Can social media data be useful in spatial modelling? A case study of 'museum Tweets' and visitor flows

R. Lovelace, N. Malleson, K. Harland, M. Birkin

School of Geography, University of Leeds, Leeds, LS2 9JT, UK
Tel: 0044 113 3430779
R.Lovelace@Leeds.ac.uk
http://robinlovelace.net

This paper explores the potential of volunteered geographical information from social media to inform geographical models of behavior. Based on a case study of museums in Yorkshire, we created a spatial interaction model of visitors to 15 museums from 179 administrative zones to test this potential. Instead of relying on limited official data on the magnitude of flows from different attractions we used volunteered geographic information' (VGI) to calibrate the model. The method represents the potential of VGI for applications beyond descriptive statistics and visuals and highlights potential uses of georeferenced social media data for geographic models. The main input dataset comprised geo-tagged messages harvested using the Twitter Streaming Application Programming Interface (API). We successfully calibrated the distance decay parameter of the model and conclude that social media data have great potential for aiding models of spatial behavior. However, we also caution that there are dangers associated with the use of social media data. Researchers should weigh up the wider costs and benefits of harnessing such 'big data' before blindly harnessing this low quality, high volume resource. Our case study also serves as the basis for discussion of the ethics surrounding the use of privately harvested VGI by publicly funded academics.

1. Data and Methods

Data were collected during 445 days between 2011-06-22 and 2012-09-09 from an area of around 2,000 km^2 surrounding Leeds and Bradford. 992,423 geo-referenced Tweets were collected in total, equating to roughly 0.5 tweets per inhabitant in the study area, with a highly skewed distribution of tweet frequency. Each record in the dataset represents one tweet and includes a timestamp for the generation time, a user id allowing tweets from the same account to be linked together and geographical coordinates of the location where the tweet originated (Russell, 2011). The Tweets were converted into a MYSQL dataset to allow for fast preliminary filtering: automated accounts and empty messages were removed, leaving a dataset that was geographically and semantically richer.

The next stage was to sample to select 'museum Tweets': messages sent during or about museum visits. Semantic and spatial filters were employed for this: the former strategy involved selecting only messages containing character strings closely related to museum visits, resulting in a sample of 1,553 Tweets (figure 1). Regular expressions were used here, to ensure that terms encapsulated by either blank space or a combination of blank space and punctuation characters were included – more complex semantic filters were also considered.

