The Classification of Space-Time Behaviour Patterns in a British City from Crowd-Sourced Data

Mark Birkin, Kirk Harland, and Nicolas Malleson

School of Geography, University of Leeds, Leeds LS2 9JT, UK
{m.h.birkin,k.harland,n.malleson06}@leeds.ac.uk

Abstract. The use of social messaging as a means to represent activity and behaviour patterns across small geographical areas is explored. A large corpus of messages provides the source from which a range of interesting marker words are identified. Profiles of the variations in language across neighbourhoods can then be constructed. Areas are classified on the basis of the types of messages which they tend to generate. The resulting patterns are interpreted as suggesting that variations in behaviour and activity over time within an urban area are an important adjunct to well-established spatial variations. It is asserted that further elaboration of these promising investigations within appropriate analytic frameworks could extend our understanding of movement and behaviour patterns in cities in important ways.

Keywords: Geodemographics, social messaging, crowd-sourced data, cluster, activity, movement pattern.

1 Introduction

The increasing abundance of spatial and socio-demographic data has attracted widespread attention in recent years. In the UK, as in many countries, potential sources include commercial data (e.g. from store loyalty cards, telecommunications records, market research data), Open Government (e.g. national mapping data, censuses and surveys), and perhaps, most importantly, crowd-sourced data or volunteered geographical information ranging from OpenStreetMap to social media. Here the focus is on the Twitter messaging service as a source of information about spatial behaviour and movement patterns. Twitter messages ("tweets") have received some attention along with other forms of social media for the purposes of: classifying 'sentiment' relating to political or civil circumstances [1]; comparing virtual social networks to real ones [2, 3]; predicting election results [4]; exploring how opinions form on social networks [5]; and exploring human mobility patterns at a national scale [6]. However, to date there have been relatively few attempts to undertake detailed spatial analysis of the content of these messages or their implications.

In this paper the possibilities for clustering tweets will be explored. Neighbourhood classifications have been commonly used by geographers and others for the purposes of spatial analysis of urban structure, resource allocation and service delivery [7,8]).

B. Murgante et al. (Eds.): ICCSA 2013, Part IV, LNCS 7974, pp. 179–192, 2013.

Commercial organisations in particular have sought to develop individual level classifications using either large scale surveys or pooled data [9] or more recently transactional data from store loyalty cards [10]. The potential benefits of clustering messages as a means for understanding behavioural patterns, and then in linking these to more wide-ranging modelling approaches to studying variations in space and time, will be considered in the discussion.

The possibilities arising from the aggregation of messages for individual users are extensive. For example, possible links between the language adopted by individuals from different socio-demographic groups could be explored. Even more intriguingly perhaps, the possibility that individuals can be characterised not by who they are (i.e. demographics), or by where they live (i.e. geodemographics) or even by what they do (lifestyle profiling or psychographics) but by 'what they say'. In the first instance, however, a less ambitious analysis will be adopted in which messages will be aggregated across small geographical units, but also within discrete intervals of time, thus showing variations in language across different areas of the city in a typical week. Through this route, it could be possible to build an understanding of how variations in language relate to either behavioural or demographic variations across a city.

2 Data

The data were collected from the Twitter messaging service between the months of July 2011 and December 2012. A total of 992,423 messages are included in the data originating within the boundary of the city of Leeds, UK (the study area). In addition to a text string, each tweet in the sample is geo-referenced, contains an accurate time stamp and includes the name and userid of the contributor, which makes it possible to connect different messages from the same user.

For the purpose of this analysis, each message was decomposed into a series of constituent 'words'. A word is defined as any continuous group of characters which are bounded at each end by a blank space, or by a punctuation mark – a full stop, comma, semi-colon, colon or apostrophe. Words of one or two letters are excluded from subsequent analysis.

The dataset includes 17,110,484 words contained in tweets originating from within Leeds. A cursory analysis of tweets originating with the study area by user account identified a small number of high activity user accounts, as shown in Figure 1. Most accounts registered less than 1,000 tweets over the study period whereas 274 accounts registered activity much higher with one particular account posting in excess of 12,000 tweets.

However, it is not just the frequency of tweet that is unusual about some of these accounts but also the spatial concentration of the tweets originating location. This is demonstrated in Figure 2 below showing the locations and count of tweets by individual user account and geographical location within the study area. Each cylindrical prism represents the count of tweets from that location by an individual user account. It is clear to see that several of accounts with high activity are also geographically static.

