Please cite as: Birkin, M., Harland, K., Malleson, N., Cross, P., Clarke, M. (2014) An Examination of Personal Mobility Patterns in Space and Time Using Twitter. *International Journal of Agricultural and Environmental Information Systems* 5, 55–72. doi:10.4018/ijaeis.2014070104

An examination of personal mobility patterns in space and time using Twitter

Mark Birkin, Kirk Harland, Nicolas Malleson, Philip Cross, Martin Clarke
School of Geography, University of Leeds, Leeds LS2 9JT, UK

ABSTRACT

New sources of data relating to personal mobility and activity patterns are now providing a unique opportunity to explore movement patterns at increasing scales of spatial and temporal refinement. In this paper, a corpus of messages from the Twitter social networking platform are examined An elementary classification of users is proposed on the basis of frequency of use in space and time. The behaviour of different user groups is investigated across small areas in the major conurbation of Leeds. Substantial variations can be detected in the configuration of individual networks. An interpretation of the patterns which result is provided in terms of the underlying demographic structures, and the basic form and function of the urban area.

KEYWORDS

Mobility; twitter; social network; spatial analysis; geodemographics.

INTRODUCTION

The widespread availability of increasingly diverse and rich sources of data about the world around us has attracted significant interest in media as widespread as the daily press (Ganesh, 2014), professional magazines (Anderson, 2008) and the scientific literature (Bell et al, 2009). Claims that better understanding of social trends will follow naturally or automatically from the analysis of such data require qualification (Mayer-Schönberger and Cukier, 2013). Analytical frameworks are certainly required in which further empirical investigations can be undertaken in relation to established behavioural or theoretical frameworks. In this paper, an investigation of daily mobility patterns in a British city will be conducted using data from the Twitter social media platform.

It will be argued that established approaches are primarily geared towards the interrogation of long-term relocation patterns (i.e. migration). Short-term movement patterns have received much lower priority, and this reflects the poverty of available data more than a lack of intrinsic interest in the subject matter. This argument is amplified in Section 2 of the paper. In order to further the understanding of daily mobility patterns, the Twitter social messaging platform presents itself as a candidate. A substantial corpus of data for the city of Leeds has been extracted from Twitter, using the methods articulated in Section 3. A straightforward but robust approach in the literature to the analysis of small area variations in social and economic structures is geodemographics, which provides a framework for the classification of individual neighbourhoods, households or individuals. In Section 4 of the paper an elementary classifier is presented for application to Twitter users. The sub-division which follows can be used as a means for further exploration of mobility patterns in space and time. Some preliminary results from the application of this procedure are reported in Section 5. The paper concludes with a discussion of the value and potential in the approach, as well as a consideration of enhancements which may be necessary if more substantial insights are to be acquired.

POPULATION MOVEMENTS IN SPACE AND TIME

Research Themes

This paper aims to link together two themes of population mobility and the character of urban and rural neighbourhoods.

Existing research on population mobility is heavily weighted towards long-term movements of individuals and households through the process of migration. High quality sources of data, particularly from government censuses and surveys, have been used to evaluate migration decisions in relation to local factors governing both push factors (e.g. unemployment rates, Hamalainen and Bockerman, 2004; house prices, Chan (2001) and pull factors (e.g. employment opportunities and wages, Di Cintio and Grassi (2011), environmental amenities, Argent et al (2010), educational opportunities, DeBrauw and Giles (2008). Increasingly the weaknesses of conventional data sources have been recognised through the adoption of registry sources, for example through primary healthcare (Bell et al, 2002; Raymer et al, 2011) and such data have the advantage of being both continually updated and assessing the frequency of movement (i.e. the number of moves, and not just a transition over a period of time). More recently, work has begun to emerge connecting movement to a wide range of behavioural and lifestyle characteristics of the population at very fine spatial scales (Thomas et al, 2014) and this is made possible through access to massive data sources through commercial organisations in the private sector (Thompson et al, 2010).

Whilst migration is an important social process, the question of short-term movement patterns is also of extreme significance e.g. to those trying to understand problems of accessibility, service provision or even questions relating to emergency planning and disaster mitigation. The ability to address such questions has clearly been advanced by the availability of new data sources, including crowd-sourced mapping (Batty et al, 2010) and social media data (Qualman, 2013). Of particular interest here are two substantial bodies of work which are intrinsically concerned with daily movement patterns; these are the Pop 24/7 project at the University of Southampton, and Landscan, under development at Oakridge National Laboratory, Tennessee.