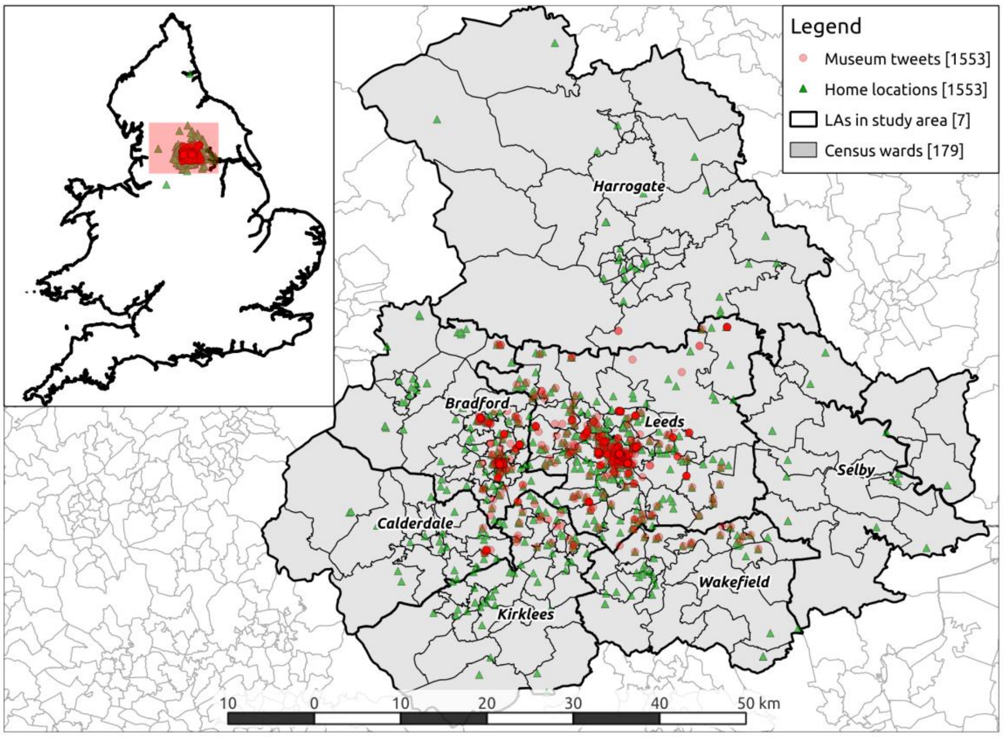

Figure 1. Overview of the geographical distribution of the semantically filtered museum tweets (red dots) and home locations (green triangles). The shade of points corresponds to density, illustrating high densities in Leeds and Bradford city centres.

The spatial filter was more straightforward: a layer of 15 museums in the case study area was created by filtering raw of Open Street Map (OSM) data (Figure 2). 10 m buffers surrounding the floor plans of the museums were created based on polygons of floor-plans extracted from OSM with the help of the R package osmar (Eugster & Schlesinger, 2013), see Figure 3. Comparative analysis of these two filtering strategies revealed that many false positives were included in the spatial filter, especially in centrally located museum sites; semantic and spatial filters were eventually used in tandem, resulting in a sample of just less than 1000 Tweets.

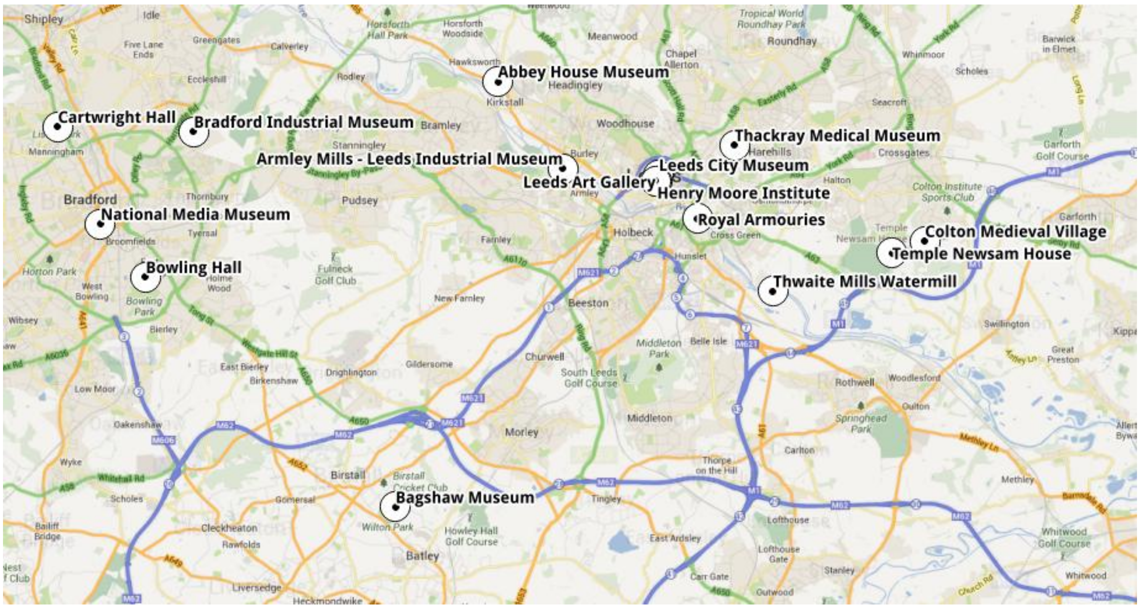

Figure 2. Locations of the 15 museums used in the case study. Basemap: Google.

