Generating Individual Behavioural Routines from Massive Social Data for the Simulation of Urban Dynamics

Nick Malleson and Mark Birkin
School of Geography, University of Leeds, Leeds LS2 9JT

School of Geography, University of Leeds
n.malleson06@leeds.ac.uk

Abstract. This paper presents recent methodological developments aimed at establishing the daily spatio-temporal behaviour of individual people from their activity on social-networking services. Ultimately, the methods will be used to provide supplementary data to a complex agent-based model of urban dynamics. This work will review recent developments in the use of crowd-source data for understanding society, outline novel methods for capturing the spatio-temporal behaviour of individual users and discuss how the this information can be incorporated into a model of urban dynamics. Finally, we conclude with a discussion about current challenges and future research.

Please cite as: Malleson, N. and M. Birkin (2014) Generating Individual Behavioural Routines from Massive Social Data for the Simulation of Urban Dynamics. *Proceedings of the European Conference on Complex Systems 2012*. Springer Proceedings in Complexity 2014, pp 849-855

1 Introduction

This paper presents recent methodological developments aimed at establishing the daily spatio-temporal behaviour of individual people from their activity on social-networking services. Ultimately, the methods will be used to provide supplementary data to a complex agent-based model of urban dynamics. This work will review recent developments in the use of 'crowd-source' data for understanding society (Section 2), outline novel methods for capturing the spatio-temporal behaviour of individual users (Section 3) and discuss how the this information can be incorporated into a model of urban dynamics (Section 4). Finally, Section 5 will conclude the article with a discussion about current challenges and future research.

2 The Social Data Deluge

Recent work recognising the character of cities as complex systems suggets that aggregate models have the effect of smoothing out a city's underlying dynamism (Batty, 2005). Hence individual-level modelling approaches (such as agent-based modelling) have become a popular means of representing various urban phenomena such as disease spread (Johansson et al., 2012), crime (Malleson et al., 2010) and land-use (Matthews et al., 2007). However, individual-level models are often criticised because they lack proper validation, which stems from a shortage of suitable data. Models often use population censuses and social surveys that have four major drawbacks:

1. they tend to deal with aggregate groups rather than individuals;
2. they occur infrequently;
3. they are usually focused on the attributes and characteristics of the population, rather than their attitudes and behaviours;
4. they provide a snapshot view rather than a dynamic and continuous perspective.

New sources of social data – commonly referred to as "crowd-sourced data" (Savage and Burrows, 2007) and "volunteered geographical information" (Goodchild, 2007) – are becoming available that resolve many these drawbacks. Services such as Facebook, FourSquare, Flikr and Twitter provide large-scale social network data that have the potential to revolutionise the study of social phenomena and our approach to social simulation. These sources potentially contain a wealth of information about peoples' spatio-temporal behaviour, if relevant information can be extracted from the vast amounts of noise that are also present. Numerous researchers are developing new methods for data collection (Cheng et al., 2010; Russell, 2011; Stefanidis et al., 2011) and making use of crowd sourced data in diverse fields such as social network analysis (Davis Jr et al., 2011), crisis

management/evaluation (Gelernter and Mushegian, 2011), disease spread (Gomide et al., 2011) and election forecasts (Tumasjan et al., 2011). However, the application of crowd-sourced data to understanding urban dynamics is a relatively under-researched field.

The use of crowd-sourced data in the social sciences shares a number of similarities with the "fourth paradigm" (Bell et al., 2009) data intensive activities that are usually limited to the physical and biological sciences. Under this new paradigm, the vast amounts of data that are often generated by physical or biological experiments can be accessed through publicly available interfaces which effectively provides a much larger number of researchers with the opportunity to use the data. This new paradigm is not an unrealistic goal for the social sciences.