The Landscan programme uses a wide variety of sources ranging from published surveys to real-time sensors in order to estimate fine scale population movement at a resolution of single km^2 or less across the whole of the planet. Suggested applications include the development of evacuation plans for an entire city (Liu, 2012) and the analysis of ambient (non-residential) crime patterns (Andresen et al. 2012; Malleson and Andresen 2014). POP 24/7 uses an extensive combination of economic, demographic and land use inventories and surveys to model the ebb and flow of the population around a city through the day. Planning for the impact of environmental misadventures, such as flooding, is seen as an application of consequence (-), although the implications for retail and service provision have also come into consideration (Leung et al, 2010).

An issue with the POP24/7 approach in particular is that it is heavily driven from the 'supply-side' i.e. it is predicated on the availability of jobs, pupil places, hospital beds, and other elements of physical infrastructure. It does not necessarily reflect demographic attributes and behaviour patterns. One consequence is that cross-sectional detail about the nature of daily populations is omitted (Birkin et al, 2013b). In earlier work (Birkin et al, 2013a) this limitation is explored using social network data from the twitter messaging service. The authors draw on ideas from neighbourhood classification, or 'geodemographics', which has been a popular method in Europe and the US for assessing the overall character of small spatial units (Harris et al, 2005; Weiss, 2000). In common with migration patterns (see previous discussion), geodemographics has typically been used to assert long-term regularities in spatial pattern (e.g. Dorling et al, 2000), What is clearly novel about the current work is that it seeks to shift the analysis to a diurnal time frame in which the changing character of areas can be monitored from one time period to another at multiple occurrences across a daily or weekly cycle. This is achieved using social messages which are mined for both content and spatial information. These data are described in more detail in the next section of the paper. Whilst the previous work has connected to geodemographics, or segmentation of areas through spatial clustering, in this paper the method is now extended to consider the classification of individual service users. Through the identification of key parameters for each individual twitter user, the generation of distinct individual profiles is facilitated. The methods for achieving this are described. This permits further analysis for a variety of timescales and spatial geometries which are outlined further. In particular, it proves possible to map out concentrations of different sub-groups across a region in space and time, and this is a significant and novel advance in our understanding of the flux in populations for local geographies and short time periods. The discussion presents reflections on the extent of these achievements, as well as some caveats and limitations, and suggested priorities for further development and extension.

Before discussing the data used in this research, the paper begins with a review of relevant social media applications to understanding mobility patterns.

Social Media Data for Understanding Urban Dynamics

A data "revolution" (Mayer-Schönberger and Cukier, 2013) is underway which has the potential to transform our understanding of urban mobility. In particular, new social media services are emerging that are being used by individuals to publish information about their daily spatio-temporal behaviour. In this manner, it is becoming possible to supplement traditional data sources, such as censuses, with information about individual charateristics and daily behaviour at

an extremely fine resolution. To this end, a number of projects are emerging that challenge traditional residential geodemographic classifications. Examples include: the classification of neighbourhoods based on visitors rather than residents using Foursquare check-in data (e.g. the Livehoods project, Cranshaw et al., 2012); the development of area profiles using social media (Kling at al., 2012; Frias-Martinez at al., 2013); and perceptions of place using Foursquare data (Colombo et al. 2013). Furthermore, a small number of studies are attempting to identify individual spatio-temporal behaviour at even finer scales (Isaacman, 2011; Qu, 2013; Malleson and Birkin, 2014).

Although there is great potential offered by these new data and research directions, there are substantial ethical and methodological issues that must be considered. From a methodological perspective, it is important to consider factors that might limit participation in social media such as the digital divide. For example, Schradie (2011), writing on web participation among different social groups, finds that "the poor and working class have not been able to use these production applications at the same rate as other uses or users, creating a growing production divide based on these elite creative functions". Problems even arise *within* social media data sets as small groups of over-active users introduce bias through the shear volume of data that they create. From an ethical position, care must also be taken, particularly around issues of privacy and informed consent. Although most social media contributions are by their very nature public, it is not always clear that users are fully informed about the volume of information that they reveal about themselves. Nevertheless, the use of social media data has great potential to inform our understanding of urban mobility and dynamics, on the condition that assumptions/biases are well publicised even if not fully understood.

