New Insights into Individual Activity Spaces using Crowd-Sourced Big Data

Nick Malleson
School of Geography, University of Leeds
LS2 9JT, UK
+44 (0) 113 343 5248
Email: n.malleson06@leeds.ac.uk

Mark Birkin
School of Geography, University of Leeds
LS2 9JT, UK
Email: M.H.Birkin@leeds.ac.uk

ABSTRACT

This article discusses ongoing work towards an accurate model of daily urban movements, calibrated in part by data from social media services. In particular, it investigates the connections between the language that is used in Twitter messages, the relative location of the user, and the significance of the location to an individuals' geographical awareness, in order to elucidate possible activity patterns from spatial and textual social media data. The results show that there are words which are closely associated with home, others with the local community, and some with more remote locations including workplaces and cultural centres. By identifying important locations for individual users, and associating these with the words that are commonly used in such places, this research contributes to a better understanding of how spatially-attributed social media data can be used to derive useful intelligence about daily urban movement patterns.

I INTRODUCTION

Crowd-sourced data are becoming ubiquitous. Coincident with an explosion in the use of social media services is the proliferation of location-aware devices that allow the geographical attribution of social media contributions. These new sources of data – that are user-generated, geo-located and contain varying degrees of contextual information (text, videos, images, etc.) – show potential as reliable sources of accurate information about daily urban mobility patterns and social perceptions of place.

Traditional large-scale data sources, such as population censuses, provide detailed information about the night-time population, but are much more limited in their descriptions of *day time* populations and associated behaviour. Smaller surveys that attempt to capture these features are limited by their sample size and geographical resolution. Although these data provide a useful starting point for research, more detailed spatio-temporal information is required to untangle the complex web of spatial interactions that ultimately drive urban systems.

This article describes research that is part of a larger project whose aim is to utilise 'big' data sources to create an accurate model of daily urban dynamics. In particular, this paper discusses the work towards establishing geographical awareness spaces for individual users of social media services and identifying key places (termed 'anchor points') that constitute an individual's conceptual map of the city. We then attempt to identify the functions of these different anchor points (e.g. workplace, leisure, education, socialising, etc.) as a means of understanding daily individual movements.

The aims of this research are to:

- evaluate the extent to which crowd-sourced data can be used to gain insight into the individual activity spaces (the places that individuals visit on a regular basis);
- identify the functions of different anchor points.

II BACKGROUND

1 THEORETICAL FRAMEWORKS

Recent work recognising the character of cities as complex systems has made it clear that at a disaggregate level cities are in a constant state of flux and aggregate data have the effect of smoothing out this underlying dynamism [1]. To address this difficulty, urban scientists can draw on seminal theoretical work to better understand how individual movements are spatially and temporally constrained, and how aggregate patterns emerge from the multitude of these individual interactions. For example, Hägerstrand's famous 'space time prism' [2], illustrated in Figure 1, provides a means of conceptualising the possible locations that an individual can reach under given spatio-temporal constraints. Using this formalism, it becomes possible to explore the spatio-temporal dynamics of peoples' behaviours to better understand, for example, who visits particular locations, which other locations have those people visited and who might they have met during their journey [3].

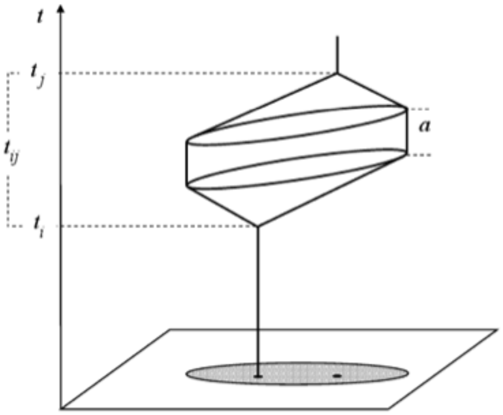

Figure 1: The 'space-time prism' [2]. Source: [4]

A related concept is that of an 'activity space'. Activity and awareness spaces form a central pillar for various academic fields including environmental criminology [5] and geography [6]. In general, a distinction is made between *activity spaces* – spaces in which normal activity occurs – and *awareness spaces* – spaces that a person is aware of beyond the areas of normal activity. For the purposes of this research, the aim is to first identify activity spaces (the places that people commonly visit) and, from these, identify the individual locations whose positions ultimately determine the structure of activity spaces.

Although such frameworks have been of interest for some time, data availability and computational limitations have made it difficult to apply these concepts to more than a few individuals. 'Big' crowd-sourced data provide, for the first time, the opportunity to test these ideas at a much larger scale – such as an entire city.

