Exploring vernacular perceptions of spatial entities: Using Twitter data and R for delimiting vague, informal neighbourhood units in Inner London, UK.

Luke Thomas Clasper
UNIGIS UK, The University of Salford,
43 Crescent, Salford, UK, M5 4WT.
l.t.clasper@edu.salford.ac.uk

Abstract

The informal nature of how citizens discuss and conceive geographic entities, such as neighbourhoods, has traditionally be difficult to capture. Social media services offer researchers an opportunity to collect large amounts of geo-referenced information concerning vernacular geography. Twitter data was harvested and analysed in R statistical software in order demonstrate whether geodata from social media is a feasible method for spatially defining vague, vernacular neighbourhood units in Inner London, UK. The results suggest that social media data can be a valuable source for capturing vernacular geography from which vernacular neighbourhoods could be delimited. The study also revealed factors which may have contributed to vernacular neighbourhood unit demarcation. Twitter data was seen to both mirror the physical form of the underlying topography and reflect the social character of the city's land use. This work builds upon previous attempts to investigate vernacular geography which used more traditional methods, such as sketch maps and interviews. It also examines how qualitative coding can improve data quality and how R statistical software can be used to capture, analyse and present geospatial data.

Keywords: Vernacular Geography, VGI, Twitter, R Project, Qualitative GIS, Urban Morphology.

1 Introduction

The replication of physical urban structure by social activity and the interactions between them have long be theorised (Tonkiss, 2013). Citizens can also offer us an insight into geographic form and the way that space is used, perceived and referred to. This bottom-up, unofficial discourse concerning geographic place is fundamental to our understanding of cities (Lynch, 1960). The question of how to reproduce these colloquial, vague notions of space in its rigid representations is also central to the advancement of GIS software (Goodchild, 2011).

This study is concerned with individual's awareness of fuzzy, abstract geographic regions in relation to their own location. These regions could be informal or defined by an official government boundary. This type of vernacular geography has important applications. The emergency services find these types of colloquial indications of place invaluable when locating reported incidents. There are commercial applications for deliveries and in-vehicle navigation and government uses for the allocation of services and collection of census data.

Collecting and aggregating this type of qualitative, casual, and often ambiguous information has proved challenging (Montello *et al,* 2003). In this paper VGI from social media will be harvested, processed and analysed in R in order and capture vernacular indications of which neighbourhood individuals think they are located in.

The qualitative nature of social media VGI is widely acknowledged, for this reason qualitative coding techniques will be employed to test for dataset veracity. Cope (2003) describes the coding of qualitative textual data as a way of interpreting and filtering data in order to categorise or classify it into themes.

Qualitative GIS (or mixed-methods GIS) is the integration of qualitative data with the quantitative analysis capabilities of GIS (Cope & Elwood, 2009). Elwood & Cope (2009) describe Qualitative GIS as an extension of GIS which includes non-numerical data, the mixing of methodologies, technologies and data, citizen participation and social practices.

After this quality testing, the point-based dataset will be validated against government administrative polygons and place name seeds. The delimitation of precise neighbourhood boundaries will then be attempted, along with an investigation into the origins and demarcation factors affecting these vernacular neighbourhoods.

2 Related Work

Vernacular geography has traditionally been captured from participants who are consciously aware that they are involved in the study. This was mainly accomplished by asking the subjects to draw sketch maps (Coulton *et al*, 2012) or using interviews and questionnaires (Vallée *et al*, 2014).

There is a growing body of work where vernacular geography has been captured from unconscious participants. These techniques have involved the use of web-scraping, (Brindley *et al*, 2014) and geo-tagged Flickr photographs (Hu *et al*, 2015).

Twitter data has been used in varied GIScience research fields. The most notable of these are for exploring social issues (Shelton *et al*, 2015) and for event detection, particularly with regards to natural disasters (Crooks *et al*, 2013).