Data

A large corpus of Twitter messages were collected for the city of Leeds between July 2011 and July 2012. The search was restricted to location-enabled messages, within a rectangular area bounded by the coordinates (x,y) to the south-west and (x2,y2) to the north-east. Messages that had some location information (e.g. a 'place' chosen by the user) but no GPS coordinates were not included in the analysis. Estimates suggest that approximately 1-2% of all Twitter messages have a coordinate appended to them (Leetaru et al. 2013, Gelernter and Mushegian 2011), as the majority of mobile telephones users at this time either lack the facility to track locations, or the user chooses not to enable this capability.

In total, more than 3 million messages were captured. A substantial number of these messages were identified as unusable in view of their content and method of generation. These include the outputs of services and devices such as traffic reports, weather reports and advertisements. Representative content for messages of this type would be 'Stationary traffic, A1 North, Wetherby to Northallerton' (traffic); 'wind 25mph, south-east, light rain' (weather); 'Used car for sale, Ford Focus, 18,000 miles' (advert). These messages are from fixed installations – they do not move around with individual users, and their content is topic focused (i.e. to traffic, weather, or product sales). The tweets are readily identifiable not just from their content but by repeated and regular occurrence at the same location, and were removed. Also, an artefact of the collection process meant that some messages located outside the bounding box were captured, although they should be ignored. Altogether 2,836,661 messages remained after the removal of these spurious messages.

Each entity in the dataset contains the following attributes – unique codes for the user and message, message content (up to 140 characters), xy coordinates, time and date of message. Individual names are not retained, but the availability of the unique code allows a linkage between messages for a common user, allowing simple queries such as the number of messages created by each user over discrete points of time. The distribution of message volumes is skewed – see Figure 1 - with a small number of prolific users, and a large number of occasional users. For example, 50% of messages are generated by just 2% of users (and indeed the top 10% of users account for fully 81% of messages).

Figure 1. The Lorenz curve of messages per user. The Gini coefficient is 0.87, indicating high inequality between the most and least prolific users.

Methods

The paper seeks to address a number of questions as follows: is it possible to detect spatial variations in messaging patterns from area to area; is it possible to characterise users according to characteristic traces in their activity profiles; can one make inferences about individual activity spaces from the location, content and frequency of twitter messages?

In order to address these questions the data are first analysed to segment the spatio-temporal profiles of each user according to the time spent at home, and at various distances from the place of usual residence.

Those users sending less than 20 tweets are removed as it was found that at least this number of tweets are required to obtain a useful amount of data for the second stage of analysis.

Analysis of the remaining analysis has been performed in Matlab as follows. Firstly, the time stamps have been converted to Unix time to allow easier time calculations. Next, the twitter data are sorted by the user ID and then by Unix time, and the time between tweets sent by each user is calculated. The Euclidean distance between these tweets is also obtained, and while this distance is less than 250m, the total time spent in the location is calculated. The data are discarded if the total time is less than one hour or greater than three days.

These data are sorted again, this time by user ID then by location (actually, by longitude), and the Euclidean distances between consecutive tweets are once more obtained. While this distance is less than 250m, the total time spent in the location is calculated. This results in a range of locations for each user and the total time spent in each. It is assumed that the location in which the user spends most time represents their home. A selection of these locations were validated by reading the content of tweets sent from home (i.e. “I’m in the bath” or “I’m watching TV”) .

The next step of this stage involves finding the day of the week and time of day when each tweet was sent. This allows the time when each tweet is sent to be divided up into early mornings, late mornings, afternoons and evenings for both weekdays and weekends. A histogram for each of these time periods has been created for all users showing the distribution of the distances from the user’s home of the locations from which tweets are sent.

The electoral ward from which each tweet was sent has been obtained by performing a spatial join. An aggregate of the tweets sent in each electoral ward for the time periods described above

has been obtained by adding all tweets sent in the ward and dividing by the number of users in the ward.

A characteristic output from this process is shown at Figure 2, showing tweet activity patterns for users in Armley, a ward in the inner west of Leeds. Profiles are presented in four six hour blocks of time for both weekdays and weekends. These time periods might be considered roughly as night-time, morning, afternoon and evening. The bars in the chart represent the shares of twitter activity in each time slice within single kilometre bands of distance from the home of the user. In this case one can detect a steady increase in activity throughout the day and this pattern is consistent between weekday and weekend; levels of activity for any slice of time are typically in the order of twice as high in the week as at weekends (except during the night-time at weekends when activity levels are extremely low); and there is a very strong concentration of volumes at the place of residence, with a noticeable spatial interaction effect as the user moves away from home.

