# The Spatial Analysis of Short-term Population Movements with Social Media Data

M. Birkin[1], N. Malleson[2]

[1]University of Leeds, Leeds LS2 9JT
Telephone: +44 113 3433307
Fax: +44 113 3433308
Email: m.h.birkin@leeds.ac.uk

[2]University of Leeds, Leeds LS2 9JT
Telephone: +44 113 3433307
Fax: +44 113 3433308
Email:n.malleson06@leeds.ac.uk

## 1. Background and Introduction

In recent years, significant traction has been gained through the notion of big data. For example, according to Bell, Hey and Szalay (2009) research in the natural sciences is now entering a 'fourth paradigm' in which empirical investigations are underpinned by prolific data sources, which could include human genetic codes, massive astrophysical data from space telescopes, or satellite images of land surfaces and the atmosphere. Similar trends have been noted in geocomputation, ranging from the availability of crowd-sourced or volunteered geographical information (Goodchild, 2007) to increasing freedom of access to public datasets (www.data.gov.uk).

Telecommunications is an especially interesting domain in which new streams of data are emerging. Increases in the use of mobile telephone technologies have been spectacular and persistent, while the spatial codes associated with telephone traffic are becoming more prevalent and robust. This paper considers data generated by the social messaging service twitter. It will be shown that many users are now in the habit of tweeting messages throughout the day and night, and these messages are often spatially tagged. These data can be randomly sampled, providing unique insights into the short-term movement patterns of individual tweeters across and around the city.

## 2. Description of the data

Messages have been captured from the twitter service over a two month period in July and August 2011 for the city of Leeds (see also Malleson and Birkin, 2012). Each of these tweets contains a message (for example, 'I love geocomputation'), a spatial reference (xy coordinate) and a unique user identifier. Under the terms of service, messages have a maximum of 140 characters. Interested listeners can simply 'eavesdrop' on twitter conversations, with the (significant) restriction that only a limited sample of the total universe of messages is retained. Furthermore, in order to capture tweet locations the device from which each message is sourced must have a geographical positioning capability, and that capability must be enabled. The vast majority of messages (more than 95%) are not spatially encoded and therefore disregarded in this study. The capture of messages has continued since August 2011; hence although the extract to be considered here has 290K messages, the full corpus now runs to nearly four million. Part of the agenda here should therefore be seen as a desire to create automated methods for the analysis of quite extensive corpora.

## 3. Language variations in space and time

As a preliminary step in the data analysis, a complete enumeration of individual words has been undertaken. In this context, a ‘word’ is simply defined as a continuous string of characters bounded at each end by spaces. The corpus contains 3.38 million occurrences of 263,778 unique words. One hundred of the most widely used nouns and substantive words are shown in Figure 1.

