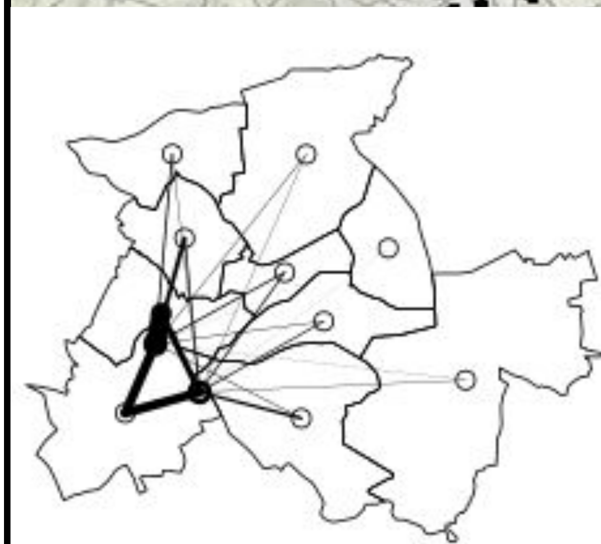
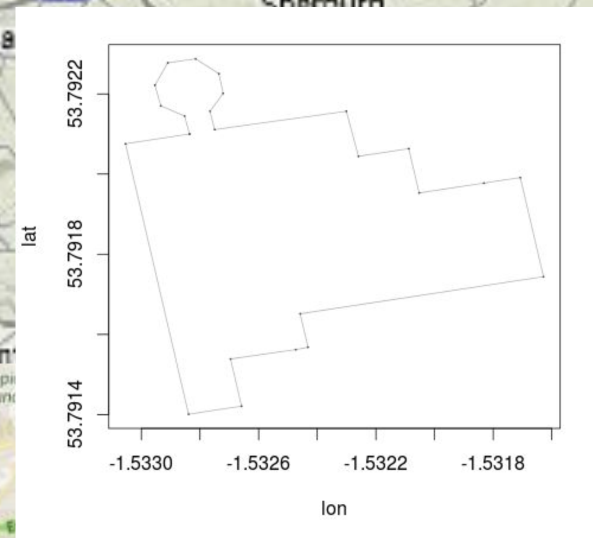
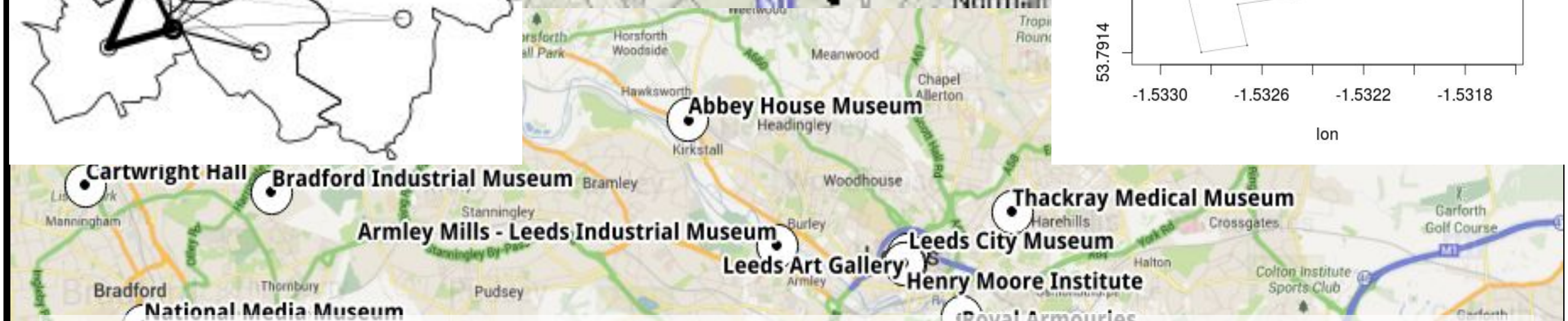
Can social media data be useful in spatial modelling?

A case study of 'museum Tweets' and visitor flows







Robin Lovelace, University of Leeds

17th April, GISRUK2014

Slides Available: robinlovelace.net

NCRM National Centre for Research Methods

UNIVERSITY OF LEEDS

What I’m going to talk about

1. Introduction + perspective
2. The case study + data
3. The method + results
4. Discussion

Part I: Spoiler + perspective

Yes, obviously they can...

