School of Geography
FACULTY OF ENVIRONMENT

UNIVERSITY OF LEEDS

Applying geographical clustering methods and identifying individual behaviour with geo-located open micro-blog posts

Nick Malleson and Andy Turner
School of Geography, University of Leeds

Outline

UNIVERSITY OF LEEDS

Introduction / Motivation

The Data and Study Area

Clustering Methods

Analysis of Individual Behaviour

Creating Behavioural Profiles

Conclusions and Future Work

Introduction

UNIVERSITY OF LEEDS

A copy of these presentation slides and our accompanying paper are available online via Andy's conference notes at the following URL:

http://bit.ly/HWOe27

Crowd-Sourced Data for Social Science

UNIVERSITY OF LEEDS

"Crisis" in empirical sociology (Savage and Burrows, 2007)

- Traditional surveys are small and occur infrequently
- Often focus on population attributes rather than behaviour
- Often spatially / demographically aggregated
- http://www.guardian.co.uk/p/33p85

Surveys are being superseded by massive, "crowd-sourced" data

- "*knowing capitalism*" (Thrift, 2005)
- Amazon.com purchasing suggestions
- Supermarket reward cards
- Strong spatial dimension (Goodchild, 2007) e.g. OpenStreetMap – Volunteered Geographic Information

But academia is slow to take advantage

Potential Uses:

- Data could be used to calibrate models *in situ (*e.g. meteorology models use daily weather data)
- …

Cr
So

"Cris

-
-
-
-

The

-
-
-
-

But

Data

- e

Datamining Twitter | Technology | guardian.co.uk

Datamining Twitter | Technolog...

www.guardian.co.uk/technology/2011/dec/05/monday-note-twitter

Google

Datamining Twitter

Making sense of the Twitter noise is about to get easier

Frédéric Filloux
guardian.co.uk, Monday 5 December 2011 10.36 GMT
Article history

Twitter on a smartphone. Social network intelligence is poised to become a big business. Photograph: Jonathan Hordle/Rex Features

On its own, Twitter builds an image for companies; very few are aware of this fact. When a big surprise happens, it is too late: a corporation suddenly sees a facet of its business – most often a looming or developing crisis – flare up on Twitter. As always when a corporation is involved, there is money to be made by converting the problem into an opportunity: Social network intelligence is poised to become a big business.

Cr
So

"Cris

-
-
-
-

The

-
-
-
-

But

Data

- e

Datamining Twitter | Technology | guardian.co.uk

Datamining Twitter | Technolog...

www.guardian.co.uk/technology/2011/dec/05/monday-note-twitter

Google

Monday Note

Datamining Twitter

Making sense of the Twitter noise is about to get easier

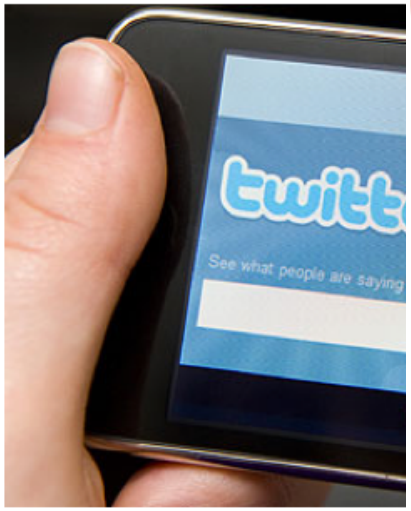
Tweet 188

Recommend 19

reddit this

Comments (4)

Frédéric Filloux
guardian.co.uk, Monday 5 December 2011 10.36 GMT
Article history



Twitter on a smartphone. Social network i
business. Photograph: Jonathan Hordle/R

> I have recorded literally everything over the last few months about people checking in to Starbucks. They don't need to say they're in Starbucks, they can just be inside a location that is Starbucks, it may be people allowing Twitter to record where their geolocation is. So, I can tell you the average age of people who check into Starbucks in the UK. Companies can come along and say: "I am a retail chain, if I supply you with the geodata of where all my stores are, tell me what people are saying when they're near it, or in it." Some stores don't get a huge number of check-ins, but on aggregate over a month it's very rare you can't get a good sampling.

On its own, Twitter builds an image for companies; very few are aware of this fact. When a big surprise happens, it is too late: a corporation suddenly sees a facet of its business – most often a looming or developing crisis – flare up on Twitter. As always when a corporation is involved, there is money to be made by converting the problem into an opportunity: Social network intelligence is poised to become a big business.

On Technology

Most viewed | Zeitgeist | Latest

Monday Note

Related

30 May 2011
Trifling Twitter

19 Jun 2009
The revolution will not be

guardianbookshop

This week's bestsellers

1. Simply Computing for Seniors
by Linda Clark
£10.99

Motivation

UNIVERSITY OF LEEDS

A better understanding of urban dynamics through the use of novel social-network data

Calibration / validation of individual-level model

Method

- Determine what a person is doing / talking about
- Develop advanced spatio-temporal (and textual?) clustering tools

Data

- ~1.2M geo-located tweets in the Leeds area

Data and Study Area

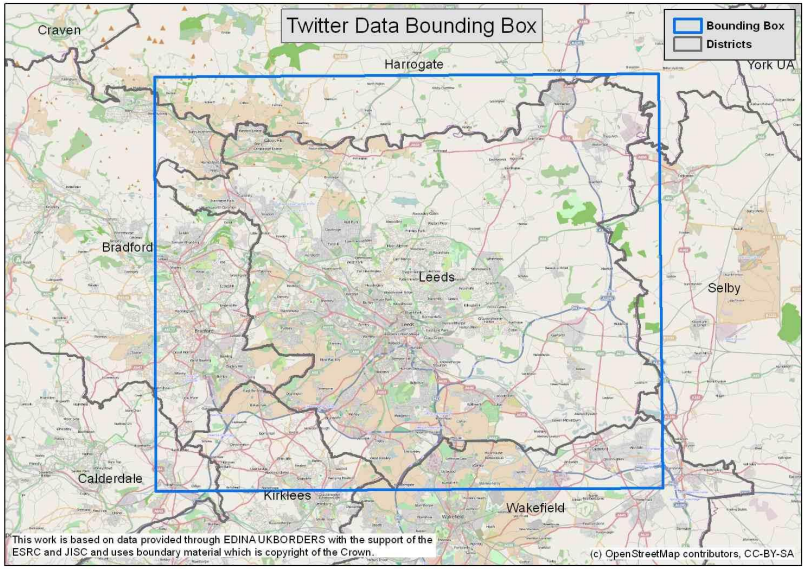
UNIVERSITY OF LEEDS

Twitter

- Social networking / microblogging service
- Users create public ‘tweets’ of up to 140 characters
- For the most part, tweets are publicly available
- Include information about user, time/date, location, text etc.
- ‘Streaming API’ provides real-time access to tweets

