Computers, Environment and Urban Systems 61 (2017) 108–118

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Deriving retail centre locations and catchments from geo-tagged Twitter data

Alyson Lloyd, James Cheshire *

Department of Geography, UCL, London, UK

ARTICLE INFO

Article history:
Received 13 January 2016
Received in revised form 20 August 2016
Accepted 28 September 2016
Available online 20 October 2016

Keywords:
Social media
Retail
Twitter
Consumer data
Human mobility

ABSTRACT

This investigation offers an initial foray into the application of geo-tagged Twitter data for generating insights within two areas of retail geography: establishing retail centre locations and defining catchment areas. Retail related Tweets were identified and their spatial attributes examined with an adaptive kernel density estimation, revealing that retail related Twitter content can successfully locate areas of elevated retail activity, however, these are constrained by biases within the data. Methods must also account for the underlying geographic distribution of Tweets to detect these fluctuations. Additionally, geo-tagged Twitter data can be utilised to examine human mobility patterns in a retail centre context. The catchments constructed from the data highlight the importance of accessibility on flows between locations, which have implications for the likely commuting choices that may be involved in retail centre journey decision-making. These approaches demonstrate the potential applications for less conventional datasets, such as those derived from social media data, to previously under-researched areas.

1. Introduction

In the UK online retailers have achieved a market share of 16.8% (Retail Research, 2016). This growing percentage represents one of a number of challenges facing traditional British high-streets and town centres, with consumers increasingly substituting certain types of town-centre retailers for these online alternatives (Weltevreden, 2007). Those most affected have been music, video and book retailers, travel agents and a wide range of other retailers that have been susceptible to the effects of online options that are increasingly reliable and easy to use (Wrigley et al., 2015). In order to remain successful, high-streets and shopping centres – or the stores within them – need a greater understanding of the digital footprints of their customers in order to better engage with them.

In light of this, many major retailers are seeking to harness digital platforms in order to attract customers to physical stores and develop closer relationships with them. However, more robust evidence is needed to understand the effectiveness of adopting on-the-go technologies as a means to boost town centre vitality (Wrigley et al., 2015). Here we present two examples where social media data offer a proxy for the digital engagement of consumers in the UK. First we seek to identify areas of high retail activity based on the content and location of consumers' Tweets. Second we combine Twitter data with an established map of retail centres in the UK to discern their digital footprints. It is hoped that these analyses will be of interest not just to the retailers themselves, but also to local authorities and policy makers seeking to re-invigorate many high streets to better meet the challenge of online retail.

The use of geo-tagged social media data as indicators of human activity has received considerable attention in recent years as researchers and businesses seek out alternative data from which to derive insights into population dynamics. Twitter has been most widely utilised since it benefits from an extensive online community of around 15 million active users within the UK (estimated by Twitter in 2015) and it offers a public application programming interface (API) that enables anyone to request a sample of Tweets according to a particular search criteria. Crucially in the context of geographical analysis, this API provides the location of where the Tweet was sent if a user has consented to revealing such information.

Within the retail sector, the value of social media data are well recognised (McKinsey & Company, 2011), with an estimated 62% of Twitter users following their favourite brands, and major retailers receiving an average of 821 direct mentions and 114 replies per day (Brandwatch, 2015). Retailers frequently use such data to improve brand awareness, listen to customer sentiments and improve customer services (Brennan & Schafer, 2010). However, the geographical components of these data have received less attention. Yet, it is estimated that 80% of Twitter users access the platform via a smartphone (Twitter, 2015) and that 25% use the service whilst shopping (Nielsen Media Research, 2014). Furthermore, Cheng, Caverlee, Lee, and Sui (2011)

* Corresponding author.
E-mail addresses: alyson.lloyd.14@ucl.ac.uk (A. Lloyd), james.cheshire@ucl.ac.uk (J. Cheshire).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.09.006

found shops and restaurants amongst the top 5 places that people are likely to 'check in' on Twitter. It's clear there is the potential to provide insight into the activities and mobility patterns of consumers within the Twitter population.

We argue that the use of the geographical component of the Tweets will serve to inform knowledge of the digital footprints of online customers and engagement with online platforms. In addition, existing retail centre location data are sparsely available and have been primarily constructed using centroid locations of formerly derived retail cores. For example, commonly used centre and boundary data developed by the Department of Community and Local Government (DCLG) State of the Cities Report were primarily defined using the underlying economic activity and locations of anchor stores in 2004. Therefore, contemporary geographical definitions of centre locations require exploration. Furthermore, the data could provide insights in the study of retail centre catchment analysis, which refers to the areal extent from which the main patrons of a store or retail centre are typically found (Birkin, Clarke, & Clarke, 2010). Notably, there is already a large body of literature exploring retail catchments, which have applied methods such as drive-time (where patrons are expected to go to the closest or most logistically convenient location), or incorporated measures of centre attractiveness into more complex models (i.e. Dolega, Pavlis, & Singleton, 2016). However, there has been little exploration of data-driven methods into flows between retail centre locations.

Despite the range of positive applications, Twitter data have a number of disadvantages that make them inherently hard to interpret and analyse - particularly in relation to their representativeness of the broader population. For instance, it is estimated that 23% of the UK population use the service (Pew Research, 2015) and only an estimated of these 1% opt to share their locations. Still, research has demonstrated that when obtaining these data using Twitter's public API, it is possible to extract over 90% of all geo-tagged posts (Morstatter, Pfeffer, Liu, & Carley, 2013) due to the rate of geo-tagged Tweets roughly corresponding to the limitation rate of the stream. The data are also susceptible to demographic biases such as an over-representation of younger cohorts between 15 and 30 (Longley & Adnan, 2016) and suffer from contribution bias, meaning that a small proportion of users generate a large percentage of the Tweets (Nielsen, 2006). Nevertheless, the data do offer a number of advantages, such as having a high temporal granularity that is international in scale and a selective but numerically large representation (Adnan, Lansley, & Longley, 2013). In addition, data-driven methods can provide benefits in comparison to traditional approaches (such as surveys), as they are able to provide unique information about the social dynamics of places that are not easily and inexpensively obtainable on such a large scale (Li, Goodchild, & Xu, 2013).