Figure 2. Twitter behaviour profile for Armley

In the second part of the analysis, the focus moves from ward profiles to individual users of the service. For each of these individuals three characteristics are sought – the proportion of time spent at home, the overall level of activity, and the balance of tweets between weekdays and weekends. For each of these dimensions, activity levels are characterised as either 'high' or 'low' by initially obtaining the median of the number of tweets sent by the users, the median distance from the users' homes of the locations of where tweets are sent, and the median number of tweets sent during week days. Those users sending a greater number of tweets than the median are given a value of '1', with the remaining users given a value of '0'. Similarly, users sending more tweets at a greater distance from their home than the median are given a value of '1', and the remaining users are given a value of '0'. Finally, those users sending more tweets during the week than the median are given a value of '1' and the remainder given a value of '0'. This analysis has also been performed for each electoral ward.

Following this procedure, users may be 'classified' into one of eight categories based on an exhaustive analysis of the combinations, as shown in Table 1 – so for example, individuals in type 1 spend most of their time at home, they are most active at weekends, and are low intensity users. Individuals in type 8 spend most of their time away from home, are most active in the week, and are high intensity users.

	Time away from home	Weekday activity	Intensity of use

Group One	Low	Low	Low
Group Two	Low	Low	High
Group Three	Low	High	Low
Group Four	Low	High	High
Group Five	High	Low	Low
Group Six	High	Low	High
Group Seven	High	High	Low
Group Eight	High	High	High

Table 1. Categorical analysis of twitter users

RESULTS

Activity profiles

As a first level of analysis, the variations in activity profiles of Leeds wards are considered. A map of these wards is provided for reference at Figure 3. In Figure 2 above an example of the ward of Armley was reviewed in which the major features appeared to be strong spatial interaction effects, a weekday bias in twitter activity, and growth in volumes through the day. In Figure 4 a different kind of area is considered. While Armley is a central ward with relatively low wages and higher than average unemployment, Rawdon and Guiseley are affluent suburban towns around ten kilometres from the city centre. A marked contrast is observed – in the weekdays night-time activity is still low, but volumes are quite balanced through the day. At weekends the propensity to tweet is very low, while tweeting distances are now quite dispersed. This could be a reflection of greater mobility amongst more prosperous individuals, in addition to the greater likelihood of commuting significant distances, especially to the city centre.

Figure 3. Map of Leeds

Figure 4. Twitter behaviour profile for Rawdon and Guiseley

A third example – Headingley - is shown in Figure 5. In this case a bias towards weekdays is still observed, but bearing in mind that there are five days in the week and only two at the weekend this pattern is reversed if we were to consider the rates e.g. number of messages per hour. The most noticeable feature here, however, is a fairly even split of messages through the day, especially at night, probably reflecting the fact that Headingley has an extremely high concentration of University students who tend to be active (on social networks) around the clock.

Figure 5. Twitter behaviour profile for Headingley

In Figure 6 the percentage of tweets away from home on weekdays during the night-time hours of midnight and six o'clock in the morning is considered as a single indicator is mapped with a view to highlighting some differences across the entire area. This map looks to have an interesting interpretation in relation to the social and economic geography of the city. In particular, the outlying towns to the east and north-east e.g. Wetherby and Tadcaster are typically either a magnet for retirees, or small freestanding communities. Domestic life seems a stronger focus here than in the busy commuter suburbs (Horsforth, Rawdon, Guiseley) to the west and north-west.

Figure 6. Tweets away from home, weeknights

A classification of users and their movement patterns

In the second part of the analysis individuals are assessed, using the method introduced in the previous section. This segmentation of the population can be viewed as a hierarchical split, rather like a decision tree (Dobra, 2009) as shown in Figure 7. Here we can see that the allocation of individuals across the groups is quite regular, bearing in mind that there are only around half as many high intensity users as low, but again these are quite evenly distributed. In view of the rules which are adopted in the formation of the groups, it is possible to suggest idealised profiles or 'pen portraits' in the style widely adopted by geodemographic classifications (for example, in the Office for National Statistics Output Area Classification for the UK; Vickers and Rees, 2007).

Types 1 & 2 – activity patterns are skewed towards the weekend over short distances. This is suggestive of individuals with community or family interests (Type 1 – Family and Friends). The higher intensity users in this category are often motivated by enthusiasm for a particular activity e.g. soccer or a reading club (Type 2 – Local Hobbyists).

Types 3 & 4 – activity patterns are skewed to the weekday over short distances. In the case of low intensity users this may reflect a concern with the necessities of daily life, e.g. getting the children to school, the domestic economy (Type 3 – Homemaker). For higher intensity users, daily life may provide an endlessly continuing stream of local rumour and gossip (Type 4 – Neighbour).