Collected Data

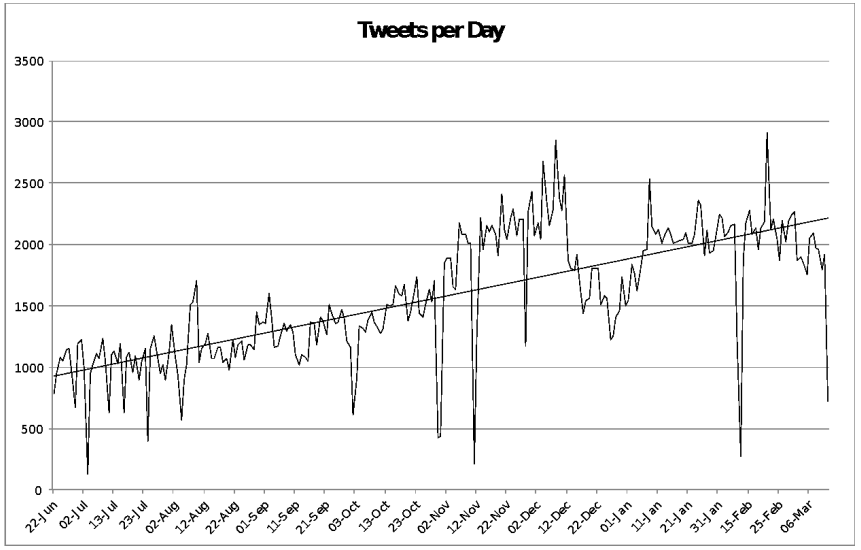
- 1.2M *geo-located* tweets around Leeds (June 2011 – March 2012).
- 403,922 Tweets within district
- 2,683 individual users
- Highly Skewed (10% of all tweets from 8 most prolific users)
- Filtered non-people

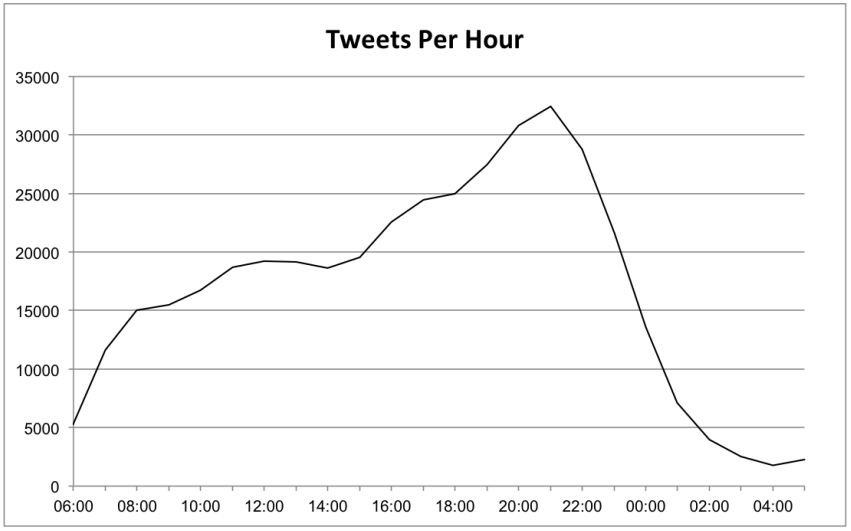


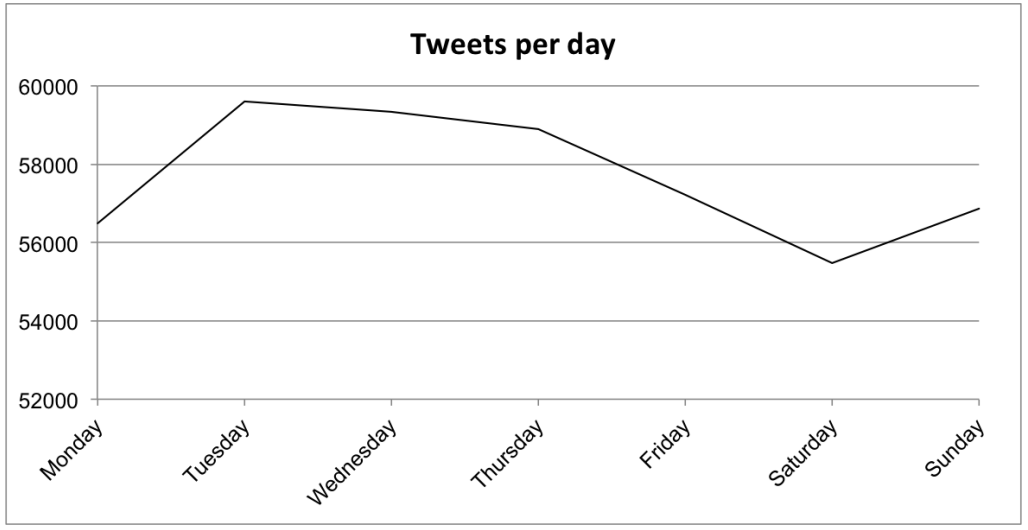
A-Spatial Temporal Trends

UNIVERSITY OF LEEDS

- Hourly peak in activity at 10pm
- Daily peak on Tuesday - Thursday
- General increase in activity over time