2 'BIG' GEOGRAPHICAL DATA

The recent emergence of powerful 'big' data sources, particularly those centred around social media, has fostered a substantial amount of new research. However, examples of the application of social media data to the study of social phenomena are more limited, and those that explore anchor points and awareness spaces even less so.

A body of literature is evolving around the concept of classifying neighbourhoods based on local social media use. These include the development of neighbourhood boundaries based on Foursquare data (the 'livehoods' project[1]) [7]; the use of Latent Dirichlet Allocation (LDA) to establish topics from social media data and subsequently develop functional profiles of areas [8]; the characterisation of land use through the application of the Self Organising Map (SOM) to classify areas based on volume of geolocated tweets and the subsequent calculation of unique activity vectors based on common spatio-temporal tweeting behaviour [9]; and perceptions of place using Foursquare data [10].

A small body of research, of direct relevance, attempts to explore the places that people commonly visit and hence their awareness spaces. This includes the identification of "patches" – defined as spatial areas that people routinely visit [11]. The authors use a density-based clustering algorithm (OPTICS) to identify regularly visited patches and discuss how these could be linked to different purposes (e.g. shopping, working, etc.). The proposed method is substantially different to that used here, but as both projects are in early stages a formal analysis of the impacts of the different choice in methodology is not appropriate. Similarly, others have attempted to identify important places (particularly 'home' and 'work') in mobile phone data [12], although preliminary results indicate inaccuracies of up to 3 miles for 88% of volunteer users. These studies are relevant here because they attempt to construct awareness spaces from social media data, but the authors are unaware of any research that aims to use the data to model individual activity spaces as this research will do.

Although there is a degree of optimism associated with these research directions, there are substantial drawbacks that must be addressed. In particular, it has been argued that social media (and related) sources suffer from "content poverty", a lack of clarity in whether a contribution contains espoused theory or theory-in-use, and positivist assumptions [13]. The optimist hopes that some of these drawback will be offset by the sheer volume of data [14], although this will be highly dependent on the data used and the application area.

[1] Livehoods: `http://livehoods.org/`

III DATA OVERVIEW

1 THE STUDY AREA: LEEDS

The research has been conducted in the city of Leeds, UK. The Leeds local authority district is one of the largest in the UK with a population estimated by the Office for National Statistics to be 757,655 in 2012. Leeds' urban form exhibits strong similarities to similar European and British cities; in particular it centres on a core activity area with high concentrations of shops, businesses and entertainment facilities. A feature that is more unique is the two large universities situated to the north of the city centre. These features support some regularity in urban movements and will be used to broadly evaluate the results of our analysis.

2 SOCIAL MEDIA DATA

The data used in this research consist of messages posted to the Twitter service that originate within Leeds during the period 22nd June 2011 – 14th April 2013. The parameters of the collection routine restricted the search only to messages with associated GPS coordinates. Such messages are commonly created using mobile devices by users who have explicitly opted to publish their location with their message. A number of extremely prolific user accounts were identified as belonging to organisations and used for advertising or disseminating information (weather forecasts, car advertisements, etc.). These were removed from the data set, leaving $N = 2,812,332$ individual messages. As well as the location, each individual message contains information about the user account, the message text and the creation time.

Figure 2 illustrates the geography of message locations. As might be expected, the central business district, as well as the university campuses, exhibit most substantial message density.

Although the volume of information that can be collected, trivially, from social media services is unparalleled, research that make use of these data must carefully consider the drawbacks. These include skewness, representation, accuracy and bias. Methods to address these drawbacks are discussed in the conclusions, although fully understanding the implications, and resolving them, is left for future work.

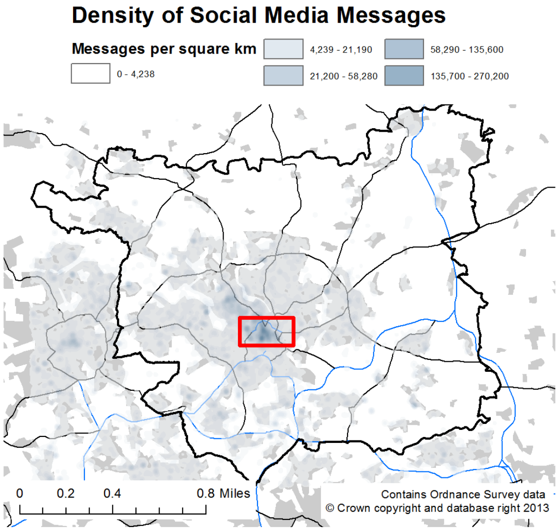

Figure 2: The density of social media messages calculated using Kernel Density Estimation [15] with a 300m kernel radius and 10m^2 cell size.

