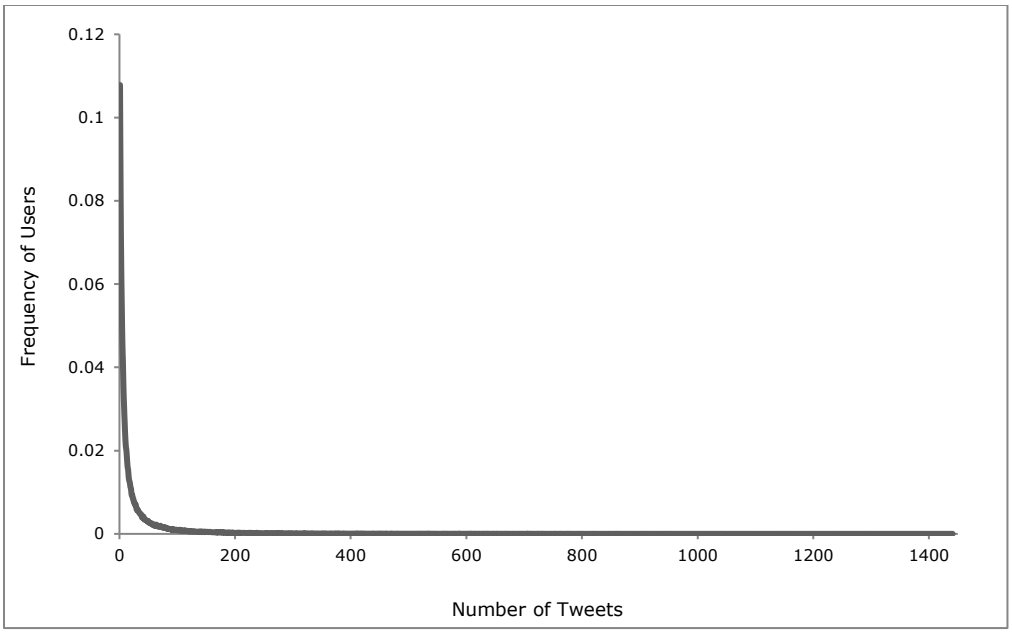


Figure 1: Proportion of the Users by number of Tweets

Figure 2, below, shows a map of the tweets sent from the Greater London. This map shows that more Tweets were sent by users located in the central part of the city than the surrounding areas of Outer London. This is consistent, inter alia, with Batty's (2013) exposition of the focus of city networks and interaction patterns, and provides a preliminary indication of the focus of information flows upon the heart of the city in terms of employment, shopping and tourism. Areas of public open space and lower settlement density towards the Greater London boundary record lower occurrences of tweeting activity.

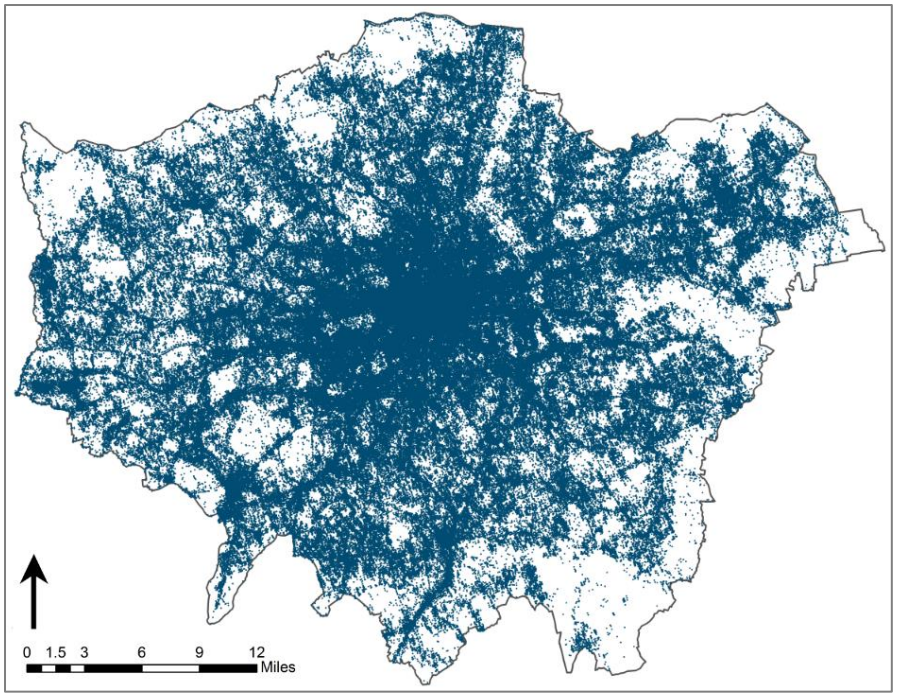


Figure 2: The Greater London geography of the 7.6 million tweets

Social media commentators use services in a wide range of settings in addition to the home, such as at employment locations, or while travelling to work or other engagements. Many users provide very large numbers of tweets, providing a valuable indication of activity patterns and also potentially of the characteristics of those that made them. Additionally, it is also possible to characterise the flows of users, and hence investigate the connectivity of places.

Tweets differ from the conventional sources used to create geodemographic classifications, in that most all user and location characteristics must be inferred or be obtained with reference to other secondary sources. The geographic precision of location information is very high, and this makes it possible to use point-in-polygon procedures to ascertain land use (commercial, industrial, public open space, etc) using UK national land use data which are available at similar levels of precision. The social milieu in which each tweet is made can be similarly ascertained by reference to conventional geodemographic classifications of residential areas (Singleton and Longley 2015). Other characteristics of the tweet – notably the age, gender and ethnicity of the person who made it – can be inferred through classifications of names extracted from Twitter

user identifiers. Finally, georeferencing can be used alongside time-stamping in order to ascertain whether the user is normally resident abroad, in the UK but outside London, or in London itself, and if the latter, which of tweets identify the usual domicile of the user.

In the analysis that follows, a total of 46 variables were created, falling into the eight domains shown in Table 1. Together, these domains provide a generalized representation of the characteristics of Twitter users and the settings in which tweeting activity occurs. They thus provide the ingredients for a general purpose classification of information generated at point locations across London.

Table 1: The 42 variables used to create the Twitter geodemographic classification

1. Residence V1: Tweet made near probable London residence V2: Tweeter lives 'outside the UK' V3: Tweeter lives in the rest of the UK outside London	**6. Number of countries visited** V28: Number of countries the tweeter has visited during the study period
2. Total Number of Tweets V4: Total number of tweets made by the user	**7. London Land Use Category** V29: Tweet sent from residential location V30: Tweet sent from non-domestic buildings V31: Tweet sent from transport links and locations V32: Tweet sent from greenspace or water locations V33: Tweet sent from all other land uses
3. Ethnicity V5: West European V6: East European V7: Greek or Turkish	**8. 2011 London Output Area Classification** V34: Intermediate Lifestyles V35: High Density and High Rise Flats V36: Settled Asians

V8: South East Asian V9: Other Asian V10: African & Caribbean V11: Jewish V12: Chinese V13: Other minority	V37: Urban Elites V38: City Vibe V39: London Life-Cycle V40: Multi-Ethnic Suburbs V41: Ageing-City Fringe
4. Age V14: <= 20 V15: 21 – 30 V16: 31 – 40 V17: 41 – 50 V18: 50 +	**9. Temporal Scales** V42: Morning Peak Hours (7 a.m. - 9:30 a.m.) V43: Week Day (9.30 a.m. – 4 p.m.) V44: Afternoon Peak Hours (4 p.m. – 7 p.m.) V45: Week Night (7 p.m. – 7 a.m., Monday - Thursday) V46: Weekend (7 p.m. Friday – 7 a.m. Monday)
5. Tweets outside the UK V19: Tweets sent in West Europe (not including UK) V20: Tweets sent in East Europe V21: Tweets sent in North America V22: Tweets sent in Central or South America V23: Tweets sent in Australasia V24: Tweets sent in Africa V25: Tweets sent in Middle East V26: Tweets sent in Asia V27: Tweets sent in Paris administrative area	

The rest of this section explains the derivation of each of these variables from raw Twitter content, their associated locations, and secondary data sources.

