Applying geographical clustering methods to analyze geo-located open micro-blog posts

Andy Turner[1], Nick Malleson

[1]School of Geography, University of Leeds, LS2 9JT
Tel. +44 (0) 113 343 0779
A.G.D.Turner@leeds.ac.uk, http://www.geog.leeds.ac.uk/people/a.turner/

Summary: In this paper we conduct an exploratory geographical analysis of a sample of post data from the popular micro-blogging service Twitter for the period 22nd June to 12th October 2011 in the city of Leeds. For some user accounts clear patterns of daily activity are observed, and spatio-temporal concentrations of Twitter posts (tweets) are thought likely to represent, among other things, the residential location of users.
KEYWORDS: crowd-source data, open data, twitter, clustering, urban dynamics, simulation

1. Abstract

In this paper we conduct an exploratory geographical analysis of a sample of post data from the popular micro-blogging service Twitter for the period 22nd June to 12th October 2011 in the city of Leeds. For some user accounts clear patterns of daily activity are observed, and spatio-temporal concentrations of Twitter posts (tweets) are thought likely to represent, among other things, the residential location of users.

Preliminary results suggest that the data could be extremely valuable as a means of exploring peoples' daily spatio-temporal behaviour. Geographical cluster detection methods are being coupled with text mining techniques to help identify patterns in the data and to identify what people were doing when they tweet. Further to presenting our analysis, we plan to outline our considerations about using these data in geographical modelling.

2. Introduction

Quantitative studies of social systems demand abundant, high-quality, individual-level data and a modelling approach. Most modelling efforts based on small-scale in-depth interview survey data can be criticised for being biased. Medium scale social surveys like the British Household Panel Survey (BHPS) provide data that are undoubtedly useful for social system modelling, but these are limited and constrained by design to only focus on a small number of behavioural and attitudinal variables. Larger scale registration and demographic census data tend not to include behavioural or attitudinal variables at all and focus more on population and household facts. SimBritain (Ballas et al. 2005) was an attempt to combine the BHPS and UK Census data outputs in a model that dynamically simulated urban and regional populations in Britain. Such a model can form the basis for a more complex behavioural model. Other similar foundations for this have been developed through the MoSeS and GENESIS projects funded by ESRC under the UK e-Science Programme. The resulting models are still lacking behavioural components, but the models are still being developed as part of the JISC funded NeISS project that aims to develop a National e-Infrastructure for Social Simulation making it easier for others to get involved in using and developing them.

Since the late 1990s, commercial lifestyle data has been growing in utility for social systems modelling. Lifestyle databases are known to provide a rich and complimentary source of information, but they are both expensive and difficult to obtain due to their intrinsic and strategic value to the companies that own and develop them.

Although the richness of commercial lifestyle data should not be ignored, the current latest fashion is

to make use of the increasingly open forms of data available on-line and predominantly from social networking sites. These data offer new and exciting opportunities for modelling social systems. Indeed, Savage and Burrows (2007) noted that survey-based enquiries are now potentially superseded by massively "crowd-sourced" databases which are often high-quality, individual-level and regularly/continuously updated.

In the long run, it will be a combination of data from all the aforementioned sources that will be of the highest utility, not only in commercial and academic work, but ultimately also in government planning and tackling big issues that society faces.

3. Data

For this research, crowd-sourced data from the Twitter service are being used. Twitter is a social networking / microblogging service that allows users to broadcast short public messages of up to 140 characters called 'tweets' (Wikipedia: Twitter Article 2011). Twitter provide the Streaming API which gives access to tweet data as they are generated, albeit filtered in a particular manner (Twitter 2011). The data being analysed are those that are geolocated -- they have been assigned spatial coordinates from the GPS (Geographical Position System) enabled device they were sent from. As well as being spatially located, each individual post collected is also temporally located with a time stamp detailed to the resolution of a second. Not all tweets have an associated location, however. For privacy reasons a user must enable the 'Tweeting With Location' feature (which is inactive by default) and it is possible that the device used to create the tweet has a poor idea of its location.