- The REAL question is this:
- In which cases do the benefits outweigh the costs?
- Just because we **can** do something, does not mean we should.
- "You should decide whether we need to be doing this." (Ed Snowden, 2013)

The **costs** of using VGI

- Potential for distraction
- Reduced policy relevance?
- Naval gazing
- High complexity -> time pre-processing
- "It's never enough" attitude - constant
- Loudest voice heard clearest
- Unrepresentative
- You need big computers -> inaccessible

The **benefits** of VGI

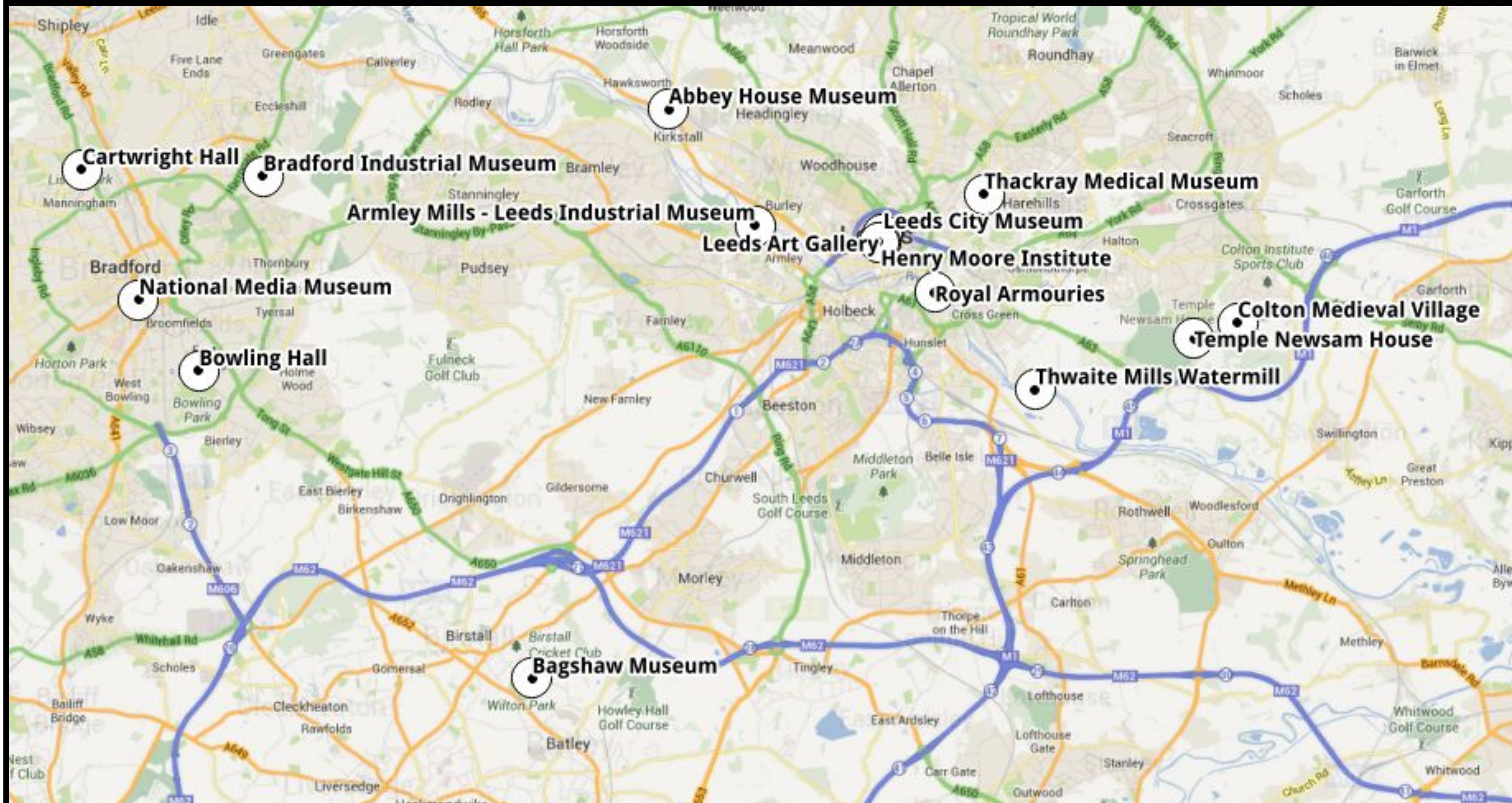
Social media data are a type of "volunteered geographic information" (VGI) (Goodchild 2007). VGI offers:

- New datasources on questions previously beyond the reach of survey
- **Constant and ever-increasing** flow of information
- Diversity, low cost, comprehensive coverage

Paper's purpose: explore these costs and benefits for modelling spatial behaviour