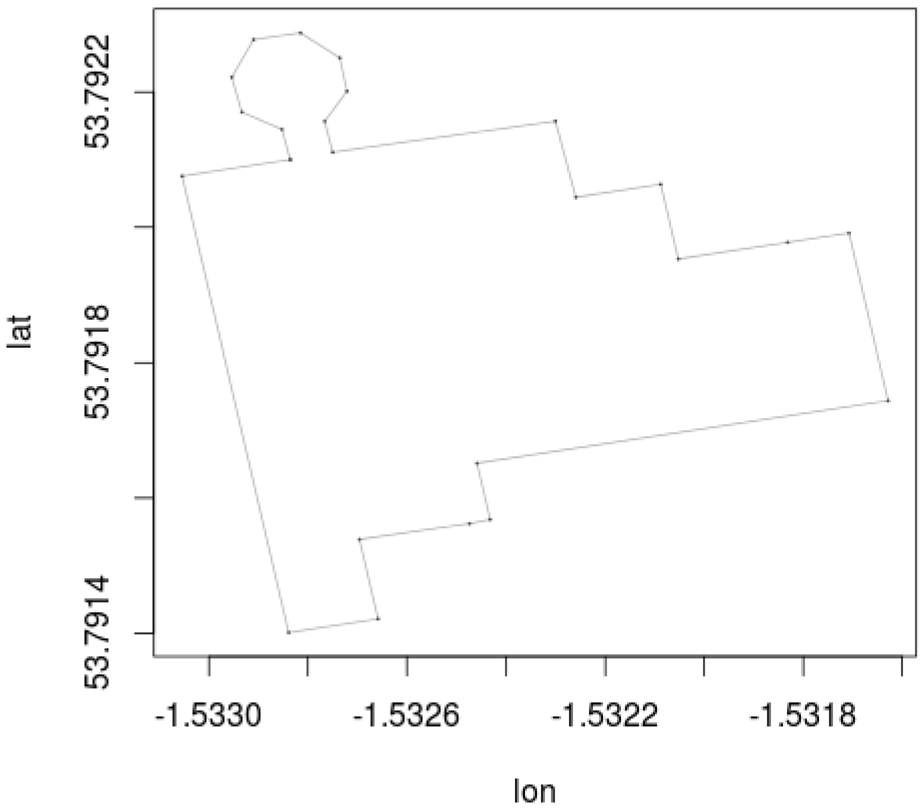

Figure 3. Floor plan of the Royal Armouries Museum, obtained from raw Open Street Map data using the osmar package in R.

The final stage was to assign each Tweet with an origin zone and destination museum then aggregate them into a flow matrix (*S*) with the same dimensions as *T*, the produced by the spatial interaction model described by eq. 1 (Wilson, 2000).

$$T_{ij} = Inc_i P_i W_j e^{-\beta d_{ij}} \tag{1}$$

In Equation 1 T_{ij} is a matrix representing the flow from origins (the rows, *i*) to each of the destinations (*j*) , *Inc* is the income-adjusted demand for museum trips per unit population (P) in each zone and β the distance-decay parameter. *Wj* is the 'attractiveness' of museum *j*, estimated based on information from Table 1.

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

The process of spatial aggregation to allocate Tweet origins to origin zones is visualized in Figure 4. The sparse matrix that resulted allowed direct comparison between the Twitter data and the model's ward-level output. Model parameters were then calibrated to optimize the fit between flows to museums inferred from Tweets and flow matrices generated by the spatial interaction model.