In this paper we explore two hypotheses. Firstly, that the content of Tweets would have an identifiable correspondence to locations of retailing activities, and secondly, that the data have the potential to inform the creation of retail catchment areas by evaluating the mobility patterns of Twitter users across different locations. Our motivations were to understand the potential applications and limitations of these data within a retail centre context and place them within the broader framework of promoting town centre resilience in this digital era of online uncertainty.

2. Data treatment

2.1. Twitter data

Tweets were obtained through Twitter's filtered streaming API between December 2012 and January 2014. The Tweet locations were predominantly recorded using the integrated Global Positioning System (GPS) on users' smartphones and typically accurate to within several metres (Li et al., 2013). Whilst the API is assumed to collect these Tweets at random, the methods that Twitter employ to sample these data are currently unknown. In total 99,139,622 Tweets sent by 1,777,873 users were collected. As is common to almost all social media datasets the frequency of Tweets per user was positively skewed, with the most active user sending 68,389 Tweets, yet a median of 7 Tweets per user.

Twitter data can also be susceptible to "bots" that characteristically send multiple spam messages (Hawelka et al., 2014). Therefore, measures were taken to clean the database. For example, one account returned 47,132 Tweets such as *"Entrepreneurs: 5 Reasons Why Your Products are Not Selling - http://t.co/BWovUzLL"*. Other users considered unrepresentative of normal Tweeting behaviour were those who had sent repeated posts to gain followers or attention (i.e. from celebrity accounts), for example, 15,081 Tweets from a single user such as *"@Real_Liam_Payne LIAM, please follow me and we love you so much x1"*. In order to filter such cases, the following procedures were applied:

1. A threshold of 3000 Tweets per user over the duration of the dataset, to avoid the large amount of contribution bias dominating the analysis and to remove prolific spam accounts (Lansley & Longley, 2016).
2. Users who had posted identical messages more than three times, as these were likely to be fake accounts (Wang, 2010).
3. Messages containing 'spam' trigger phrases.

The threshold removed 5,206,922 Tweets from 1032 users (5.25% of the full sample, but only 0.05% of users). Messages with high counts eliminated 236,208 Tweets from 173 users. The 'spam trigger phrases' then aimed to further identify the repeated messages that had been modified to avoid detection. Phrases were obtained from Mequoda (2015) and were edited so that they were relevant for the Twitter data. For example, terms such as 'credit' returned many non-spam messages within the database. However, terms such as 'no obligation', and 'ecommerce' were useful to identify spam content. There were 11 spam terms in total (see Appendix A). A total of 9467 Tweets from 33 users were removed at this stage. This left a final sample of 93,687,025 Tweets from 1,776,635 users for analysis.

2.2. Identifying retail tweets

Interactions with major retailers were identified from the cleaned data. An initial list of 366, major UK retailers were obtained from the IRUK Retailing Top 500 Annual Report (IRUK, 2016), across 11 categories (see Table 1). Only primarily high-street retailers were selected, based on the assumption that these interactions may be spatially representative of retail centre activities. This was extended manually to include as many as possible within the UK, including leisure categories (i.e. food and drink).

Due to the informal nature of Twitter, abbreviated or incorrect spelling variations needed to be accounted for. Common variations could be identified by manually observing the sample and the live Twitter feed. For example, "Marks and Spencer's" had 11 variations that produced relevant Tweets (see Appendix B). However, some major retailers could not be utilised in a general query, such as 'Next' and 'Boots', as there was no way of differentiating between relevant Tweets and general usage. For these only mentions of their official Twitter handle were included (i.e. '@NextOfficial' and '@BootsUK'). The final retailer mentions subset comprised 621,946 Tweets from 277,177 users. Therefore, of the cleaned sample, 15.61% of users had interacted with a retailer, but only 0.66% of Tweets were considered retail related.

2.3. Extracting retail centre tweets

In order to create the Twitter catchments, retail centre location and boundary data were obtained from the Local Data Company (LDC), a commercial research consultancy specialising in retail locations. The data consisted of 1287 location centroids and boundaries that defined retail centre spatial extents (see Fig. 1). These were derived from the underlying economic activity defined by the Department for Communities

Table 1
Examples of retailers used for identifying interactions in geo-tagged Twitter data.

	Category	Description	Example retailers
1	Fashion & clothing	General Clothing	Topshop, Primark, River Island
2	Department store		Debenhams, Harvey Nichols, Selfridges
3	Grocery	Groceries, Supermarkets & Food Shops	Tesco, Sainsburys, Waitrose
4	Electronics	Electrical Goods & Home Entertainment	Dixons, Currys PC World, Carphone Warehouse
5	Jewellery	Jewellers, Watches	F·Hinds, Omega, Pandora
6	Leisure	Restaurants and Cafes	Nandos, Yo! Sushi, Starbucks, Costa
7	Discounters	Discount & Surplus Stores	Poundland, Pound Stretcher, B&M Stores
8	Health & cosmetics	Chemists, Toiletries, Make up, Health	Boots, Mac Makeup, Jo Malone
9	Stationary	Books, Arts & Crafts, Stationery	WhSmith, Staples, Cards Galore
10	Toys & hobbies		Lego Store, Build-A-Bear, Disney Store
11	Homeware	DIY, Hardware & Household Goods	Laura Ashley, Habitat, Wilko

and Local Government in 2004. Geographic boundaries for the UK were obtained from the Office for National Statistics (ONS).

By spatially joining the LDC boundaries with the Tweet locations, all Tweets sent from within a retail centre could be extracted. This resulted in 57,028,470 Tweets from 643,575 users, 60.8% of the cleaned sample and 36.2% of users. Mobility patterns for individual users could then be delineated by extracting all unique user ID's from the subset and collecting all further instances where they had shared a location. Aggregating these patterns by retail centre locations and applying density thresholds allowed delineation of a spatial 'catchment' for retail centres across the UK.

Centres contained an average of 198,251 Tweets from 966 unique users. However, the range was large (see Appendix C) with the most active centres being major cities, such as Manchester City Centre (54,250 Tweets), to no data in two centres - Hunstanton and Downham Market in Norfolk. Both of these centres were close in proximity and had been documented as a mobile Internet black spot (The Guardian, 2013). However, there were still 63 centres with fewer than 50 users. All of

Fig. 1. The LDC defined retail centre locations and boundaries in the UK and Greater Manchester.

these were rural centres and likely subject to spatial biases in the data, such as lack of mobile internet and an effect of demographic attributes characterising rural areas. To illustrate this, Fig. 2 shows the average age from the 2011 Census per Output Area and retail centre locations within the UK.