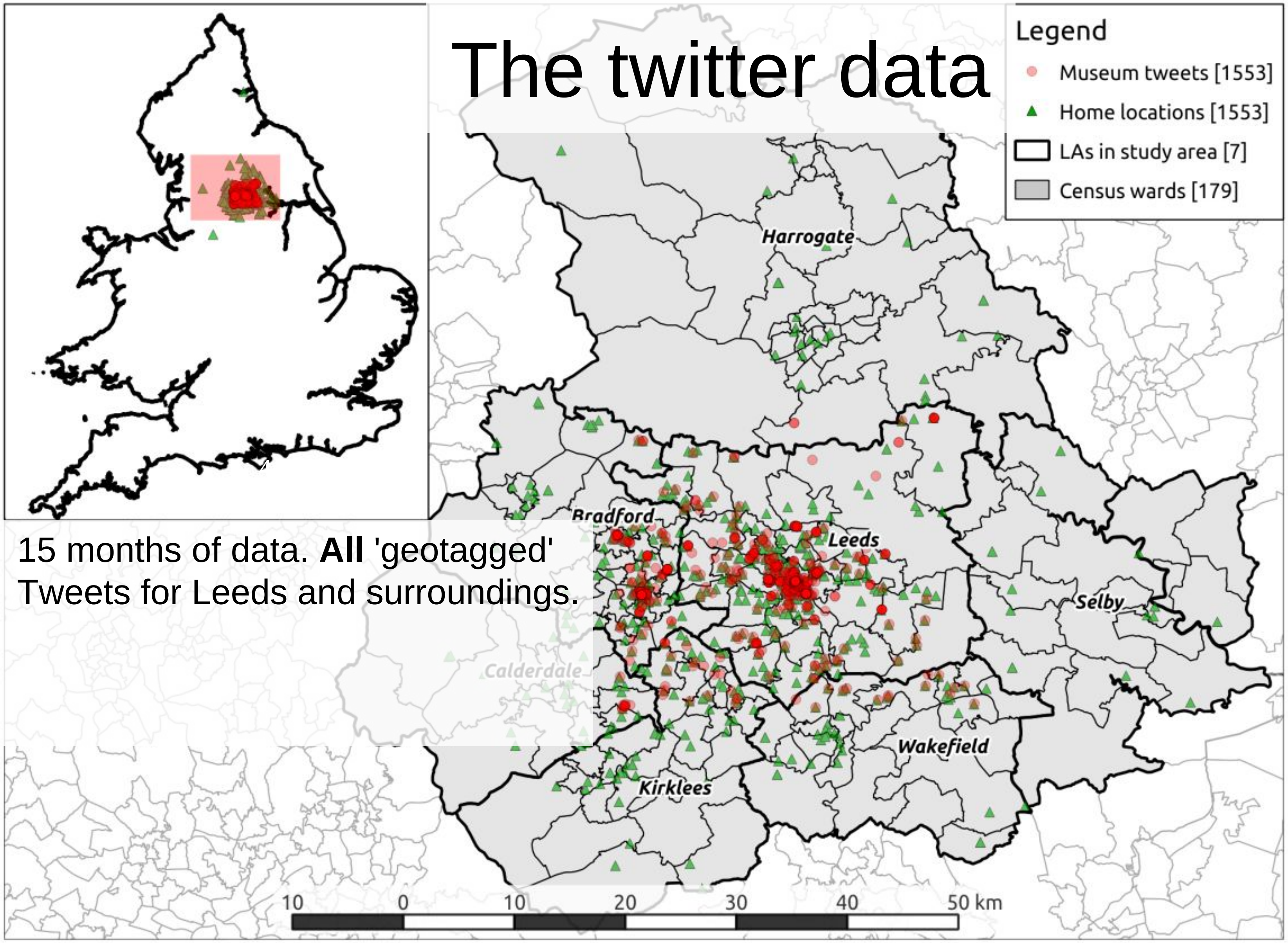
Part II: The case study

We decided to look at museums: not much official data, often 'tweeted' about



15 museums in case study area (west Yorks). OSM dataset with 'museum' tags

The twitter data

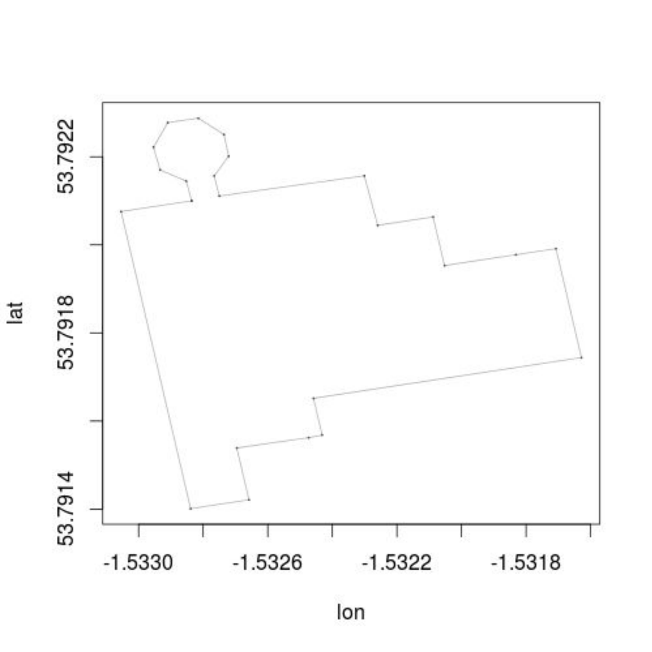


15 months of data. **All** 'geotagged' Tweets for Leeds and surroundings.