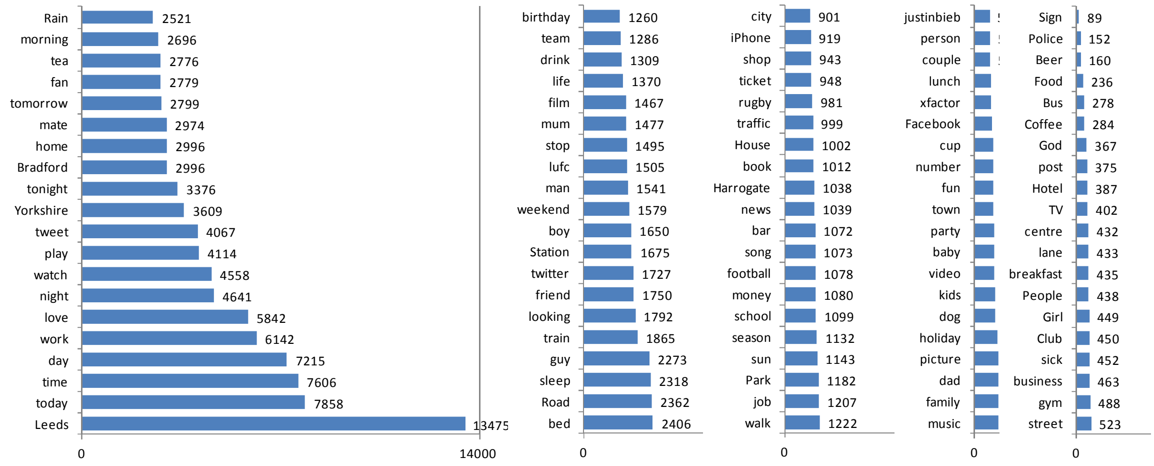


Figure 1. Substantive words favoured on twitter

An interesting question which we wish to consider now is whether individual words have their own particular traces in space and time. For example, are there some words which most commonly present themselves in the city centre, and others in the suburbs? Do some words or expressions only come out at night?! The top ten urban words are shown in Table 1 for locations within 2km of the city centre. In addition to three obvious location words (‘Leeds’, ‘Centre’ and ‘City’) the list portrays the role of the city as a hub for transport (‘Leeds’, ‘Bus’, ‘Street’, ‘Station’) and the commercial heart of the region (‘business’, ‘hotel’). The word ‘picture’ seems to hint at another function of the central area as a nexus for leisure and entertainment, as this word is commonly used in the act of uploading or inspecting pictures from social networking sites.

Table 1 also shows the major words associated with time classified as home (night-time), work (daytime) and play (evenings and weekends). Here it appears that home is largely associated with hygiene activities (‘morning’, ‘breakfast’, ‘sleep’, ‘food’) while the pressures of commuting are high amongst the concerns of the working day (‘traffic’, ‘bus’, ‘road’, ‘ticket’). There is a pleasing correlation between playtime and recreational pursuits (‘rugby’, ‘football’, ‘shopping’, ’drinking’ and of course the unmissable ‘Xfactor’).

| City words | Home words | Work words | Play words |
|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Station | 494 | morning | 206 | traffic | 180 | xfactor | 228 |
| Leeds | 319 | breakfast | 203 | Coffee | 170 | bed | 146 |
| Hotel | 244 | rugby | 165 | lunch | 168 | rugby | 145 |
| street | 197 | sleep | 164 | Bus | 167 | football | 141 |
| centre | 182 | Station | 160 | Road | 165 | drink | 138 |
| business | 181 | Rain | 157 | Harrogate | 155 | person | 136 |
| train | 180 | Food | 156 | lane | 151 | shop | 129 |
| picture | 179 | Hotel | 152 | ticket | 136 | picture | 124 |
| city | 178 | Club | 150 | House | 136 | tonight | 122 |
| Bus | 170 | today | 149 | business | 132 | song | 120 |

Table 1. Space-time exploration of twitter words

## 4. Investigation of spatial clustering

In order to promote further investigation of the twitter messages, data was prepared for input to the GAM software. The Geographical Analysis Machine (GAM) is a cluster detection algorithm created by Stan Openshaw and colleagues in the early days of geocomputation (e.g. Openshaw et al, 1987). Multiple runs of the GAM is a time-consuming process requiring manual intervention, so complete analysis of the 100 word sample was not possible at this juncture. The search for 'interesting' words was qualified by the calculation of an index of dissimilarity for each word, which compares the spatial distribution across all census wards in the Leeds area (see Table 2). The appearance of the word 'rain' at the top of this list is spurious, as it turns out that the majority of tweets are produced by a fixed point weather station (typical tweet:. "Wind 4.5 mph W. Barometer 994.4 Falling slowly. Temperature 17.9 °C. Rain today 0.0 mm. Humidity 65%"). This is not necessarily completely uninteresting, as it suggests the possibility of simple spatial analysis to detect and eliminate bogus messages of this type.

| Words and frequencies | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Rain | 0.480 | Harrogate | 0.309 | lufc | 0.252 | Coffee | 0.219 | lane | 0.197 |
| justinbie | 0.349 | Sign | 0.292 | Food | 0.238 | today | 0.215 | Beer | 0.180 |
| World | 0.336 | Hotel | 0.272 | House | 0.221 | Bradford | 0.200 | business | 0.180 |
| Station | 0.332 | Bus | 0.268 | Club | 0.220 | rugby | 0.198 | xfactor | 0.178 |

Table 2. Index of dissimilarity for twitter words

Figure 2 shows outputs from the GAM for the words 'police' (top left), 'today' (top right), and 'lufc' (bottom left). Figure 2d (bottom right) shows the pattern of all tweets which is used as the base population for the GAM clustering process. At this stage it is unclear without further investigation why the words police and today should cluster in this particular fashion. LUFC is a tag favoured by supporters of the Leeds United football, and its concentrations are in and around the football ground itself which is to the south-west of the city centre. This probably reflects messaging activity from fans to and from the games, as well as clusters of supporters living relatively close by.