The study of how invisible social factors affect and mirror the physical fabric of cities has successfully been explored using social media data. Batty *et al* (2013) saw street networks and population densities replicated virtually by Tweets and Ferrari *et al* (2011) extracted urban mobility flows from social media data. These studies show us how VGI is affectively linking the virtual world with the underlying physical urban structure by revealing geo-located virtual traces of processes and activities (Steiger *et al*, 2015).

3 Data and Methods

3.1 Data

Twitter (2017) states that it has 313 million active monthly users, Lansley & Longley (2016) calculate that this results in 500 million daily Tweets. Twitter data will be collected for a number of neighbourhoods within Inner London. The dataset will consist of individual Tweets, each containing numerous fields of information. Two of these fields are a longitude and latitude which the Tweet was sent from, this will be used to form a point geometry object. Other fields include; the status update text of the Tweet, creation date and time, the source of the Tweet (e.g. from a linked social media site such as Instagram), the screen name of the user and a unique identifier for the Tweet.

3.2 Software

The open-source statistical software environment R and the R language will be used for all data collection and analysis. R offers excellent research reproducibility and self-documentation due to its command line format. It is also efficient at analysing large datasets, repeating tasks and has the ability to draw down base mapping through internet calls. R includes packages which provide functions and code libraries for statistics, visualisation, data handling and data collection.

3.3 Data collection from the Twitter API

The Twitter API will be accessed from R using the twitteR library. Tweets will be filtered based on the searchTwitter() function's geocode argument and a query for keywords and hashtags which reference an Inner London neighbourhood (e.g. Soho, #Soho). The geocode argument specifies a geographic location (a latitude/longitude) and a search radius, which will both remain constant. The geographic location chosen is Charing Cross (traditionally thought of as the centre of London) and a search radius of 5 miles, which is the extents of the study area (the current Congestion Charge Zone https://tfl.gov.uk/modes/driving/congestion-charge/congestion-charge-zone).

The keywords and hashtags will be changed each time the query is run depending on the neighbourhood that is being researched, the query will be run multiple times for each neighbourhood. Tweet retrievals from the Twitter API are limited to 1500 each time the query is run. Longley *et a*l (2015) suggest that this is roughly 1% of a random selection of Tweets, however Lansley & Longley (2016) advise that this small percentage can still obtain over 90% of all geo-tagged Tweets. 28 neighbourhoods will be studied, selections were based on place name seeds from OpenStreetMap and Ordnance Survey to give an even spread throughout the study area. Table 1 shows the neighbourhoods selected and if they have a current official administrative boundary namesake.

Table 1: Neighbourhoods selected for keywords and hashtags.

Neighbourhood	Official Administrative Boundary?
Aldgate	Aldgate Ward
Barbican	No
Bishopsgate	Bishopsgate Ward
Blackfriars	No
Bloomsbury	Bloomsbury Ward
Brick Lane	No
Clerkenwell	Clerkenwell Ward
Covent Garden	Holborn and Covent Garden Ward
Elephant and Castle	No
Euston	No
Farringdon	Farringdon Within Ward, Farringdon Without Ward
Fitzrovia	No
Holborn	Holborn and Covent Garden Ward
Hoxton	Hoxton Ward
Kings Cross	Kings Cross Ward
Lambeth	Lambeth London Borough, Lambeth and Southwark GL Assembly Const
Leicester Square	No
Marylebone	Marylebone High Street Ward
Mayfair	No
Paddington	No
Shoreditch	No
Soho	No
Southbank	No
Southwark	Southwark London Borough, Lambeth and Southwark GL Assembly Const
Spitalfields	Spitalfields and Banglatown Ward
Strand	No
Vauxhall	No
Waterloo	No

3.4 Qualiataive thematic coding

Lovelace *et al* (2016) concede that social media data suffers from a lack of veracity. To improve data quality a methodology of quantitative coding will be used. The manual scrutinising of geo-tagged social media data for locational errors was implemented by Hollenstein & Purves (2010). The qualitative examination of Tweets for topic related errors was considered by Albuquerque *et al* (2015). A combination of these two approaches will take place as soon as the Twitter dataset is collected to produce a derived, quality controlled, dataset ready for analysis. Tweets will first be visualized geographically and assessed for outlying Tweets in unexpected locations. All Tweet's text will then be examined for off topic subject matter that indicate the user is not located within the neighbourhood that they are Tweeting about. Finally, Tweets will be filtered by assigning each one with a textual code depending on its content, as was implemented by Jung (2015).