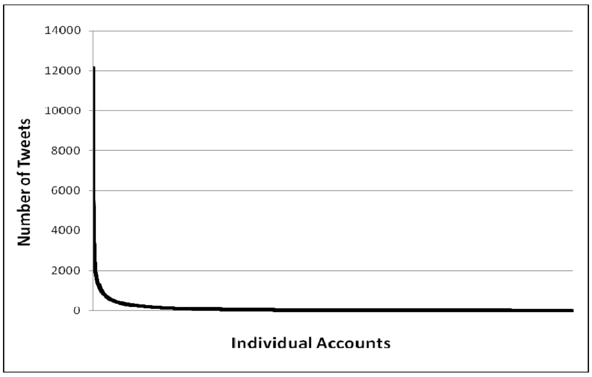

Fig. 1. Tweet frequency by user account

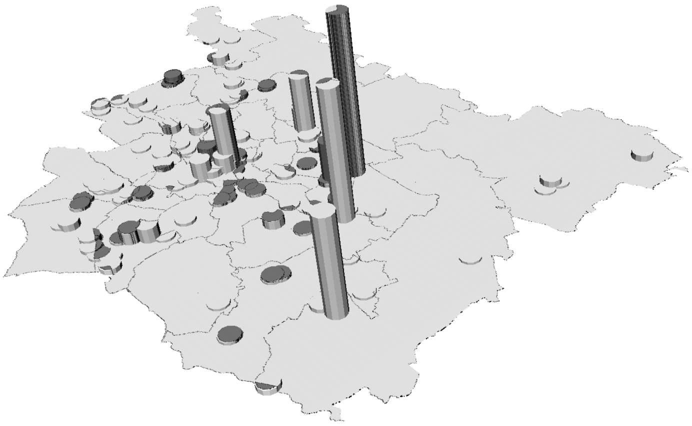

Fig. 2. User accounts and locations with over 1,000 tweets

On further investigation of the 274 highly active accounts, 12 were found to follow regular linguistic patterns containing words such as Temperature along with ‘wind’, ‘barometer’, ‘pressure’ and similar words which are typically associated with fixed weather stations which tweet atmospheric conditions at regular intervals – these tweets are therefore artificially concentrated in particular spatial locations Car – unfortunately automotive vehicles are the subject of regular advertisements in the twittersphere – the message ‘Used car recently added’ is a common banner for a promotion. Although the spatial concentration of these messages are more dispersed the linguistic pattern is repetitive and potentially introduces interference into the analysis.

Severe – often associated with regular traffic warnings, i.e. ‘severe delays’ again spatially concentrated and linguistically repetitive.

http:// – followed by an internet location denoted accounts advertising news stories associated with online magazines etc.

These 12 accounts are removed from subsequent analysis. A number of individual accounts had high numbers of tweets with a standard format ‘I’m at <X location>’ which are automatic tweets associated with location logging applications on mobile devices. These automated tweets were interspersed with manually created tweets and therefore no action was taken with these accounts. In all 11,505,719 words remained for analysis across 27,999 individual user accounts.

The density of tweets for wards in Leeds is by far the highest for areas around the central business district (City and Holbeck 19,273 tweets per km2) and student areas (Headingly 16,392 tweets per km2 and University 15,800 tweets per km2). This is to be expected with both the demographic bias of twitter users, younger users such as the student age groups are more prevalent than older age groups, and the high volume and concentration of workers commuting to City and Holbeck on a daily basis. For the remaining wards the density generally reflects the overall population density with urban areas having higher density suburban and rural areas having significantly lower tweets per km2.

In order to qualify the level of interest or utility of each word, an investigation of spatio-temporal distributions was undertaken with the filtered account information. To each word in the dataset, the following qualifiers are retained from the original tweet – the time, exact location, and userid of the sender. For each word two profiles are developed:

i. A spatial profile – for each of 33 census wards in Leeds, a count of the frequency of use of each word;
ii. A time profile – for each of eight time periods (night, morning, daytime, and evening; for both weekdays and the weekend). On the 24 hour clock, night is defined as the hours from 0000 to 0600; morning is from 0600 to 1200; afternoon from 1200 to 1800 and evening from 1800 to 0000; and weekends as either a Saturday or a Sunday.

For each word (k), the spatial profiles are articulated as X_j^k and the temporal profiles as $X_j^k(t)$ for ward j (j=1,...,33) and period t (t=1,...,8) respectively. The joint profiles $X_j^k(t)$ are considered further below.