2.1 Ethnicity

Our analysis uses given and family names of Twitter users to infer their ethnicity (see Mateos et al 2011). As a prelude to ethnicity assignment, forename-surname pairs of Twitter users were extracted from their user identifiers. In many cases, users enter tokens other than their given and family names in the user identifiers, as in 'JustinBieberHome', 'What is Love', 'MadMind' and so forth. Some users enter a part of their name in the user identifier, as in 'Vanessa', 'John138', 'Amir_Hello', and so forth. These user identifiers cannot be used for defining the ethnicity of Twitter users. The name extraction algorithm developed by Longley et al (2015) nevertheless suggests that a large number of users (more than half) enter plausible forename-surname pairs in this field. We used the same algorithm to extract forename-surname pairs of Twitter. Results suggested probable forename-surname pairs for 4,449,323 of the 7.6 million Tweets. These 4.4 million Tweets were sent by 107,551 of the 155,249 unique users.

Mateos et al (2011) describe the development and use of the Onomap classification to assign users into different cultural, ethnic and linguistic groups on the basis of distinctive forenames and surnames, using the results of a cluster analysis of names extracted from electoral registers and telephone directories from different countries. The classification used here was created from a version of the 2007 Electoral Register for Great Britain, enhanced from consumer files to include the names of non-electors and other individuals whose names did not appear in the public version of the Register. The classification technique has been used for research purposes by a variety of organisations, with notable success in health care where it has been applied to patient records to ascribe probable ethnicity when monitoring the success of care interventions such as breast screening. The classification operates by assigning a probable ethnicity for a forename-surname pair: using information from both elements is considered to be useful for accommodating issues arising from self-assignment of ethnicity as a token of individual identity, and may also be helpful when considering issues of mixed race, although this issue was not

addressed in this case study. User feedback has suggested that the software is successful in ascertaining ethnicity in a range of case studies, albeit that it is difficult and error prone to confirm success in predicting a characteristic that is sensitive to some individuals – who may be disproportionately concentrated in some ethnic categories of the classification.

The Onomap software was applied to each of the forename-surname pairs that could be identified from the 4.4 million Tweets: approximately 3.8 million (3,896,731) million Tweets, sent by 98,607 unique users, were successfully classified. 552,592 Tweets were not classified. Individuals were assigned to a total of 67 ethnic groups. Table 2 provides a lookup of how Onomap assigned ethnic groups were aggregated for purposes of building our classification. The 'British and Irish' variable was not used in any of the analysis of this paper as it is the largest population group in Greater London, and they appear to be over-represented on Twitter – accounting for 77% of the 3,896,731 classified tweets.

Table 2: Onomap-to-Ethnicity variables lookup table

Ethnic group category	**Onomap ethnic group**
British and Irish	English, Celtic, Scottish, Welsh, Irish
V5: West Europeans	Spanish, Italian, Portuguese, German, French, Swedish, Dutch, Finnish, Danish, Norwegian
V6: East Europeans	Polish, Romanian, Hungarian, Albanian, Serbian, Czech, Ukrainian
V7: Greek & Turkish	Greek, Turkish
V8: South East Asian	Indian Hindi, Sikh, Pakistani, Pakistani Kashmir, Bangladeshi, Sri Lankan
V9: Other Asian	Vietnamese, East Asian & pacific, South Asian, Japanese, South Korean, Malaysian, Labanese, Muslim Middle East, Iranian, Armenian,
V10: African & Caribbean	Black Caribbean, Nigerian, Ghanaian, African, Sierra

	Leonean, Black Southern African, Ugandan, Somali, Ethiopian, Congolese, Eritrean
V11: Jewish	Jewish
V12: Chinese	Chinese, Hong Kongese
V13: Other	All other Onomap categories

2.2 Age

Given names can be used to infer additional characteristics about an individual, since choice of a baby's name is driven by social and cultural influences that vary across time as well as space. In the present context, given names were used in order to estimate the likely ages of their bearers, using an enhanced version of CACI's (London, UK) Monica system. This uses c. 7 million records drawn from UK consumer dynamics files to identify the frequencies of 11,700 different given names within five year age bands. The majority of the names are also identified as gender specific. The nature of the source data means that younger adult cohorts are under-represented relative to the UK population as a whole, and there are no records pertaining to individuals below the age of 18 at all (despite their making up 22% of the UK population).

In order to mitigate this bias, all names from birth certificates with frequencies of two or above, representing 9.7 million individuals, were acquired from the Office of National Statistics for the years 1994 – 2011. (It was not, however, possible to identify a source that would have enabled the names of international immigrants under the age of 18 to be identified.) The birth certificate data were disaggregated into five year age group bands for consistency with the Monica classification, and both sources were reweighted to fit the age distribution of the 2011 Census of Population for England and Wales. The average ages of bearers of the names in the resulting dataset ranged from just 2 years to 83 years of age.

Using the above datasets, each Twitter user was allocated to one of five age bands (<=20 years, 21-30, 31-40, 41-50 and 50+). Of the 98,607 Twitter users for whom names identifiable using Onomap, it was possible to model the probable ages of 85,798 users, who sent a total of

3,296,330 tweets. As with the Onomap software for ethnicity, the Monica system has been successfully used in a range of case studies, but no attempt was made to validate the age assignments in our case study.

2.3 Usual residence, commuting, international travel and tweeting activity level

The international travel behavior and usual national residence of tweeters was identified for the 85,798 users for whom ethnicity and age could be modeled. A point in polygon operation was performed in order to assign each tweet to a country. Any country from which a user was observed to send more than 50% of their tweets was designated the user's country of usual residence. This procedure suggested that 5,642 of the 85,798 users were likely resident outside the UK. A similar procedure was used to identify usual residences inside or outside Greater London, and to assign probable residential locations to the former category. This procedure identified 4,634 users that were likely resident in the UK outside Greater London, while the remaining 75,522 users were resident within it.

A third procedure was used to pinpoint the probable residence of the 75,522 twitter users in Greater London to a 170m x 170m grid square. Figure 3 illustrates the use of the grid to identify the probable residence of a single user who sent a total of 3,297 georeferenced tweets.

(A) (B) (C)

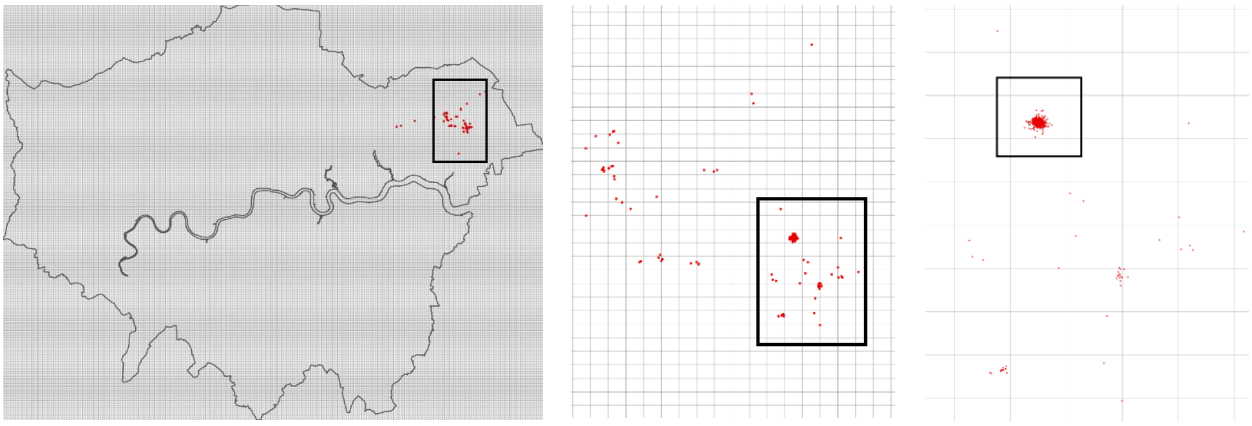


Figure 3: (A) The Tweets sent by a single user in Greater London, overlaid on a 170m x 170m grid, with details shown in (B) and (C).

This procedure was automated and repeated for every Twitter user for whom five or more Tweets were collected during our study period, using grid assignment and a point-in-polygon operation. Each user was assigned to the centroid of the 170x170m cell in which the maximum number of his or her Tweets took place. Where there was a tied value (e.g. because no more than a single Tweet occurred in any cell) the procedure was repeated at the 2011 Census Output Area scale, with remaining users assigned to the Output Area where they made the maximum number of Tweets in each of the time windows. Using this procedure we were able to find the probable residence of all 75,522 London-based twitter users.