The qualitative coding is designed to find and categorise Tweets that may be sent from outside the neighbourhood, e.g. Tweets sent travelling to or from a neighbourhood. Coding will also find a neighbourhood keyword used in the wrong context, e.g. a Tweet about a person or entity named after the neighbourhood or a Tweet about an event that took place in the neighbourhood or will take place in the future.

3.5 Point clustering and polygon delimitation

A combination of methods will be used to investigate the point clusters of Tweets, both before and after qualitative coding. The Tweet point patterns' mean centres will be calculated and the dispersions of neighbourhood Tweets around official place name seeds will be recorded. Standard Distance Deviation (SDD) of Tweet dispersals will be calculated in order to validate the dataset. Standard Deviational Ellipses (SDE) will then be used to look for any underlying directional factors affecting Tweet dispersal. Kernel Density Estimation (KDE) will be employed to determine neighbourhood point concentrations. These will be geographically compared to official administrative boundaries (where they exist) to again validate the dataset. 2D contours will be used to research the basis of neighbourhoods, this will look for centres, or origins, of neighbourhoods in order to draw conclusions about reasons for vernacular neighbourhood demarcation. Finally, convex hulls will be built to spatially determine neighbourhood extents and create discrete vernacular neighbourhood polygons.

4 Results

4.1 Results for Tweet collection

Over a period of two months 31,692 Tweets were collected, sent by 14,832 individual Twitter users. Figure 1 shows the uneven distribution of Tweets collected between the neighbourhoods. Shoreditch and Soho have by far the greatest number of Tweets, followed by Covent Garden and Mayfair.

When the Tweets are viewed spatially (Figure 2) denser point clusters of Tweets can be found to the west and east of the study area. As well as highlighting the areas of high Tweet intensity the 2D density estimation contours (Figure 3) highlight the areas of sparse Tweet coverage. These can be seen around the City of London, Westminster, Hyde Park, Regent's Park, Green Park and large swathes to the north and south of the study area.

Figure 1: Distribution of collected Tweets between neighbourhoods.

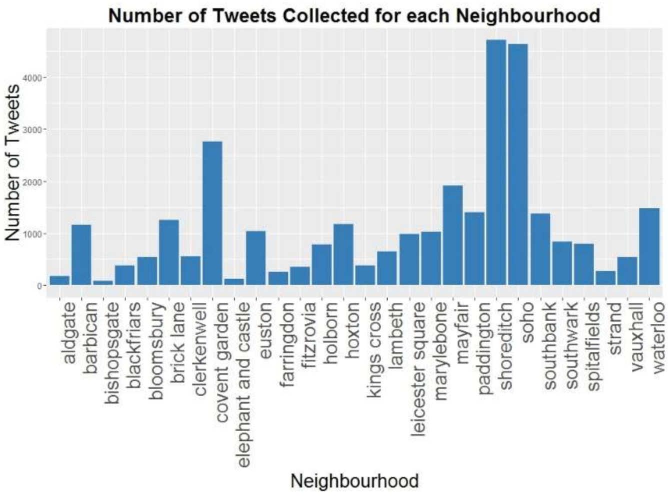


Figure 2: All Tweets collected for each neighbourhood. © OpenStreetMap contributors (http://www.openstreetmap.org/copyright).