Next, two traces are established for each word using the well-established concept of an Index of Dissimilarity (IoD). Firstly a spatial trace θ (j,x) and secondly a temporal trace φ (k,t):

$$\theta(j,x) = 0.5 \times \sum_j \left| \frac{X_j^k}{X_*^k} - \frac{X_j^*}{X_*^*} \right| \qquad (1)$$

$$\varphi(k,t) = 0.5 \times \sum_j \left| \frac{X_j^k(t)}{X_*^k(t)} - \frac{X_j^*(*)}{X_*^*(*)} \right| \qquad (2)$$

Table 1. Traces for a Sample of Marker Words

		IoD				IoD				IoD	
	Count	Spatial	Temporal		Count	Spatial	Temporal		Count	Spatial	Temporal
Leisure				Sports				Emotions			
watch	7321	0.103	0.128	Lufc	6443	0.317	0.095	haha	28763	0.159	0.067
park	4644	0.236	0.150	win	5746	0.143	0.067	lol	26580	0.183	0.072
hair	4398	0.132	0.028	play	5439	0.072	0.028	o	18008	0.217	0.066
club	3531	0.190	0.084	team	4464	0.103	0.072	hate	6724	0.106	0.036
gym	3402	0.153	0.136	Football	3522	0.124	0.110	omg	5164	0.171	0.088
tv	3342	0.108	0.119	fan	2663	0.143	0.078	ya	4856	0.134	0.039
song	3143	0.095	0.108	united	2561	0.208	0.087	lovely	4784	0.107	0.046
film	3102	0.122	0.185	league	1971	0.176	0.084	funny	4414	0.099	0.117
iphone	2731	0.123	0.026	match	1818	0.135	0.084	excited	3916	0.104	0.030
facebook	2559	0.100	0.038	Games	1769	0.120	0.050	brilliant	2609	0.115	0.081
bought	1992	0.102	0.106	cup	1711	0.125	0.060	bored	2284	0.111	0.074
video	1967	0.121	0.083	goal	1627	0.180	0.203	poor	2207	0.100	0.081
shop	1941	0.109	0.131	Olympics	1426	0.156	0.121	perfect	2051	0.105	0.045
walking	1799	0.107	0.098	Rugby	1406	0.218	0.105	stupid	1995	0.130	0.051
xfactor	1385	0.212	0.480	score	1218	0.162	0.165	sweet	1926	0.105	0.041
band	1359	0.118	0.099					lucky	1891	0.108	0.036
garden	1188	0.157	0.177	Food and Drink				yay	1777	0.134	0.023
cinema	1155	0.381	0.156	food	4161	0.110	0.106	laugh	1757	0.113	0.096
channel	1053	0.121	0.127	bar	3726	0.325	0.091	gay	1757	0.105	0.108
movie	1035	0.159	0.149	drink	2628	0.100	0.040	aha	1750	0.376	0.134
website	1022	0.133	0.103	coffee	2595	0.220	0.274	gutted	1732	0.107	0.048
gig	1021	0.165	0.107	beer	2515	0.148	0.143	love	1723	0.107	0.094
				breakfast	2359	0.258	0.464	annoying	1578	0.120	0.068
Family and friends				lunch	2310	0.225	0.403	boring	1572	0.123	0.069
mate	6647	0.130	0.051	dinner	1956	0.140	0.207	joke	1445	0.119	0.060
house	3810	0.112	0.058	wine	1857	0.162	0.187	fantastic	1431	0.125	0.055
mum	3486	0.152	0.083	pub	1769	0.127	0.153	gorgeous	1371	0.152	0.045
dude	3143	0.132	0.108	pizza	1454	0.153	0.183	hilarious	1203	0.120	0.173
girls	2866	0.109	0.053	meal	1238	0.186	0.114	excellent	1048	0.131	0.043
family	2700	0.148	0.052	cake	1227	0.133	0.055				
boys	2650	0.113	0.064								
dad	2553	0.155	0.086	Work and Travel				Others			
lad	2500	0.127	0.051	station	9979	0.613	0.241	Please	9506	0.109	0.054
kids	2402	0.149	0.057	railway	5694	0.671	0.261	Bed	9123	0.134	0.254
pal	2288	0.205	0.040	bus	5158	0.177	0.223	Sleep	7101	0.107	0.342
dog	1930	0.126	0.023	city	4929	0.305	0.133	Help	4514	0.110	0.052
woman	1864	0.101	0.050	train	4589	0.350	0.175	Free	4464	0.105	0.087
young	1844	0.118	0.047	school	4342	0.186	0.122	Cheers	3855	0.113	0.051
babe	1816	0.143	0.134	course	2765	0.126	0.040	Pic	3603	0.181	0.097
Social	1773	0.374	0.032	class	2244	0.101	0.047	Waiting	3569	0.133	0.112
Brother	1494	0.132	0.102	business	2223	0.335	0.266	News	3341	0.103	0.062
Couple	1494	0.114	0.102	college	1804	0.233	0.245	Photo	2487	0.222	0.050
Cat	1280	0.208	0.050	office	1786	0.238	0.264	Holiday	2460	0.110	0.084
Sister	1254	0.134	0.067	university	1729	0.608	0.260	Ill	2273	0.126	0.046
				meeting	1387	0.139	0.135	Dream	1559	0.106	0.156
				exam	1344	0.258	0.127	Driving	1390	0.127	0.107
				email	1121	0.130	0.064	Church	1338	0.526	0.218
								Sex	1320	0.120	0.159
								Euro	1125	0.277	0.281
								Rich	1125	0.365	0.079
								Shower	1123	0.171	0.130
								Police	1009	0.165	0.057

in which the asterisk (*) denotes summation across a missing index. As with an IoD, the values for each trace are automatically a standardised between zero (uniform distribution) and one (the word is concentrated within one time interval or spatial location). However the number of locations is quite large relative to the number of time intervals.