It was thus possible to identify whether each tweeter was usually resident: (a) in the UK but outside Greater London; (b) outside the UK and in one of the country aggregations described in Table 1; or (c) in a particular neighbourhood in Greater London.

For each user, the total number of tweets sent all around the world was also enumerated.

2.4 Land Use

Land cover data were drawn from the UK government Generalised Land Use Database for 2005 (GLUD). This data source was created using an algorithm that reassigned features from Ordnance Survey MasterMap into nine classes of land use, to a claimed 1m precision (Department for Communities and Local Government, 2006). For purposes of this paper the original nine classes of the data were amalgamated into residential (including gardens), non-domestic building, transport, green space and water: the first of these categories provides a basis to comparison with Census data, while the others facilitate analysis of different activities. The following Figure 10 shows a map of the different type of land-use categories available.

For each of the 3,296,330 tweets, a point in polygon operation was performed to join every tweet to its land-use category – defined as residential, transport links and termini, greenspace or water, non-domestic buildings or other.

2.5 London Output Area Classification

The 2011 London Output Area Classification is an open and free-to-access geodemographic classification which has been created by the cluster analysis of 2011 Census data. Singleton and Longley (2015) document the procedures used to create it and the nomenclature of the eight Super Groups that it comprises. Each Census Output Area in Greater London is assigned to one of the eight Super Groups shown in Figure 4.

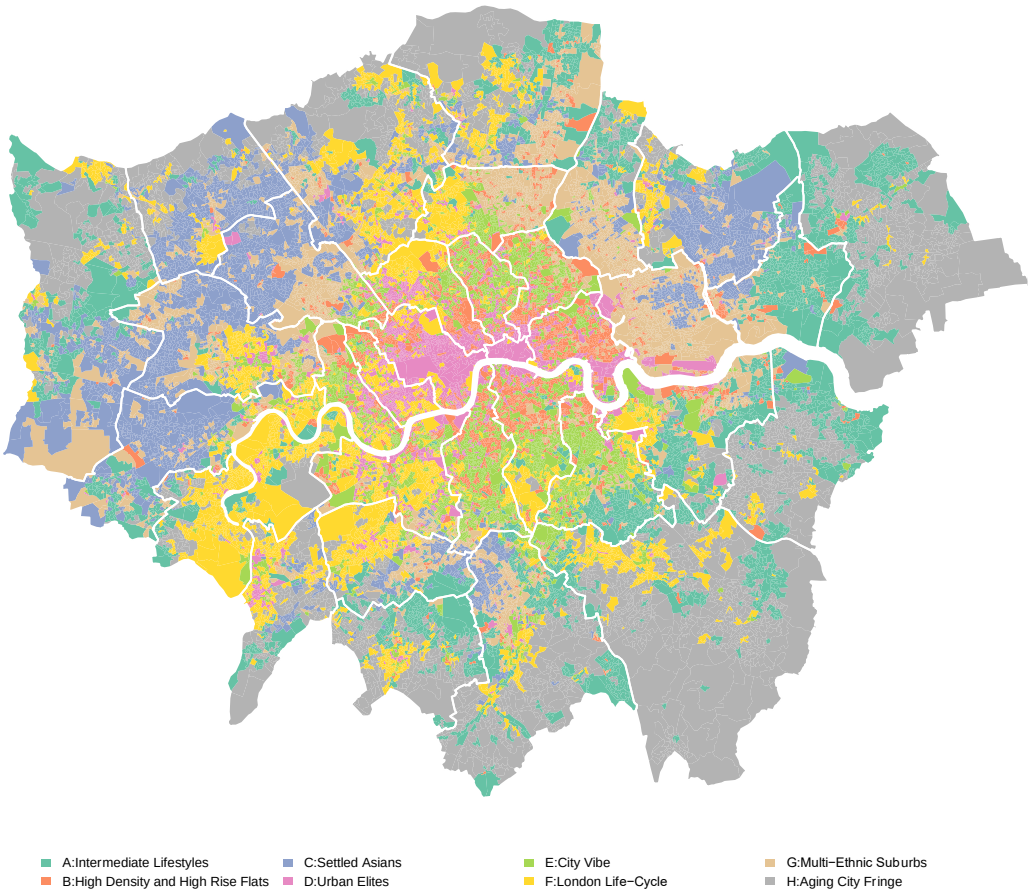


Figure 4: The 2011 London Output Area Classification

A point in polygon operation was performed to assign every tweet to the Output Areas from which it was made. Eight variables, corresponding to the Super Groups, were thus created – defined as Intermediate Lifestyles, Settled Asians, City Vibe, Multi-Ethnic Suburbs, High Density and High Rise Flats, Urban Elites, London Life-Cycle, and Aging City Fringe.

2.6 Temporal Scales

To identify any temporal patterns from the data, variables corresponding to five time intervals were created: morning peak hours (Monday-Friday, 7 a.m. – 9:30 a.m.), afternoon peak hours (Monday-Friday, 4 p.m. – 7 p.m.), weekday (Monday-Friday, 9.31 a.m. – 3.59 p.m.) week nights (Monday-Friday, 7.01 p.m. – midnight and midnight to 6.59 a.m.) and weekends.

3. Calculating the Evenness of Twitter usage

Shannon Entropy (Batty 2010) provides a useful summary measure that can be used to describe the evenness of tweeting activity across Greater London. Entropy is a measure of the uncertainty in a random variable (Shannon, 1948) X with n outcomes $\{x_1, x_2, \dots, x_n\}$. Shannon entropy $H(X)$ is defined as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \quad \text{.. (1)}$$

where

$$\sum_{i=1}^{n} p(x_i) = 1$$

where $p(x_i)$ is the probability mass function of outcome x_i , the number of Twitter users in a given Lower Super Output Area i. Computing the Shannon Entropy for different areas of Greater London allows us to identify the areas that are the major sources of people and information: high values identify that there are high flows of people and information to and from a given location, and vice versa.

Figure 5 presents the Shannon Entropy across the 4,765 Lower Super Output Areas (LSOA) that make up Greater London. Each of the 7.6 million tweets was assigned to a LSOA using a point in polygon operation and the total for each LSOA was summed. The frequency distribution of each

individual's tweets across all LSOAs was computed in order to calculate the Shannon Entropy measure. In Figure 5, darker shading identifies areas where there is evenness of social media usage, arising because a large number of Twitter users visit those areas. These areas include the centre of the city and Heathrow Airport at the western extremity of Greater London.

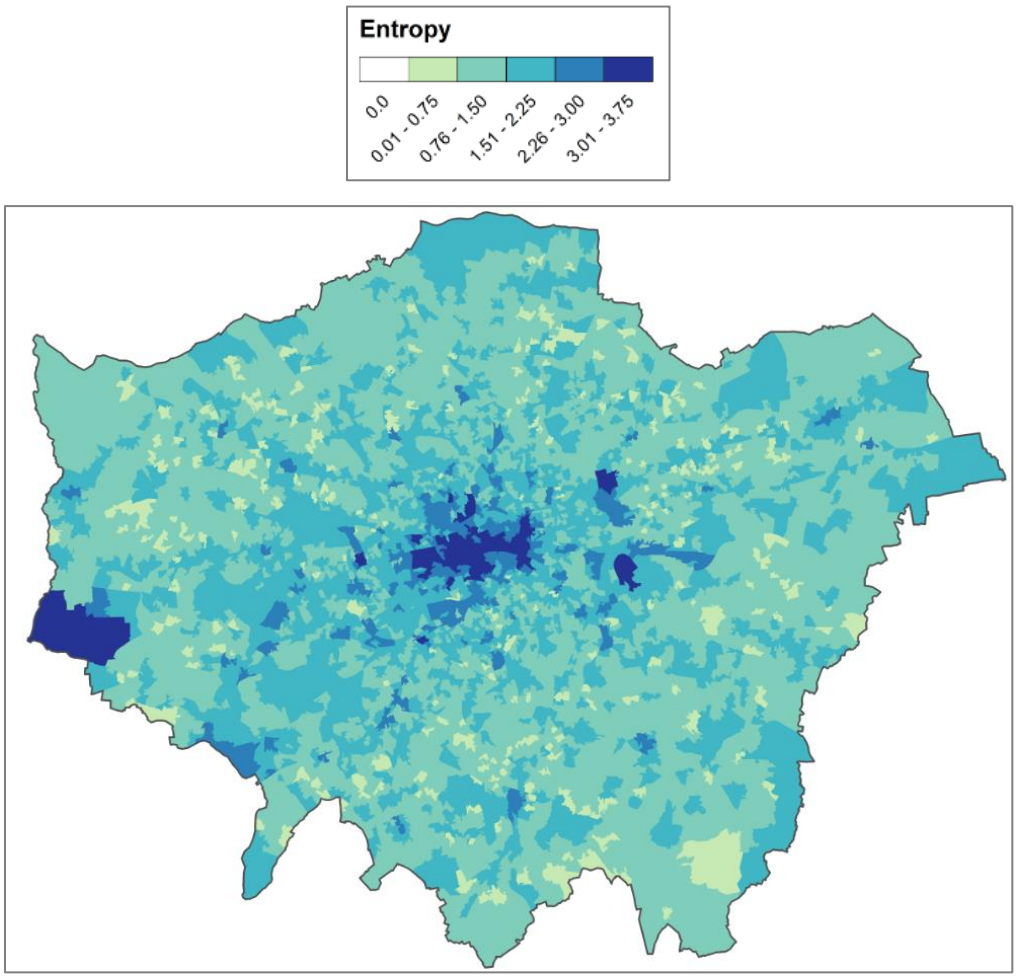
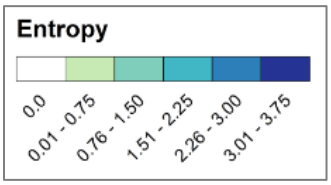