The growing use of online social media also has the potential to lead to a paradigm shift in the design and delivery of traditional social surveys. For example, the Mapiness (MacKerron, 2012) project has used a mobile phone application to collect data on peoples' perceived levels of happiness. Importantly, the geographical location of the respondent is known (using the phone's GPS equipment) so the researchers are able to relate their happiness to environmental conditions. At the time of writing, over 50,000 people had installed the application and the project had collected approximately 3.5 million individual data points. Clearly this amount of data would be extremely difficult to collect using a traditional survey; suggesting the potential for a more concerted effort in developing applications that are attractive to end users. In a similar vein, Birkin et al. (2011) utilised data collected through a large online survey which was publicised on UK local news. The survey sought peoples' views on how their travel behaviour would change following the implementation of a road charging scheme and Birkin et al. used this information to calibrate a model of future traffic flows. Again, the number of responses to the survey (more than 15,000) would have been difficult to gather using traditional methods.

3 Establishing Individual Behaviours

The overarching aim of the research is to build an accurate agent-based model of the daily ebb and flow of a city. At present, the research focusses on commuting behaviour, but future work will incorporate additional key behaviours (shopping, leisure activities, school, etc.). Therefore in the current research iteration we attempt to identify two specific behavioural states: 'at home' or 'at work'. This section will discuss how data from Twitter can automatically be classified into these two categories.

Initially, a person's base location (referred to as their 'home') is identified by calculating the most dense location of all their tweets. In future, it will be possible to generate more accurate estimates of behaviour through the incorporation of the textual content of the message as well as spatial location – see, for example, Eisenstein et al. (2010). The Kernel Density Estimation (KDE) algorithm, as formulated by Silverman (1986), was applied to every distinct tweet location to identify the point with the highest density which was assumed to be the user's home. This process can be repeated for every user in the data.

Following the identification of 'home' locations, it is possible identify where each tweet was sent from relative to the location of the user's home. At present, it is assumed that all tweets within 50 meters of a user's home indicate that they are 'at home', otherwise they are assumed to be 'away from home'. By temporally aggregating this information for every user into a single day, it is possible to perform a cursory validation of the behaviour that has been identified. To this end, Figure 1 presents the proportions of each behaviour in each hourly time period for all users. Tweets on days Friday to Sunday have been ignored at this stage in order to capture likely commuting behaviour. It is clear that there are more 'away from home' behaviours during the day and a greater proportion of 'at home' tweets in the evening which is an encouraging preliminary result.

4 Towards an Individual-Level Model

The identification of messages that originate from 'home', or elsewhere, has laid the groundwork for later use in an agent-based model of urban dynamics (which is currently under development). The aim is to use these data in conjunction with other sources, such as population censuses, to provide novel ways of seeding, calibrating and validating the model.

The first stage of the modelling process is to use *microsimulation* to disaggregate the 2001 UK census (and the 2011 census when it becomes available) to generate a synthetic population of individuals who will occupy the model (Birkin et al., 2006; Harland et al., 2012). Each individual is created with number of personal attributes which can be used in later research to tailor their behaviour. The individuals are grouped into households during the microsimulation process and these are geo-referenced to the resolution of the smallest census area boundary (called an 'output area'). With the incorporation of additional building data they can be assigned to

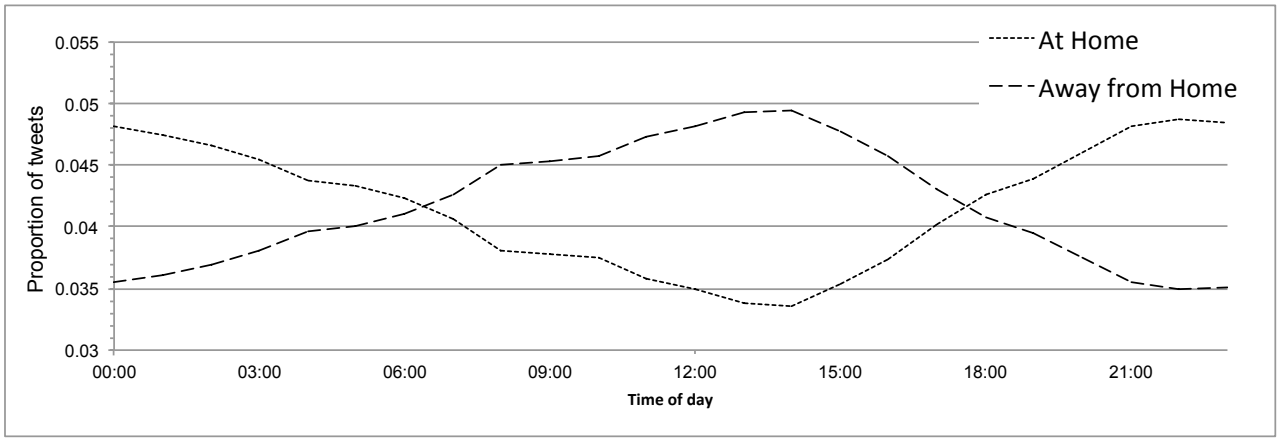

Fig. 1. The proportions of tweets corresponding to 'home' or 'work' activities for an average weekday (all Monday – Thursday tweets) for all users.