The spatial and temporal concentration of each word was used as a means to assess the value of that word as a potential discriminator of behaviour. Any word which appeared less than 1,500 times in the database was excluded, and any word with a 'spatial trace' (IoD) of less than 0.1 was also excluded. It has not yet proved possible to generate a completely rule-based approach to the selection of a word set, however. Large numbers of words were excluded for a number of reasons, for example:

Aaron, Zoe – a person's name, probably used in tweeting or re-tweeting messages within a particular social network of friends.

North, Pudsey – points of the compass and place names were generally regarded as unhelpful since they tend to be trivially related to specific geographical areas, and therefore tend to hinder rather than help in the identification of patterns.

Monday, morning – references to days of the week or times of day are also unnecessary since these concepts are already embedded in the time series for the data.

Nandos, Asda – specific names of bars, restaurants or shops were also regarded as too closely tied to outlets at particular geographical locations.

Dirty, fast, window – all examples of words which appear to be spatially concentrated but which have no obvious interpretation in terms of the state of mind, behaviour or activity patterns of the sender of a message. In general, verbs and adjectives were also excluded unless these are connected to emotions which might potentially be of interest.

Ultimately 136 words were selected for inclusion in the cluster analysis. These are grouped into seven categories in relation to: i) sports; ii) recreation; iii) food and drink; iv) family; v) work and travel; vi) emotion; and vii) others. Counts for the occurrence for each word, along with the spatial and temporal trace, are shown in Table 1.

3 Methods

A cluster analysis was undertaken using the 136 key words identified in Section 3, disaggregated across 33 electoral wards of the city of Leeds and split by the eight time periods outlined previously. As noted above, k-means cluster analysis has been widely used as a basis for geodemographic classifications in order to establish the principle that 'birds of a feather flock together'– in other words, that neighbourhoods within a city can be strongly differentiated in relation to their distinctive demographic profiles.

In the present investigation, the purpose of the analysis has some similarities to conventional geodemographics, and also some differences. As a starting point, it is interesting in its own right to explore whether traditional geodemographic relationships are reproduced in the language of the twittersphere. For example, do the deprived populations of the inner city communicate (with one another) in a different way than their more prosperous suburban cousins? The temporal dimension adds a significant extra twist here. To what extent is there a variation in language between different periods of the day, and does this begin to suggest the varying character of areas through longer time cycles? It can be argued that a failure to capture the dynamics of neighbourhoods over both the long-term and shorter time frames is a significant limitation to geodemographic approaches (in particular, the distinction between working and domestic populations in both the employment centres and residential parts of the city). Thirdly, and most important perhaps, the study is an exploration of a new data set which seeks to establish whether further and more detailed investigations might be worthwhile as a means to understand the movement and behaviour patterns of an urban population. This theme will be reviewed in the discussion section, following a presentation of key results.

Prior to the classification process, the counts for each word were transformed to z-scores. Starting with an array of counts X_j^k(t) for 33 wards j, 8 time periods t, and 136 words k, the mean μ(k) and standard deviation σ(k) were calculated for each word. Counts were translated into z-scores using the routine formula:

$$z_j^k(t) = \frac{X_j^k(t) - \mu(k)}{\sigma(k)} \tag{3}$$

These z-scores were used as the inputs to k-means cluster analysis which was performed using the ‘Classify’ analysis procedure within IBM SPSS Version 19. For a general review of this procedure, see [11]; specific applications in the context of spatial data and geodemographics can also be found [8,12]. It is not unusual for a certain amount of ‘expert intervention’ to be needed to add to the structure of the computational procedure and that was the case here, as a four-step hierarchical sub-division of the data set was needed to create six output clusters, as shown in Table 2 (column ‘Clus’ represents cluster ID within the hierarchical stage and ‘Alloc’ is the allocated output cluster or stage identifier for further processing). In particular, it was necessary to override the inclination of the procedure to produce small clusters, often with only a single case. For example, after Step 1 clusters 3 and 4 are all quite similar and therefore merged together on the basis of similarity in their ‘z-profiles’. At this stage, eleven clusters are allocated or merged, but the largest cluster (number 7, with 219 cases) is split into four further clusters at the second stage in the hierarchical procedure, which continues through two further iterations. The word counts and z-scores for the Leeds Twitter data set can be accessed online by readers with an interest to reproduce or extend this analysis.[1]