2.1 SKEWNESS

With respect to skewness, although the total volume of data is large, the distribution of messages per user is highly skewed; see Figure 3. In the data used here, 90% of all messages were generated by only 17% of the users. Only 8,209 (14%) of the users created more than 50 messages which is the cutoff used in later analysis.

2.2 REPRESENTATION

A second drawback, representation, relates to the extent to which a user's messages correspond to their actual activity. It is likely that this correspondence will increase with the number of messages that a user creates, but even the most prolific users might use social media only in some specific circumstances. In these cases the data only represent a portion of their usual daily behaviour. Furthermore, as the data used here span a period of approximately two years, it is possible that spatial activity patterns will vary over time. In this analysis behaviour is assumed to be consistent throughout.

2.3 ACCURACY

A third drawback relates to the geographical accuracy of the data. Although each message has associ-

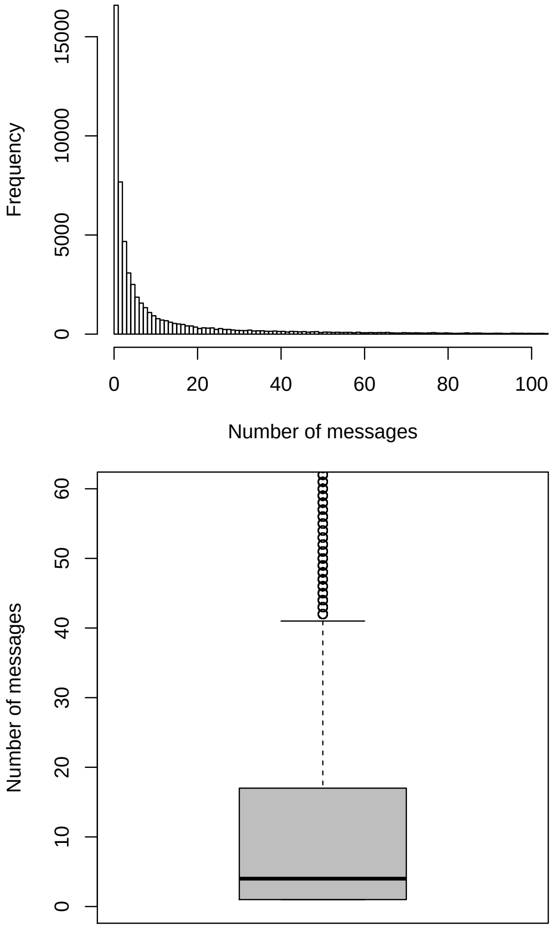

Figure 3: The number of messages created per user.

ated geographical coordinates, the precise spatial accuracy is difficult to ascertain. It is possible to assess geographical position accuracy by linking to other social media data sets with more accurate geographical coordinates (e.g. foursquare[2]) or by comparing message locations to *likely* spatial locations (residential buildings, shopping centres, etc.). This accuracy validation will be applied in future research iterations.

2.4 BIAS

The fourth drawback, bias, brings in to question how well the data represent the behaviour of society as a whole. Although N is very large, the number of active users in the data set (60,293) is small in comparison to the size of the Leeds residential population of nearly 800,000. This is reduced further when only users with 50 or more messages are included in the analysis (8,209 users).

Finally, there is also the potential for participation inequality to introduce bias into the data set as the use of social media services will vary across social groups – see, for example, 'digital divide' research [16,17]. It is common to find that younger, well educated, and more affluent groups of people are the most likely to engage with digital media [18,19] although, interestingly, Twitter usage in the United States has been found to contradict many of these findings [19, 20]. This signifies the need for more nuanced research into the uptake of social media services specifically.

IV METHOD

In this section, we present a method for estimating a user's activity space (the spaces in which normal activity occurs) and subsequently the main anchor points that drive a individual's daily urban movements.

1 GENERATING A MESSAGE DENSITY SURFACE

The preliminary stage in the analysis is to create a two-dimensional surface representing the density of messages created by each user that will later be used to elucidate anchor points. The analysis has been restricted to users who created a sufficient number of messages. Here, 50 messages has been chosen as a threshold because it corresponds closely with the

[2] https://foursquare.com/

definition of an upper outlier (see Figure 3).