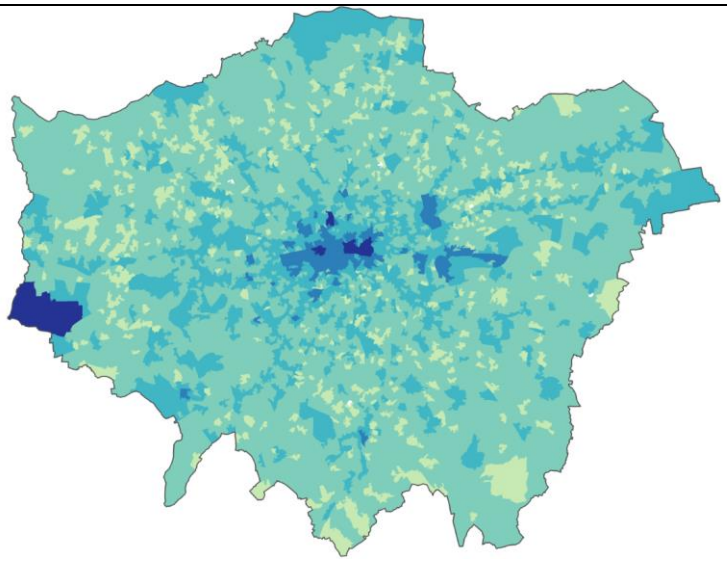
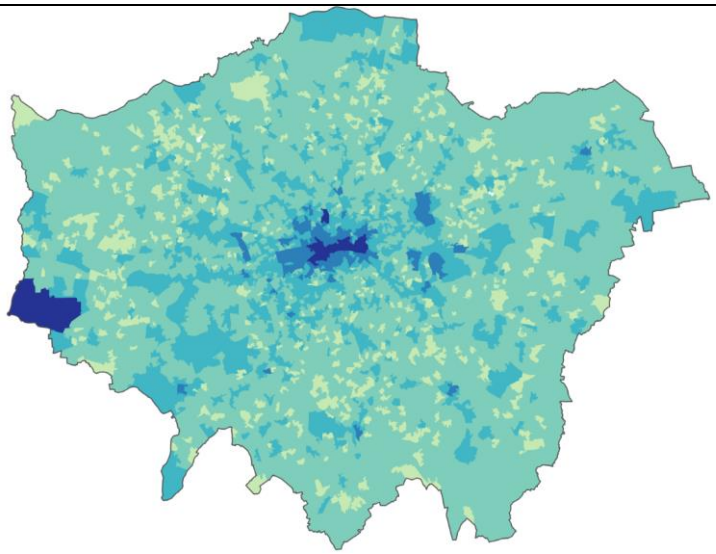
Figure 5: Shannon Entropy values for the 4,765 LSOAs in Greater London

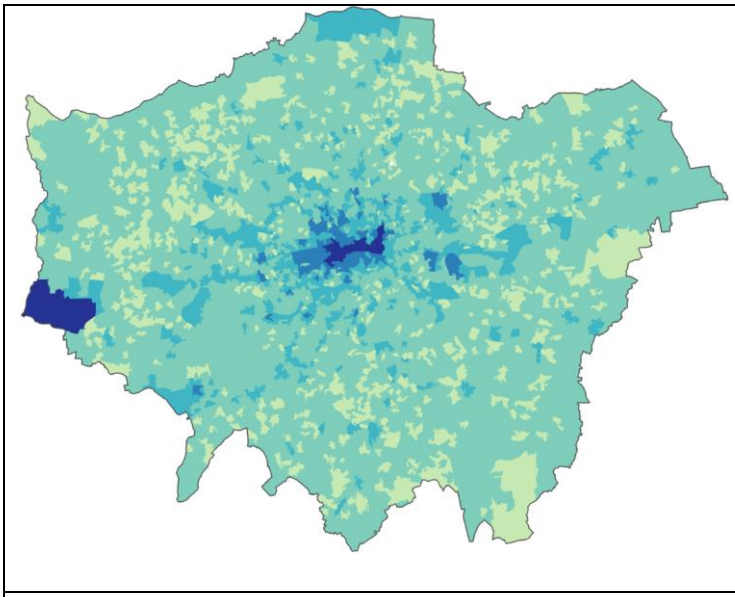
The centre of the map reveals a general concentric pattern with some additional secondary concentrations of high entropy values towards the edges of the Inner London area. This indicates that the central part of the city is the most important source and probable destination of information flows for Londoners as well as those who travel to it for business and leisure

purposes. When viewed at higher granularities, this map also identifies the transport links connecting the central part of the city to its hinterland.

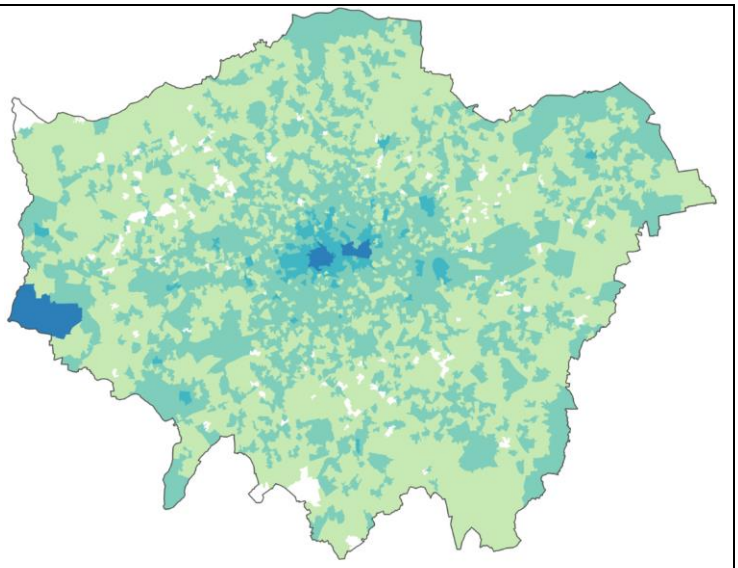
The time-stamping of the tweets makes it possible to calculate Shannon Entropy for different times of day, aggregated in Figure 5 to 'mornings' (6.00 a.m. to 11.59 a.m.), 'afternoons' (12.00 p.m. to 05.59 p.m.), 'evenings' (6.00 p.m. to 11.59 p.m.) and 'nights' (12.00 midnight to 5.59 a.m.).



	
(A)	(B)



(C)



(D)

Figure 6: Shannon Entropy values for (A) mornings, (B) afternoons, (C) evenings and (D) nights.

Figure 6 provides a general picture of the daily rhythm of activities in the city. In the mornings the pattern indicates high levels of tweeting activity along the tentacular arrangement of transport arteries, and high concentrations of activity in the City area and at Heathrow Airport. However, in the central part of the city, activity reaches its peak in the afternoons and evenings. There is less activity on transport links in the afternoons. Activity diminishes much earlier in the suburban areas of the city, and creases abruptly after midnight.