The resulting clusters were profiled and labelled in relation to the distribution of z-scores. For example, detailed analysis of the 136 words for Cluster 1 revealed scores of greater than one (an occurrence of more than one standard deviation above the mean value for all ward-time period combinations) for words such as ‘Score’ , ‘Match’, ‘Pub’ and ‘Xfactor’. The picture which emerges is a set of activities and

[1] `http://www.geog.leeds.ac.uk/people/m.birkin`

behaviours associated with sports and recreation within the home or local community. This cluster was labelled as ‘Rest and Relaxation’, and looks most likely to be associated with a weekend afternoon or evening in suburban neighbourhoods.

Table 2. k-Means Classification Steps for Twitter Data

Cluster H1			Cluster H2		
Clus	Count	Alloc	Clus	Count	Alloc
1	4	5	1	161	H3
2	2	5	2	30	3
3	1	1	3	1	5
4	1	1	4	3	2
5	1	5	5	17	4
6	5	2	6	7	4
7	219	H2			
8	2	1			
9	4	4			
10	1	5			
11	23	6			
12	1	5			

Cluster H3			Cluster H4		
Clus	Count	Alloc	Clus	Count	Alloc
1	140	H4	1	20	3
2	10	4	2	112	6
3	7	2	3	2	6
4	4	6	4	2	6
			5	1	6
			6	3	3

4 Results

The 136 words are quite useful for the profiling exercise, but provide rather a lot of detail as regards documentation. For descriptive purposes, the scores were aggregated across the six major categories introduced at Table 1 (excluding the ‘Other’ category, which features some interesting words but which by definition lack any particular coherence which might bear further interpretation). These aggregated scores are presented in Figure 3 as a series of radar charts. In general, the charts bespeak a series of distinctive clusters, wherein the first group shows a dominant concern with concepts relating to work and has therefore been labelled ‘Daily Grind’. Group 4 also shows a relatively one-dimensional concern with sporting events, and is called ‘Match of the Day’, with reference to the weekly UK Premier League football show of the same name. Groups 2 and 5 have similar radar charts and both show a focus on issues relating to food and drink. However, the two clusters differ as the latter contains phrases where the emphasis veers towards bars and restaurants and greater reference to work, in the former the context is more likely domestic. These are labelled 'Family Life' and 'Rest and Relaxation'. Clusters 3 and 6 both show a more balanced distribution of interests, although particular words favoured in Cluster 6 include 'Sleep' and 'Bed' and is therefore characterised as 'Sweet Dreams' in the expectation that late nights are probably implied. Cluster 3 is classed as 'Idle Chatter', in which individuals are content to issue messages with a wide variety of content and less in the way of dominant themes and concerns.

It now remains to be seen whether the patterns revealed in this typology of tweeting patterns makes sense in space and time. The various clusters are mapped in Figure 4 and spatial reference points discussed below are mapped on Figure 5. It is gratifying to observe that the city centre areas seem most likely to originate messaging which relates to work and transport issues during the daytime. These areas are also the most likely to become lively in the evenings, with higher levels of eating, drinking and social activity, a characteristic also shared with the neighbouring student areas. It is characteristic of the latter that relative activity levels are also highest late into the night, but not necessarily in the mornings! Thus for example while most communities are enjoying 'sweet dreams', students may still be engaged in lively discussions on sporting matters, or just in 'idle chatter'.

Outside the central areas, the most striking feature of the activity patterns is perhaps their relative homogeneity rather than spatial differentiation. Thus it tends to be the case that many areas follow a standard pattern in which behaviour changes over time in a similar way for different areas. Outside the central areas, the night is typically a quiet time – for sleep and dreams – while during the working day those who have time for social messaging are mostly concerned with trivia. Towards the late afternoon and early evening, then a more social pattern begins to emerge, and this pattern is reinforced at the weekends, often with a more family-oriented emphasis. The concern with sporting activities is perhaps more marked in some of the less affluent central suburbs (some of which, such as Middleton and Wortley, may be influenced by their close proximity to the major football ground at Elland Road in the south-west of the city).

One of the most striking trends in Figure 4 is that the transformation of behaviour types is at least as noticeable over time that it is in spatial terms. For example in many areas at the weekend an early morning of Sweet Dreams gives way to Family Life later in the day, before sporting of leisure events (Match of the Day/ Rest & Relaxation). During the working week, dreams still dominate the night time before giving way to a combination of daily grind or idle chatter, before the evenings again dissolve into family or leisure-oriented activities. This seems to reinforce the notion that at the very least social media data may refine traditional approaches to neighbourhood classification by identifying short-term variations in activity and behaviour patterns.

