Mining Consumer Insights from Geo-Located Social Media Datasets

Alyson Lloyd[*1], James Cheshire[†1].

[1]Department of Geography, UCL, London, UK.

January 4th, 2016

Summary

Twitter's geo-referenced and opinion rich data offers the potential to determine when, where and what people Tweet about retail. However, the utility of Twitter to extract consumer insights specific to retail centres remains relatively under-investigated. Retail related Tweets were identified and their spatial attributes examined in relation to centre locations. An advanced form of kernel density estimation revealed that retail related Tweets might be a good indicator of areas of elevated retail activity within the UK. The work further highlights that to examine Twitter data, heuristics that account for the underlying geographic distribution of users are necessary. The approach demonstrates potential methodologies and applications for less conventional datasets, such as those derived from social media data, to previously under-researched areas.

KEYWORDS: *Social Media; Twitter; Geo-Reference; Retail; Consumer.*

1. Introduction

The mining and analysis of social media datasets has received considerable attention in recent years, as researchers and marketing companies seek new insights into consumer dynamics. Twitter's 'micro-blogging' nature enables its users to instantly distribute their thoughts via multiple Internet connected devices. Twitter also allows for 'geo-referenced' messages (adding latitude and longitude coordinates) and accurate time stamps. The platform has become widely used in research, in comparison to the likes of Facebook for example, since it offers a public API that enables anyone to request a sample of Tweets according to particular search criteria. With 15 million active users within the UK alone (estimated by Twitter in September 2014), Twitter has an extensive online community already mined for consumer insight (Brennan & Schafer, 2010).

In spite of its high uptake and widespread use by advertisers, the utility of Twitter to extract consumer insights specific to retail centres remains relatively under-investigated. Empirical evidence is first required to investigate the appropriateness of social media data to this context – particularly in relation to its representativeness of the broader population. For instance, it is estimated that only 12% of the general UK population use the service (Peerreach, 2013) and only an estimated 1% of these are geo-referenced (Morstatter et al., 2013). Furthermore, it could be susceptible to demographic biases (Li et al., 2014; Pavalanathan & Eisenstein, 2015). For example, Twitter is particularly popular among those under 50, with the highest percentage of users being 18-29 year olds (Pew Research, 2015). Its content can also be unstructured and informal, often making it hard to utilise in research (Andrienko et al., 2013) and contribution bias is often apparent (Nielsen, 2006) meaning that few users contribute a large percentage of the Tweets. It is clear that more robust evidence is required in order to understand the uses of geo-referenced Twitter data within the retail sector. This exploratory investigation offers a starting point of examining such issues.

2. Data

[*] alyson.lloyd.14@ucl.ac.uk
[†] james.cheshire@ucl.ac.uk

A total of 99,139,622 Tweets were obtained through Twitter's API between December 2012 and January 2014. However, 16,173,077 Tweets were removed by the cleaning process. A set of 170 major retailers was assembled from those that reside in high-street retail centres and shopping centres across the UK. A Tweet subset of retailer 'mentions' was created from this, consisting of 177,635 Tweets (0.21% of the cleaned sample). A proportional control sample was randomly selected from the whole Tweet sample to be used in the KDE analyses.

Retail centre location and boundary data were provided by the Local Data Company Ltd (LDC, http://www.localdatacompany.com), who are a commercial research consultancy that specialise in physically documenting locations, boundaries and attributes of retail centres across the UK (see Figure 1).

Figure 1. LDC defined retail centre boundaries for the UK and Greater Manchester.

3. The Spatial Distribution of Geo-referenced Tweets

The raw spatial distribution of the Tweets in the UK can be observed in Figure 2. At this scale, the Tweet sample unsurprisingly maps population centres in the UK. The retailer mentions exhibit a similar pattern. At finer scales, Tweet patterns can be delineated by places of probable high human activity such as roads and train networks and any inhabited space.