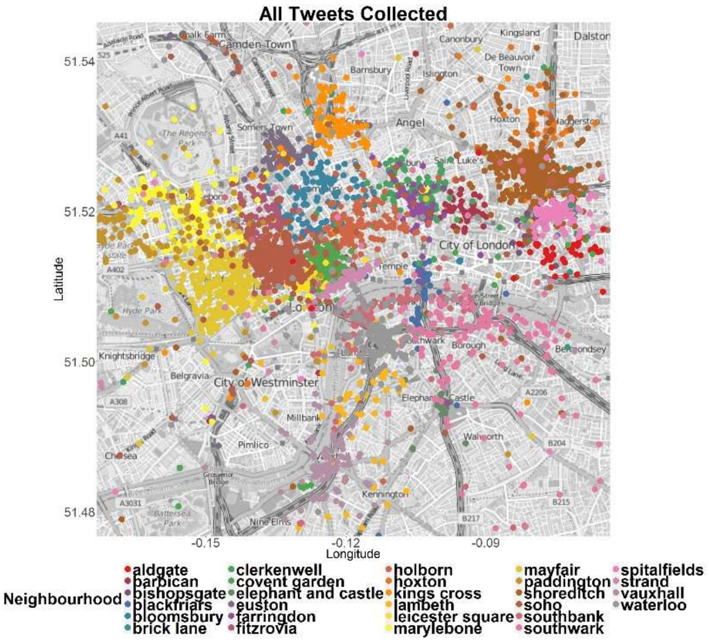
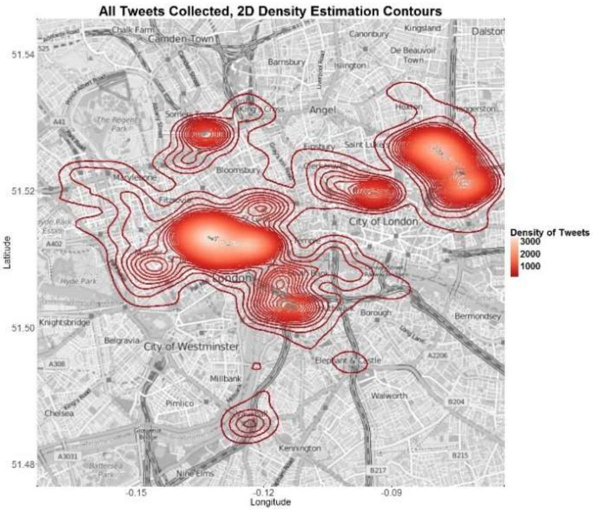


Figure 3: 2D density estimation contours for all Tweets collected. © OpenStreetMap contributors.



4.2 Qualitative coding results

The results of the qualitative coding exercise are presented in Table 2. What is clear from this is that the Well Located Tweet category includes by far the greatest number of Tweets, with around 95%.

Table 2: Results of the qualitative coding.

Qualitative Coding Category	Number of Tweets
GPS positional error	393
Travelling to or from neighbourhood	69
Truncated coordinates	49
Tweet about person or entity named after neighbourhood	108
Tweeting from a venue named after neighbourhood	286
Tweeting from other location about neighbourhood	430
Uncertain outlier	47
Well located Tweet	30,210

4.3 Cluster analysis results

The standard distance between official place name seeds and mean centres of clusters decreased after qualitative coding. The SDD results saw dispersal of Tweets around the means decrease for all neighbourhoods after qualitative coding, they also gave an indication to how dispersed or compact a neighbourhood is. In all cases the KDEs of the neighbourhood Tweets are within or in close proximity to their official boundary. Figure 4 shows the KDE for Marylebone. The results of the SDE analysis demonstrated the directional tendencies of the neighbourhood Tweets. Figure 5 illustrates this for the Brick Lane neighbourhood, showing a linear directional trend along its namesake thoroughfare.