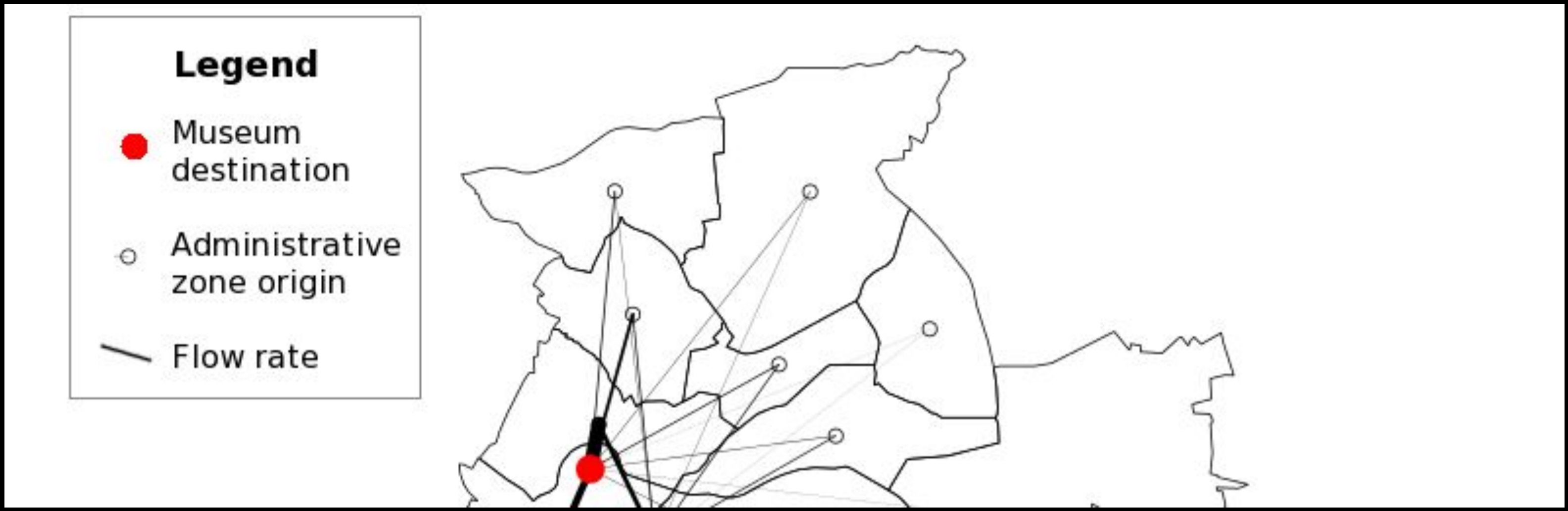
Filtering the tweets

- **Semantic filters**
- Basically "regex"
- Search terms
- Overall just under 1,000 'museum Tweets' resulted from filters

- **Spatial filters**
- A buffer around each museum with osmar



Part III: The model and results



In R code:

```
for(i in 1:nrow(w)){
  for(j in 1:nrow(m)){
    S[i,j] <- inc * P[i] * W[j] * exp(-beta * D[i,j])
  }
}
```

```
D <- gDistance(m, pops, byid=T)/1000
inc <- 0.1
beta <- 0.3
P <- pops$totpop # zone population
W <- A <- rep(1, times=nrow(m))
S <- D^0
```