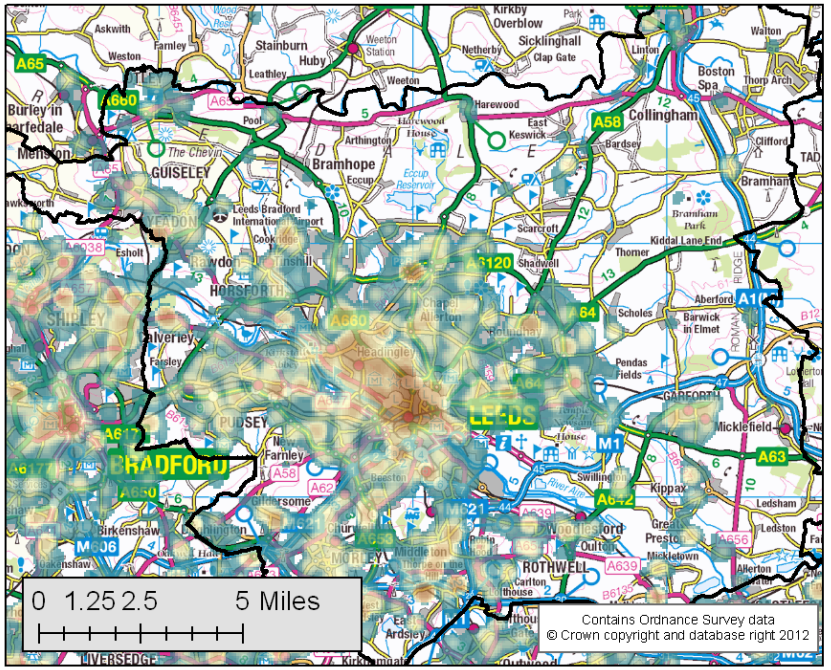
Spatial Overview

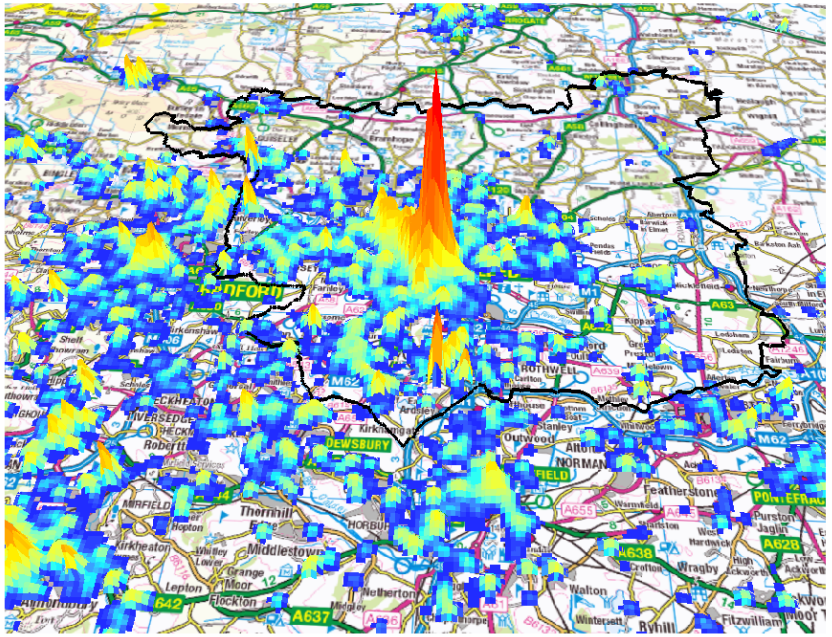
Point density appears to cluster around urban centres.

Also able to distinguish roads in non-urban areas

General pattern somewhat distorted by locations of prolific users

Density of All Tweets





Geographical Clustering

UNIVERSITY OF LEEDS

Spatial

Might also be temporal

Takes into account two variables

Essentially where one variables is unusually high, given the value of the other variable (which is assumed to have a positively correlated distribution of values)

Geographical Clustering

UNIVERSITY OF LEEDS

"The simplest way of defining a cluster is as a localised excess incidence rate that is unusual in that there is more of some variable than might be expected. Examples would include: a local excess disease rate, a crime hot spot, an unemployment black spot, unusually high positive residuals from a model, the distribution of a plant or surging glaciers or earthquake epicentres, pattern of fraud etc. [...] Pattern detection via the identification of clusters is a very simple and generic form of geographical analysis that has many applications in many different contexts..." Openshaw and Turton (1998)

http://web.archive.org/web/20040316070705/http://www.ccg.leeds.ac.uk/smart/gam/gam3.html

Geographical Concentration

A special sort of Geographical Clustering where the expected incidence or the variable for comparison (denominator for the rate) is evenly distributed

Geographical Clustering of Twitter posts

There are some prolific posters and the locations at which they post are concentrated

Where does the pattern of posting density differ most in both absolute and relative terms for different types of posting?