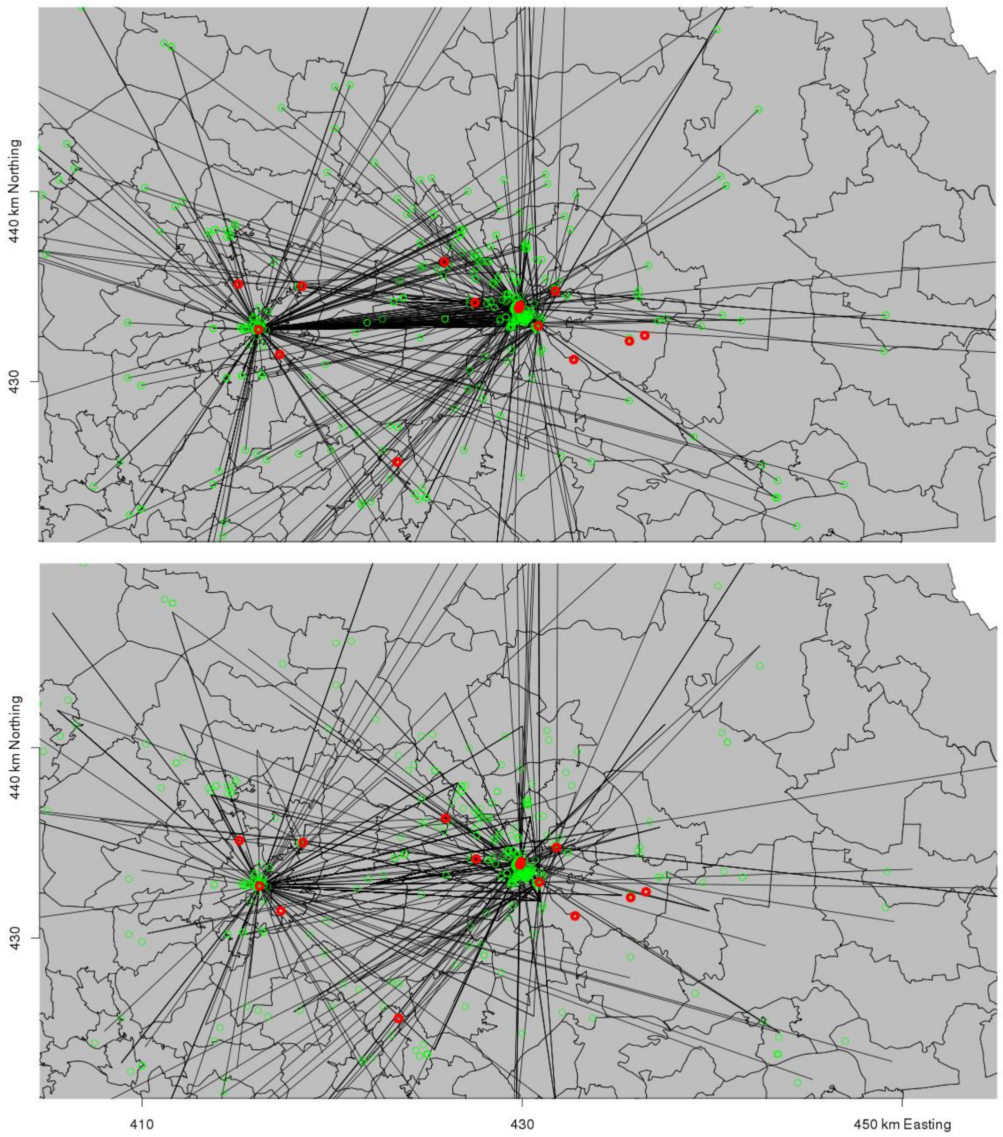

Figure 4. Flow maps of inferred museum visits from raw tweets (above) and from spatially aggregated tweet home locations (below). Green dots are home locations; red dots are museums.