In maths:

$$T_{ij} = Inc_i P_i W_j \exp(-\beta \, d_{ij})$$

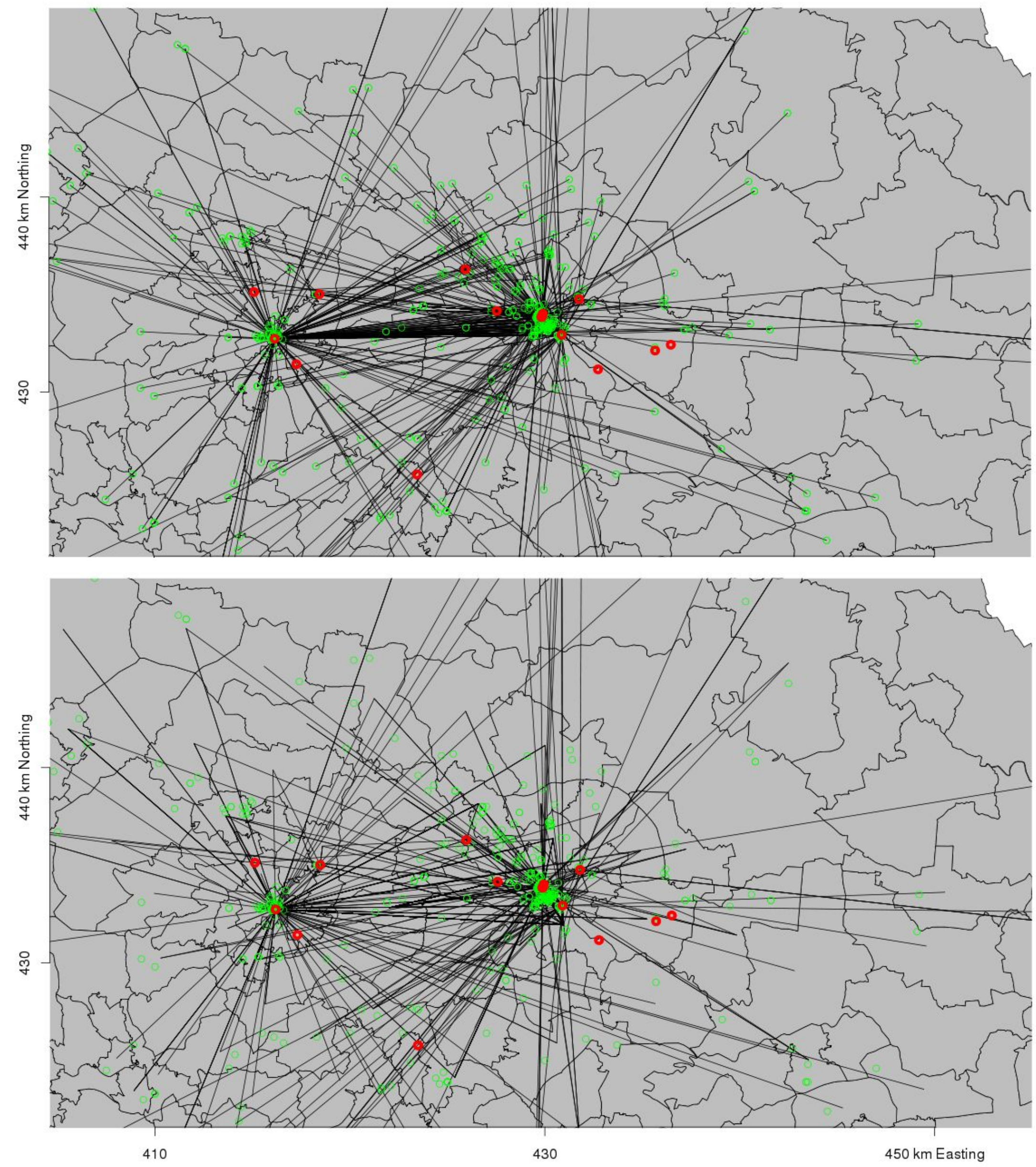
Inc: income proxy
P: population
W: museum attractivenes
beta: dist. decay constant
d: Euclidean distance
i, j: Origins and destinations