4. Geo-Temporal Demographics

Prior to the classification process, variables v4 and v19-v28 (as described in Table 1) were assigned normalized values between 0 and 1, consistent with the range of all of the other variables (see Adnan et al, 2010 for a full rationale). A *k*-means clustering algorithm was then used to assign the tweets to homogeneous groups. K-means is the standard clustering algorithm used to created geodemographic classifications, including the LOAC classification

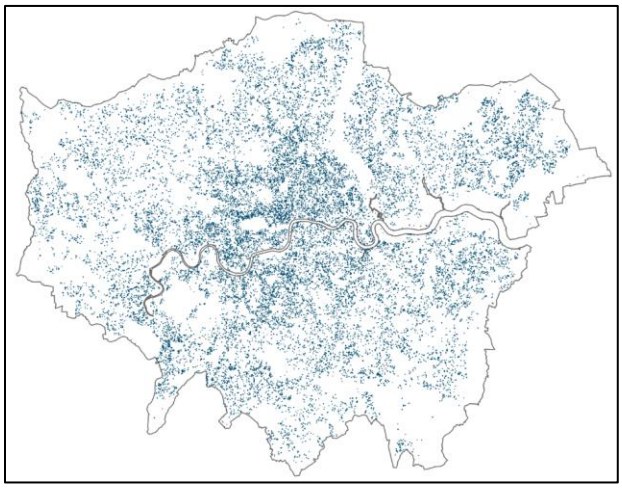
shown in Figure 4. The *k*-means algorithm seeks to find the set of cluster centroids that minimises

$$V = \sum_{x=1}^{n} \sum_{y=1}^{n} (z_x - \mu_y)^2 \qquad (2)$$

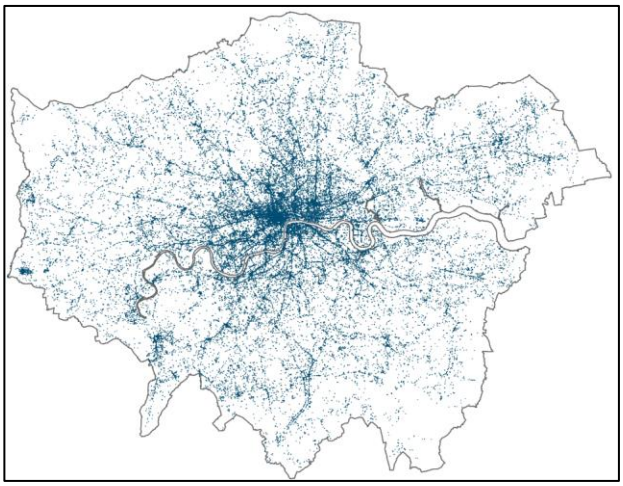
where *n* is the number of clusters, and μ_y is the mean centroid of all the points z_x in cluster *y.* The *k*-means algorithm works by assigning a set of *n* seeds within the dataset and then proceeds by assigning each data point to its nearest seed. Cluster centroids are then created for each cluster, and the data points are assigned to the nearest centroid. The algorithm then re-calculates the cluster centroids and repeats these steps until a convergence criterion is met (usually when switching of data points no longer takes place between the clusters). There is evidence that the outcome of the classification process can be influenced by the initial random positioning of initial seed points. In previous research (see Singleton and Longley 2015) we have undertaken extensive sensitivity analysis of these effects. However, the results of preliminary sensitivity analysis suggested that this was not an issue with the classification developed here.

The within cluster sum of squares values were calculated for all solutions requiring 2 – 12 clusters. Consistent with the findings of Vickers & Rees (2007), the results suggest the most parsimonious number of clusters to be seven. The mapped outcome of the seven cluster solution is shown in Figure 7. Radial plots for each cluster, on which the standardized scores of each of the 46 variables are presented relative to their grand mean scores, were used to define the profiles of the clusters (see Figure 8).