A kernel density estimation (KDE) algorithm was applied to the spatial (two-dimensional) data by overlaying a regular grid and calculating the density of each cell. A Gaussian kernel was used with a bandwidth and cell size of 236m and 33m^2 respectively. These values were chosen through a process of trial and error to identify the configuration that most successfully combined incident points without merging disparate high-density areas. The density of a given cell (x, y) is calculated using the formula:

$$\sum_{i=0}^{n} \frac{1}{2\pi 2b} \exp\left[-\frac{1}{2}\left(\frac{x-x_i}{b}\right)^2 - \frac{1}{2}\left(\frac{y-y_i}{b}\right)^2 \right] \quad (1)$$

where b is the bandwidth and (x_i, y_i) are the coordinates of all of a user's messages, n. The formula produces a smooth two-dimensional kernel, as illustrated in Figure 4. For efficiency, no points beyond the bandwidth of a given cell were evaluated given their negligible impact on the density. Finally, the data were transformed using the natural logarithm to reduce the impact of a small number of extremely high density areas and normalised to the range 0–1 to allow comparisons across users.

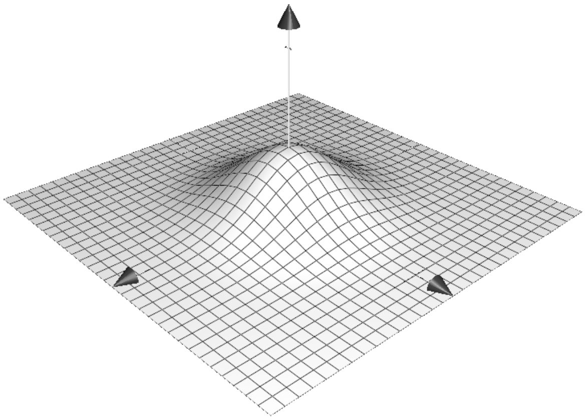

Figure 4: The structure of the kernel used to calculate message density (see Equation 1).

2 IDENTIFYING ANCHOR POINTS

Having identified the areas that are most commonly used to publish messages from, the next task is to identify the individual *anchor points* that create a person's awareness spaces. This can be accomplished by identifying the cells with a substantially higher density than their surroundings. Algorithmically separating genuine peaks from their surroundings is a complicated process. For this research, a routine that is more commonly used for identifying peaks in geographical digital elevation data [21] has been used. The most successful parameter configurations (those that allowed the algorithm to most accurately identify peaks from their surroundings) were identified using trial and error and are provided in Table 1.

Table 1: Parameters used to configure the LandSerf algorithm. Note: all densities have been normalised to the range 0-1.

Parameter	Description	Value
MinHeight	The minimum height for a peak	0.001
MinDrop	The minimum drop in height for cells surrounding a potential peak	0.1

Results of the anchor point identification process vary. A manual inspection of a sample of results reveals instances where peak identification works well – the algorithm is able to identify genuine areas of substantially higher density – and other instances where the results are poorer. Figure 5 illustrates a number of such results. The main drawbacks occur in the cases where two distinct anchor points are situated close together and are classified as a single point (e.g. Figure 5c). Future work will attempt to develop a more nuanced algorithm that is able to modify the KDE kernel bandwidth (b) and the anchor point classification parameters (Table 1).

3 INTERPRETING ANCHOR POINTS

The anchor points have been used to construct a dictionary of twitter narratives at each location. Consider the example of Figure 5b – here three different anchor points have been identified for the user, and these are normalised so that the most intense centre has an index of 1. It is hypothesised that this modal point will usually be the home location. This analysis is reproduced for all of the twitter users with 50 messages or more.

A catalogue of marker words has been created [22], comprising 134 words with distinctive spatial and/or