2. Results

The baseline scenario consisted of the simplest implementation of an unconstrained SIM for this scenario. Thus, the only variable to optimize was β, which was initially set to 0.3, resulting in a positive correlation of 0.31 between the modeled and 'observed' (Twitter-inferred) flows. Iterating through 200 model-tweet observation comparisons (step size = 0.01) it was found that model fit was optimal with a β value of 0.95, with the correlation peaking with an r value of 0.39.

The next model tested added the W variable – see Equation 1 – to the model, to make larger and more frequently mentioned museums more attractive than small museums that few people had heard of. The impact of this change was dramatic, with r values peaking at 0.60. The optimal distance decay function in this model specification was found to be steeper (β = 1.32).

In the final model test the added refinement of variable demand from the origins was set, by altering Inc values according to a combination of income and geodemographic data acquired from the consultancy Acxiom. The performance of this model specification was found to be intermediate compared with the other two, with a maximum r value of 0.55. These results are displayed in Figure 5 below.

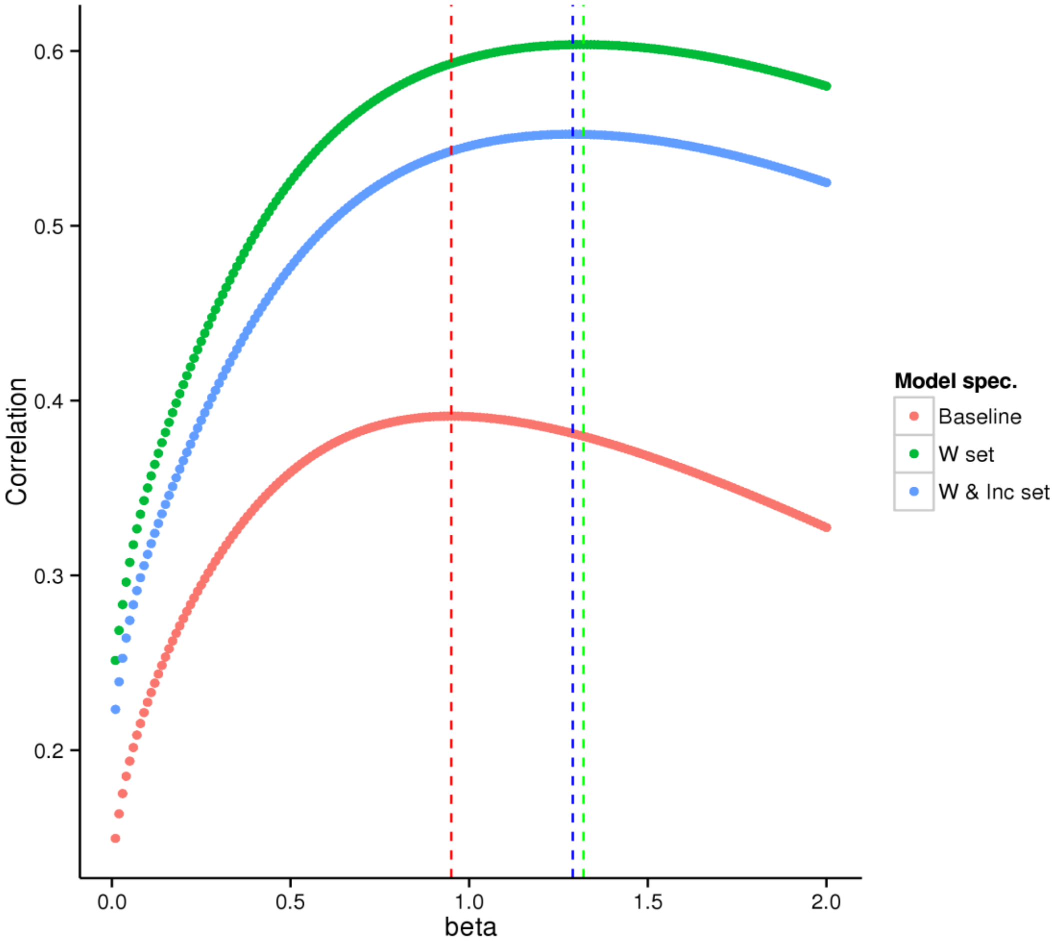

Figure 5. Correlation between the spatial interaction model (S) and flows inferred from tweets (S') against β values, for three different model specifications.