A first look at posts for different times of day

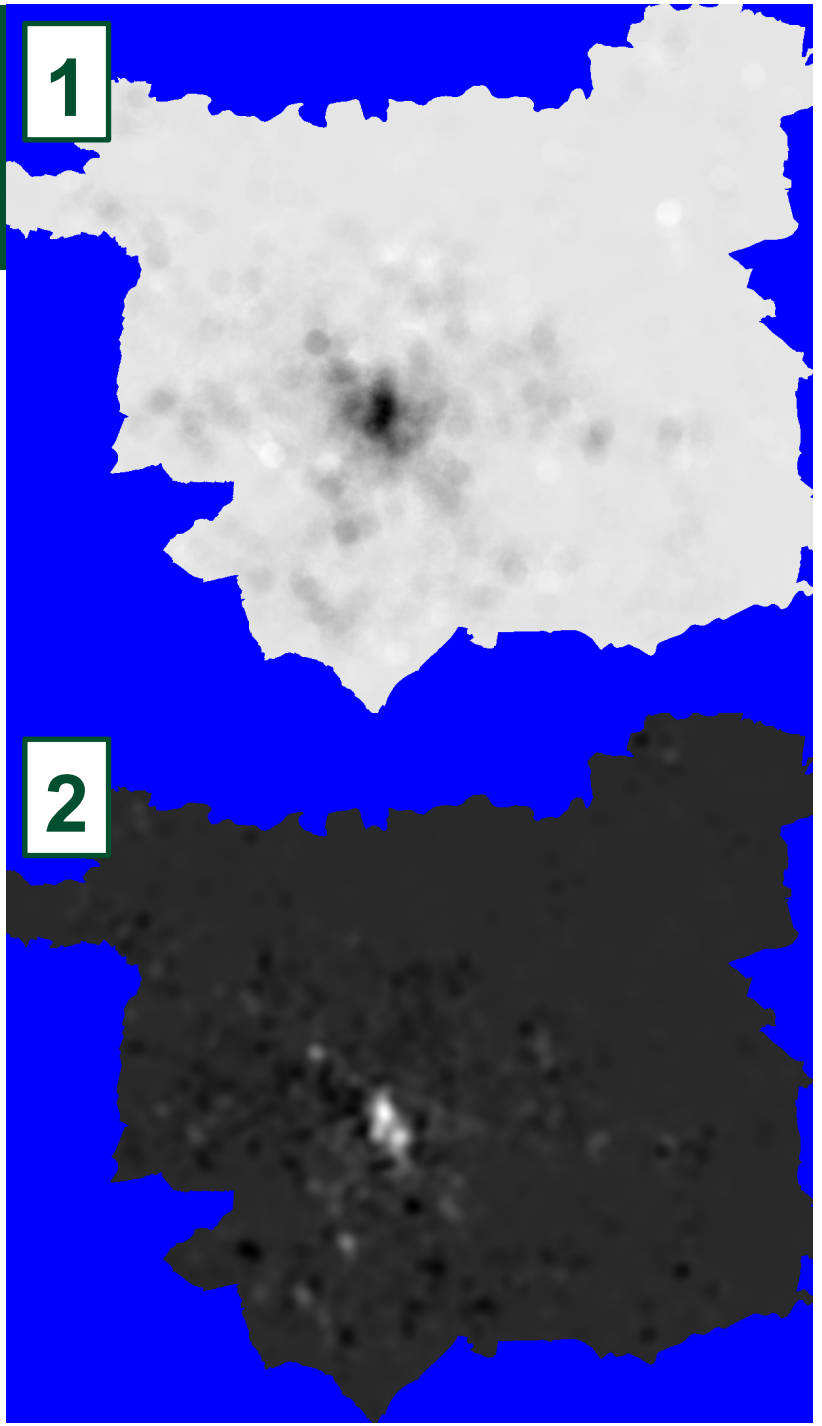
Clustering Methods (I)

Geographical clustering

Geographical concentration

Geographical Clustering of Twitter posts

1. absolute difference between *weekend and weekday* posts
2. absolute difference between *afternoon and evening* posts



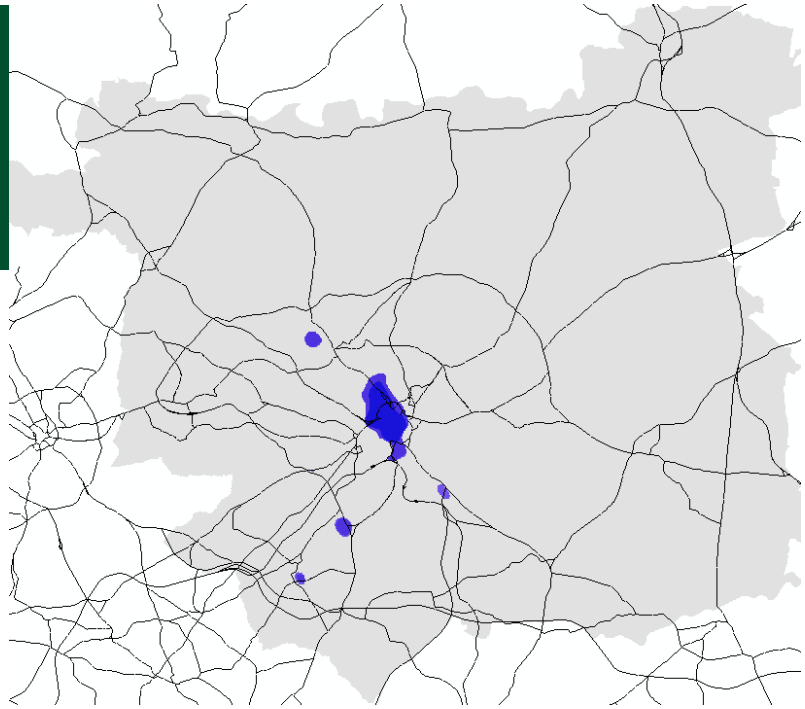
Clustering Methods (II)

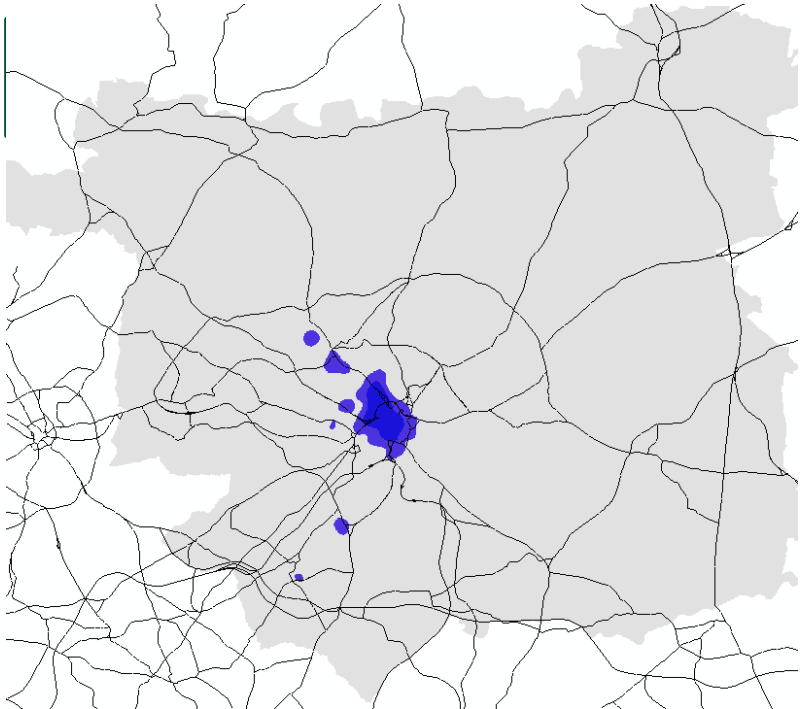
Geographical Clustering of Twitter posts

- identify places with many more postings in the evening than in the afternoon
- relative difference and mask off where the absolute difference is above some threshold filter value
- absolute difference and mask off to only show where relative difference is high

Further work

- some of the most identifiable concentrations and clusterings are a result of a single prolific use
- being able to identify them is potentially useful...