Figure 4: KDE for Marylebone. © OpenStreetMap contributors

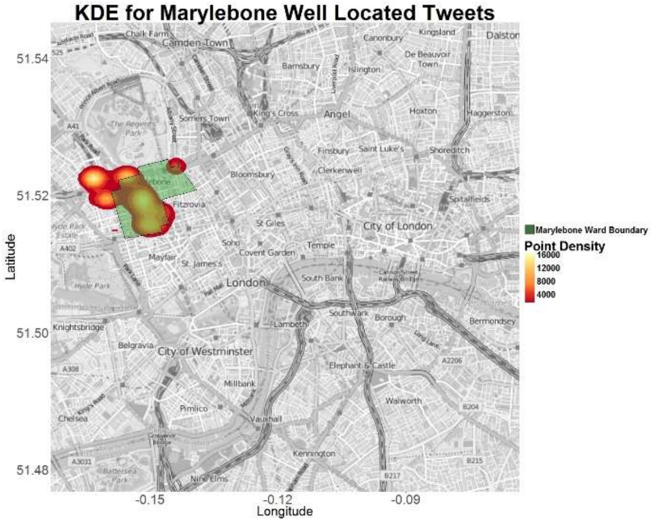
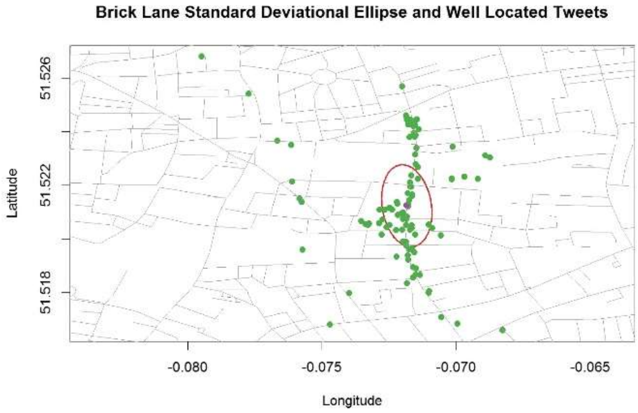


Figure 5: SDE result for the Brick Lane neighbourhood. Contains OS OpenData © Crown Copyright/database right 2017.



4.4 Results for delimiting vernacular units and centres

Each neighbourhood's Tweets were delimitated into vernacular neighbourhood polygons using convex hulls. The results for the neighbourhoods in West London are presented in Figure 6. The epicentres and origins of neighbourhoods were examined with 2D density contours (Figure 7).

Figure 6: West London vernacular neighbourhood boundaries. © OpenStreetMap contributors

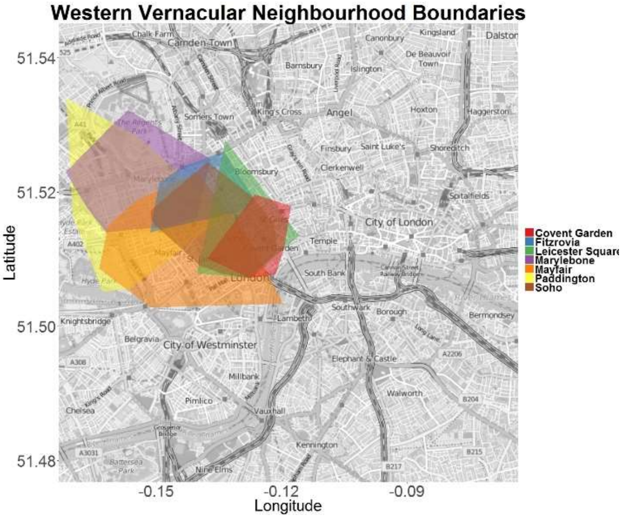
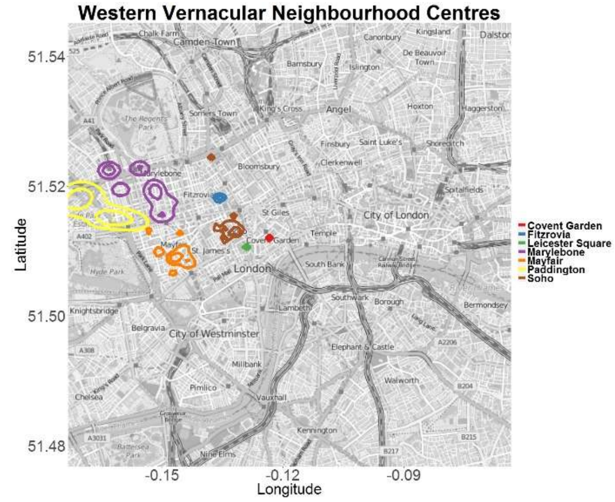


Figure 7: West London neighbourhood centres. © OpenStreetMap contributors



5 Interpretation of Results

A point based dataset of Tweets concerning vernacular perceptions of place was successfully compiled. The accuracy, and overall certainty of this dataset was then enhanced by qualitative coding.