Rural areas are more likely to house more elderly populations, which may go some way to explaining the relative lack of Tweets. This may have implications for using on-the-go technologies to boost town centre vitality, as such technologies may not be suitable for the specific local demographics of some centres.

3. Inferring retail locations with twitter

3.1. Spatial distributions

The spatial distributions of the retail and general Tweet samples were compared to explore initial variations. Unsurprisingly, both Tweet distributions delineated the urban geography of the UK, clustering in major towns and cities and other places of probable high human activity such as roads and train networks. To verify any areas of a greater than expected proportion of retail-related Tweets, the data were aggregated to 2 km × 2 km grid and the location quotient calculated (see Fig. 3). Higher values indicated a higher concentration of retailer interactions.

Fig. 3 highlights areas with higher concentrations of retailer mentions, primarily in town and city centre locations. To better isolate areas of higher than expected retail-related Twitter activity, an advanced form of kernel density estimation (KDE) was utilised.

3.2. Adaptive KDE

In its simplest form, a KDE consists of placing a probability density (kernel) over each observation in a sample. A grid is then overlaid and an estimate of density is obtained at the intersections of the grid. The density estimate is then the average of the densities of all kernels that overlap at a given point. The traditional kernel density estimator for bivariate data can be defined as:

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

where K is the kernel function, and h_i is the smoothing parameter (or bandwidth) for the $_i th$ observation. However, an adaptive KDE method was applied for this investigation since it offers a number of advantages

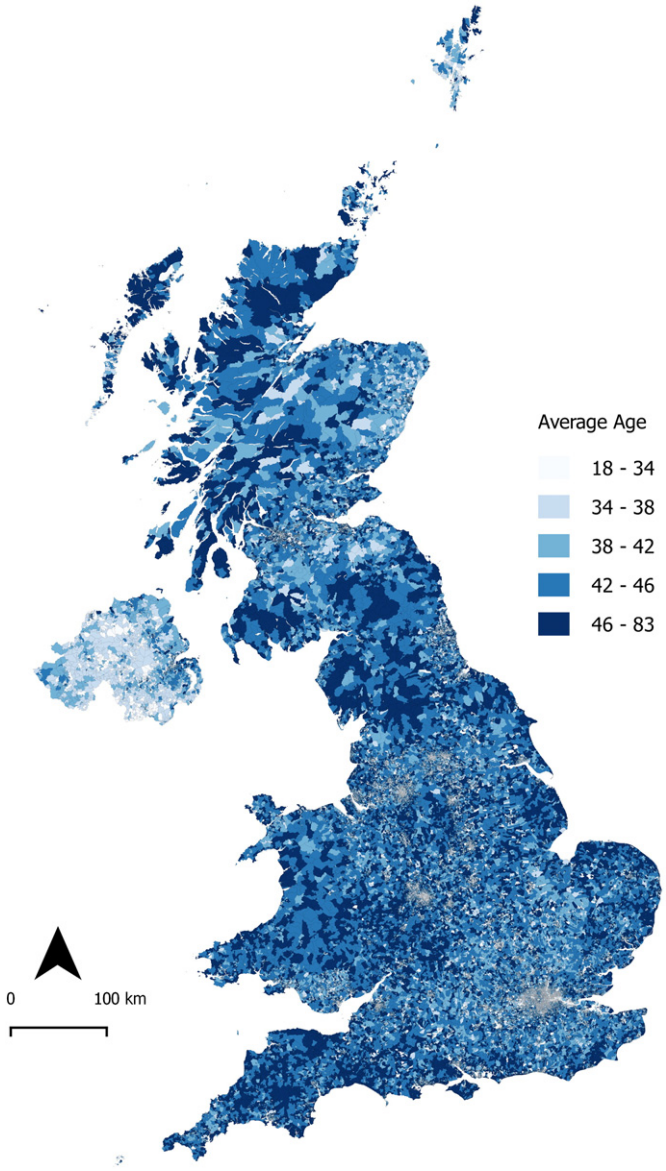

Fig. 2. Average age per Output Area within the UK.

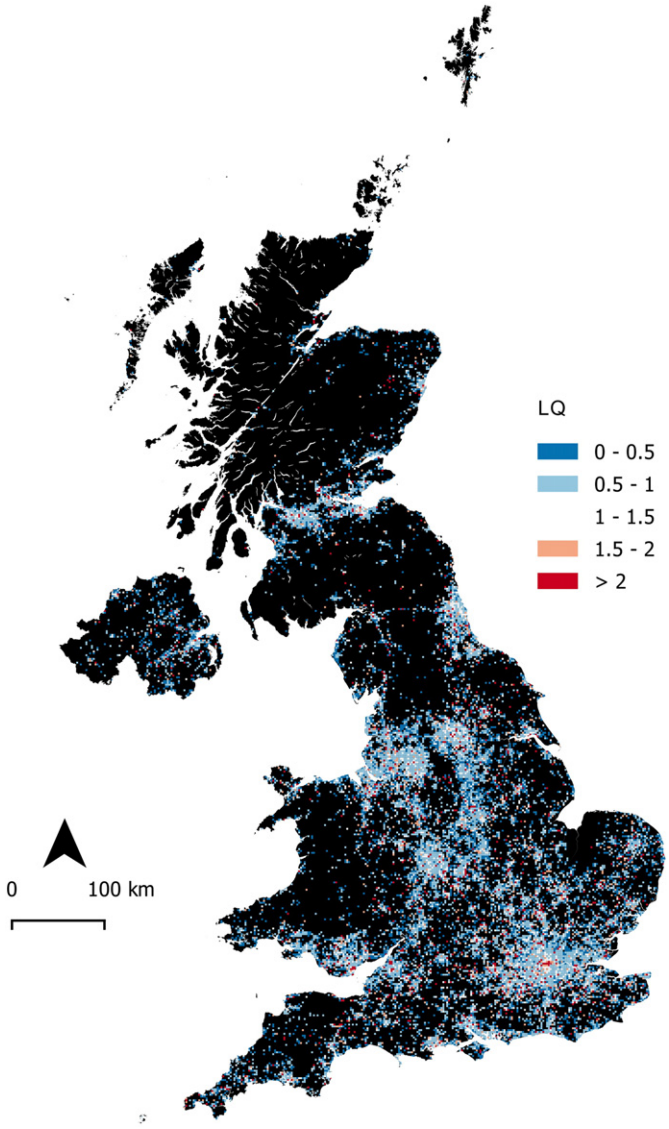

Fig. 3. The location quotient showing concentrations of retailer mentions.