Clustering collaboration

UNIVERSITY OF LEEDS

Some work on Geographical Clustering was revisited recently (Turton and Turner, 2011)

https://docs.google.com/document/d/1rX9XHhAittUF4aMFBKfzqYXkWqrCPK1mdN8bcpT_gO4/edit

We aim to add the methods used to produce the figures shown here as part of the Spatial Cluster Detection Tool made available via:

http://code.google.com/p/spatial-cluster-detection/

Analysis of Individual Behaviour (I)

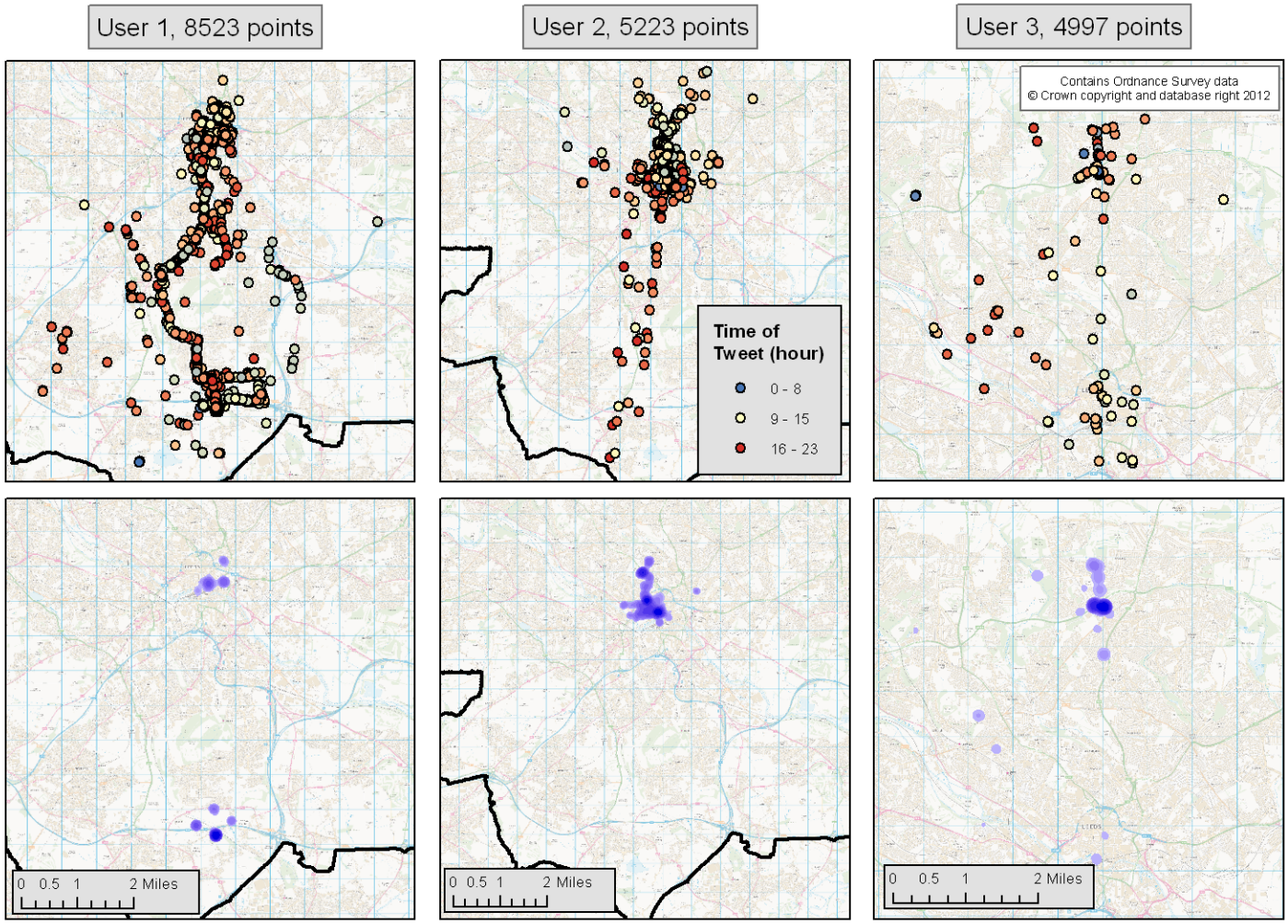
UNIVERSITY OF LEEDS

As well as analysing aggregate patterns, we can try to identify the behaviour of individual users

Some clear spatio-temporal behaviour (e.g. commuting, socialising etc.).

Estimate ‘home’ and then calculate distance at different times

- Journey to work?