Tweets that can be geolocated to an area surrounding Leeds have been collected for the time period 22nd June - 12th October 2011. This consists of 290,215 individual tweets from 9,223 different users. The parameters chosen and the levels of activity were such that the rate limit of the stream (the amount of data transferred in any short time period) was not exceeded for the period in which the data were collected. Had it been, then there would have been further filtering of the data. It is possible to receive more data from Twitter using multiple accounts and machines and by liaising with the company providing the service directly.

Tweets often include special identifying text called 'tags'. Special tags include the usernames of other users; group names; and, tags are commonly set up and used for events, places and concepts. URLs to web content may also be included in Tweets.

The profile information available about each user is limited. A small amount of profile text is sometimes provided by users, but this is not structured data and it is often not available or it is hard to discern any characteristics of the user from it. Subsequently, it is not straightforward to group all postings of a particular geodemographic group such as students. However, it is possible to map the density of tweets for individual and grouped accounts and those that have (or have not) posted in a particular place and that have (or have not) used a particular tag in a post.

4. Mapping the data

Figure 1 illustrates the density of all tweets in the dataset, generated using a Kernel Density Estimation (KDE) algorithm. As would be expected, areas of high population density such as central Leeds, Bradford and surrounding smaller towns (e.g. Otley, Weatherby, Guiseley) exhibit the highest tweet density. This is encouraging as it suggests that the data could be used to estimate urban populations at different times of day.

Figure 1. The density of all Tweets collected around Leeds using the KDE algorithm.