There are many refinements that can be made to the basic models specified above and some more experiments were conducted. A major change would be to use a constrained spatial interaction model, whereby the flows from each origin are set (see Figure 6). It was found that the model fit declined greatly in this scenario, however, seemingly due to forced flows to distant museums from wards in the periphery of the study region.

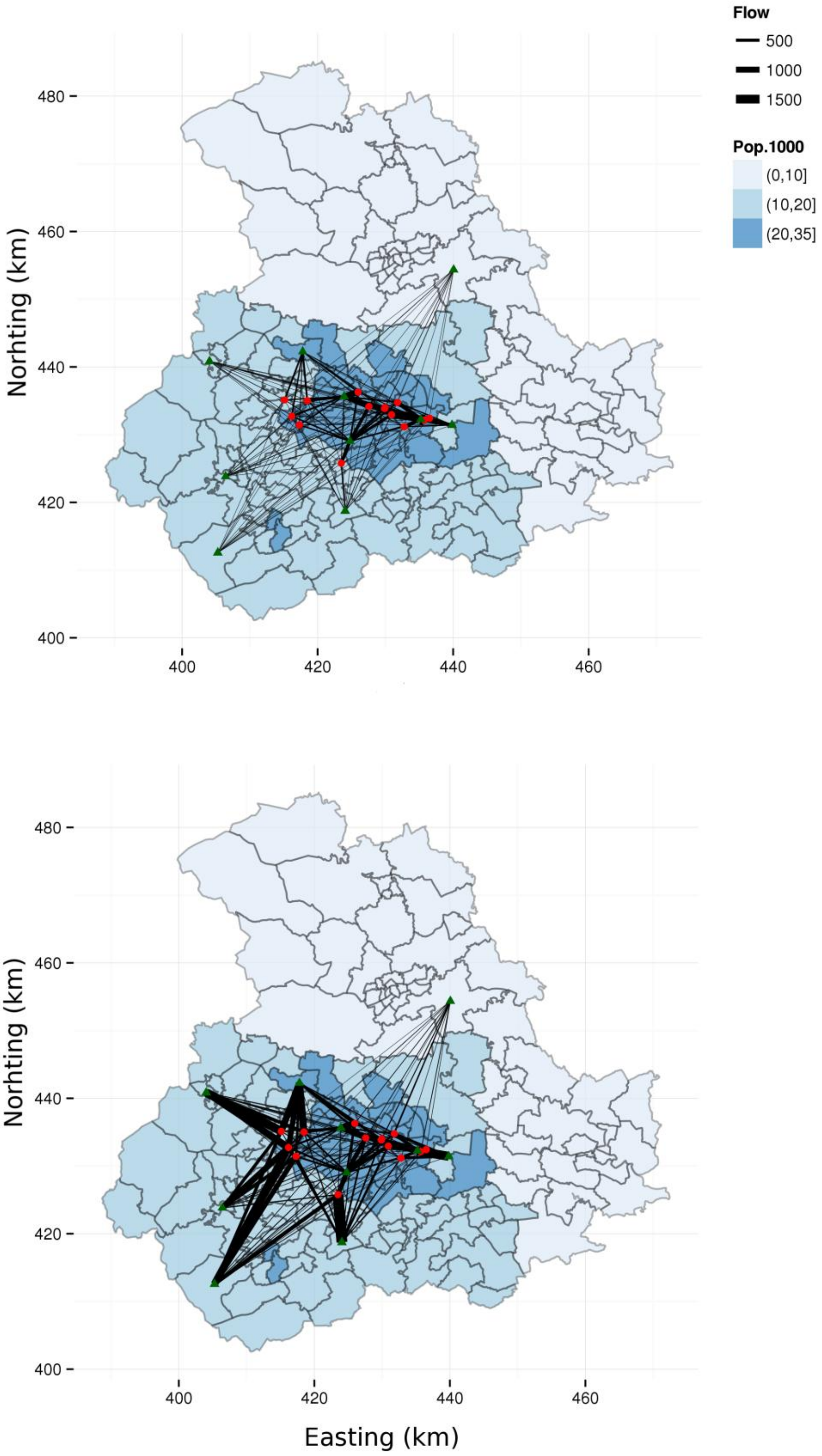

Figure 6. Constrained and unconstrained version of the model with W set. The green triangles are 20 randomly selected origins; the reds dots are museums. Blue shading is proportional to population in each ward.

3. Discussion

The three main contributions of this paper to geographic are:

- The proposition that increasingly ubiquitous social media data can improve quantitative models of spatial behavior.
- A practical demonstration of this idea based on georeferenced Twitter data to calibrate the distance decay parameter in a spatial interaction model.
- A platform for the wider discussion of the relative merits and drawbacks of VGI.

The areas of academic study where no official data collection occurs, or where datasets are unavailable to public academics is vast: datasets on buying habits, online search terms, location (collected by mobile operators via the triangulation of phone signals to the nearby masts) and even the location of house searches are all collected by private companies but seldom harnessed by academics for they may regard as 'the greater good'. It is safe to say that datasets on more obscure areas of knowledge, such as the most frequently visited sea kayaking areas in Britain, the routes most commonly taken by cyclists into work and the distance decay functions of sports are either very rare or absent. Before the online social media phase of the ongoing digital revolution, academics could only speculate about such questions or conduct expensive surveys to try to uncover the basics. Now, in every instance, there is the feeling that the data is somehow 'out there' on a server, just waiting to be harnessed.