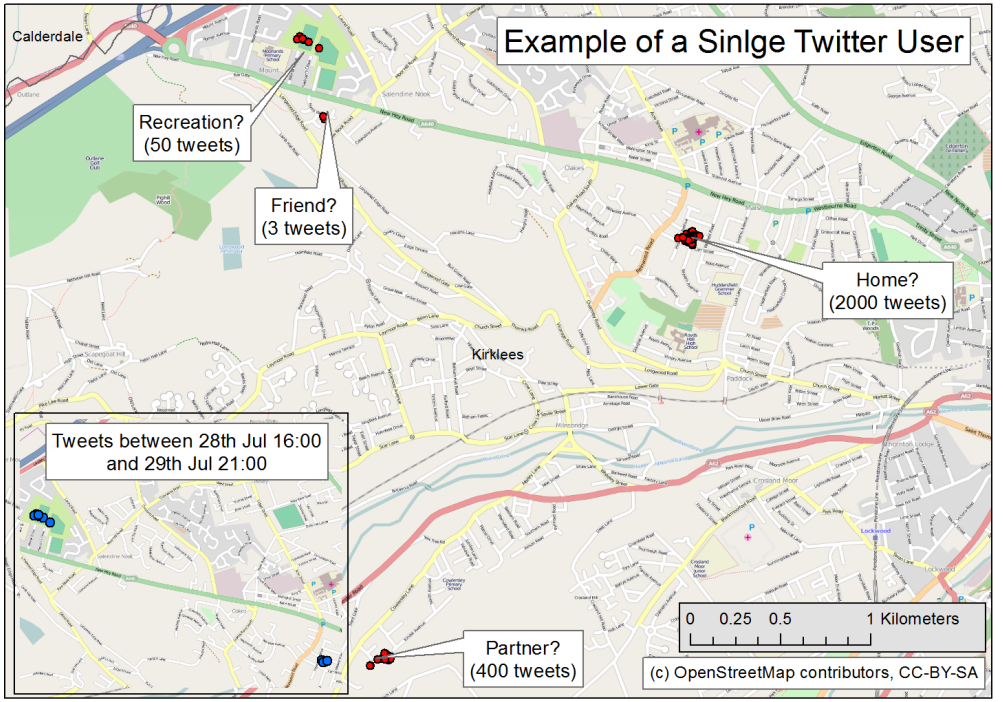
Analysis of Individual Behaviour (II)

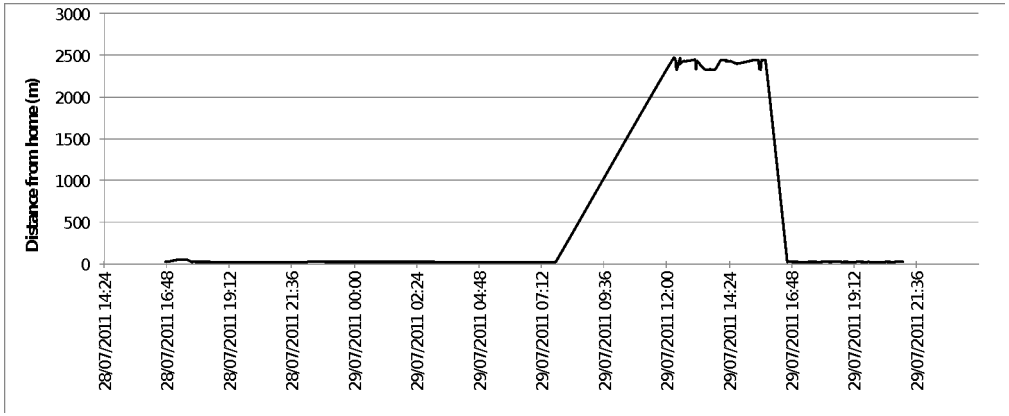
UNIVERSITY OF LEEDS

Possible to construct profiles of individual behaviour at different times of day?

Could estimate journey times, means of travel etc.