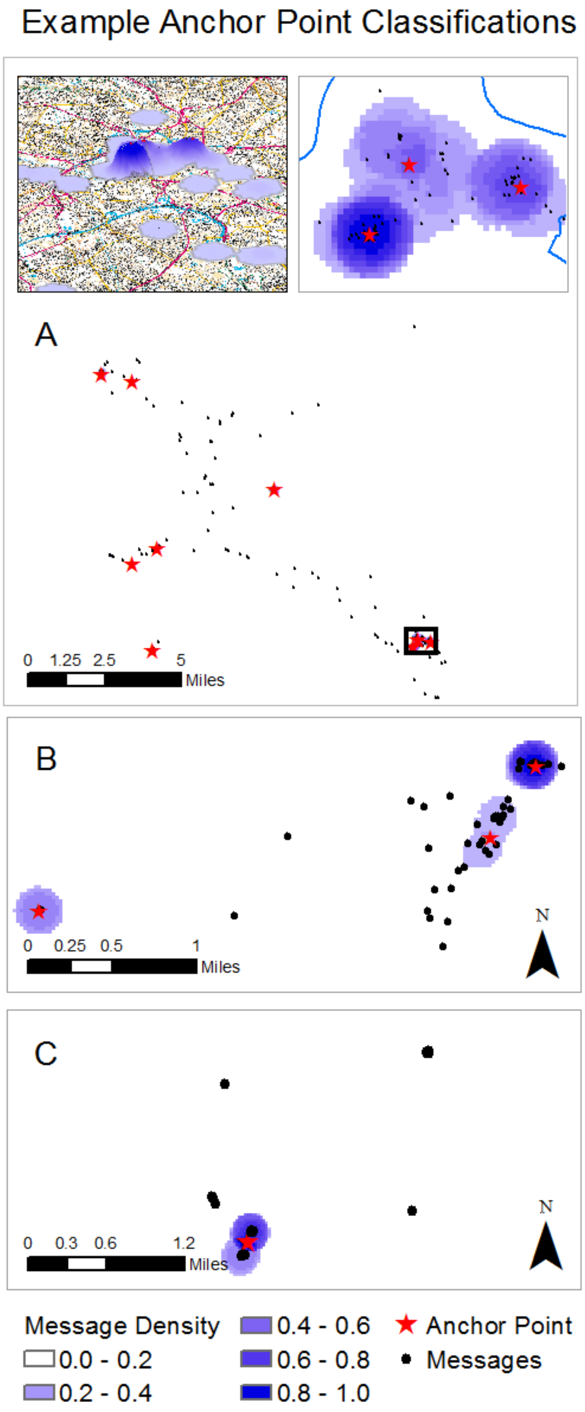

Figure 5: Examples of individual user anchor points for three users (A, B, C) resulting from the anchor point classification using LandSerf [21].

temporal profiles. These words are in turn grouped into seven activity categories on the basis of their context and use. In this section, the locational narratives are explored to investigate patterns in the use of language across the anchor points.

First, the anchor points are separated into two groups – those indexed at unity, the most popular location for each user and hypothecated as 'home' locations; and those others, termed simply 'away' for the present. For each word, a quotient which will be referred to as 'domestic skew' is calculated as follows:

$$ds(k) = \frac{[N(k,1)/N(*,1)] * 100}{N(k,*)/N(*,*)} \qquad (2)$$

where $N(k, m)$ is the count of occurrences of word k at a home location ($m = 1$) or an away location ($m = 2$), and $*$ denotes summation over the missing superscript. Thus the indicator compares the relative importance of each word at a home location with its overall importance.

This analysis is summarised in Table 2, in which the marker words are ranked by skew for each of the seven activity types. A distinctive and not unexpected contrast is clearly observable between 'friend and family' with a focus on the home; and 'food and drink' which concentrates elsewhere, probably with an orientation to bars and restaurants in the city centre. The exceptions here are interesting, for example 'mate' and 'dude' appear as remote words, so may be more common in social than domestic environments; 'cake' and 'wine' appear to be consumables that are more likely enjoyed in the home than at recreation.

A further question might be whether there is any spatial decay effect; for example are there certain types of narrative prevalent in a local community or neighbourhood. This is explored at four distance bands around the home (0-100m, up to 2km, 2-10km, more than 10km) for just six examples in Fig 6. The word *sleep* has a strong domestic focus: other words (not included here) showing a similar focus are activities like *watch(ing) xfactor* on *tv* and words expressive of family relationships such as *dad*, *sister*, or *kids*. The word *station* is in marked contrast, and is characteristic of travel by both *bus* and *train*, often to an *office*, or (albeit with a slightly flatter distribution) for *shopping*, which may be associated with a stop for *coffee*, a trip to the *cinema* , and perhaps later a *meal* and a visit to a *bar* or *nightclub*. Documented interactions with a *college* appear to be somewhat more localised and community-based; the same is true for

Table 2: A catalogue of marker words [22], ordered by residential skew (see Equation 2).