In cases where sufficient quantity and quality of data can be found, there is little doubt that this new digital information (which we term 'Big Data', that includes VGI) can yield new and important insights. It is doubtful that the particular dataset analysed in this article is of sufficient quality to add significantly to human knowledge about leisure time – in other words we do not feel that the results, in themselves, of this paper are particularly interesting. Yet it would not be hard to imagine that, with an increase in the volume of data (perhaps by a factor of 10), useful insights such as spatial and temporal variability, and the types of zone which tend to visit certain types of museum, could be generated. Of course, this would require further processing power, which is currently available in top-end computers, but whether or not it is the *optimal* use of social media datasets is open to debate. It is hoped that this paper serves as a catalyst for further discussion of the issue and prioritization. For now, the analysis of Twitter data suggest that the kinds of areas well suited to analysis via VGI include the following:

- Areas of knowledge where publicly available knowledge is lacking.
- Subjects about which behavior can be reliably inferred from social media (e.g. location – inferring thought processes is a far more challenging area).
- Phenomena about which small additional insights that can realistically be gleaned from relatively small (and accessible) VGI, with the potential for large social benefit.

This final point is critical and has been rarely commented on in the literature: the word *social* in social media should not be ignored. Being provided by the public to the world, surely the results should be made available to the world in a way conducive to social benefit? We therefore suggest that publicly funded researcher explicitly weigh up the social costs and benefits of using social media to inform their analysis, rather than using whatever datasets are available. Using publicly viewable social media data is hardly equivalent to massive digital surveillance by the likes of the USA's National Security Agency (NSA) the UK's Government Communications Headquarters (GCHQ): the datasets are much smaller and the level of invasion into the private sphere is much smaller. However, there are parallels. We contend that it is useful to frame the debate surrounding the use of VGI in academic research not only in academic terms, but in academic terms also. In particular, we advocate a pro-active and transparent decision making process regarding whether or not it is worth using social media data. "You should decide whether we need to be doing this." (Ed Snowden, 2013)