relevant for this context. Firstly, traditional KDE uses a fixed bandwidth value across the data, which can result in under smoothing in areas with few observations and over smoothing in others. In contrast, adaptive KDE allows the bandwidth to vary with the sample data, which reduces this bias. This is achieved by varying the bandwidth inversely with the local volume of data, therefore using a broader kernel over observations in regions of low density (Davies, Hazelton, & Marshall, 2011). The adaptive estimator, as suggested by Abramson (1982), is calculated by:

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

where h_0 refers to the global bandwidth, which is scaled by the product of the inverse square root of the pilot (local) density ($f(x_i)^{-1/2}$) 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 created using the least-squares cross validation (LSCV) approach (Bowman & Azzalini, 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).

Secondly, traditional KDE alone suffers from an inability to normalise data based on an underlying spatial distribution. For example, fixed bandwidths struggle to capture important finer details in densely populated areas when a large amount of smoothing is applied in order to control the noise where the data are sparse (Davies & Hazelton, 2010). However, by applying the relative risk function, adaptive KDE allows us to identify areas of statistically significant fluctuations between density estimations, by taking into account the underlying population density between samples. This is achieved by differentiating case (points of interest) and control (general population) distributions. The resulting relative risk function describes the difference in spatial variations and highlights 'risk' areas (a commonly used tool in describing disease risk across populations; Kelsall & Diggle, 1995b).

3.2.1. Method

The sparr package in R (Davies et al., 2011) was used to carry out the analysis. This contained the necessary functions to identify significant clusters of retail Tweets and calculate contours at a given significance level. A proportional sample of general Tweets (621,946) was randomly selected from the cleaned sample to be used as the control data in the estimation, ensuring no retail Tweet duplicates. To test that this random sample would not affect the resulting density estimations, the frequencies of control Tweets per cell were compared across three samples, using a 1 km × 1 km grid. Frequencies per cell were almost completely positively correlated with coefficients ranging from 0.98 to 1 and p-values of <2.2e-16. This indicated a very minimal difference in distributions between randomly sampled control Tweets.

Overall, three sets of pilot and global bandwidths and density estimations were required; for the pooled (all), case (f) and control (g) data. For the density estimations, a grid size could be specified of which the optimum resolution was considered to be 300 m × 300 m since higher resolutions substantially increased computation times, yet did not significantly improve results. Pilot bandwidths were computed using the LSCV approach and global bandwidths using the OS principle (Terrell, 1990), which utilises the maximum smoothing that is consistent with the estimated scale of the data. Pooled estimates (pilot, global and density) were computed initially as the result from the density estimation was required in the proceeding analyses. This was so that the global bandwidth worked on the same scale in all estimations (see Davies & Hazelton, 2010). Pilot bandwidths were calculated separately for the case and control data in order to assist in preserving the spatial heterogeneity of the samples (Davies & Hazelton, 2010). The global bandwidth then acted as a secondary smoothing multiplier, based on the whole sample.

Finally, the relative risk function was applied in order to identify areas of significant retailer Tweet density. This computed the probability of an event occurring by calculating the ratio of the case and control densities (Bithell, 1991). In order to ensure that the treatment of the two density estimates was identical, the ratio was then log-transformed. To highlight areas of significant retailer Tweet densities, tolerance contours were calculated using the z-statistic-based asymptotic normality test (Davies & Hazelton, 2010) at significance levels of $\alpha = 0.01$ and $\alpha = 0.05$. This determined whether or not a given peak in an estimated surface reflected truly heightened risk or was simply a product of random variation. Therefore, in this context, the tolerance contours delineated areas that had a significantly elevated probability of a retailer interaction occurring in comparison to the distribution of general tweeting. Areas of high retailer Tweet probability could then be identified and analysed.

3.2.2. Results

Fig. 4 shows the adaptive log-relative risk function results for retailer interactions within the UK, with asymptotic tolerance contours delineating areas of significantly elevated density. Table 2 shows the bandwidths used for the computation.

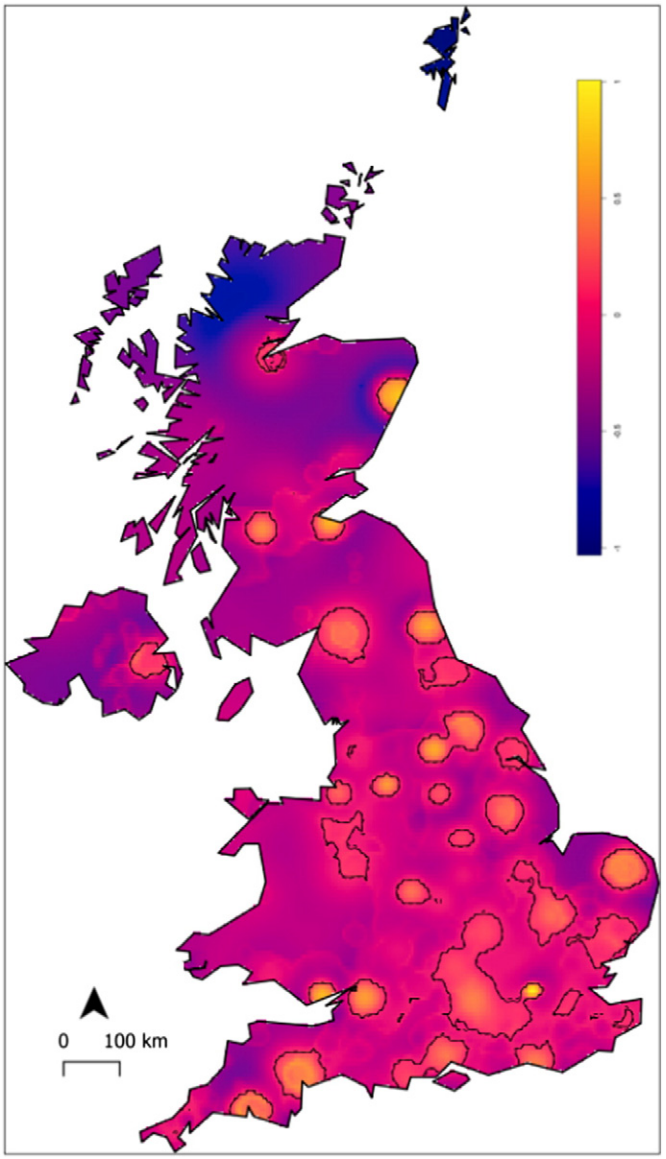

Fig. 4. Adaptive log-relative risk function showing significant fluctuations of retailer mentions on Twitter, within the United Kingdom between December 2012 and January 2014, with asymptotic tolerance contours at 0.05 (dashed) and 0.01 (solid).