The delimitation of discrete vernacular neighbourhood polygons from the fuzzy, validated Tweet dataset proved effective. The overall visualisation of neighbourhood polygons (Figure 6) provides a very overlapping picture, reflecting the underlying discord between individuals' spatial perceptions. There were also positive correlations between vernacular neighbourhood polygons and the official administrative boundaries, highlighted in Figure 4. Figure 8, demonstrates the overlap between the Bishopsgate vernacular neighbourhood polygon and the official Bishopsgate Ward polygon. To provide a quantitate measure of overlap between the vernacular neighbourhood polygons and the official boundary polygons the percentage of the vernacular polygons which intersect with their official boundary polygon (where they exist) was calculated. The results are illustrated in figure 9.

Figure 8: Bishopsgate vernacular neighbourhood and Bishopsgate Ward. Contains OS OpenData © Crown Copyright/database right 2017.

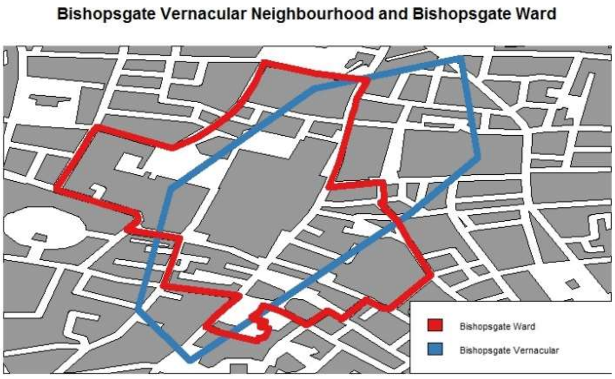
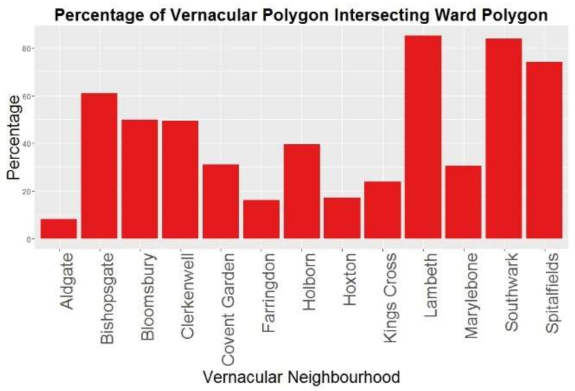


Figure 9: Percentage of vernacular polygons intersecting with their namesake ward polygons.



There are few Tweets in the City of London, in parks and in the large swathes of residential land to the north and south of the study area. Here we see areas of open green spaces, rivers, roads and zones of lower social activity effectively acting as edges (Lynch, 1960). Conversely, areas of social functionality are exhibiting higher social media activity and mirroring the underlying topography and density of London. An example of an edge delimiting a vernacular neighbourhood is presented in Figure 10, where the River Thames is acting as a perimeter, or a physical boundary, to the Southbank vernacular neighbourhood. Landmarks and transport hubs were also seen to form the basis of neighbourhoods, illustrated in Figure 11 with Waterloo.

Figure 10: River Thames and Southbank vernacular neighbourhood. Contains OS OpenData © Crown Copyright/database right 2017.