distinct houses (e.g. Malleson and Birkin, 2011), although this has not been attempted at present. Individuals' destinations (their workplaces) are also estimated from UK census.

Once the synthetic population has been initialised, an agent-based model is used to simulate their behaviour. At present, individuals travel once from home to work each day. A simplifying assumption at this stage is that agents travel at a constant speed and in a straight line from home to work, although constraining and allocating the flows to transport networks is planned as a future objective. At each iteration, agents decide whether to go to their home or to their work depending on the simulated time. Agent's sample from normal probability distributions to determine the times that they will start to travel home or to work. Thus the model can be configured by altering the mean and standard deviation of the probability distributions. There are four parameters overall: the mean and standard deviation of the time the agents leave their homes (μ_h and σ_h) and the times that the agents leave their work (μ_w and σ_w). Figure 2 illustrates the distribution of agents in the running model with default (uncalibrated) values for μ_h, σ_h, μ_w and σ_w. Agents travel from within Leeds and the outskirts of the city towards the central business district for work in the morning and then return home again in the afternoon.

Crowd-sourced data will then used for model calibration. A genetic algorithm will be applied in order to search for optimal values for the model parameters and model error will be calculated by comparing the overall times that agents spend at work or at home to the data from Twitter (e.g. the data presented in Figure 1). Hence the research will generate a temporally realistic, spatially-explicit model of individual commuting behaviour, calibrated using novel crowd-sourced data, which will help us to understand the spatio-temporal dynamics of urban areas.

5 Challenges, Conclusions and Ongoing Research

Crowd-sourced data have the potential to revolutionise the ways that social scientists collect data and how simulation models make use of data. However, there are considerable challenges that must be addressed before the data will be truly useful.

Bias

There are potentially insurmountable biases associated with crowd-sourced data. For this research, only a small percentage of all tweets – found to be 1% by Gelernter and Mushegian (2011) – that are sent from mobile devices also have accurate GPS coordinates associated with them. In addition, approximately only a 1% sample of all activity is actually revealed by Twitter without charge. Perhaps more substantially, there are large sections of society that do not use social-networking services and hence will not be captured in the analysis. However, this does not need to be an insurmountable drawback if precautions are taken. By analysing crowd-sourced data with other sources, such as population censuses and geo-demographic products, it will be possible to estimate which sections of society are not captured in the data. Supplementary data can then be used to capture the