Table 2
Estimated bandwidths for UK density estimations.

Tweets	N Nf = Ng	N inside window	Pooled bandwidths (metres)		Estimation bandwidths (metres)	
			LSCV h	OS h	LSCV h	OS h
Retail Tweets (f)	621,946	595,681	7646.12	19,333.06	4883.22	19,333.06
Control Tweets (g)	621,946	576,170	7646.12	19,333.06	6094.09	19,333.06

This methodology was able to delineate significantly elevated areas of retailer interactions by taking into account the control distributions. The denser areas of retailer interactions identified across the UK were able to locate major retail centre locations, primarily major towns and city centres. However, other more minor retailing locations such as Exeter also demonstrated significant activity, on a seemingly wider scale than some major centres. This could have been an effect of the proximity of competing destinations in surrounding areas. For example, the contour surrounding Exeter also incorporated the neighbouring towns Exmouth and Torquay. However, at this scale it was difficult to assess results at a retail centre level of granularity. Therefore, Fig. 5 shows the same estimations on a finer scale for the area of Greater Manchester. Table 3 shows the bandwidths used in this computation.

At this level of granularity, the data were able to delineate retail centre locations on a finer scale. For example, the Intu Trafford Centre demonstrated the most significant fluctuation of retailer interactions, followed by the city centre. Other significant areas were a major town centres such as Bury, Stockport and Leigh. Densities were also then explored across different retail categories (see Fig. 6).

These results demonstrated that patterns varied in line with expectations as to where we would expect different types of retailing activity to occur. For example, fashion and clothing retailer interactions were most prevalent in the Intu centre and the city centre, whilst homeware related activity delineated areas such as out of town retail parks where such retailers reside, most prominently in Ashton-Upon-Lyne and numerous other parks of a similar format. Grocery retailer interactions demonstrated more sporadic distributions of acitivity, as we may expect from the physical distribution of supermarkets.

Overall patterns suggested that the data were able to identify major retail centre locations and elevated areas of retailing activity across different retail categories, to some degree of accuracy. However, a consistent pattern was the inability to identify all centre locations, primarily more rural centres. This is most likely due to the uneven distribution of data across centres due to the biases acknowledged in Section 2.3.

4. Twitter user mobility patterns

The second area of analysis aimed to understand the contributions of geo-tagged Twitter data to mobility flows between centres. Therefore, the spatial footprints of Twitter users were explored using the full dataset, rather than only retail related interactions.

4.1. Method

The package adehabitatHR in R was used to compute user 'catchments' (see Calenge, 2011). This package was originally created to examine the use of space by wildlife, however it offers a number of tools appropriate for this context. For example, it allows for 'home-range' estimation (assessing the area in which an animal lives and moves), which can similarly be applied to estimate mobility 'catchments' for individual users. This can be achieved by documenting all relocations (any alternative geo-tagged location that an entity was documented at) and applying the utilisation distribution method (UD; van Winkle, 1975) to assess these locations at different points in time. This method applied the kernel function principle described in Equation 1, which provided a probability density for an individual to visit any location. The common choice of the "reference bandwidth" (default) was used to estimate the size of the kernels (see Calenge, 2011).

The method estimated the UD in each pixel of a grid superimposed on the locations an individual's Tweets. Each grid cell was 500 m × 500 m, as although the resolution may not have a large effect on the estimates (Silverman, 1986), it created smoother contours. Once the UD was calculated, the density could then be converted into a 'home-range', or catchment, estimate. To do this, contours were calculated to delineate areas of equal density at a given threshold. In this case, catchments were defined as the smallest area containing 70% (primary), 80% (secondary) and 95% (tertiary) of the UD. The primary catchments delineated the more refined, higher probability relocation areas whereas the tertiary catchments aimed to show the full extent of mobility patterns, but remove the top 5% of extreme or anomalous cases.

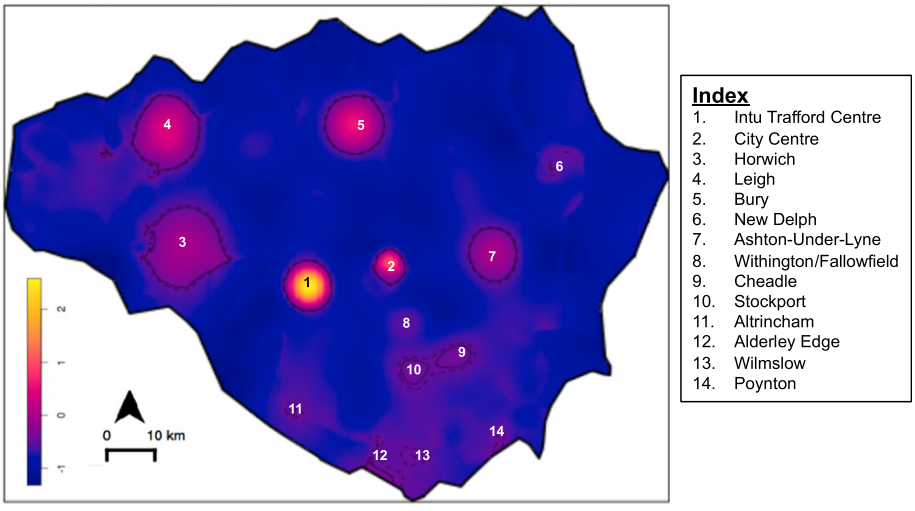

Fig. 5. Adaptive log-relative risk function showing significant fluctuations of retailer mentions on Twitter within Greater Manchester between December 2012 and January 2014, with asymptotic tolerance contours at 0.05 (dashed) and 0.01 (solid).