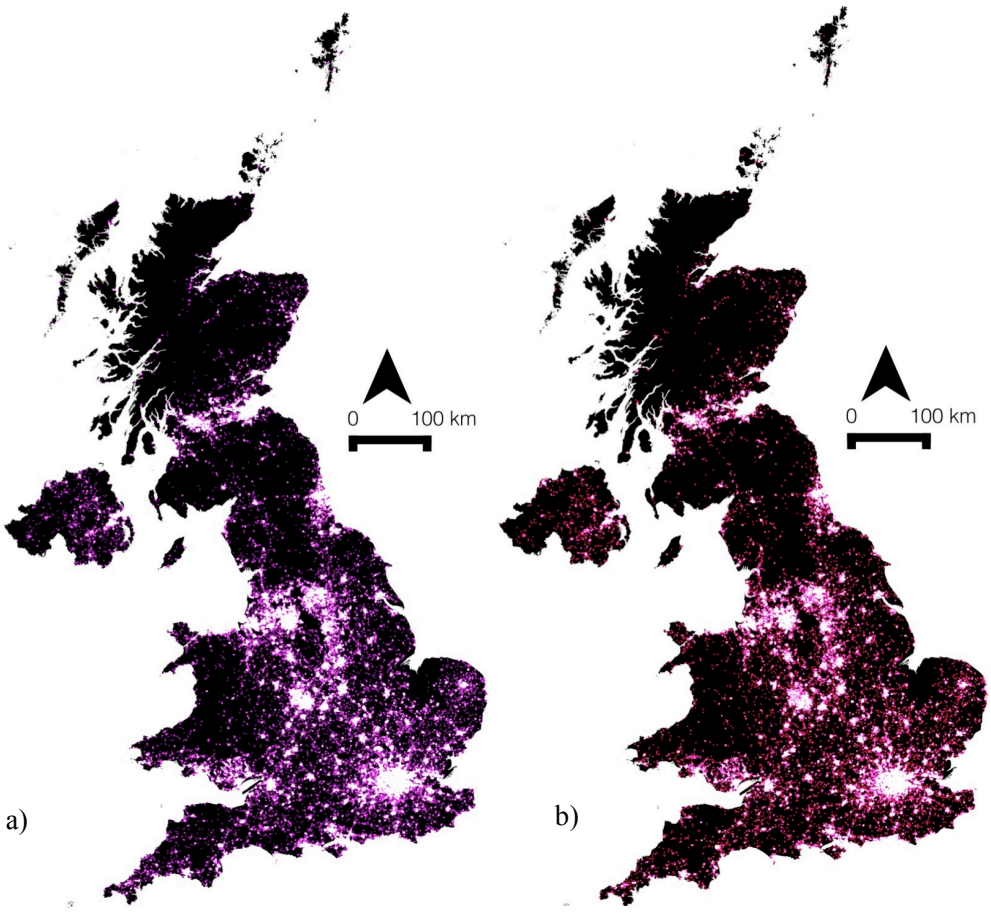

Figure 2. The raw spatial distribution of a) the sample of general tweets and b) retailer mentions, sent within the UK between December 2012 and January 2014.

This suggests little overall variation in spatial distribution between the general and the retail Tweets and therefore it is unlikely that a point pattern analysis would extract anything meaningful. Nevertheless, the fact that retailer mentions are apparent in major town centres and shopping centres indicates possible links to retail.

4. The Spatial Distribution of Retail Tweets

In order to extract any significant differences in spatial variation between the global (as represented by the control sample) and the retail Tweets, an advanced form of KDE was deployed. The kernel density estimator for bivariate data can be defined as:

$$\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n} h_i^{-2} K\left(\frac{z - X_i}{h_i}\right) \tag{1}$$

where K is the kernel function, and h_i is the smoothing parameter (or bandwidth) for the $_i th$ observation. However, traditional KDE suffers from an inability to normalise data based on an underlying spatial distribution, in addition to subjecting analyses to fixed bandwidth biases such as under smoothing in areas with few observations and over smoothing in others. In contrast, adaptive KDE allows for the bandwidth to vary with the sample data, which reduces this bias. The adaptive estimator, as suggested by Abramson (1982), is calculated by:

$$h_i = h_0 f(x_i)^{-1/2} \gamma^{-1} \quad (2)$$

where h_0 refers to the global bandwidth, which is scaled by the product of the inverse square-root of the pilot (local) density ($f(\boldsymbol{X}_i)^{-1/2}$) and the geometric mean (γ) of this term. Therefore, for adaptive estimations, two bandwidths must be selected: a pilot and a global bandwidth. The pilot density is itself a fixed kernel density estimate. This was applied using the least-squares cross validation (LSCV) approach (Bowman, 1997), which examines various bandwidths and selects the one that gives a minimum score $M_1(h)$ for the estimated error (the difference between the unknown true density function and the kernel density estimate).

4.1. Method

The sparr package in R (Davies, Hazelton & Marshall, 2011) was used to carry out the analysis, which contained functions necessary to identify significant clusters of retail tweets and superimpose contours at a given significance level using the relative risk function (a commonly used tool in describing disease risk; Kelsall & Diggle, 1995b). This aims to identify areas of statistically significant fluctuations between density estimations by taking into account the underlying population density. This is achieved by differentiating case (points of interest) and control (underlying population distribution) data. The resulting relative risk function describes the difference in spatial variations and highlights 'risk' areas. Tolerance contours (at significance levels of $\alpha = 0.01$ and $\alpha = 0.05$) were calculated using the z-statistic-based asymptotic normality test (Davies & Hazelton, 2010). Areas of high retail Tweet density could then be identified, analysed and compared to LDC retail locations.