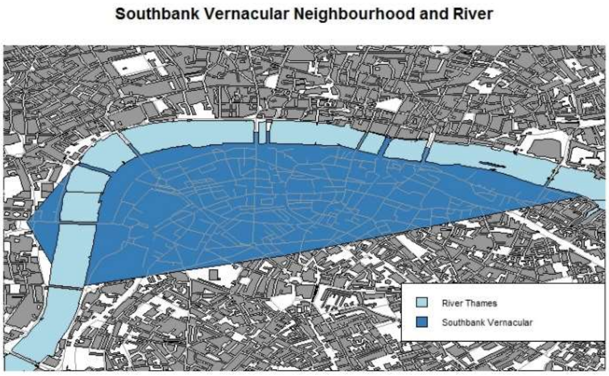
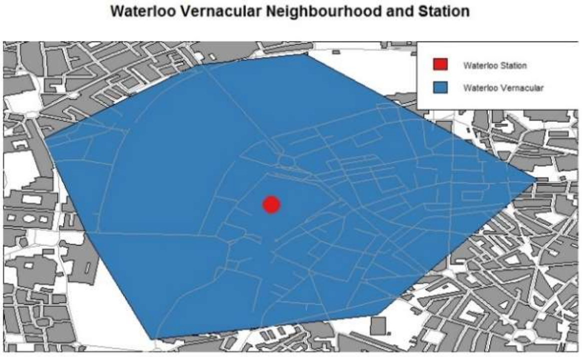


Figure 11: Waterloo vernacular neighbourhood. Contains OS OpenData © Crown Copyright/database right 2017.



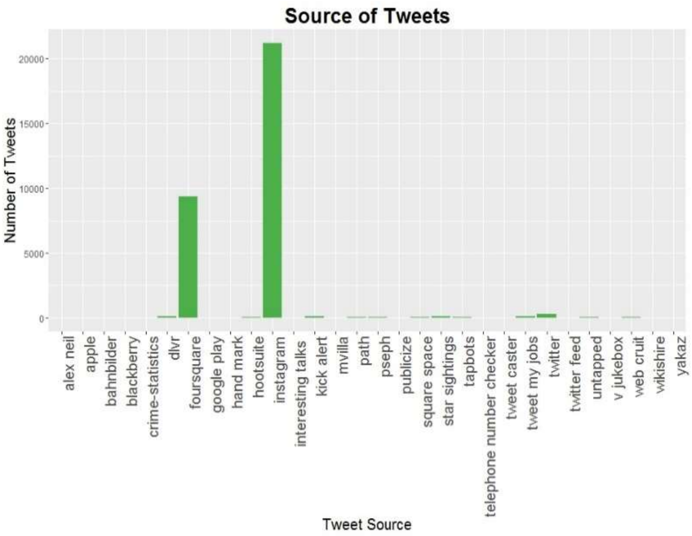
6 Conclusions and Further Work

31,692 Tweets were collected; this is considerably more responses than it would be feasibly possible to collected from traditional questionnaire or sketch map techniques. Slightly concerning is that these Tweets were sent by only 14,832 Twitter users, which means only 14,832 individual perceptions from which to study vernacular geography. This phenomenon of a few highly active users dominating Twitter output, and affecting research, was also observed by Shelton *et al* (2015).

The subjective and time consuming nature of the manual qualitative coding process could be a limitation if the study was to be expanded. To negate these concerns a process of machine learning could be devised (in R). However, Hahmann *et al* (2014) found that in the context of their study human Tweet classification proved to be more accurate than automated text detection techniques.

Twitter allows users to Tweet from other utilities, which links data between websites. Within the Source field of the Tweet is the name of the utility used to post a Tweet. When plotted (Figure 12) it becomes evident that a very high proportion of the Tweets collected were also posted on Instagram and Foursquare. In future work it would be

Figure 12: The source utilities of the Tweets collected.



This research has demonstrated the feasibility of capturing vernacular geography from Twitter. The next step will be to link the Twitter-derived vernacular neighbourhood boundaries to OpenStreetMap. Thus adding a new level of detail and granularity to the collaborative mapping project and also connecting two forms of VGI. It would also be interesting to test data quality further by comparing the vernacular neighbourhood polygons form this study against polygons derived from other VGI sources such as Flickr.

Refrences

Albuquerque, J, P. de., Herfort, B., Brenning, A. & Zipf, A. (2015) A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. International Journal of Geographical Information Science, 29 (4): 667-689.

Batty, M., Gray, S., Hudson-Smith, A., Milton, R., O'Brien, O. & Roumpani, F. (2013) Visualising spatial and social media. CASA working papers series, paper 190: London: University College London.