Types 5 & 6 – activity patterns are skewed to longer distance interactions at the weekend. At lower intensity, this could reflect a more mature demographic which is mobile and active (Type 5 – Socialite) but less attuned to social media than the more active users (Type 6 – Student).

Types 7 & 8 – activity patterns are dominated by long distance movements in the week. At low intensity this might be typical of professional workers and businessmen for whom a busy working day provides a little time for significant reflections (Type 7 – Executives). High intensity users may have greater opportunities for social messaging activity in the course of travel by bus, train or private modes (Type 8 – Commuter).

It is interesting now to conduct further investigations about where we find the highest concentrations of these tweeting groups. To examine this question, we calculate an index of concentration for each group in each census ward:

$$CI = 100 * (x(i,k) / x(I,*)) / (x(*,k) / x(*,*))$$

Where x(i,k) is the number of people of group k in area i – in other words this index compares the proportion of each group in a local area with their proportion in the study region as a whole, and presents the result as an index around 100.

On the basis of this calculation, strong concentrations of the groupings can be found as follows:

Type 1 – Family and Friends; Headingley and Armley

Type 2 – Local Hobbyists; Burmantofts and Middleton Park

Type 3 – Homemaker; Bramley and Chapel Allerton

Type 4 – Neighbour; Temple Newsam and City

Type 5 – Socialite; Bramley and Hyde Park

Type 6 – Student; Ardsley and Armley

Type 7 – Executives; Headingley and City

Type 8 – Commuter; Killingbeck and City

Figure 7. Categories and counts of twitter users

These patterns are in accordance with expectations to some degree. For example, the Hyde Park area of Leeds has amongst the highest concentrations of student populations in the city, and is also associated with diverse movement patterns at irregular times. Weekday activity with long distance trips (Groups 7,8) are unsurprisingly associated with the City Centre. Overall, there is a tendency for most groups to concentrate in the central areas of the city, which could reflect the relatively high incidence of both younger and more well-educated populations towards the urban core.

It should also be noted, however, that once twitter users have been filtered into regular users, and then disaggregated across both locations and activity groups, then the number of observations in each category is starting to become rather unreliable, and exposes the analysis to extreme variations from the behaviour of a small number of high volume users who may be untypical of the population as a whole. The breakpoints adopted in this analysis are also somewhat arbitrary, and a more refined investigation could perhaps explore more sophisticated distinctions in time, location, message volume, or other characteristics not considered here.

Spatial Networks

In a third piece of analysis, the characteristics of the spatial networks of individual users will be assessed. Recall that the location analysis of the previous section is predicated on an ability to cluster the activity spaces of individuals into nodes. Thus we can ask questions relating to both the density and spread of nodes for each person. Consider an individual whose daily routine takes them from home in the suburbs on a long journey to work through the daytime, and then back at home again in the evening. This person would be characterised by a small number of activity nodes (maybe only two) with a high spread. Now think of an individual who lives in the centre of the city, maybe goes to some classes through the day, meets for coffee with some

friends during the day, has an exercise session at a local gym in the early evening, does regular top-up shopping, and regularly takes in an evening of theatre or live music. This individual would be characterised by a large number of activity nodes with a low spread (see Figure 8).

Figure 8. Individual Activity Spaces

Figure 8a – A hypothetical individual with a small number of activity nodes

Figure 8b – A hypothetical individual with a large number of activity nodes

This intelligence can be used to investigate a range of questions about the character of personal activity spaces, specifically: is there a relationship between network density and spatial location; is there a relationship between network spread and spatial location; and is there a relationship between person type (as identified above) and either spread or density?

The evidence relating to locations is shown in Table 2. This appears to provide clear support to the idea that the activity spaces of individuals in the suburbs are more diverse (higher spread) than those in the centre of the city. However the situation with respect to density is more difficult to generalise – in spite of the greater distances involved, it seems that residents of the outlying areas are likely to move between just as many locations, if not more, than their urban counterparts. This could reflect greater personal mobility e.g. through higher car ownership, or the patterns associated with busy family lives combining education, shops, after-school activities as well as work and domestic leisure.