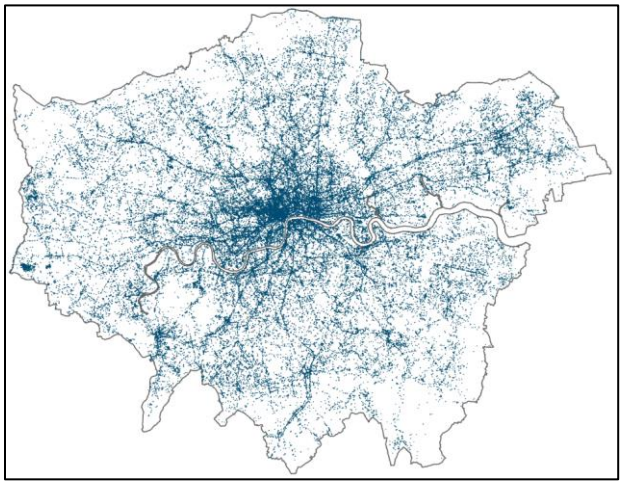
Cluster A



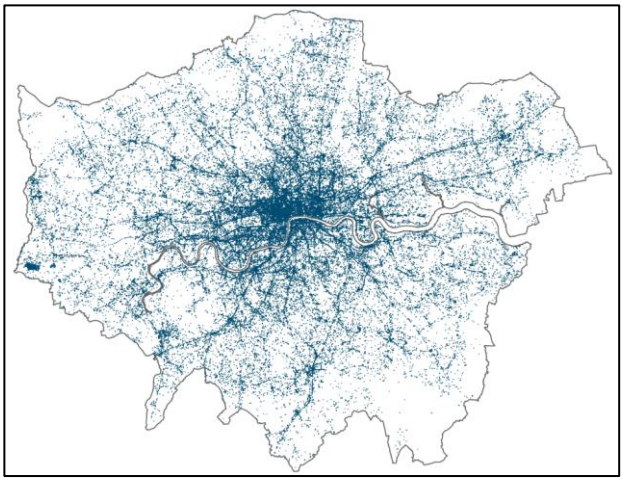
Cluster B



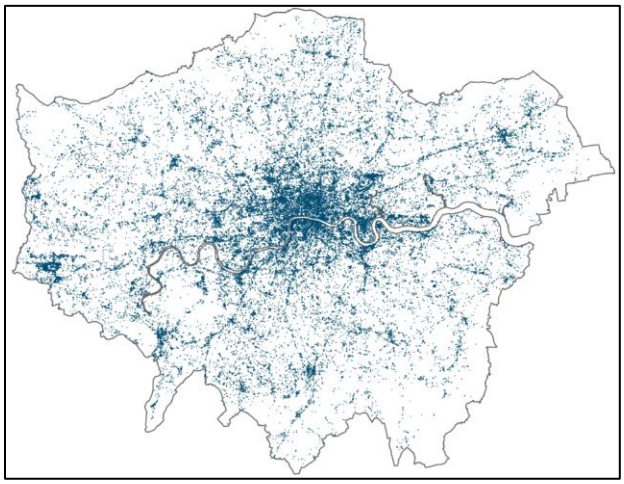
Cluster C



Cluster D



Cluster E



Cluster F

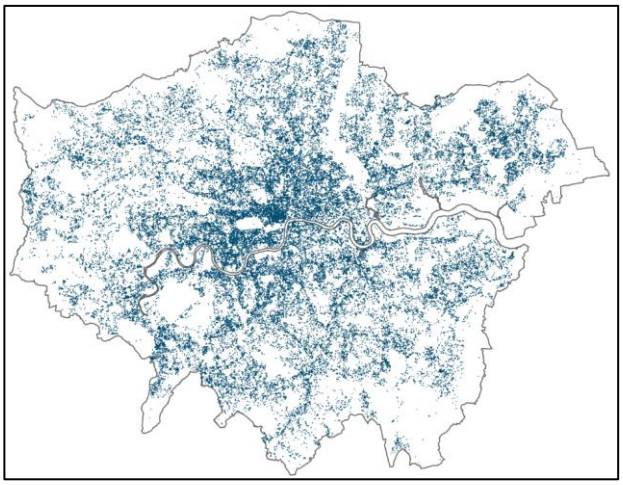
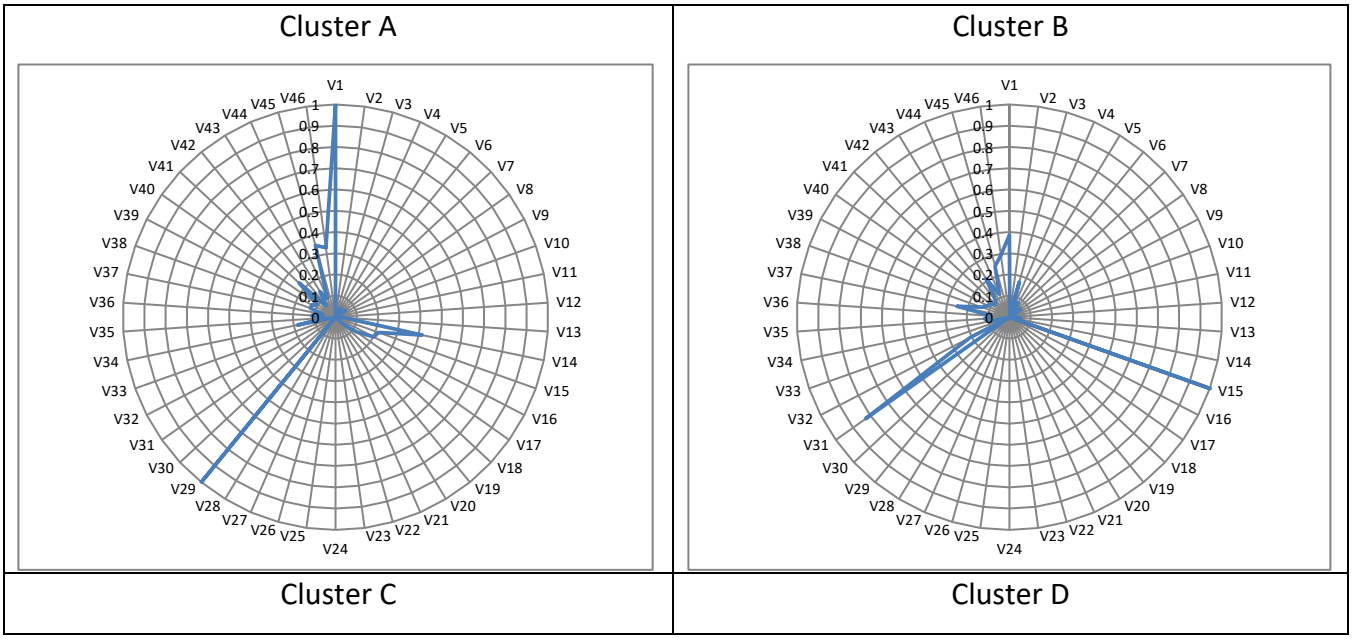


Figure 7: The spatial distributions of the seven clusters (A) through (G)



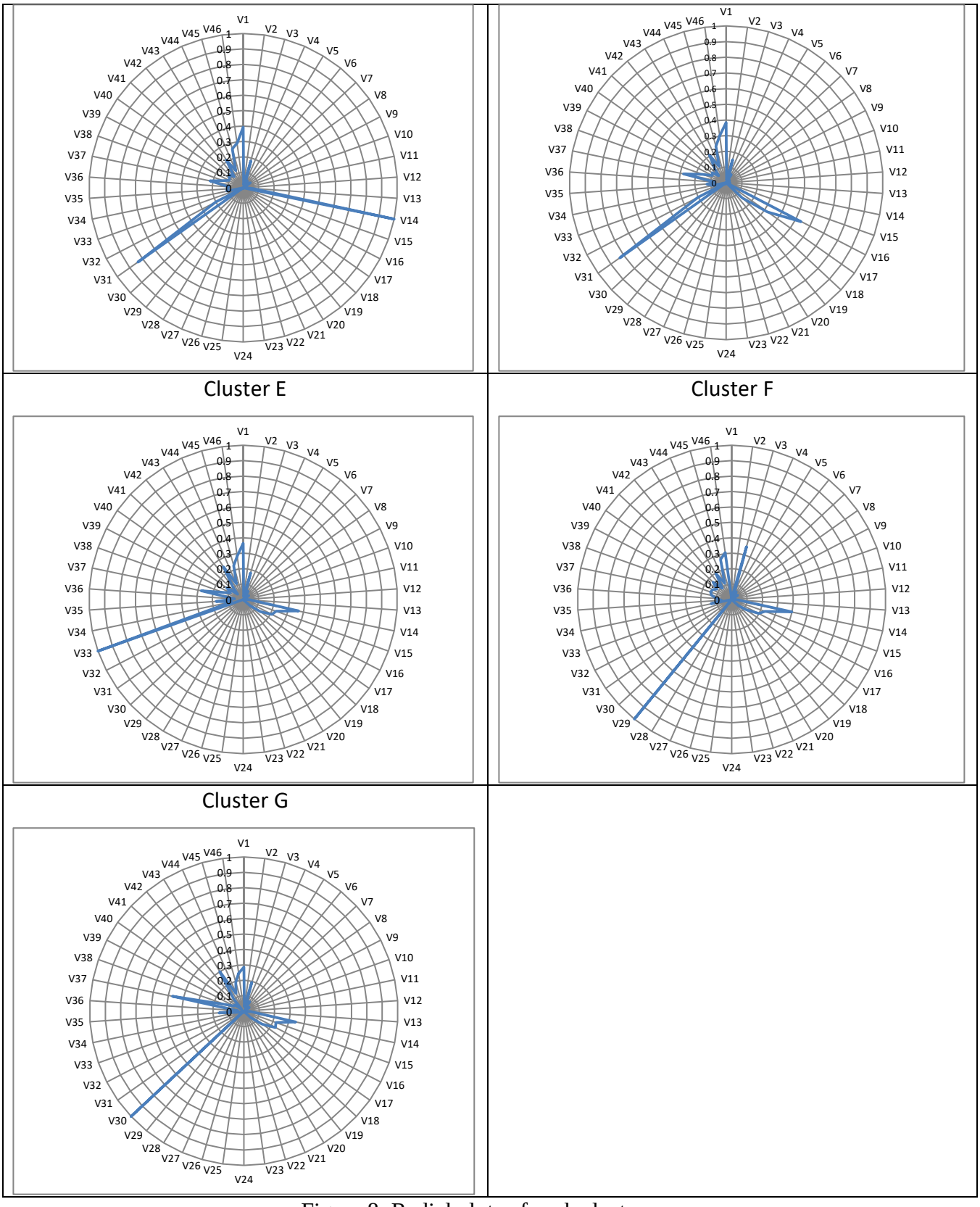


Figure 8: Radial plots of each cluster

Verbal ‘pen portraits’ of cluster profiles are widely used to summarise the dominant variables in a cluster, in our case as follows:

Group A: London Residents:

Tweets clustered into this Group were made by users estimated to come from a full age range of Twitter users. Very high proportions were made from residential locations, which were likely to be primary residences. Tweets tended to be made on weeknights or at weekends.

Group B: Commuting Professionals: Tweets tended to be made by individuals of intermediate (21-30) age, and from locations that were associated not only with transport functions but also by the high status ‘Urban Elites’ LOAC classification. A large proportion was sent on weeknights.

Group C: Student Lifestyle: Many of the tweets assigned to this Group were made at or near probable residence (V1) or from transport-related locations, but outside of the early morning and late afternoon commuting peak. Inferred names were indicative of youth (V14).

Group D: The Daily Grind: Tweets classed into this Group we made predominantly during peak weekdays and nights, and were sent from probable residences or in transit. Names were drawn from throughout the age spectrum, suggesting a higher than average age for Twitter users.

Group E: Spectators: The most distinctive characteristic of this cluster was that they were made from ‘other’ land uses, which principally comprise large leisure facilities such as the Olympic Stadium and other major sports facilities. Some of these tweets were made at or near probable London residences, but others were made from probable UK residences outside London. High density and high socio economic status residential land use is also prominent. Weekend and weekday tweeting behavior is common, though not during the morning commuting peak, indicating inter-urban and long distance commuting behavior. “All other land uses” category contains the locations of airport runways (specifically at Heathrow) which are not classified as transport land use.

Group F: Visitors: Tweeters in this Group live outside Greater London, and indeed significant numbers live outside the UK. When in London, they nevertheless tweet from residential land uses, and these are drawn from a wide range of LOAC classes. They are drawn from the core Twitter age groups. Tweeting activity takes place at weeknights and weekends. This cluster

represent Twitter users who do not live in London and tweet from residential locations. This might be the case of people travelling from other locations of the UK and staying in London or visitors from other cities/countries visiting residential areas of London.