Table 3
Estimated bandwidths for Greater Manchester density estimations.

Tweets	N Nf = Ng	N inside window	Pooled bandwidths (metres)		Estimation bandwidths (metres)	
			LSCV h	OS h	LSCV h	OS h
Retail Tweets (f)	621,946	13,110	689.84	2006.11	755.31	2006.11
Control Tweets (g)	621,946	10,234	689.84	2006.11	828.49	2006.11

4.2. Results

Fig. 7 shows 4 examples of the catchments produced for retail centres in the UK. Generally, the primary catchments incorporated surrounding towns and cities, whereas the tertiary catchments picked up nationwide movements, illustrating that users had likely travelled during the time period of the sample. There were consistent catchment overlaps between centres, indicating the extent of flows between urban locations in the UK. In particular, Greater London was documented in almost all centres' tertiary catchments. This may be anticipated when considering it is a central destination for multiple types of visits, including business and tourism. However, it is important to note that this type of analysis is bi-directional, meaning that we are unable to ascertain the direction of flow between given locations. For example, catchments including London indicated either that an individual had visited London, or alternatively that they resided in London and visited the given retail centre. Nevertheless, the catchments provided us with interesting insights into the flows between locations. All catchments are available to explore online via the Consumer Data Research Centre web-mapping platform (http://maps.cdrc.ac.uk/).

A consistent finding across catchment patterns was that the extents could be delineated primarily by transport links such as major roads and railways. For example, Banbury's primary catchment could be described by the M40 motorway between London and Birmingham, whereas Llandelio's primary catchment by the M4 across South Wales. Falkirk's catchment was also delineated by the surrounding motorway and major road links and Holyhead demonstrates even if no major roads are available, catchments could be delineated by train links such as between Holyhead and Liverpool. These findings show that flows between centres may be heavily influenced by available transport links in the surrounding areas. For example, the Twitter catchments were able to appreciate the impact of the motorways that may speed up travel times and therefore increase the probability of travel.

These observations could also explain why it appears that many geographically distant areas fall into centre catchments and why they are particularly efficient at picking up large urban centres. For example, although geographically distant, the travel time between centres may be relatively short. This is evident in the Holyhead catchment, where easy accessibility of a ferry port can likely explain why this catchment extends into Ireland. Furthermore, the Twitter data allow for appreciation of the fact that there may be regions with no customers, but probability then increases again in other locations. Finally, the extent of the contours suggests that patterns may also depend on the proximity of surrounding areas. For example, Falkirk exhibited a relatively small overall catchment, which may be a result

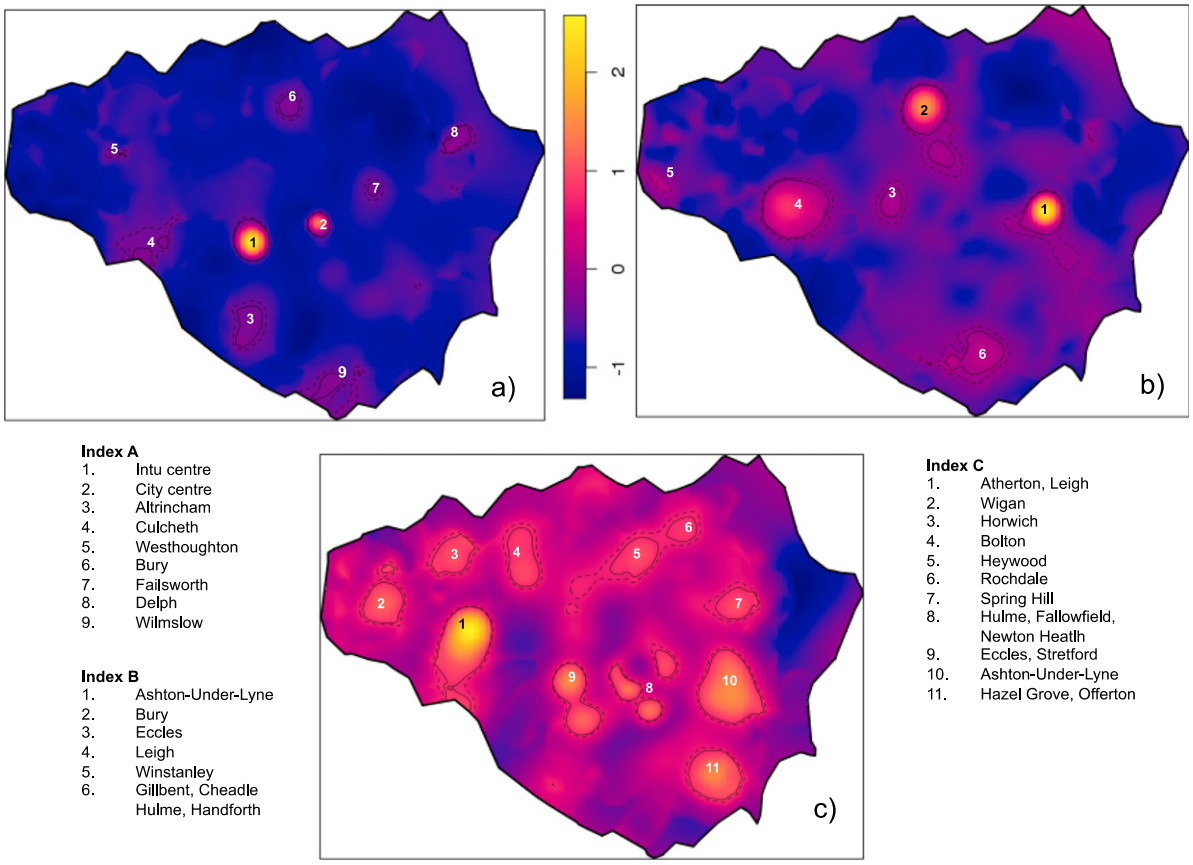

Fig. 6. Significant fluctuations of retailer mentions for a) Fashion and Clothing b) Homeware and c) Grocery retailers.