4.2. Results

Figure 3 shows areas of significantly high probabilities for retailer mentions to occur, within the UK. However, at this scale it is difficult to assess densities based on a retail centre level of granularity. Therefore, Figure 4 shows the results conducted on a finer scale, for the area of Greater Manchester.

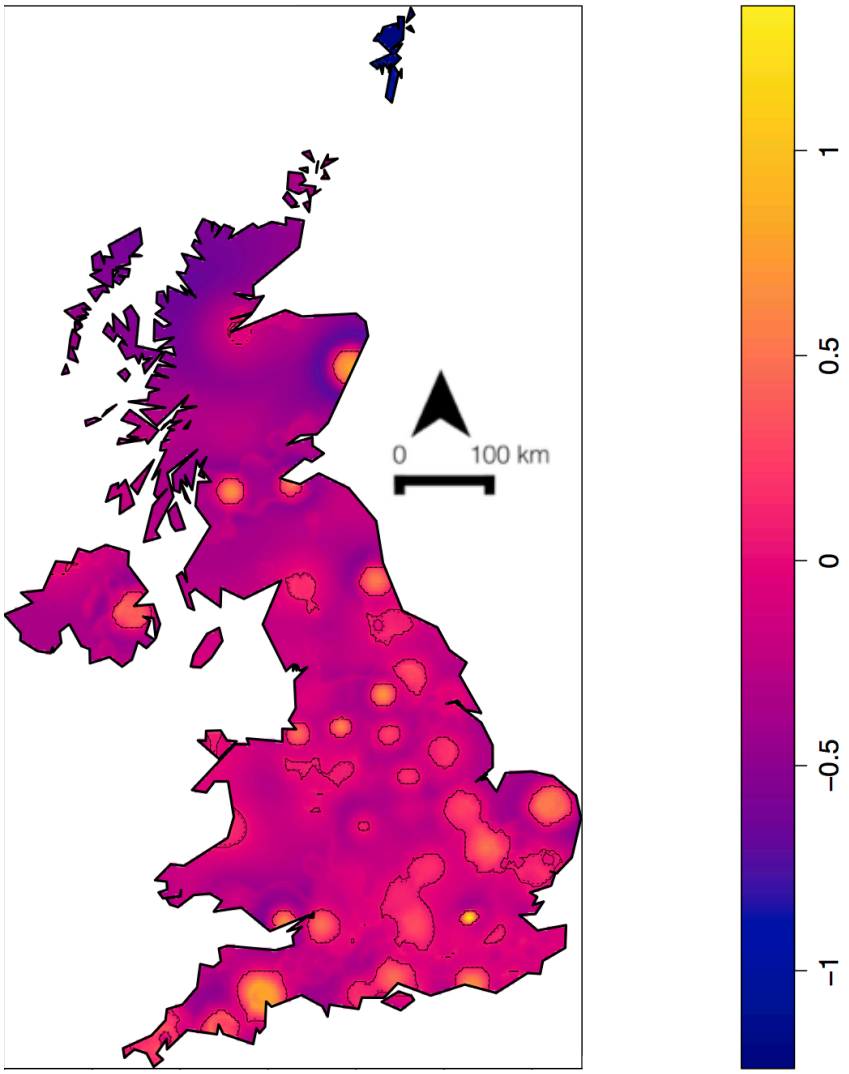

Figure 3. Significant fluctuations of retailer mentions on Twitter, within the United Kingdom between December 2012 and January 2015, with tolerance contours at 0.05 (solid) and 0.01 (dashed).

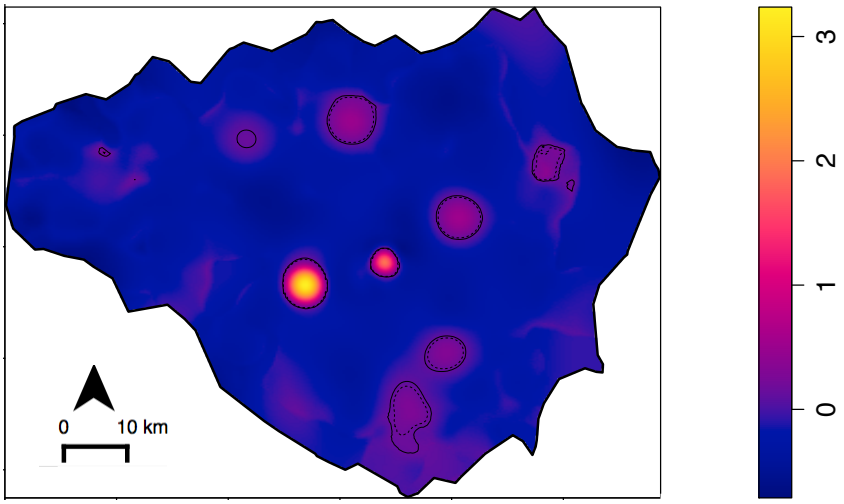

Figure 4. Significant fluctuations of retailer mentions on Twitter, within Greater Manchester between December 2012 and January 2015, with tolerance contours at 0.05 (solid) and 0.01 (dashed).