Group G: Workplace and tourist activity: The most obvious characteristic of Tweets in this Group is that they are sent from non-domestic buildings. The residential context to these buildings is a mixture of high density and high status. The tweets originate from a mix of residents and international visitors. Weekend and weekday activity (the latter outside the commuting peaks) predominates, and a full range of Twitter age cohorts is represented.

5. Connectivity of places

Our final analysis considers the connectivity of different areas within Greater London, by performing the following steps using the 7,609,574 tweets:

a) Every tweet was assigned to one of the 4,657 LSOAs that make up Greater London, using a point in polygon operation. A symmetric matrix of 4,657 rows and 4,657 columns was then created, recording the intersection of activities of the Twitter users between each of the LSOAs. An intersection of activity between any two LSOAs is defined by a single user sending tweets from both LSOAs.
b) Ward's hierarchical clustering (Ward, 1963) was applied to the full matrix. This popular method of hierarchical agglomeration forms hierarchical groups of mutually exclusive subsets in attribute space. The algorithm began by assigning the n initial observations to $(n - 1)$ exclusive sets by considering the union of all possible $[n(n - 1)/2]$ pairs for the functional relation that matches an objective function chosen by the investigator, and then proceeds by successive iteration.
c) Visual interpretation of the resulting dendogram suggested that k=7 was a parsimonious cluster solution.

Each individual cluster represents the areas which are highly connected together based on the interactions revealed by the Twitter data. When mapping the resulting classification (Figure 9) it is noticeable that the connectedness of the areas is divided by the River Thames.

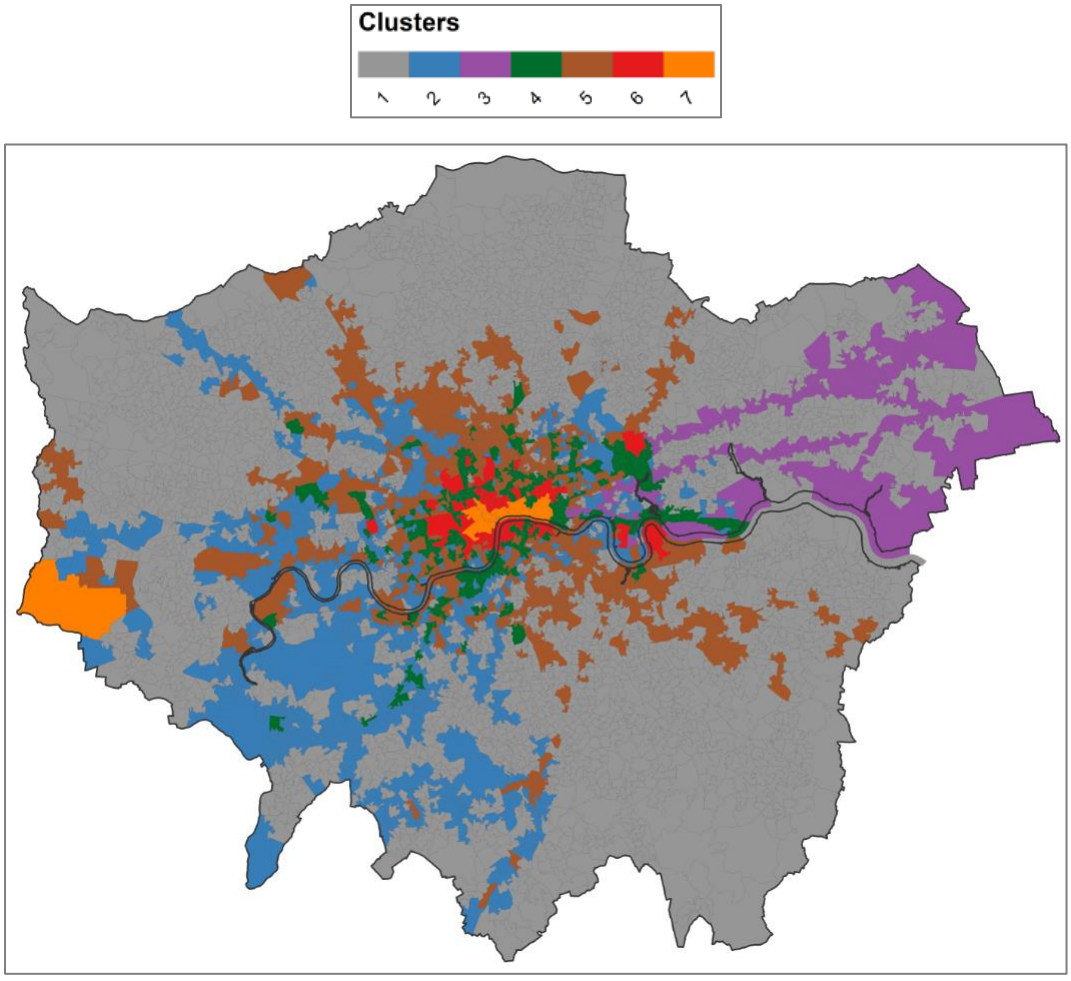


Figure 9: Mapped outcome of Ward's hierarchical clustering

Cluster 1 represents the areas with low tweeting activity i.e. low flows of people and information between different areas. Cluster 2 represents the areas mostly to the South West, West, and North West of Greater London. The areas within this cluster link outer London to the center of the city as visible from the continuous lines coming from outer to the inner areas of London. Cluster 3 comprises areas along transport links for travelers to the central city from the

eastern side of London. Cluster 4 largely comprises areas of Urban Elites and City Vibe geodemographic classes scattered principally around Inner London, and characterized by high Twitter usage. Cluster 5 represents a number of transport corridors, plus some residential areas in the north and south east parts of London. Clusters 6 and 7 are areas characterized by the highest volumes of tweeting activity. These clusters are concentrated in Central London although the inclusion of Heathrow Airport in Cluster 7 also testifies to the strong functional interdependence between these areas

6. Discussion and conclusion

The three stages to the analysis reported here make three related contributions to understanding the geo-temporal demographics of social media usage: assessing the spatial evenness of Twitter usage at different times of day; measuring the functional connectedness that this implies and is associated with; and ascribing characteristics to the individuals who participate in social media activity and the locations in which this activity takes place. All of these analyses are explicitly spatial, and are based upon the activities of a small self-selecting subset of Twitter users who have opted in to make their locations publically available. Additionally, a threshold was used to remove very occasional Twitter users from the analysis, and the process of inferring individual characteristics from user identifiers led to removal of further records. Figure 10 shows the cumulative effect of these selection criteria upon the set of observations that were analysed, along with the methods that were used to display and summarise the data.