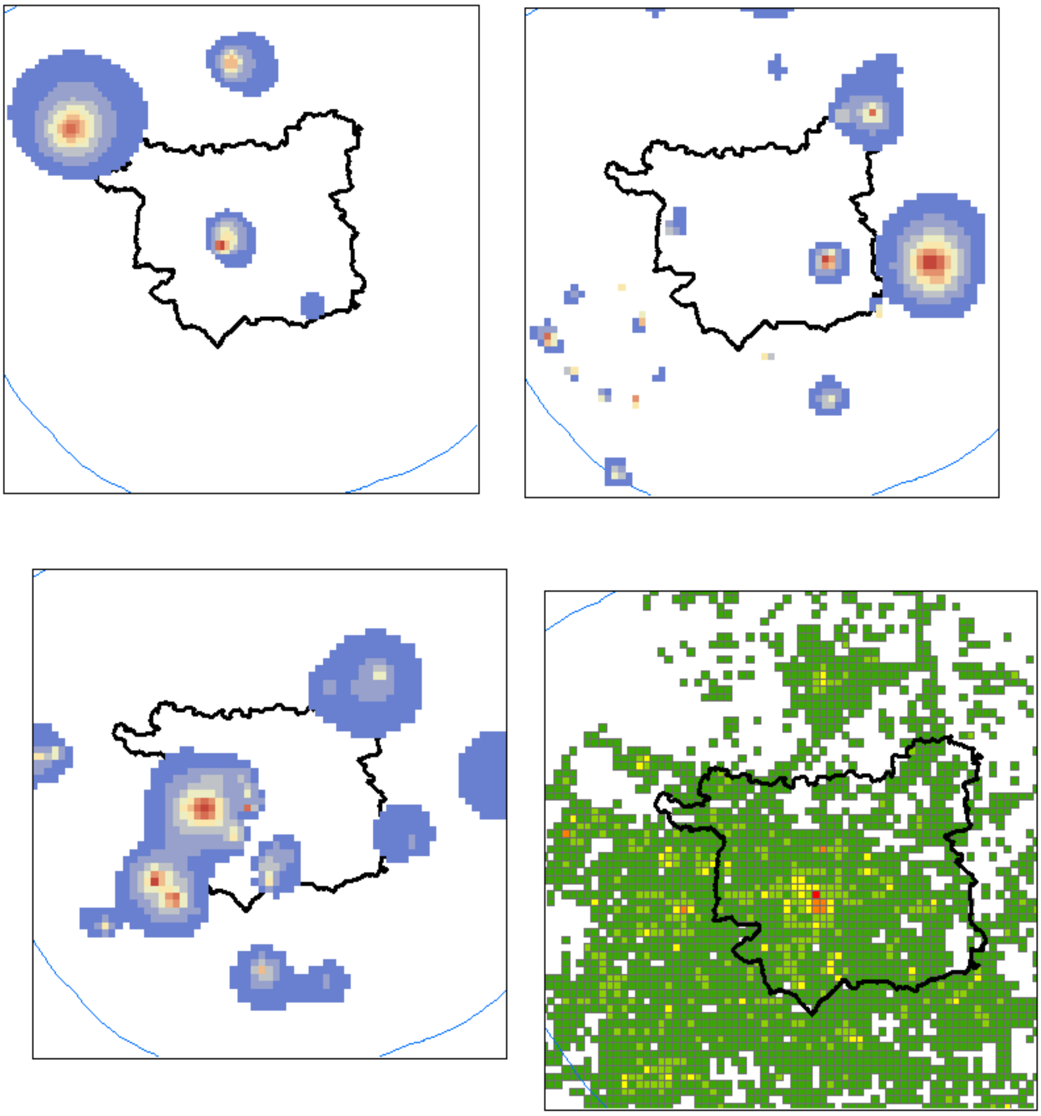
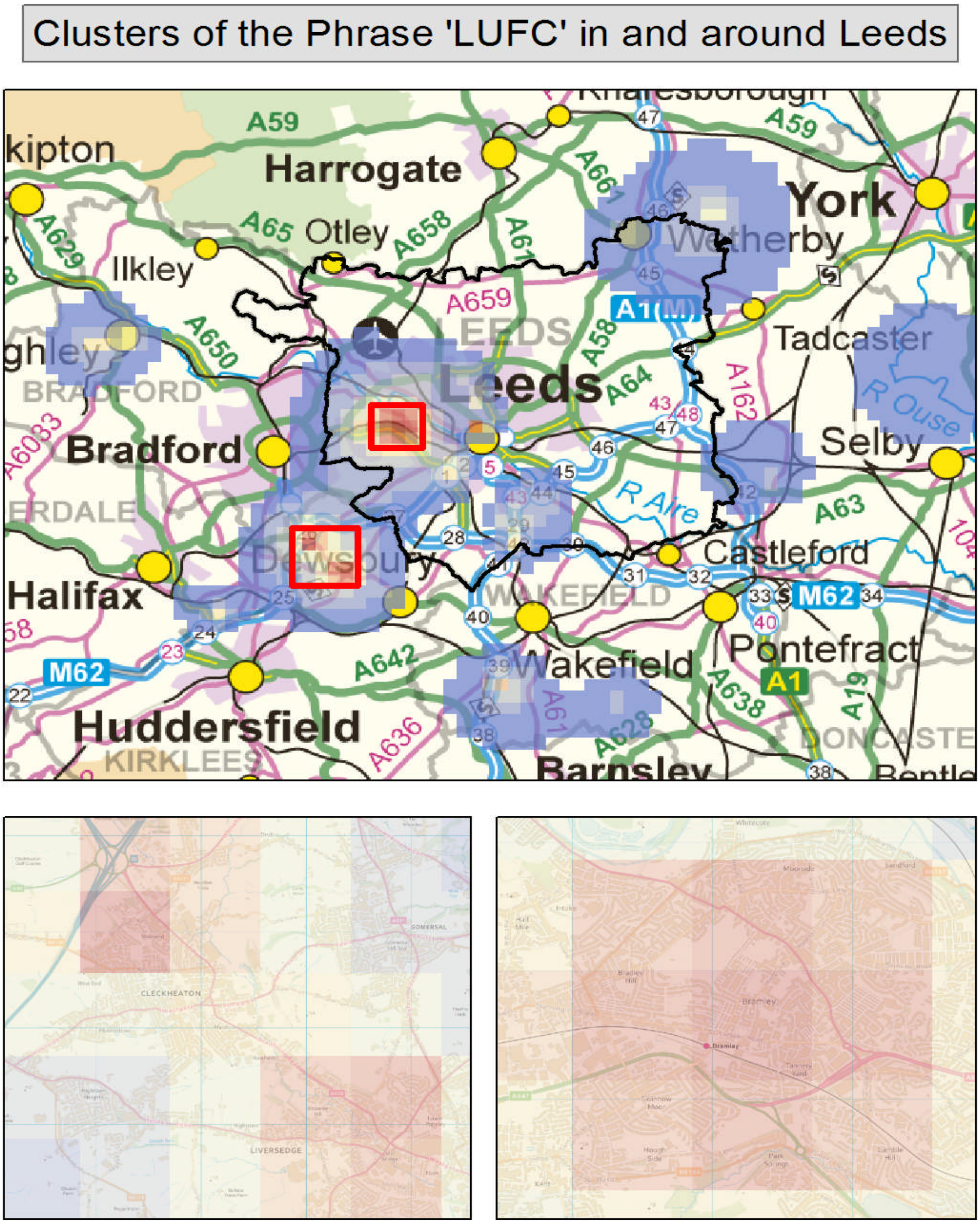