Friend & family		Food & drink		Leisure		Work	
mum	173	cake	113	xfactor	180	exam	116
sister	171	wine	109	channel	162	course	96
dad	145	drink	90	tv	139	school	87
babe	134	food	88	movie	136	email	83
family	130	meal	85	song	135	class	79
brother	129	dinner	82	watch	134	city	69
boys	126	pub	75	film	132	meeting	69
kids	122	beer	74	hair	124	college	49
girls	122	pizza	73	faceboo	112	train	45
young	118	breakfas	72	iphone	107	busines	45
dog	114	coffee	47	video	106	bus	39
pal	106	lunch	46	garden	97	office	38
woman	103	bar	44	bought	96	station	14
lad	102			band	79	universit	13
couple	90			walking	78	railway	11
mate	88			gig	75		
cat	87			website	70		
dude	87			shop	68		
house	83			gym	67		
social	61			club	52		
				cinema	50		
				park	36		
Sports		**Emotions**		**Emotions**		**Other**	
league	131	aha	169	annoyin	120	shower	151
goal	131	omg	166	boring	119	please	144
score	128	gorgeou	140	joke	118	sleep	144
win	124	hilarious	139	brilliant	117	euro	141
football	123	poor	133	fantastic	116	bed	141
olympic	121	funny	131	sweet	115	ill	126
Lufc	117	perfect	130	bored	113	dream	124
fan	117	laugh	128	yay	110	sex	117
team	113	love	125	excited	108	police	112
match	112	lucky	124	gay	108	help	111
cup	111	stupid	124	excellen	106	holiday	102
Games	110	haha	122	gutted	104	rich	86
play	77	ya	121	lovely	104	driving	85
rugby	73	lol	121	o	89	cheers	85
united	47	hate				free	80
						waiting	77
						photo	73
						news	70
						pic	63
						church	44

University and (even more so for *schools*) where *examinations* are a likely topic of conversation. Other neighbourhood pastimes include drinking *beer* in a *pub* (an interesting contrast in terminology and outlook to city centre bars), working out at the *gym*, or *walking* in a nearby *park*. Amongst the expressions considered, *i-phone* is the least spatially clustered, presumably reflecting the portability and increasing ubiquity of this technology; unlike *e-mail*, which seems much more likely to be a work-related activity.

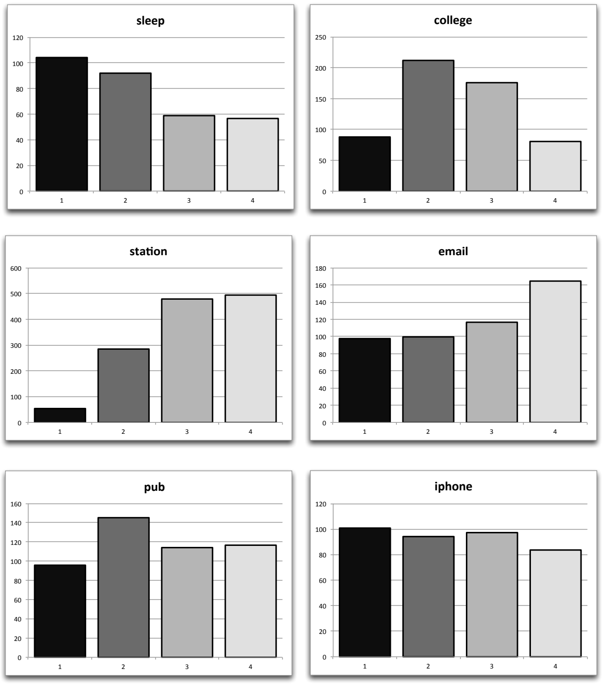

Figure 6: Residential skew for chosen words at four distance bands: 0-100m, up to 2km, 2-10km, more than 10km.

If a home location attaches to a particular lexicon of words in common use, then there is a further possibility to profile each anchor point according to the words used there. A domestic 'score' has been computed for each anchor point by weighting the domestic skew quotient across the words which appear there. Amongst 8,195 users identified in this study, 6,218 exhibit the highest domestic score at the anchor point previously identified as 'home'. For another 418 users there are less than ten keywords on which to make this judgement; leaving 1,559 users for whom the home location could plausibly be other than the most common tweet frequency. This could be seen as a preliminary test of the robustness of the simple assignment of 'homes', indicating that this procedure probably has an effectiveness of around 80%, but could be refined by further consideration of tweet narratives, or by some other means.

V DISCUSSION

This article contributes to ongoing research working towards an accurate model of daily urban dynamics through the interrogation of 'big' social data sources. Here, the investigation centres around the language which is used in Twitter messages and its relation to the locations of places that are important for individual Twitter users (termed 'anchor points'). Ultimately, the overall aim is to use these findings to better understand and model 'normal' daily urban movement patterns (shopping, commuting, social activities, etc.).