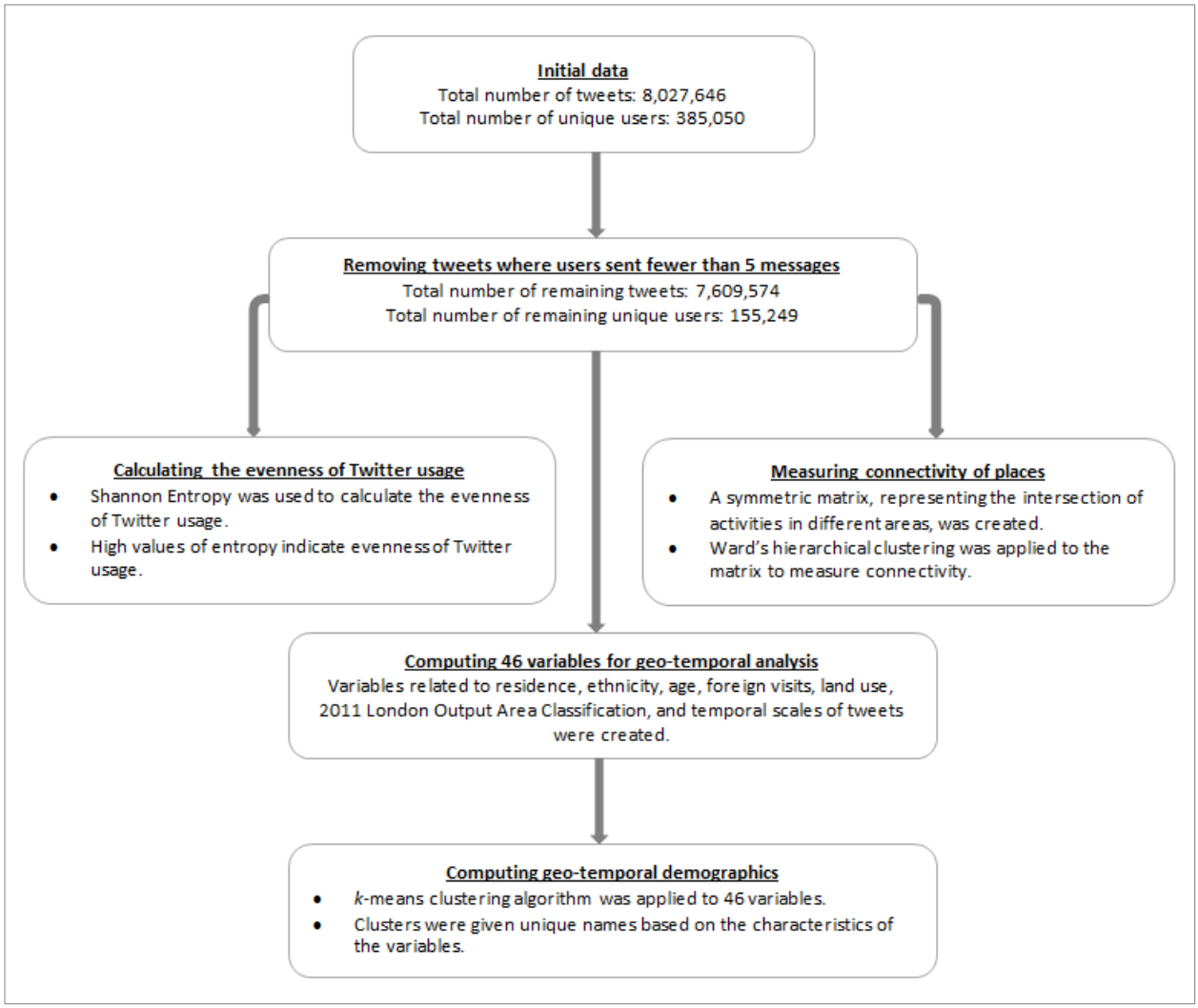


Figure 10: Flow chart of the analysis

In the first part of the analysis we computed the evenness of Twitter usage, measured as Shannon Entropy. Our segmentation of this analysis into time periods associated with different dominant work and social activities provides insight into the daily rhythm of activities in London, while also providing benchmarks in terms of level of social media usage and evenness of use across the city.

A characteristic of previous research using georeferenced Twitter data is that it is very detailed in terms of spatial resolution, but that the attributes of social media users are restricted to the very few details that are provided upon user registration. The focus of the second stage of our

analysis was thus to use onomastic (proper name) analysis and data linkage to infer ethnicity, age, socioeconomic and residence characteristics of users, along with the dominant economic land use at the tweet location. Cluster analysis was used to compute the demographic classification and verbal 'pen portraits' of cluster profiles were used to name the clusters. This provided an important extension of conventional neighbourhood geodemographics into the temporal domain, at fine levels of spatial and temporal granularity, providing insight into the different types of Twitter users and the variance in Twitter activity over time.

Our final analysis measured the connectivity of different areas within Greater London based on observed levels of social media activity amongst users. The analysis revealed the areas which are highly connected together based on the interactions extracted from the Twitter data.

The innovation of social media has brought the availability of large volumes of data that are highly disaggregate in terms of temporal granularity and (with respect to users who opt in) spatial location. Although much richer in these terms, it is nevertheless very difficult to assess how representative such users are of residential populations, or indeed any clearly defined population at all. An important contribution of the analysis reported here is the use of a range of procedures to infer user characteristics that can be linked to conventional census residential and workplace statistics, and establish some basis to generalization of the results. This inference at the level of the individual is the only granularity at which it can be appropriate to infer characteristics of social media users, given that the data exhaust of social media interactions represents user interactions when undertaking a wide range of activities.. The representations that we have begun to formulate here are much richer, in that they pertain to human individuals rather than spatial aggregations, and that there is much more detailed time-stamping of events and occurrences. However, whilst they enable detailed depictions of occurrences and flows, they do not pertain to any clearly defined population. There is a clear need for further research to identify whether it is possible to establish a basis to generalisation of social media data to night-time and day-time populations.

If successful, such efforts would advance our understanding of the ebbs and flows of the various activities that are integral to the form and function of cities (Batty and Longley 1994).

Time and space define the locus of human interactions, and geo-temporal demographics have the potential to represent interactions at granularities that complement those available for built form. They provide a framework for representing and hence understanding the geo-temporal interactions through which the morphology of urban land use and social networks mutually reinforce one another, at spatial and temporal scales that are much more fine-grained than hitherto. The motivation for all of this is consistent with a central quest of relating urban form to function through the spatial processes that result in observed forms and patterns of interaction (Batty 2013).

Acknowledgements

This work was completed as part of the EPSRC research Grant "****" (EP/***). We are very grateful to our collaborator **** for suggesting the analysis for Section 6, and (as so often) to Mike Batty for helpful ideas and comments.

References

Adnan, M., Longley, P.A., Singleton, A.D., Brunsdon, C. 2010. Towards real-time geodemographics: clustering algorithm performance for large multidimensional spatial databases. *Transactions in GIS*, 14 (3): 283-97.

Al Zamal, F., Liu, W., and Ruths, D. 2012. Homophily and latent attribute inference: inferring latent attributes of twitter users from neighbors. In *Proceedings of the Sixth International AAAI conference on Weblogs and Social Media*.

Batty, M. 2010. Space, scale, and scaling in entropy maximizing. *Geographical Analysis* 42: 395–421

Batty, M. 2013. *The New Science of Cities*. Cambridge, Mass., MIT Press.

Batty, M., Longley, P.A. 1994. *Fractal Cities: a Geometry of Form and Function*. London, Academic Press.

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

Birkin, M., Harland, K., Malleson, N. 2014. The classification of space-time behaviour 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**: 179-192.

Cox, K. R., Golledge, R. G. (eds.) 1981. *Behavioral Problems in Geography Revisited*. New York: Methuen.

Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. 2012. The Livehoods Project: utilizing social media to understand the dynamics of a city. In *Proceedings of the Sixth International AAAI conference on Weblogs and Social Media.*

Department for Communities and Local Government 2006.Generalised Land Use Database Statistics for England 2005. London. http://webarchive.nationalarchives.gov.uk/20120919132719/http://communities.gov.uk/documents/planningandbuilding/pdf/154941.pdf

Gale, C. 2014. *Creating an Open Geodemographic Classification Using the UK Census of Population*. Ph.D. thesis, University College London.

Goodchild, M. F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69: 211-21.

Harris, R., Sleight, P., Webber, R. 20015. *Geodemographics, GIS and Neighbourhood Targeting*. Chichester, Wiley.

Longley, P.A., Adnan, M., Lansley, G. 2015. The geo-temporal demographics of Twitter usage. *Environment and Planning A* 47: 465-84.

Mateos, P., Longley, P. A., O'Sullivan, D. 2011. Ethnicity and population structure in personal naming networks. *PloS ONE (Public Library of Science)* 6 (9) e22943.

Pennacchiotti, M., Popescu, A. 2011. A machine learning approach to Twitter user classification. In *Proceedings of the Fifth International AAAI conference on Weblogs and Social Media.*

Quercia, D., Seaghdha, D.O., Crowcroft, J. 2012. Talk of the City: Our Tweets, Our Community Happiness. In *Proceedings of the Sixth International AAAI conference on Weblogs and Social Media.*

Shannon, C. 1948. A mathematical theory of communication. *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, July and October.

Shelton, T., A. Poorthuis, and M. Zook. 2015. Social media and the city: rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*. In press.

Shelton, T., A. Poorthuis, M. Graham, and M. Zook. 2014. Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'. *Geoforum* 52: 167-179.

Singleton, A.D., Longley, P.A. 2015. The population structure of Greater London: a comparison of national and regional geodemographic models. At *Papers in Regional Science*.

Stephens, M., Poorthuis, A.. 2014. Follow thy neighbor: connecting the social and the spatial networks on Twitter". *Computers, Environment, and Urban Systems*. In press.

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

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

Vickers, D.W., Rees, P.H. 2007. Creating the National Statistics 2001 Output Area Classification. *Journal of the Royal Statistical Society, Series A*. 170(2), 379-403.

Ward, J. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* 58, 301: 263 – 244.

Zook, M., Poorthuis, A. 2014. Offline brews and online views: exploring the geography of beer Tweets". In The Geography of Beer, eds. M. Patterson and N. Hoalst-Pullen. Springer. pp. 201-209.