The evidence for different user groups is shown in Table 3. A notable feature here is that as one would expect the high activity users (groups 2,4,6,8) all have more dense networks than the low activity users. However this is not reflected in the spread of their networks, which are typically not substantially larger. The average distances tend to be quite low amongst groups 7 and 8, which could reflect the tendency of these groups to cluster in the city centre as was revealed in the earlier analysis; however it is also notable that the activity spaces of groups 7 and 8 are actually less diverse than those of groups 1 and 2, and this could be seen as reinforcing the conclusions from the spatial analysis – that the city dwellers do not have denser networks of activity and interaction, despite the higher concentration and greater access to opportunities in the urban core.

Area	Spread (km)	Area	Average contacts
Guiseley and Rawdon	6.3	Ardsley & Robin Hood	12.0
Middleton Park	5.1	Armley	11.2
Otley and Yeadon	5.0	Otley & Yeadon	10.1
Garforth & Swill'n	4.9	Rothwell	10.1
Kippax and Methley	4.8	Killingbeck	9.8
Cross Gates & Whinmoor	4.7	Middleton Park	9.6
Alwoodley	4.6	Hyde Park & Woodhouse	9.3
Temple Newsam	4.6	Calverley & Farsley	9.3
Ardsley and Robin Hood	4.1	Burmantofts & Richmond Hill	8.2
Rothwell	4.0	City and Hunslet	8.2
Adel and Wharfedale	4.0	Beeston & Holbeck	7.9
Harewood	3.9	Morley North	7.9
Killingbeck and Seacroft	3.6	Farnley & Wortley	7.8
Calverley and Farsley	3.5	Kippax & Methley	7.7
Farnley and Wortley	3.4	Guiseley & Rawdon	7.4
Morley North	3.3	Bramley & Stanningley	7.1
Chapel Allerton	3.3	Chapel Allerton	7.1
Moortown	3.2	Horsforth	6.9
Gipton and Harehills	3.1	Roundhay	6.9
Bramley and Stanningley	3.0	Pudsey	6.8
Pudsey	3.0	Temple Newsam	6.8
Burmantofts and Richmond Hill	2.9	Cross Gates & Whinmoor	6.7
Wetherby	2.9	Garforth & Swillington	6.6
Weetwood	2.8	Alwoodley	6.6
Morley South	2.7	Weetwood	6.5
Armley	2.5	Morley South	6.4
Horsforth	2.4	Kirkstall	6.2
Roundhay	2.3	Moortown	6.0
Beeston and Holbeck	2.1	Headingley	5.6
Kirkstall	2.1	Adel & Wharfedale	5.5
Headingley	1.9	Harewood	5.3
City and Hunslet	1.8	Gipton and Harehills	5.2
Hyde Park and Woodhouse	1.5	Wetherby	4.7

Table 2. Spread and Density of Activities across the Neighbourhoods of Leeds

Group	Group Name	Spread (km)	Average contacts
One	Family and Friends	3.5	7.3
Two	Local Hobbyists	4.4	11.3
Three	Homemaker	2.6	5.5
Four	Neighbour	2.6	10.4
Five	Socialite	3.2	6.0
Six	Student	4.0	12.7
Seven	Executives	2.2	5.6
Eight	Commuter	2.6	10
Total		**3.6**	**7.5**

Table 3. Travel distance and activity spaces for the eight user groups

DISCUSSION

Twitter messaging patterns have been used as a means to explore spatial and temporal patterns in daily urban mobility. Clear variations have been uncovered which can be interpreted in relation to the underlying demographics and infrastructure of the city. This spatial networks for individual users are typically more compact in the centre of the urban area; on the other hand high intensity users are not necessarily active across a bigger physical space. However tentative, these results begin to add to our understanding of routine movement patterns, which is somewhat lacking at present, and this work has significant importance for various policy applications e.g. in emergency planning and resource allocation.

The work is still limited by vagaries in the underlying data. Although messages were collected over a long period of time from a substantial population of users, strong distortions have been introduced through the impact of a small number of prolific users. The overall value of the dataset is restricted by the fact that location-enabled messages are still in the minority, although it is likely that the trend towards increasing uptake of the service as a whole, and the growing likelihood that devices are location-enabled will ameliorate this problem over time. Furthermore, it is well-known that users of social media such as Twitter are not a representative sample of the population. For example, only 8% of social media users in the US are currently aged 55 and over (www.pingdom.com, accessed 18/11/2014), compared to 25% of the population. The skews in both the Twitter users, and the sub-group of users who allow their tweets to be geolocated, therefore reveals more about the behaviour of Twitter users than about the population as a whole.