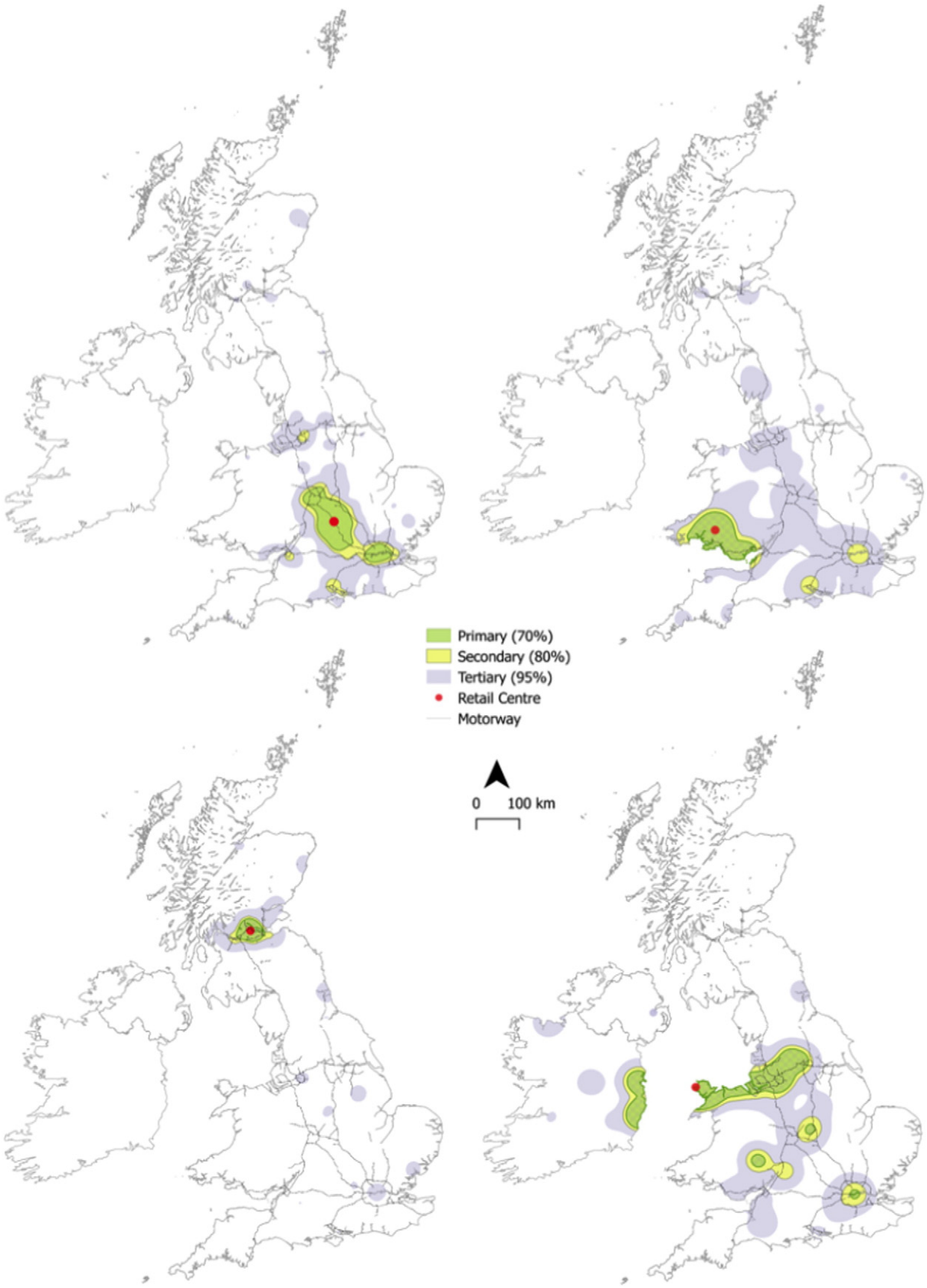

Fig. 7. Primary, secondary and tertiary retail centre catchment estimations for a) Banbury b) Llandelio c) Falkirk and d) Holyhead.

of having considerably less easy accessibility to southern areas. Conversely, in the south, urban centres are much more densely located, which most likely affects the probability of flows between locations.

Fig. 8a and 8b demonstrate these accessibility effects on a finer scale for two centres that are close in proximity; Coleford, Gloucestershire and Monmouth, South Wales. It can be observed that the

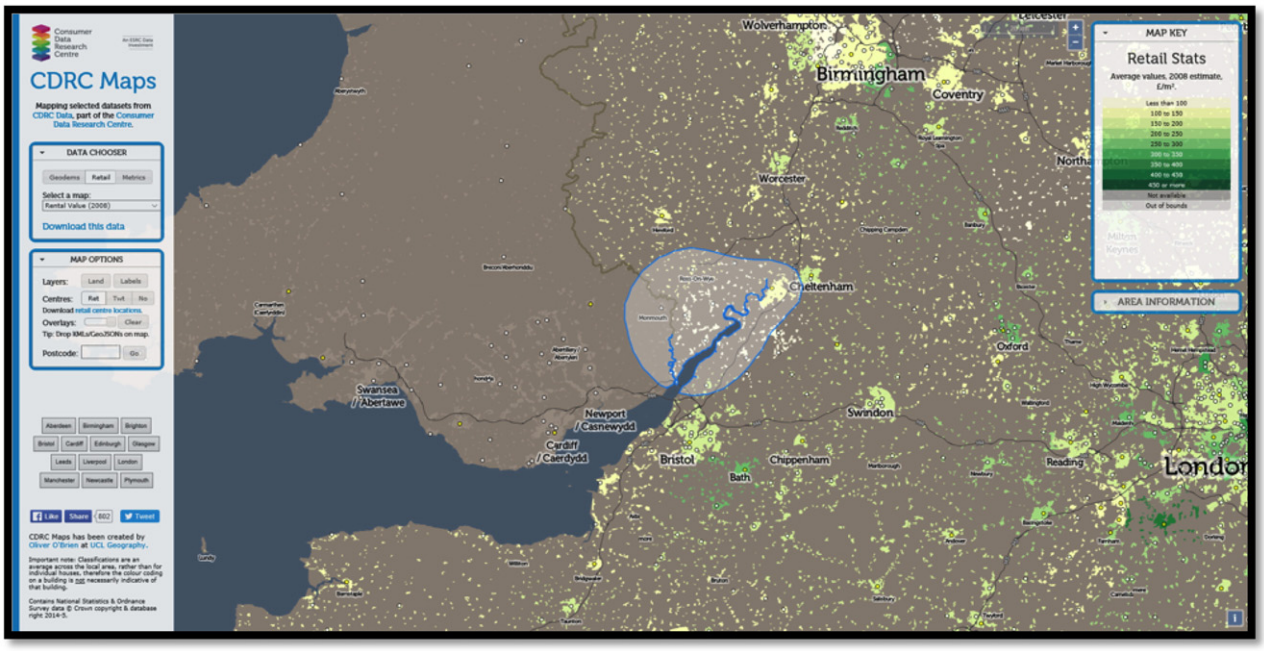

Fig. 8a. Primary catchment estimation for Coleford, Gloucestershire.