Aggregation

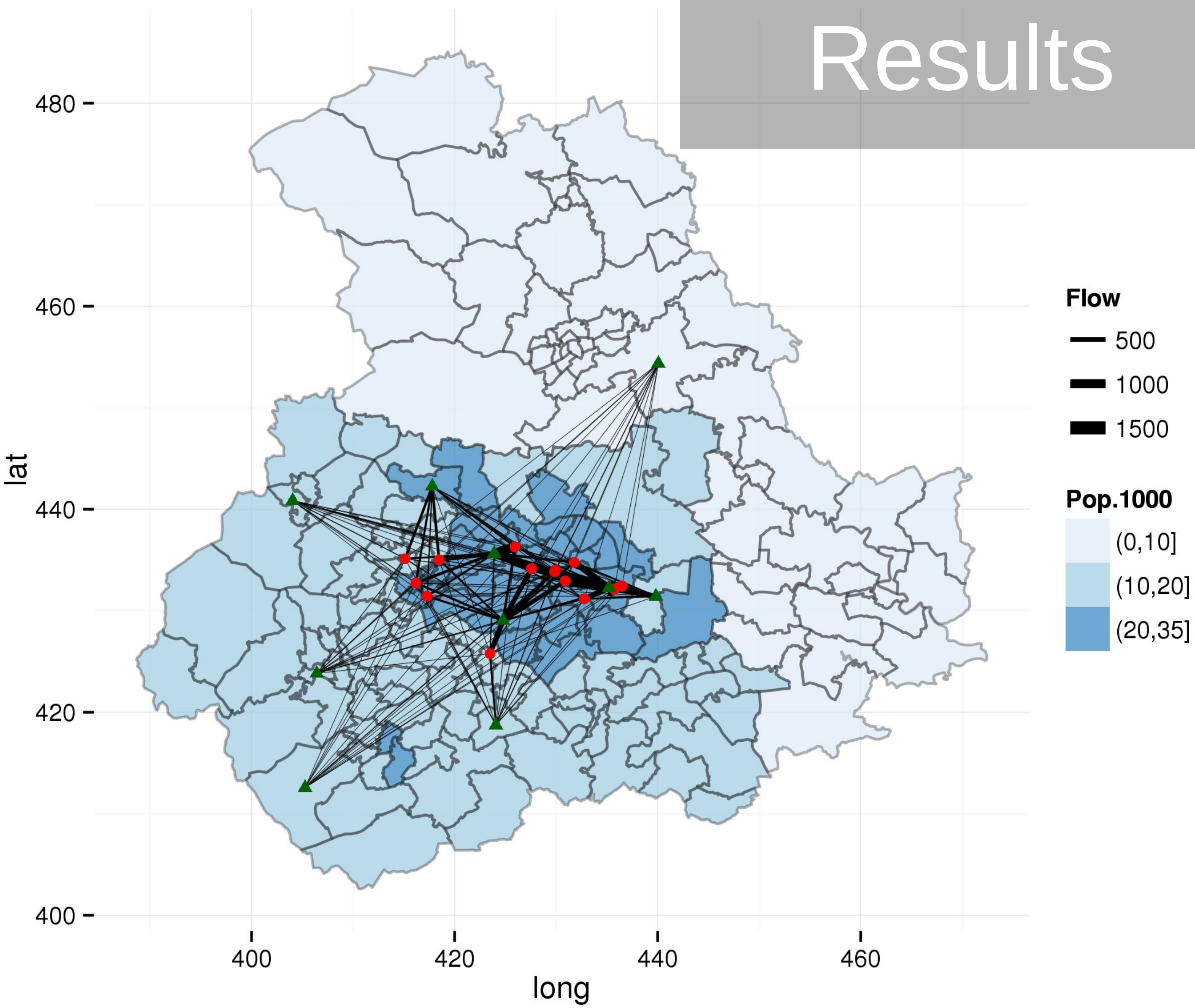
Necessary to compare aggregate flow model with individual Tweets