An alternative approach towards greater robustness in the data may be simply to exploit other sources. For example there is evidence that mobile telephone operators may be interested in coordinating massive transactional data as a window into such problems (Telefonica, 2014). The very large customer bases of the multi-national market leaders means that they are capturing many millions of events every day, and although they may lack the narrative depth of the twitter messages may have greater long-term potential as a reliable and voluminous source of intelligence about mobility.

The classifier adopted in this research was a straightforward decision-tree based on three user characteristics. Segmentations which have been adopted in other forms of urban and regional analysis, as well as for practical applications e.g. in direct marketing and service location, are typically multi-dimensional. The imputation of further characteristics such as behaviour (e.g. the purpose of a trip, or the mode of transportation) or demographics (e.g. age, occupation, income or family status) would allow for a much more sophisticated approach. One way to achieve this would be to link individual users to another dataset with names and individual characteristics. This procedure would raise problems of its own relating to privacy, confidentiality and the ethics of use. This issue is neatly recognised in the observation that 'just because content is publicly accessible does not mean that it was meant to be consumed by just anyone' (Boyd and Crawford, 2012, 672). Nevertheless similar applications are already possible under controlled circumstances – for example the Northern Ireland Longitudinal Survey (NILS) employs a linkage between individual census records and patient data using a trusted third party linkage which is never exposed to the research consumer of the service (O'Reilly et al, 2012).

The 'Big Data' narrative suggests that the overwhelming majority of intelligence about demographics, movement and behaviour is now being generated from unconventional sources in commercial or public service environments. Nevertheless the bias in academic research is still heavily weighted to traditional sources such as censuses and panel surveys. This will change as new sources become more well-known and easier to access, but as well as the potential it may be clear from the work presented here that there are still numerous challenges to overcome before full advantage can be made.

REFERENCES

Anderson C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired Magazine*.

Andresen, M.A., G.W. Jenion, and A.A. Reid. 2012. "An Evaluation of Ambient Population Estimates for Use in Crime Analysis." *Crime Mapping: A Journal of Research and Practice* 4(1): 7–30.

Argent, N., Tonts, M., Jones., R., Holmes, J. (2010) Amenity-led migration in rural Australia: A New Driver of Local Demographic and Environmental Change, *Demographic Change in Australia's Rural Landscapes*, Landscape Series Volume 12, 23-44.

Batty, M., Hudson-Smith, A., Milton, R., and Crooks, A. (2010) Map Mashups, Web 2.0 and the GIS Revolution, *Annals of GIS*, 16, 1, 1-13.

Bell, G., Hey, T., Szalay, A. (2009) Beyond the Data Deluge, *Science*, 323, 1297-1298.

Bell, M., Blake, M., Boyle, P., Duke-Williams, Rees, Stillwell, Hugo, G. (2002) Cross-national Comparison of Internal Migration: Issues and Measures, *Journal of the Royal Statistical Society, Series A*, 165, 435-464.

Bhaduri, B., Bright, E., Coleman, P., Urban, M. "LandScan USA: A High Resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics" *GeoJournal*. 2007. 69: 103-117.

Birkin M, Harland K, Malleson N (2013) The Classification of Space-Time Behaviour Patterns in a British City from Crowd-Sourced Data , *Proceedings of the International Conference on Computer Science and its Applications*, Lecture Notes in Computer Science, Springer, Berlin.

Birkin M, Harland K, Malleson N, Martin D (2013) Microsimulation of daily movement patterns in a British city, 20th International Congress on Modelling and Simulation (MODSIM), Adelaide.

Boyd, D., Crawford, K. (2012) Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon, *Information, Communication and Society*, 15, 5, 662-679.

Chan S. (2001) Spatial Lock-In: Do Falling House Prices Constrain Residential Mobility?, *Journal of Urban Economics*, 49(3), 567-586.

Cheng, L. (2012) A quick response emergency evacuation system for ultra large dataset, Hazards and Emergency Response, Annual Meeting of the Association of American Geographers, New York, 24/2/2014.

Colombo, G.B., Chorley, M.J., Tanasescu, V., Allen, S.M., Jones, C.B., Whitaker, R.M. (2013) Will you like this place? A tag-based place representation approach, in: *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pp. 224–229. doi:10.1109/PerComW.2013.6529486

Cranshaw, J., Schwartz, R., Hong, J., Sadeh, N. (2012) The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City, in: *Sixth International AAAI Conference on Weblogs and Social Media*. Presented at the Sixth International AAAI Conference on Weblogs and Social Media.