Very useful for calibration of an individual-level model





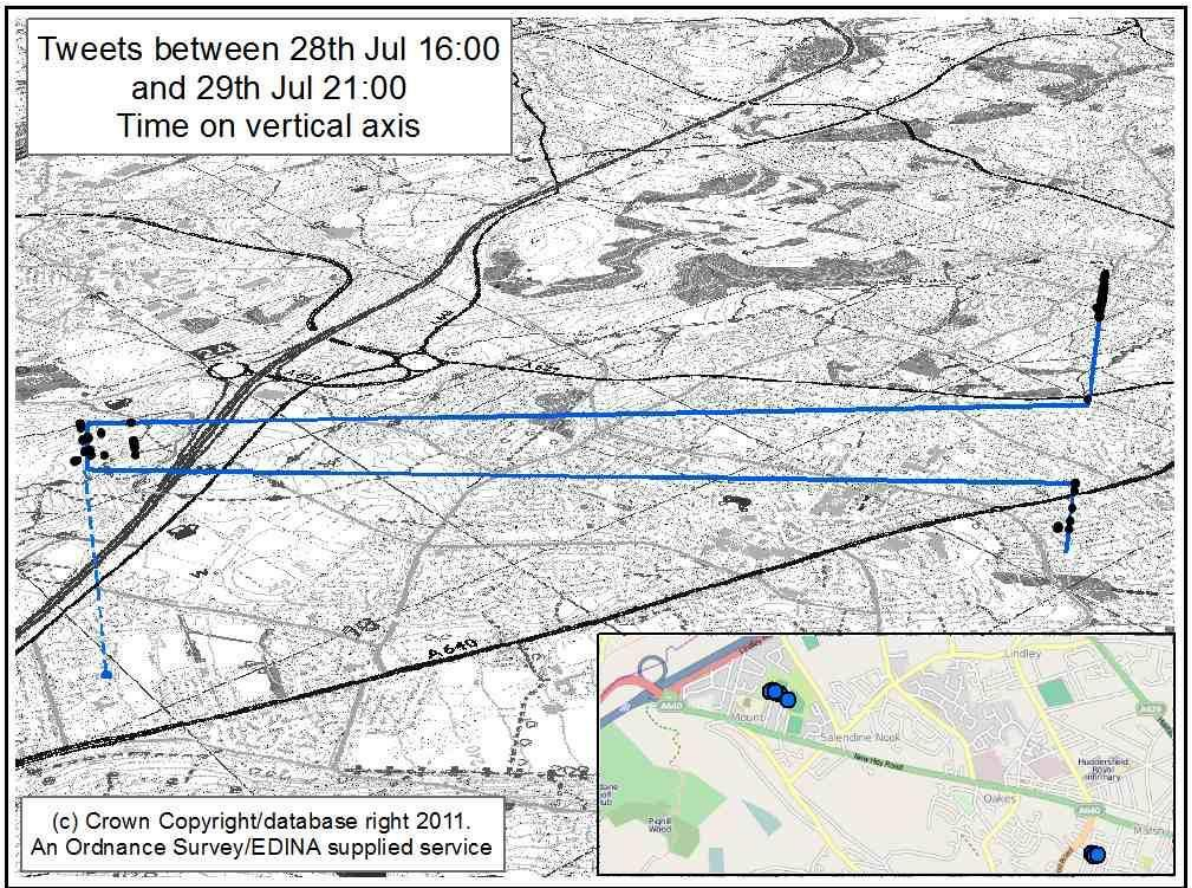
Spatial Behaviour: the Space-Time Prism

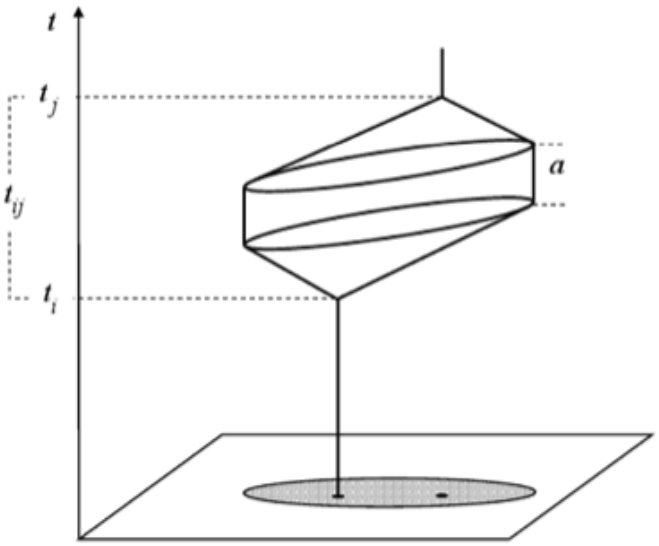
UNIVERSITY OF LEEDS

Visualise data in 3D

Clear representation of a ‘space-time path’ (Hägerstrand, 1970)

Test time geography concepts over an entire city?





Source: Miller, H. J. (2004). Activities in Space and Time. In P. Stopher, K. Button, K. Haynes, and D. Hensher (Eds.), *Handbook of Transport 5: Transport Geography and Spatial Systems*. Pergamon/Elsevier Science.

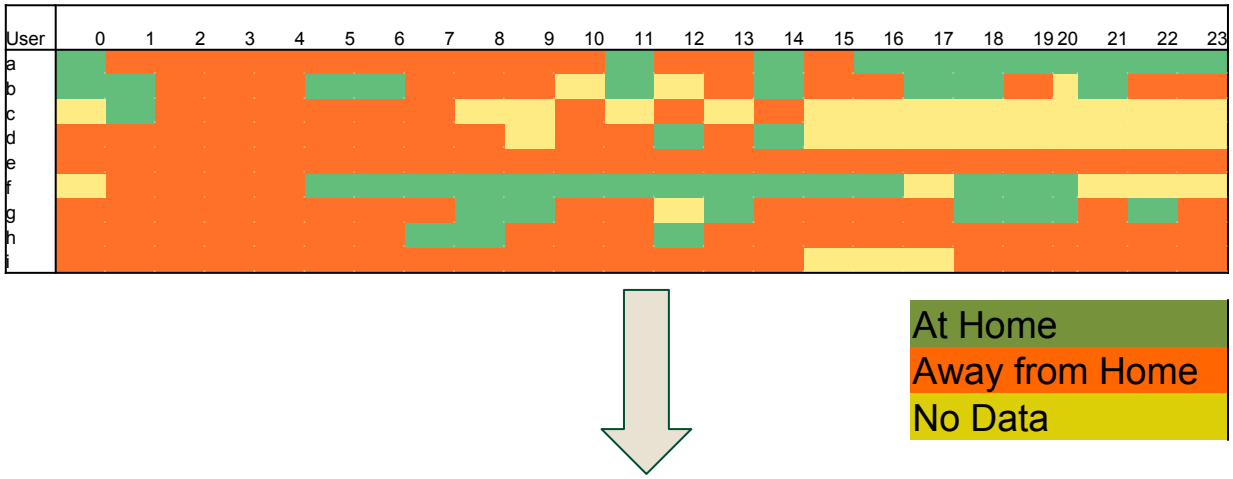
Activity Matrices (I)

UNIVERSITY OF LEEDS

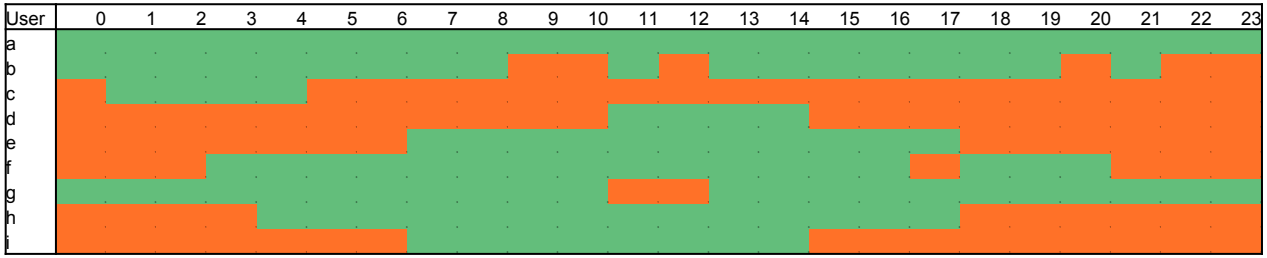
Once the ‘home’ location has been estimated, it is possible to build a profile of each user’s daily activity

The most common behaviour at a given time period takes precedence

‘Raw’ behavioural profiles



Interpolating to remove no-data



Activity Matrices (II)