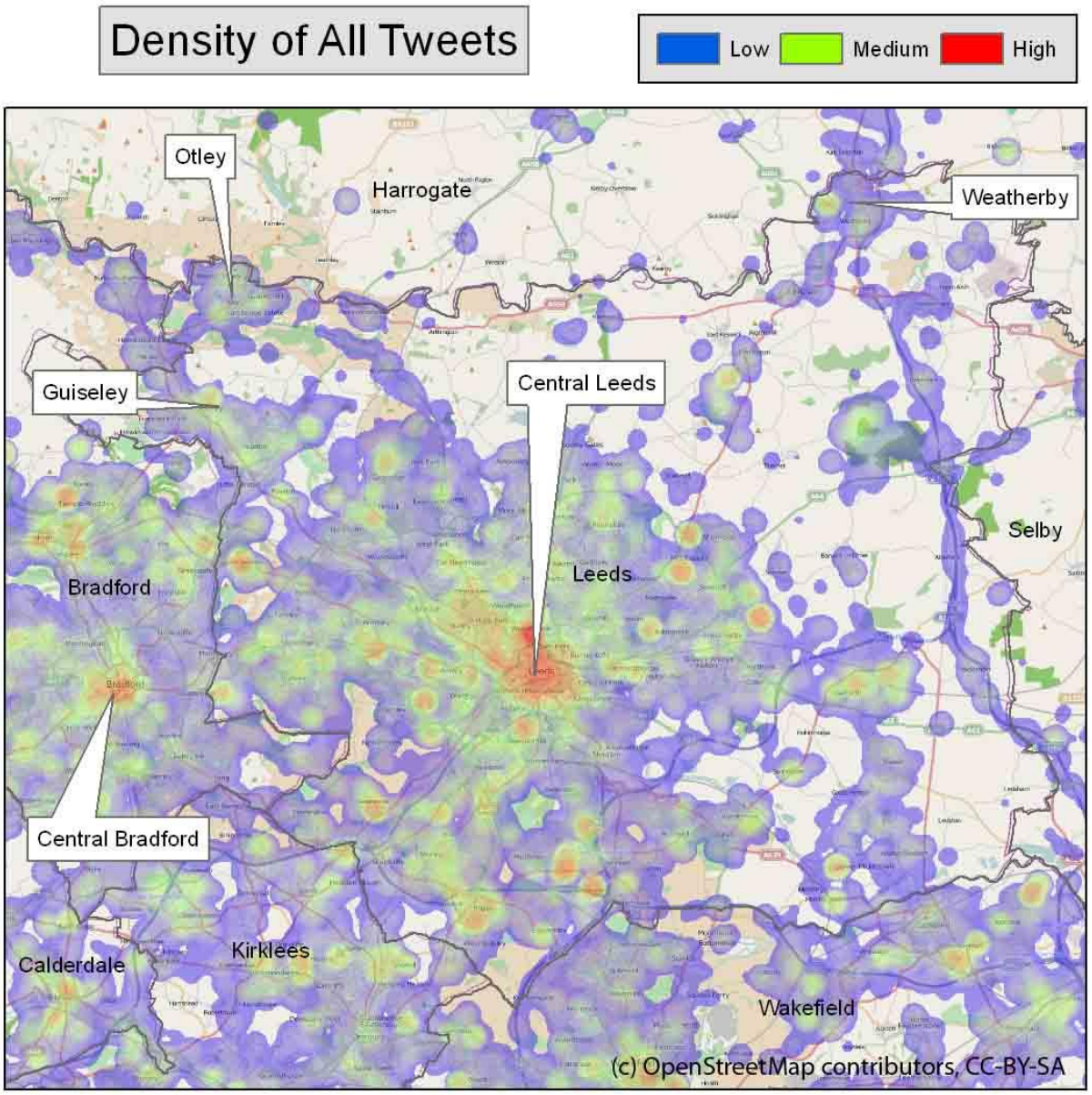

Figure 1. The density of all Tweets collected around Leeds using the KDE algorithm.

It is also possible to disaggregate the data to look for lower-level spatio-temporal patterns. Exploring the densities of tweets for individual users can be quite revealing; they are often concentrated around one or two places and only in rare cases are there many different localised concentrations. Hence it can be assumed that these high density areas represent important 'anchor points' for the users, whether these are their home, work or another location such as a friends house, a cafe, a gymnasium or any other entertainment venue they frequent.

5. Analysis methods

A spatial analysis of the tweet data in isolation provides insufficient information to assign, with confidence, labels to particular locations. There is a demand for an advanced clustering routine to analyse both the textual content of tweet data in combination with their spatio-temporal location.

A modified version of Geographical Analysis Machine (Openshaw et al., 1987) is being developed to identify the location of the single most clustered region in a set of Tweets. This will be coupled with text mining methods to identify, for each individual user, what action they are most likely performing

at the time that they published their tweet. This information will be an invaluable starting point for research that explores the spatio-temporal behaviour of people in a city. Obtaining data to this level of detail is novel for research that does not recruit human participants; traditionally a researcher would have to equip a group of people with a location-tracking device and observe their daily behaviour (e.g. Wiehe et al., 2008). Hence these new data sources have the potential to revolutionise our understanding of social phenomena and our approach to social modelling.

5. References

Openshaw S, Charlton M.E., Wymer C., Craft A. (1987) A Mark 1 Geographical Analysis Machine for the Automated Analysis of Point Data Sets. In the International Journal of Geographical Information Systems, Vol. 1, No. 4, pages 335-358. [online] http://www.informaworld.com/openurl?genre=article&issn=1365-8816&volume=1&issue=4&spage=335 [Accessed on] 2011-01-04.

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

Twitter (2011) The Twitter Streaming API [Online] https://dev.twitter.com/docs/streaming-api [Accessed on] 2011-12-02

Wiehe, S. E., A. E. Carroll, G. C. Liu, K. L. Haberkorn, S. C. Hoch, J. S. Wilson, and J. D. Fortenberry (2008). Using GPS-enabled cell phones to track the travel patterns of adolescents. International Journal of Health Geographics 7(1).

Wikipedia Twitter Article (2011) [Online] http://en.wikipedia.org/wiki/Twitter [Accessed on] 2011-12-01.

6. Biography

Andy Turner

Mr. Turner's primary research interest is in computational geography to do with developing and using geographical analysis and modelling software.

Nick Malleson

Dr. Malleson's primary research interest is in developing spatial models of social phenomena with a particular focus on crime simulation.