DeBrauw A., Giles, J. (2008) Migrant opportunity and the educational attainment of youth in rural China, Policy Research Working Paper 4526, World Bank, Washington DC.

Di Cintio, M., Grassi, E. (2011) Internal Migration and Wages of Italian University Graduates, *Papers in Regional Science*, 92(1), 119-140.

Dobra, A. (2009) Decision Tree Classification, *Encyclopaedia of Database Systems*, 765-769.

Dorling, D., Mitchell, R., Shaw, M., Orford, S., Davey-Smith, G. (2000) The Ghost of Christmas Past: health effects of poverty in London in 1896 and 1991, *British Medical Journal*, 321(7276), 1547-1551.

Frias-Martinez, V., Soto, V., Hohwald, H., Frias-Martinez, E., (2013). *Sensing Urban Land Use with Twitter Activity*. Available online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.409.878 [accessed July 2014]

Ganesh, J. (2014) Big data may be invasive but it will keep us in rude health, *Financial Times*, 22/2/2014.

Gelernter, J. and Mushegian, N., 2011. Geo-parsing Messages from Microtext. *Transactions in GIS* 15 (6), 753–773.

Hamalainen, K., and Bockerman, P. (2004) Regional Labour Market Dynamics, Housing and Migration, *Journal of Regional Science*, 44, 543-568.

Harris, R., Sleight, P., Webber, R. (2005) *Geodemographics, GIS and neighbourhood targeting*, John Wiley, Chichester.

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A. (2011) Identifying Important Places in People's Lives from Cellular Network Data, in: Lyons, K., Hightower, J., Huang, E.M. (Eds.), *Pervasive Computing, Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 133–151.

Kling, F., Pozdnoukhov, A. (2012). When a city tells a story: urban topic analysis, in: *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '12. ACM, New York, NY, USA, pp. 482–485. doi:10.1145/2424321.2424395

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A. (2013) Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18 (5).

Leung, Samuel, Martin, David and Cockings, Samantha (2010) Linking UK public geospatial data to build 24/7 space-time specific population surface models. In, GIScience 2010: Sixth international conference on Geographic Information Science, Zurich, Switzerland, 14 - 17 Sep 2010.University of Zurich, 7pp.

Malleson, N., Andresen, M.A. (2014). The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science* 1–10. doi:10.1080/15230406.2014.905756

Malleson, N. and M. Birkin (2014) New Insights into Individual Activity Spaces using Crowd-Sourced Big Data. In: *2014 ASE BigData/SocialCom/CyberSecurity Conference*, Stanford University, May 27-31 2014. Available online: http://www.ase360.org/handle/123456789/31 [accessed July 2014].

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

O'Reilly, D., Rosato, M., Catney, G., Johnston, F., and Brolly, M. (2012) Cohort Description: The Nortehrn Ireland Longitudinal Study, *International Journal of Epidemiology*, 41, 634-641.

Qu, Y., Zhang, J. (2013) Regularly visited patches in human mobility, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13. ACM, New York, NY, USA, pp. 395–398. doi:10.1145/2470654.2470711

Qualman, E. (2013) *Socialnomics: How Social Media Transforms the Way we Live and do Business*, Wiley, Chichester.

Raymer, J., Smith, P., Giuletti, C. (2011) Combining Census and Registration Data to Analyse Ethnic Migration Patterns in England from 1991 to 2007, *Population, Space and Place*, 17, 1, 73-88.

Schradie, J. (2011) The digital production gap: The digital divide and Web 2.0 collide. *Poetics* 39, 145–168. doi:10.1016/j.poetic.2011.02.003

Telefonica (2014) Dynamic Insights: Smart Steps, http://dynamicinsights.telefonica.com/488/smart-steps, accessed 3rd April 2014.

Thomas M; Gould M; Stillwell J (2012) Exploring the potential of microdata from a large commercial survey for the analysis of demographic and lifestyle characteristics of internal migration in Great Britain, Working Paper of the University of Leeds, School of Geography, 12/2 .

Thompson, C., Stillwell, J., Clarke, M. and Bradbrook, C. (2010) Understanding and Validating Acxiom's Research Opinion Poll Data, Working Paper 10/6, School of Geography, University of Leeds.

Vickers, D., Rees, P. (2007) Creating the UK National Statistics 2001 output area classification, *Journal of the Royal Statistical Society: Series A*, 170, 2, 379–403.

Weiss, M. (2000) *The clustered world: How we live, what we buy, and what it all means about who we are*, Little, Brown, Boston.