UNIVERSITY OF LEEDS

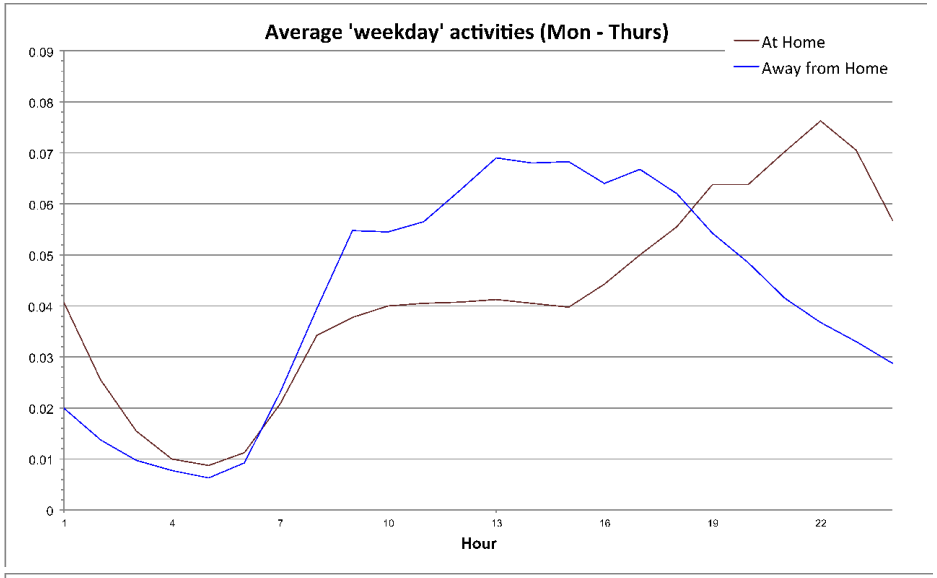
Overall, activity matrices appear reasonably realistic

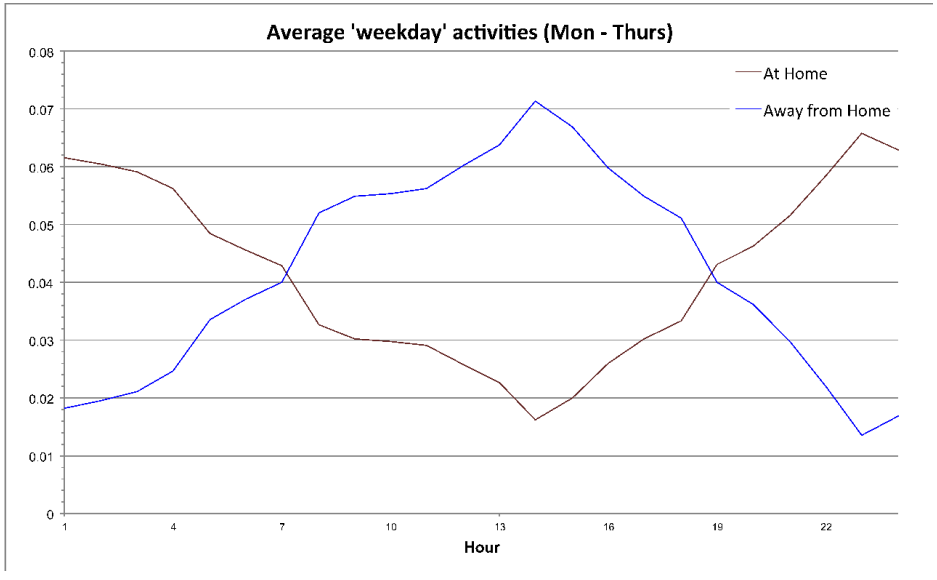
- Peak in *away from home* at ~2pm
- Peak in *at home* activity at ~10pm.

Overall, activity matrices appear reasonably realistic

Next stages:

- Develop a more intelligent interpolation algorithm (borrow from GIS?)
- *Spatio-temporal text mining* routines to use textual content to improve behaviour classification





Towards an Agent-Based Model of Urban Dynamics

UNIVERSITY OF LEEDS

1 – Generate synthetic population

- Previous research has created spatially referenced synthetic population for the study area using census data
- Richly specified attributes including gender, ethnicity, marital status, employment, etc

2 – Create agents

- Rules to determine behaviour can be parameterised from individual characteristics (e.g. employment, home location etc).

3 – Calibrate with crowd-sourced data

- Identify which agents have similar characteristics to users in the data
- Calibrate agent behaviours to match data

Conclusions

UNIVERSITY OF LEEDS

Aim: A better understanding of urban dynamics through the use of novel social-network data

Patterns in the data have a distinct resonance with theoretical concepts

- Can detect intra-urban movement patterns
- This level of data usually very difficult to obtain

Main problem: data bias

- Who isn't tweeting?

Ethical issues

- Usually participants give permission for the study – not feasible with crowd-sourced data
- Data is publicly available but do users understand what they have made available?

Future Work

UNIVERSITY OF LEEDS

More advanced clustering methods

A great deal more can be done to look at the spatial patterns in the the twitter data using clustering methods
We are only really scratching the surface as we learn about these data
There are many challenges to analysing this data, not least is the fact that some of the most identifiable concentrations and clusterings are a result of a single prolific use
Still being able to identify them is potentially useful...

Improved identification of behviour

“Spatio-temporal text mining”

- Methods to classify text based on spatio-temporal location as well as textual content

In situ model calibration

References

UNIVERSITY OF LEEDS

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

Miller, H. J. (2004). Activities in Space and Time. In P. Stopher, K. Button, K. Haynes, and D. Hensher (Eds.), Handbook of Transport 5: Transport Geography and Spatial Systems. Pergamon/Elsevier Science.

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

Thrift, N. (2005) Knowing Capitalism. London: Sage.

Thank you

UNIVERSITY OF LEEDS

A copy of these presentation slides and our accompanying paper will be available online via Andy's conference notes at the following URL:

- http://bit.ly/HWOe27

More information

- Andy Turner
  - http://www.geog.leeds.ac.uk/people/a.turner/
- Nick Malleson
  - http://www.geog.leeds.ac.uk/people/n.malleson/
  - Blog: http://nickmalleson.co.uk/blog/