The example of museum visits is rather innocuous compared with the kinds of application to which this method could be put. To pick one commercially relevant example, the kinds of shop preferred by inhabitants of people from different geodemographic could be inferred from similar technique, given a sufficient quantity of data. Yet the private sector is already far ahead of the academic sector when it comes to product targeting based on social media, for example as illustrated by its use of 'micromarketing' (Sunday et al. 2014). Another factor to consider before using this kind of data relates to data ownership: while the public can access social media datasets, this ownership is ultimately mediated through private corporations that can decide who gets which databases and, critically, the price.[1]

In agreement with Goodchild (2007), we conclude that free, geographic and semantically rich datasets derived from the harvesting of social media sites en masse should continue to be of great and growing interest to spatial analysts. Data quality remains a serious concern for all such VGI, however (Flanagin & Metzger, 2008). The sheer diversity and sporadic nature of the data poses new challenges to researchers accustomed to relatively clean official datasets. As emphasized throughout, these challenges should not be overlooked: they must be acknowledged at the outset and tackled with care. More specifically, the main limitations of the data used in this study include:

- limited data availability
- the sub national nature of the study area
- a limited set of museums being considered, and
- uncertainty about travel behaviour of.

Improved data harvesting and retrospective data collection could help overcome the first two issues; use of officially registered museums and visitor data could tackle the third and fourth points.

Despite the data limitations, ethical concerns and unknown longevity of social VGI as a free data source, it seems likely that the size and richness of available datasets will continue to grow. In parallel with this, computing power will continue to improve and computer programs will continue to develop towards greater functionality and user friendliness. This means VGI from social media will become an increasingly attractive alternative to official datasets for

[1] An industry has emerged in the sale social media data to academics and other companies. To illustrate the point, the University of Leeds has recently purchased a large dataset of tweets relating to Hurricane Sandy for approximately $3000. This raises the question: if the information was 'volunteered', who has the right to its ownership and sale? This question of ownership underlies Twitter's donation of its historical archive to the US Library of Congress and announcements by the US Library of Congress to make available Twitter's vast data archive (Rivers and Lewis, 2014). However, the practicalities of what datasets will be available and who and how to access them have yet to be clarified and the potential for private tech companies to profit from public academics interested in social media data is great.

geographical problems that are presented in this paper, where data limitations remain a major constraint.

In summary, it is hoped that this paper will lead to further discussion of the relative merits of Twitter and other volunteered social media information for informing geographical research. Ethical considerations should also guide the research: if the information is provided by the public for free, surely the benefits that accrue should be for public benefit.

4. References

EUGSTER, M. J. A., & SCHLESINGER, T. (2013). osmar: OpenStreetMap and R. *R Journal, 5*(1), 53-63.

FLANAGIN, A. J., & METZGER, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal, 72*(3-4), 137-148.

GOODCHILD, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal, 69*(4), 211-221. doi: 10.1007/s10708-007-9111-y

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

SNOWDEN, E. (2013). Interview with Glen Grweenwald - Full Transcript. http://www.policymic.com/articles/47355/edward-snowden-interview-transcript-full-text-read-the-guardian-s-entire-interview-with-the-man-who-leaked-prism

SUNDAY, E. M., & AWARA, N. F. (2014). Customer Satisfaction and Social Media Driven Micromarketing: An Empirical Evidence. International Business and management, 8(1), 32-36.

WILSON, A. G. (2000). *Complex spatial systems: the modelling foundations of urban and regional analysis*: Pearson Education.

5. Biography

I trained as an environmental geographer, exploring the interface between human and environmental systems. After completing an MSc in Environmental Science I arrived at the conclusion that energy is the 'master resource' that underlies this interaction. Since completing my PhD in the energy costs of transport I have focussed on the use of datasets, official and 'big', to better understand energy use and its drivers. My long-ter aim is to leverage this work to model post carbon futures.