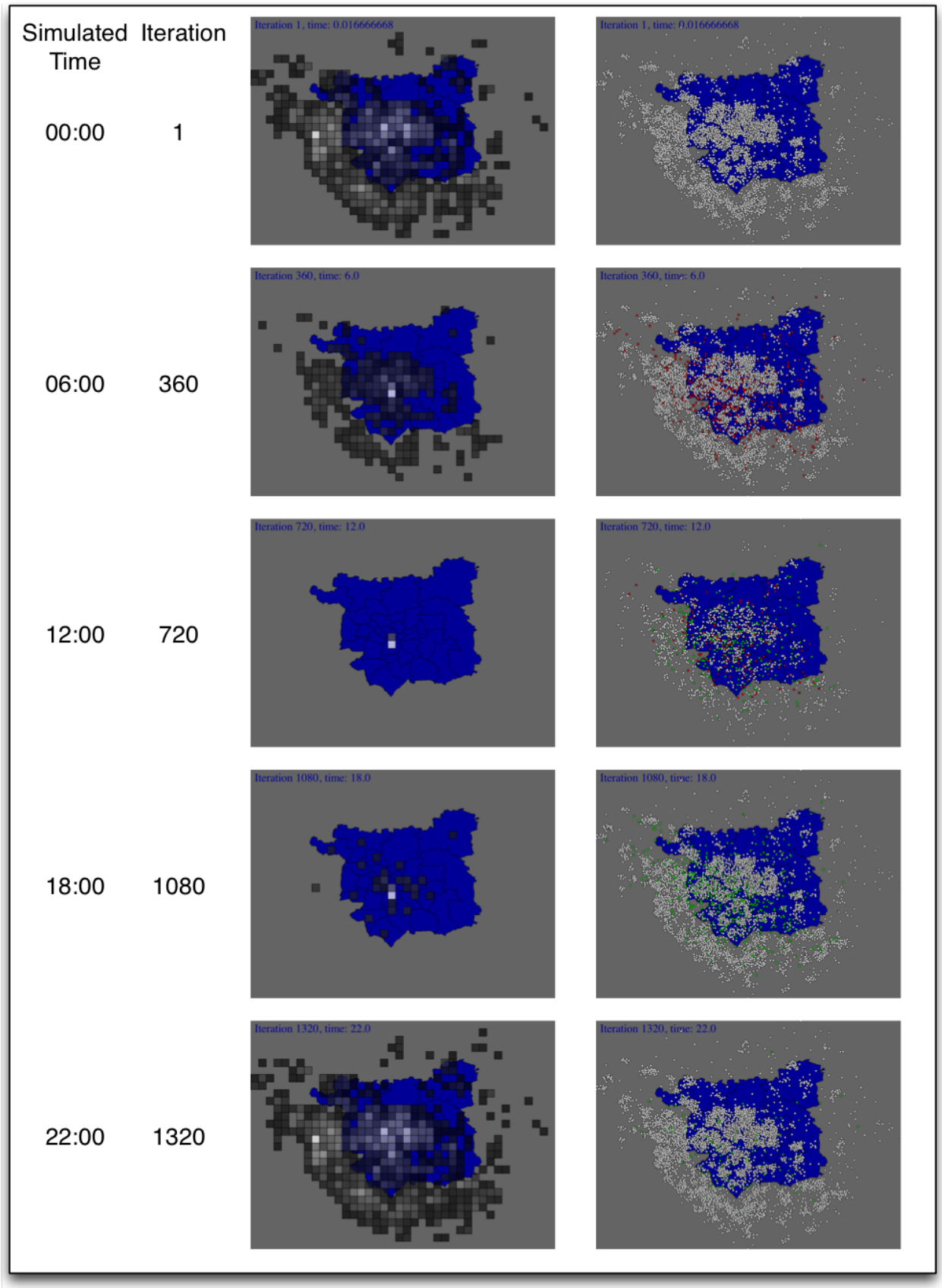

Fig. 2. The prototype model running over a single simulated day (uncalibrated). Agents travel from within Leeds and the outskirts of the city towards the central business district. The left maps show density (white cells are the most densely populated). On the right maps, white points indicate stationary agents, red signifies agents travelling to work and green shows agents travelling home.

behaviour of these missing populations. In essence, the crowd-sourced data can be applied to populations that are well represented – young people visiting night clubs might be an example of a situation where this will work – and replaced with alternative data for phenomena that involve people who do not participate in social networking.

Reliability

Accepting biases in the data, there are still questions over how reliable the process of generating behavioural profiles can be. Although the behavioural estimates appear to be reliable in aggregate (see Figure 1) there are numerous individuals for whom nonsensical behaviour is generated (e.g. individuals who appear to be away from home for 24 hours per day). However, this suggests that an analysis of the locations of twitter messages *in isolation* is insufficient to establish a person's behaviour, rather than a crippling drawback with the approach in general. It can be expected that the coupling of text mining techniques with traditional spatial analysis (e.g."spatio-temporal text mining") will provide a means of more accurately classifying a person's behaviour.

Ethical Considerations

Finally, ethical considerations must play a large part in any future research. The concept of crowd-sourced / volunteered data is relatively new and ethical frameworks for using the data in research are still under-developed. Messages posted on public forums (including Twitter) are inherently accessible to the general public and hence do not fall under the remit of traditional human-subject research. However, it is possible that users are naively unaware of the amount of information they reveal about themselves when they use social networking services. Although early work in this area suggests that users do not need to be asked for consent and university ethics committee consideration need not be sought (Wilkinson and Thelwall, 2011), more work in this area is clearly needed.