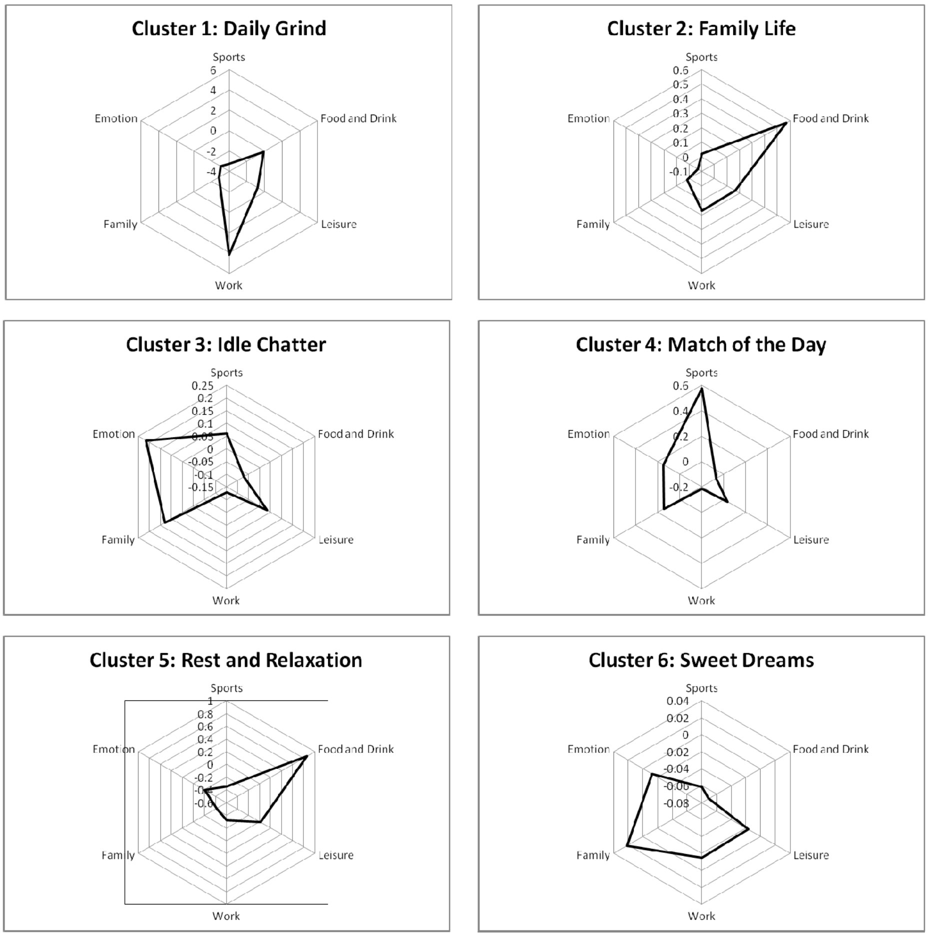

Fig. 3. Language Variations by Cluster

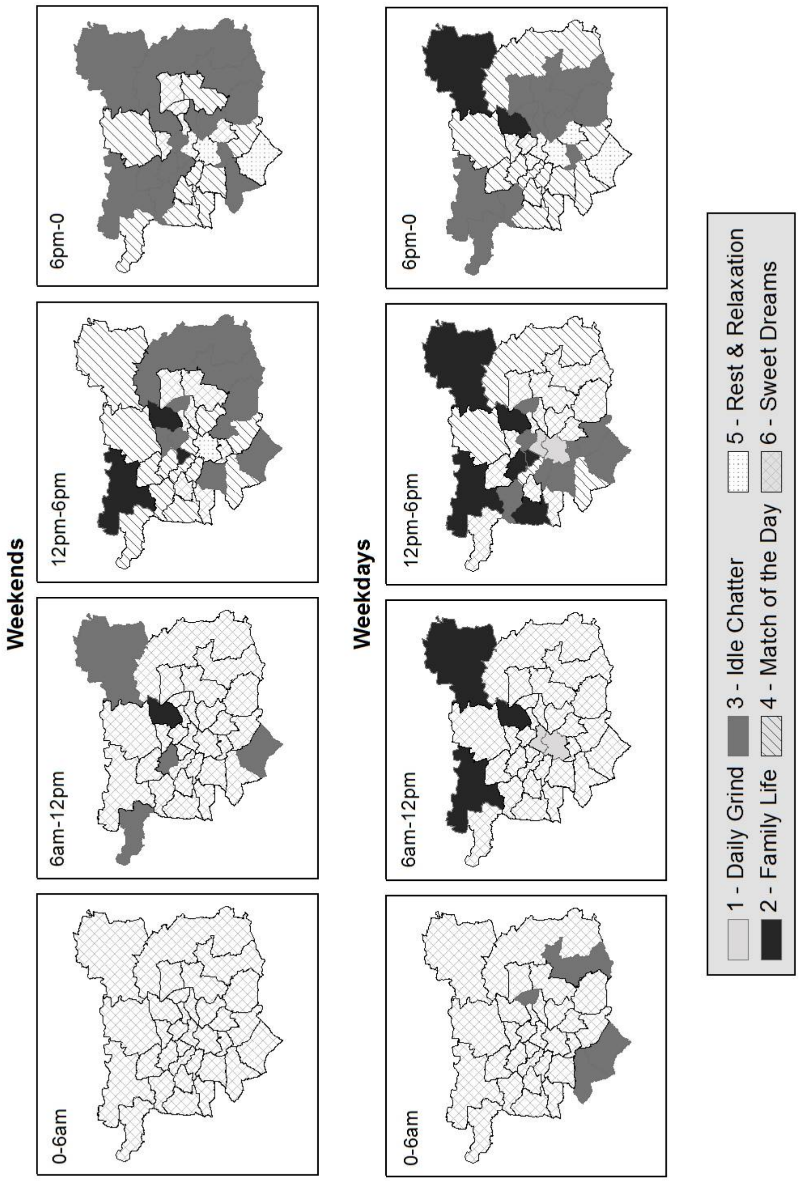

Fig. 4. Spatio-temporal occurrences of word clusters

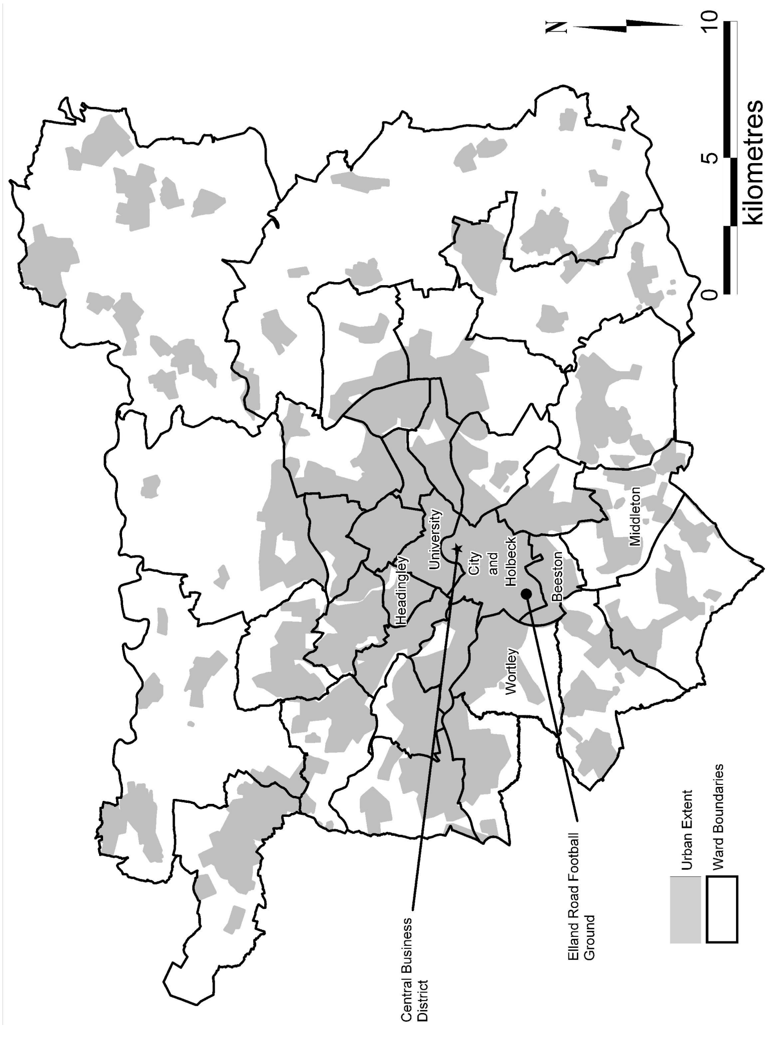

Fig. 5. Location reference points

5 Discussion

In this paper the use of social messaging data from the twitter service has been considered as a means of profiling the behaviour and activity patterns of urban residents. It has been seen that noticeable variations exist in the way that people communicate not just in geographical space, but during different time periods and from day-to-day. These results complement traditional geodemographics by emphasising that when behaviour is analysed, place-level variations (i.e. who lives where) need to be augmented with a perspective that recognises varying activities between time-periods. This could be particularly important in considering the provision of various services, such as crime prevention and health care – for example, there is no sense in providing neighbourhood policing throughout the day to cater for a residential population which is largely absent (unless perhaps the measures are targeted specifically to the protection of vacant properties). This kind of application is a staple for geodemographics [7].