Significant fluctuations in retail related Twitter activity can be observed within major town centres and shopping centres in and around Greater Manchester. The area of highest density was the Old Trafford Shopping Centre, followed by the city centre, whilst major towns such as Bury, Bolton and Stockport were also significant. This suggests that retail related Twitter content is representative of the locations of elevated retail activity.

6. Discussion and Conclusions

This work provides some evidence as to the spatial distribution of retail related social media activity, which has previously been neglected in research. It has highlighted that to examine Twitter data, heuristics that account for the underlying geographic distribution of users are necessary. In addition, it demonstrates that alternative data sources such as these, may potentially be utilised to identify areas of elevated retail activity across the UK. As Twitter is accessible internationally, this could be replicated on a global scale. These results also show that Tweets could potentially be used to detect centres of other behaviors or activities that may be less obvious. Finally, this work shows that the relative risk function can be implemented beyond the scope of disease and risk measurement. For example, it could useful applications for further datasets with potential underlying population dispersions, such as consumer transactional data. Such an approach could, for instance, highlight areas of elevated spending, independent of population density.

It is most likely that the uneven distribution of Tweets across retail centres is reflective of the demographic biases in the data, such as the higher proportions of older age groups living in rural areas (Pateman, 2011), yet youths dominating Twitter (Pew Research, 2015). In addition, geo-referencing is reliant on access to mobile Internet, of which many rural areas experience 'black spots' (The Guardian, 2013). These, along with other biases acknowledged, should be carefully considered when utilising Twitter data. However, despite these unavoidable biases, this investigation demonstrates that within major town centres and shopping centres, there is a significant amount of retail related content that could be further utilised for consumer insights.

7. References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates-a square root law. *The annals of Statistics*, *10*, pp. 1217-1223.

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic patterns in georeferenced Tweets through space-time visual analytics. *Computing in Science and Engineering*, *15*, pp. 72-82.

Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations: the kernel approach with S-Plus illustrations*. Oxford University Press.

Brennan, B., & Schafer, L. (2010). *Branded!: How retailers engage consumers with social media and mobility*. John Wiley & Sons.

Davies, T. M., & Hazelton, M. L. (2010). Adaptive kernel estimation of spatial relative risk. *Statistics in Medicine*, *29*, pp. 2423-2437.

Davies, T. M., Hazelton, M. L., & Marshall, J. C. (2011). Sparr: analyzing spatial relative risk using fixed and adaptive kernel density estimation in R. *Journal of Statistical Software*, *39*, pp. 1–14.

Dramowicz, E. (2005). Retail Trade Area Analysis Using the Huff Model. [Online] *Directions Magazine*. Avaliable at: http://www.directionsmag.com/articles/retail-trade-area-analysis-using-the-huff-model/123411. [Accessed: July 2015].

The Guardian (2013). UK mobile phone coverage: the country's signal blackspots. [Online]. Data Blog Available from: http://www.theguardian.com/money/datablog/interactive/2013/oct/29/uk-mobile-phone-coverage-interactive-map-signal. [Accessed September 2015].

Kelsall, J. E., & Diggle, P. J., (1995b). Non-Parametric Estimation of Spatial Variation in Relative Risk. *Statistics in Medicine*, 14, 2335–2342.

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*, pp. 61-77.

Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley. 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming Api with Twitter's Firehose." In *Proceedings of ICWSM*. Cambridge, MA: AAAI Press.

Nielsen, J. (2006). Participation Inequality: Encouraging More Users to Contribute.". *Jakob Nielsen's Alertbox*, *9*.

Pavalanathan, U. and Eisenstein, J. (2015) Confounds and Consequences in Geotagged Twitter Data. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. EMNLP; 2015.

Pateman., T. (2011) Rural and urban areas: comparing lives using rural/urban classifications. [Online] Regional trends, Office for National Statistics. Avaliable from: http://www.ons.gov.uk/ons/rel/regional-trends/regional-trends/no–43–2011-edition/index.html. [Accessed September 2015].

Peerreach (2013). 4 ways how Twitter can keep growing. Peerreach Blog. [Online]. Available from: http://blog.peerreach.com/2013/11/4-ways-how-twitter-can-keep-growing/. [Accessed: October 2015]

Pew Research Center (2015). Social Networking Fact Sheet [Online]. Avaliable from: http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/. [Accessed: September 2015].

8. Acknowledgements

Thanks are given to the ESRC for the funding of this research and to the LDC for supplying the retail centre data.

9. Biography

I am currently completing the first year of my PhD in Retail Sustainability and Resilience at UCL. This involves examining the potential applications of big data for insight within the retail sector.