Conclusion

This paper has presented a novel effort to classify individual behaviour from the spatial location of twitter messages and use this information to improve the accuracy of an agent-based model of urban dynamics. Although current behaviours are limited to commuting to work, the intention is to extend this to a broader range of behaviour including shopping, socialising, education, etc. (see Yang et al. (2008) for example). Immediate future work will focus on methods to generate more reliable behavioural estimates (e.g. taking the textual content into of messages into account as well as their spatial location) including machine-learning algorithms and necessary modelling improvements such as the incorporation of a transport network. Although there are considerable barriers to the use of these data in earnest, this preliminary research has shown, at least, that there is utility in continuing to use crowd-sourced data. This type of data is only going to become more prolific hereafter so the tools developed now will undoubtedly be useful, after some refinements, in the future.

Bibliography

Batty, M. (2005). Agents, cells, and cities: new representational models for simulating multiscale urban dynamics. *Environment and Planning A 37*, 1373–1394.

Bell, G., T. Hey, and A. Szalay (2009). Beyond the data deluge. *Science 323*, 1297–1298.

Birkin, M., N. Malleson, A. Hudson-Smith, S. Gray, and R. Milton (2011). Calibration of a spatial simulation model with volunteered geographical information. *International Journal of Geographical Information Science 25*(8), 1221–1239.

Birkin, M., A. Turner, and B. Wu (2006). A synthetic demographic model of the UK population: Methods, progress and problems. In *Proceedings of the Second International Conference on e-Social Science*, Manchester, UK.

Cheng, Z., J. Caverlee, and K. Lee (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, pp. 759–768. New York, New York, USA: ACM Press.

Davis Jr, C. A., G. L. Pappa, D. R. R. de Oliveira, and F. de L Arcanjo (2011). Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS 15*(6), 735–751.

Eisenstein, J., B. O'Connor, N. A. Smith, and E. P. Xing (2010). A Latent Variable Model for Geographic Lexical Variation. In *EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics.

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

Gomide, J., A. Veloso, W. Meira, V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira (2011). Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proceedings of the ACM*, Koblenz, Germany, pp. 1–8.

Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal 69*, 211–221.

Harland, K., M. Birkin, A. Heppenstall, and D. Smith (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of microsimulation techniques. *Journal of Artificial Societies and Social Simulation 15*(1).

Johansson, A., M. Batty, K. Hayashi, O. Al Bar, D. Marcozzi, and Z. Memish (2012). Crowd and environmental management during mass gatherings. *The Lancet Infectious Diseases 12*(2), 150–156.

MacKerron, G. (2012). *Happiness and environmental quality*. Ph. D. thesis, The London School of Economics and Political Science, London, WC2A 2AE.

Malleson, N. and M. Birkin (2011). Towards victim-oriented crime modelling in a social science e-infrastructure. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 369*(1949), 3353–3371.

Malleson, N., A. Heppenstall, and L. See (2010). Crime reduction through simulation: An agent-based model of burglary. *Computers, Environment and Urban Systems 34*(3), 236–250.

Matthews, R., N. Gilbert, A. Roach, J. Polhill, and N. Gotts (2007). Agent-based land-use models: a review of applications. *Landscape Ecology 22*(10), 1447–1459.

Russell, M. (2011). *Mining the Social Web*. Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites. O'Reilly Media, Inc.

Savage, M. and R. Burrows (2007). The Coming Crisis of Empirical Sociology. *Sociology 41*(5), 885–899.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman and Hall.

Stefanidis, A., A. Crooks, and J. Radzikowski (2011). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 1–20.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe (2011). Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. *Social Science Computer Review 29*(4), 402–418.

Wilkinson, D. and M. Thelwall (2011). Researching Personal Information on the Public Web: Methods and Ethics. *Social Science Computer Review 29*(4), 387–401.

Yang, Y., P. Atkinson, and D. Ettema (2008). Individual space–time activity-based modelling of infectious disease transmission within a city. *Journal of The Royal Society Interface 5*(24), 759–772.