The results presented here are relatively exploratory and unrefined. In fact, the census wards which have been used as a basis for the analysis are rather large spatial units. It might be far more interesting to explore sub-divisions to much smaller entities, which could then start to reveal the character of much more coherent local neighbourhoods, including perhaps shopping centres, zones of education and entertainment as well as conventional residential neighbourhoods. The temporal frame of reference could also be extended from the current division of eight coarse time periods, for example into an hour-by-hour framework, or even by extending across a range of seasons and holiday periods. Early experiments beyond those presented here have indicated that the incorporation of smaller units of time could be highly beneficial.

It would also be interesting to explore the idea of profiling individuals rather than areas. The way that language is used by individuals could provide a uniquely interesting perspective on their behaviour and activity patterns, using keywords relating to recreation, shopping, domestic life, and other domains. Micro-level classifications of this type have already been attempted within the market analysis industry (such as PersonicX and Prizm – see, for example, www.myacxiom.com) but as with neighbourhood classifications the emphasis tends to be on attributes rather than behaviour. If behaviours are incorporated, for example through lifestyle surveys, then these are more likely to encompass long-term preferences such as hobbies and interests than short-term activity patterns. Psychographic classifications which seek a more detailed perspective on individual attitudes and motivations are typically not data rich or spatially detailed [13]. Of course it should be noted however that the ethical implications in using individual data to construct representations of this type are potentially quite significant; as with all geodemographics spatial aggregation at least has the benefit of sidestepping such considerations.

An extension of this line of research enquiry would be towards the production of individual activity spaces which start to link together individual movement patterns with activities, behaviours and purposes, alongside the characteristics and attributes of individual actors [14]. The question of how individuals move around cities, and to what purpose, has been a subject which has intrigued geographers since at least the time of Torsten Hagerstrand [15], but until now researchers have lacked the information with which to start to unpack this problem. The long-term goals of the research which has been reported here are to incorporate much more powerful methods such as microsimulation and agent-based modelling in order to represent and perhaps predicted the flows of people around a city. Although the experiments reported here constitute only the most modest of steps towards this goal, the results are encouraging enough to suggest that further refinements would be worthy of further attention.

References

1. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment in twitter events. Journal of the American Society for Information and Technology 62, 406–418 (2011)
2. Kawamoto, T.: A Stochastic Model of Tweet Diffusion on the Twitter Network. Physica A: Statistical Mechanics and Its Applications (in press), `http://dx.doi.org/10.1016/j.physa.2013.03.048`
3. Xiong, F., Liu, Y., Zhang, Z., Zhu, J., Zhang, Y.: An Information Diffusion Model Based on Retweeting Mechanism for Online Social Media. Physics Letters A 376, 2103–2108 (2012)
4. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. Social Science Computer Review 29, 402–418 (2011)
5. Sobkowicz, P., Kaschesky, M., Bouchard, G.: Opinion Mining in Social Media: Modeling, Simulating, and Forecasting Political Opinions in the Web. Government Information Quarterly 29, 470–479 (2012), doi:10.1016/j.giq.2012.06.005.
6. Cheng, Z., Caverlee, J., Lee, K., Sui, D.: Exploring Millions of Footprints in Location Sharing Services. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM) (2011)
7. Longley, P.: Geographical Information Systems: a renaissance of geodemographics for public service delivery. Progress in Human Geography 29, 57–63 (2005)
8. Harris, R., Sleight, P., Webber, R.: Geodemographics, GIS, and neighbourhood targeting. Wiley, Chichester (2005)
9. Farr, M., Webber, R.: MOSAIC: From an area classification to individual classification. Journal of Targeting, Measurement and Analysis for Marketing 87, 681–699 (2001)
10. Humby, C., Hunt, T.: Scoring Points: How Tesco is Winning the Battle for Customer Loyalty. Kogan-Page (2003)
11. Everitt, B.: Cluster Analysis, 5th edn. John Wiley, Chichester (2011)
12. Vickers, D.: Multi-Level Integrated Classifications Based on the 2001 Census. PhD thesis University of Leeds (2006)
13. Cathelat, B.: Socio-Styles. Kogan Page, English London (1990)
14. Malleson, N., Birkin, M.: Estimating Individual Behaviour from Massive Social Data for An Urban Agent-Based Model. In: Modelling Social Phenomena in a Spatial Context GeoSimulation. LIT Verlag, Berlin (2012)
15. Hagerstrand, T.: Innovation Diffusion as a Spatial Process. Chicago University Press, Chicago (1967)