Brindley, P., Goulding, J. & Wilson, M. L. (2014) Mapping Urban Neighbourhoods from Internet Derived Data. In Proceedings of GISRUK 2014: 355-364.

Cope, M. (2003) Coding Transcripts and Diaries. In Clifford, N., French, S. & Valentine, G. (Editors) Key Methods in Geography. Chapter 27: 440-452. London: Sage.

Cope, M. & Elwood, M. (2009) Qualitative GIS. A mixed methods approach: London: Sage.

Coulton, C.J., Jennings, M.Z. & Chan, T. (2012) How Big is My Neighbourhood? Individual and Contextual Effects on Perceptions of Neighbourhood Scale. American Journal of Community Psychology 51 (1-2): 140-150.

Crooks, A., Croitoru, A., Stefanidis, A. & Radzikowski, J. (2013) #Earthquake: Twitter as a Distributed Sensor System. Transactions in GIS 17 (1): 124-147.

Elwood, M. & Cope, M. (2009) Introduction: Qualitative GIS: Forging Mixed Methods through Representations, Analytical Innovations, and Conceptual Engagements. In Cope, M. & Elwood, M. (Editors). Qualitative GIS. A mixed methods approach: (Chapter 1): 1-12: London: SAGE

Ferrari, L., Rosi, A., Mamei, M. & Zambonelli, F. (2011) Extracting urban patterns from location-based social networks. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, November 1st 2011, Chicargo, IL, USA: 9-16.

Goodchild, M.F. (2011) Formalizing place in geographic information systems. In Burton, L.M, Kemp, S.P., Leung, M., Matthews, S.A. & Takeuchi, D.T. (Editors), Communities, Neighbourhoods and Health: Expanding the Boundaries of Place (Chapter 2): 21-33. New York: Springer.

Hahmann, S., Purves, R.S. & Burghardt, D. (2014) Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. Journal of Spatial Information Science 9: 1-36.

Hollenstein, L. & Purves, R.S. (2010) Exploring place through user-generated content: Using Flickr tags to describe city cores. Journal of Spatial Information Science 1: 21-48.

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W. & Prasad, S. (2015) Extracting and understanding urban areas of interest using geotagged photos. Computers, Environment and Urban Systems 54: 240-254.

Jung, J.K. (2015) Code clouds: Qualitative geovisualization of geotweets. The Canadian Geographer 59 (1): 52-68.

Lansley, G. & Longley, P.A. (2016) The geography of Twitter topics in London. Computers, Environment and Urban Systems 58: 85-96.

Longley, P.A., Adnan, M. & Lansley, G. (2015) The geotemporal demographies of Twitter usage. Environment and Planning 47: 465-484.

Lovelace, R., Birkin, M., Cross, P. & Clarke, M. (2016) From Big Noise to Big Data: Towards the verification of Large Data sets for Understanding Region Retails Flows. Geographical Analysis 48: 59-81.

Lynch, K. (1960) The Image of the City. Cambridge (Massachusetts): The MIT Press.

Montello, D.R., Goodchild, M.F., Gottsegen, J, & Fohl, P. (2003) Where's downtown? Behavioural methods for determining referents of vague spatial queries. Spatial Cognition and Computation, 3 (2 & 3): 185-204.

Shelton, T., Poorthuis, A. & Zook, M. (2015) Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. Landscape and Urban Planning 142: 198-211.

Steiger, E., Westerholt, R., Resch. B. & Zipf, A. (2015) Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. Computers, Environment and Urban Systems 54: 255-265.

Tonkiss, F. (2013) Cities by design: the social life of urban form. Cambridge: Polity.

Twitter (2017) Twitter Usage/Company Facts (Last updated June 30, 2016). Available from: https://about.twitter.com/company [Accessed 12th March, 2017].

Vallée, J., Le Roux, G., Chaix, B., Kestens, Y. & Chauvin, P. (2015) The 'constant size neighbourhood trap' in accessibility and health studies. Urban Studies 52 (2): 338-357.