Figure 1. GAM Outputs (1a. Police; 1b. Today; 1c. LUFC; 1d. Total tweets)



© Crown Copyright/database right 2012. An Ordnance Survey/EDINA supplied service

Figure 2. Clusters of the word 'LUFC' in and around Leeds

## 5. Discussion

This paper has used the content of social messaging to present a preliminary spatial analysis and early attempts to detect clusters in space and time. For example, one of the more extravagant suggestions for the semantic analysis of social media data has been the possibility of detecting outbreaks of illnesses such as flu. Could one trace the use of words like ‘influenza’ or ‘virus’ to detect such patterns? Such work would require more intense streams of data, and perhaps an extension of the GAM technology to temporal as well as spatial analysis

The ultimate aim of this research is to support simulation models of daily activity patterns. For example if we could connect the words people use to specific locations, such as shops, restaurants or offices, then a basis for understanding local movement patterns would start to emerge. Even allowing for the unrepresentative nature of social media data, a reliable source of data about where people are, and what they are doing at what time could provide crucial information about the spread of ideas, illnesses and social phenomena such as civil unrest.

## 6. References

Bell, G., Hey, T. & Szalay, A. 2009 Beyond the data deluge. Science 323, 1297–1298. (doi:10.1126/science.1170411)

Goodchild M (2007) Citizens as Sensors: the World of Volunteered Geography, GeoJournal, 211-221.

Malleson N, Birkin M (2012) Estimating Individual Behaviour from Massive Social Data for An Urban Agent-Based Model, 8th Conference of the European Social Simulation Association (ESSA 2012), 13th September 2012, Salzburg.

Openshaw S, Charlton M, Wymer C, Craft A (1987) A Mark I Geographical Analysis Machine for the automated analysis of point data sets, International Journal of Geographical Information Systems, 1, 4, 335-358.