For the Leeds corpus, it has been shown that there are words that associate closely with home, others with the local community, and some with more remote locations including workplaces and cultural centres. The analysis was based on a simple assignment of anchor points based on frequency of activity. This method could potentially be extended to identify other kinds of location such as workplaces, sports centres or retail outlets, for example using the kinds of methods that have been used to establish connections between internet search terms and a phenomenon of interest [23].

The work could also be extended by considering a much larger vocabulary of words or phrases. The results from this research would add a great deal to our understanding of individual activity spaces, for example in understanding the configuration and spatial linkages between anchor points representing different facets of work, family life and social behaviour. The interrogation of new forms of spatial and social data can therefore be regarded as a promising avenue for a better understanding of mobility patterns and purposes within daily urban environments.

1 DRAWBACKS

Whilst there is considerable potential offered by these new data sources, results must be treated with a measure of caution. Of the drawbacks outlined earlier (skewness, representation, accuracy, bias), bias is potentially the most difficult to address. Although there is an abundance of data created using social media services – in 2011 it was estimated that approximately one billion tweets were created every 4-5 days [24] – this is a relatively small number of messages per person. For example, the study area used here houses nearly 800,000 residents, but only 8,209 of these created a sufficiently large number of messages in a two-year period to be included in the analysis. Furthermore, it is likely that some level of participation inequality will stem from an uneven engagement with social media across different socio-economic groups. However, as this research attempts to identify the home locations of individual users, it becomes possible to link social media data to well-established residential geodemographic data such as the national census. In this manner we can begin to identify the groups who are most, and least, likely to contribute using social media and re-weight the data accordingly.

Although this research has noted some of the potential drawbacks, no attempt has yet been made to thoroughly assess their impact on the results or to design methods capable of reducing their influence. Immediate future work will begin to address these issues by estimating the demographics of individual users (based on the home location) and re-weighting their contributions to reduce the influence of over-represented groups. For sections of society that are not represented in the social media data, alternative data sources will be used in an attempt to capture their typical behaviour, such as national censuses and other large-scale surveys. In this manner, different data sources can be used to model those people who are particularly well represented by them.

2 VERIFICATION / VALIDATION

Following an analysis of the potential drawbacks, the research will be in a position to begin a formal evaluation of the results. For example, through a more nuanced analysis of message text it would be possible to begin ground-truthing the content of messages to infer an activity ("I have just arrived at work"), or by considering foursquare check-in messages [11] – for example "I'm at Starbucks". Along similar lines, the inclusion of other forms of data, such as land use data, could also be considered. If an area is known to be dominated by retail outlets, offices, or residential space, then this is powerful evidence to add to the model training process. Land use data of this kind are routinely captured by national agencies (e.g. Ordnance Survey in the UK) as well as becoming increasingly prevalent as volunteered geographical information (e.g. Open Street Map).

3 ETHICAL IMPLICATIONS

Another serious consideration is that of the ethical implications of such research, and personal privacy in particualr. Although most most authors are largely in agreement that 'public' data published on the internet are suitable for research [25–27], these assumptions are predicted on the concept of *informed consent*. It would be difficult to argue that social media contributions could be considered private communications, but it is not always clear that individuals are aware of volume of personal information that they publish, and hence whether they have truly consented to its use. Whilst we do not argue against the use of such public data, care must be taken with data storage and publication to protect the anonymity of contributors.

VI CONCLUSION

The limitations of crowd-sourced social network data from twitter have been noted. As this means of communication continues to grow in popularity across a broad range of demographic segments then these limitations could become less restrictive. The volume of spatial data under consideration might also be increased by several orders of magnitude by considering a data set such as mobile telephone calls or traffic flows, but this would come at the expense of a loss of textual detail. As the 'Big Data' revolution continues unabated, it must be expected that the abundance and variety of information sources will continue to multiply. We hope and anticipate that the kinds of frameworks which we have begun to articulate here can be seen as providing initial components to a long-term platform for the creation of significant added value from these data.

References

[1] M. Batty, "Agents, cells, and cities: new representational models for simulating multiscale urban dynamics," *Environment and Planning A*, vol. 37, pp. 1373–1394, 2005.

[2] T. Hägerstrand, "What about people in regional science?" *Papers of the Regional Science Association*, vol. 24, no. 1, pp. 6–21, 1970.

[3] K. Hornsby and M. J. Egenhofer, "Modeling moving objects over multiple granularities," *Annals of Mathematics and Artificial Intelligence*, vol. 36, no. 1, pp. 177–194, 2002-09-01.