Also vital to 'smooth' the stochasticity inherent to VGI

In reality: LOTS more data needed for reliable results

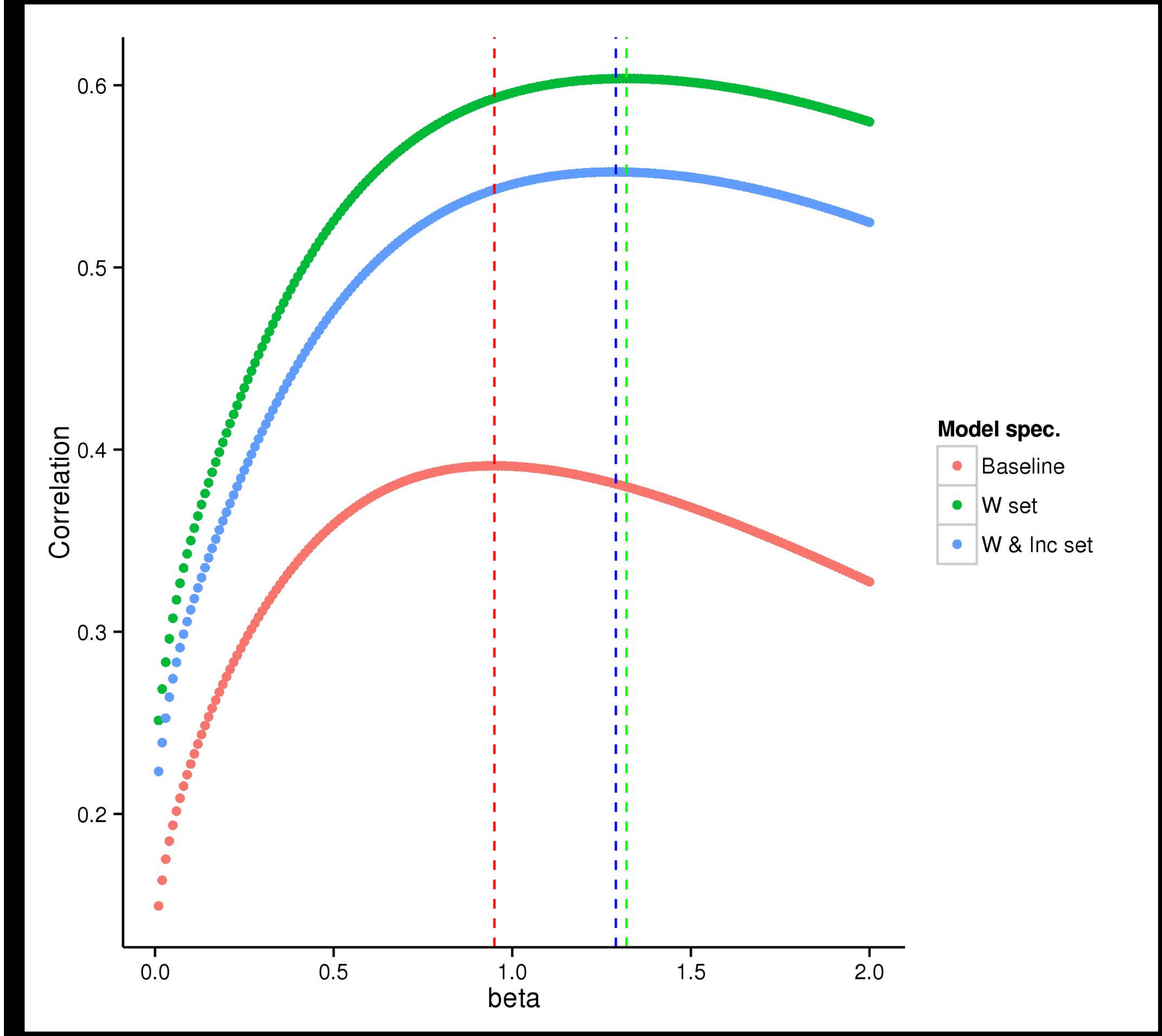


Results



480
460
440
420
400
lat
400
420
440
460
long
Flow
500
1000
1500
Pop.1000
(0,10]
(10,20]
(20,35]

Calibration



Very simple calibration procedure: reran model for many different beta values

Closest aggregated tweet/model fit selected for different model implementations

Opportunities for Bayesian approaches here

Part IV Discussion

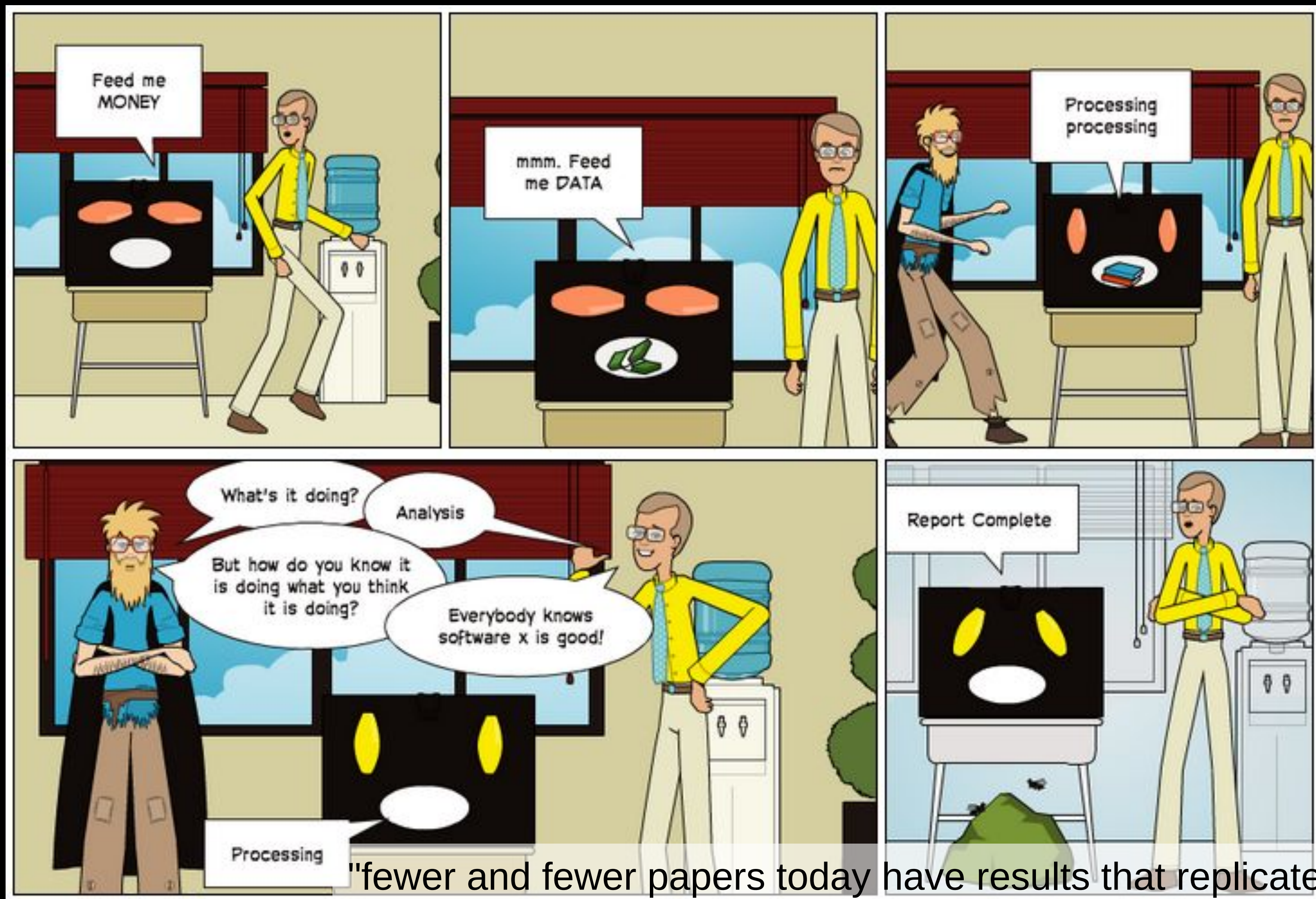
- Results in themselves not massively interesting
- **Large methodological implications**
  - New ways to corroborate theoretical models
  - Some reproducible code for using geo Tweets
- **Ethical issues raised**
  - Who created the dataset? Who **owns** it? Who will **benefit** from it?
  - Payment of public $$$ to private companies for the public's data? (CDRC Leeds)