[4] E. J. Miller, J. Douglas Hunt, J. E. Abraham, and P. A. Salvini, "Microsimulating urban systems," *Computers, Environment and Urban Systems*, vol. 28, no. 1, pp. 9–44, 2004.

[5] P. L. Brantingham and P. J. Brantingham, "Notes on the geometry of crime," in *Environmental Criminology.* Sage Publications, 1981, pp. 27–54.

[6] H. J. Miller, "A measurement theory for time geography," *Geographical Analysis*, vol. 37, no. 1, pp. 17–45, 2005.

[7] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh, "The livehoods project: Utilizing social media to understand the dynamics of a city," in *Sixth International AAAI Conference on Weblogs and Social Media*, 2012-05-20.

[8] F. Kling and A. Pozdnoukhov, "When a city tells a story: urban topic analysis," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '12. ACM, 2012, pp. 482–485.

[9] V. Frias-Martinez, V. Soto, H. Hohwald, and E. Frias-Martinez, "Sensing urban land use with twitter activity," 2013. [Online]. Available: http://www.enriquefrias-martinez.info/yahoo_site_admin/assets/docs/socialcom20121.17131525.pdf

[10] G. Colombo, M. Chorley, V. Tanasescu, S. Allen, C. Jones, and R. Whitaker, "Will you like this place? a tag-based place representation approach," in *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2013, pp. 224–229.

[11] Y. Qu and J. Zhang, "Regularly visited patches in human mobility," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. ACM, 2013, pp. 395–398.

[12] S. Isaacman, R. Becker, R. CÃ¡ceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Pervasive Computing*, ser. Lecture Notes in Computer Science, K. Lyons, J. Hightower, and E. M. Huang, Eds. Springer Berlin Heidelberg, 2011, no. 6696, pp. 133–151.

[13] R. Goodspeed, "The limited usefulness of social media and digital trace data for urban social research," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013-06-28.

[14] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work and Think.* John Murray, 2013.

[15] B. W. Silverman, *Density estimation for statistics and data analysis.* Chapman and Hall, 1986.

[16] L. Yu, "Understanding information inequality: Making sense of the literature of the information and digital divides," *Journal of Librarianship and Information Science*, vol. 38, no. 4, pp. 229–252.

[17] C. Fuchs, "The role of income inequality in a multivariate cross-national analysis of the digital divide," *Social Science Computer Review*, vol. 27, no. 1, pp. 41–58.

[18] J. Schradie, "The digital production gap: The digital divide and web 2.0 collide," *Poetics*, vol. 39, no. 2, pp. 145–168.

[19] D. R. Brake, "Are we all online content creators now? web 2.0 and digital divides," *Journal of Computer-Mediated Communication.*

[20] A. Smith and J. Brenner, "Twitter use 2012," 2012. [Online]. Available: http://pewinternet.org/Reports/2012/Twitter-Use-2012.aspx

[21] Wood, Jo, "The LandSerf manual," 2009. [Online]. Available: http://www.soi.city.ac.uk/~jwo/landserf/landserf230/doc/landserfManual.pdf

[22] M. Birkin, K. Harland, and N. Malleson, "The classification of space-time behaviour patterns in a british city from crowd-sourced data," in *Computational Science and Its Applications – ICCSA 2013*, ser. Lecture Notes in Computer Science, B. Murgante, S. Misra, M. Carlini, C. M. Torre, H.-Q. Nguyen, D. Taniar, B. O. Apduhan, and O. Gervasi, Eds. Springer Berlin Heidelberg, 2013, no. 7974, pp. 179–192.

[23] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009-02-19.

[24] A. Tsotsis. Twitter is at 250 million tweets per day, iOS 5 integration made signups increase 3x. [Online]. Available: http://techcrunch.com/2011/10/17/twitter-is-at-250-million-tweets-per-day/

[25] American Sociological Association, *Code of Ethics and Policies and Proceduresof the ASA Committee on Professional Ethics.* American Sociological Association. [Online]. Available: http://www.asanet.org/images/asa/docs/pdf/CodeofEthics.pdf

[26] G. Eysenbach and J. E. Till, "Ethical issues in qualitative research on internet communities," *British Medical Journal*, vol. 323, no. 7321, pp. 1103–1105.

[27] D. Wilkinson and M. Thelwall, "Researching personal information on the public web: Methods and ethics," *Social Science Computer Review*, vol. 29, no. 4, pp. 387–401. [Online]. Available: http://ssc.sagepub.com/cgi/doi/10.1177/0894439310378979