http://www.pixton.com/comic/xya3s212

The impact of "too much" data

[Big data is] "a version of cherry-picking that destroys the entire spirit of research and makes the abundance of data extremely harmful to knowledge." (Taleb 2012, 416)

Transparency even more important



"fewer and fewer papers today have results that replicate"

Prevents abuse, ensures **reproducibility (the cornerstone of science) + public participation**

http://www.pixton.com/comic/xya3s212

Outputs, completed and to do

- Paper under review for Geo-spatial information science (Preprint on arXiv.org)
- Discussion paper on broader issues
  - Suggestions of where to publish?
- Working paper on spatial interaction models in R building on Dennet (2012)

Conclusion: in what situations are benefits of VGISM > costs

- Where little/no official data but verification possible VGI from Social Media is useful
  - Phenomena that are ephemeral, so not conducive to standard surveys
- Situations where people actually have time to Tweet
- Subjects that can be 'geovalidated'
  - Eg: road safety perceptions, visits to cinemas, geo-behavioural demographics
- Where application is clearly for public benefit
- Use social media data for public engagement

NB: Things to follow-up on

- **Full references in conference paper**
- Check Snowden, E. (2013). Interview with Glen Greenwald - Full Transcript.
- Reproducible code available on rpubs.com/robinlovelace
- See how to set up your very own 'Twitter Listener'
- Check out the 'big data backlash' (Taleb 2012 on Wired.com)
- Slides available from robinlovelace.net

Key References

- Dennett, A. (2012). Estimating flows between geographical locations:’get me started in'spatial interaction modelling. UCL Working Papers Series, 44(0), 0–24.
- Taleb, N. N. (2012). Antifragile: things that gain from disorder. Random House LLC.
- Lovelace, R., Malleson, N., Harland, K., & Birkin, M. (2014). Geotagged tweets to inform a spatial interaction model: a case study of museums. arXiv preprint arXiv:1403.5118.

Table 1. Museum characteristics and proxies of attractiveness. Distances are averages.

Museum	Tweet count	Dist. to home (km)	'Museum tweet–museum dist. (m)	Floor plan (m2)	News Mentions
Abbey House Museum	8	2.9	132	1072	2
Armley Mills	55	3.5	194	2734	2
Bradford Industrial Museum	11	5.6	110	1382	1
Cartwright Hall	2	8.5	95	1519	4
Henry Moore Institute	25	6.6	86	562	5
Leeds Art Gallery	93	5.5	115	1322	8
Leeds City Museum	102	5.2	130	1731	7
National Media Museum	288	8.5	131	3211	252
Royal Armouries	154	6.4	134	5180	36
Thackray Medical Museum	18	13.7	136	1790	5