catchment for Coleford is primarily skewed towards the nearest neighbouring town of Gloucester. However, this catchment does not include the larger retail destination of Bristol, which is of a similar drive-time away. Alternatively, the catchment for Monmouth incorporates Gloucester, Bristol, and some surrounding areas in South Wales.

Whilst Coleford is closer to Bristol than Monmouth, Monmouth benefits from a more accessible route. For example, it closely links to the M4 motorway, which offers quick access by road to areas such as Bristol and Cardiff. Conversely, the most direct route from Coleford is via multiple B roads and there are no direct train routes, likely making Bristol a less easily accessible option. This demonstrates that the Twitter data were able to account for contextual transportation barriers to mobility flows between retail centre locations.

These findings suggest that overall, the extent of flows between areas will be dictated by the availability of transport links, and therefore ease of accessibility to alternative destinations. Whilst this may not be a surprising finding, using these data-driven catchments provides evidence of the important influence that these accessibility dynamics may have on flows between centre locations. Whilst it cannot be assumed that the alternative locations documented were retail centre oriented, these catchments implicate the likely commuting choices that may be involved in retail centre journey decision-making.

5. Discussion and conclusions

This investigation offered an initial foray into the applications of geotagged Tweets for insights within two areas of retail geography. Results

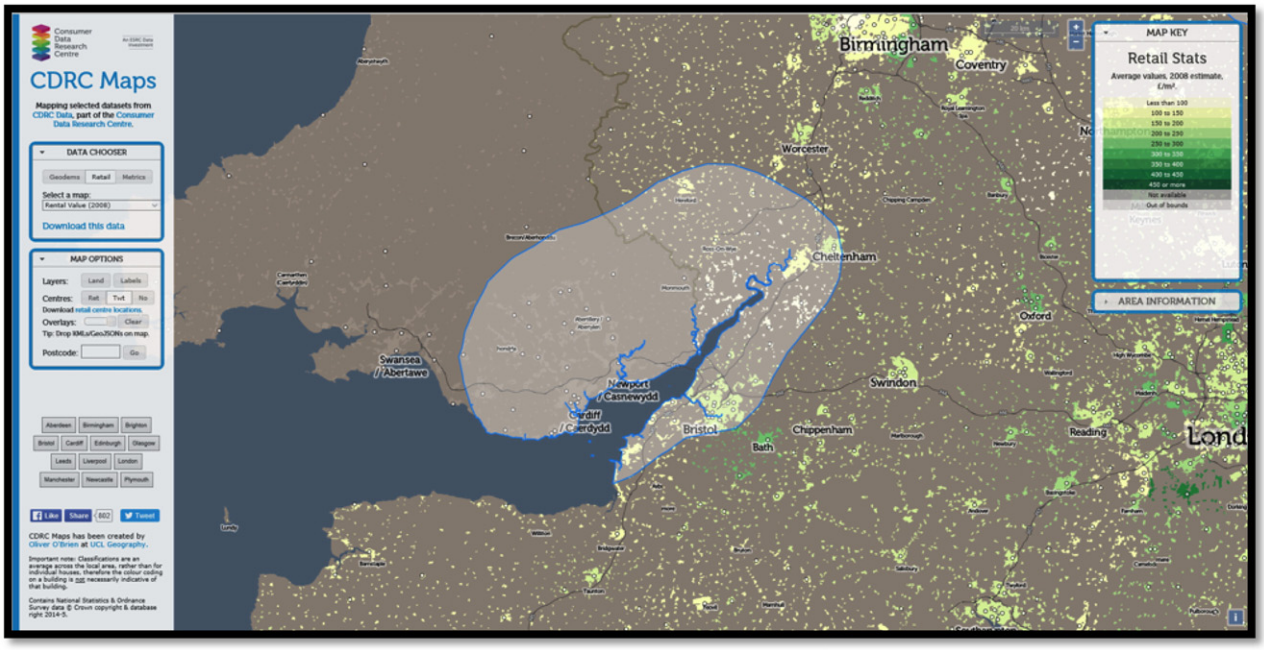

Fig. 8b. Primary catchment estimation for Monmouth, South Wales.

from the location analysis suggested that the data can be successfully implemented to identify areas of elevated retailing activity. Also, prominent locations for different categories of retailing could be identified with some accuracy. However, the data could not be utilised for ascertaining the locations and extents of all centres, likely due to biases in the data causing a substantially uneven distribution of data across different centre locations. For example, areas of density would have been subject only to users who allowed a shared location and also to the spatial and demographic biases acknowledged in Section 3.3. These biases may have implications for using on-the-go technologies to boost town centre vitality, as such technologies may not be suitable for the specific local demographics of some centres. In addition, the advanced KDE methodology applied may have implications for future analyses of population data, such as consumer transactional data. This could, for example, highlight areas of elevated spending independent of population density. This work also demonstrates that both the use of the relative risk function and home-range estimations are likely able to be successfully implemented to the study of a multitude of social phenomenon.

Results from the mobility analysis demonstrated that geo-tagged Twitter data can be utilised to examine human mobility patterns and dynamics in a retail centre context, highlighting the potential influence of accessibility and transport barriers on travel flows between locations. This data driven approach suggests that there may be regions with no customers, yet geographically distant areas such as large urban centres may still fall into catchments due to ease of travel accessibility. Outside of the retailing context, the catchments may also be useful for understanding commuter characteristics. However, it is important to acknowledge that results are only representative of a subset of the Twitter population and not the behaviours of the general population. Furthermore, although using this data-driven approach may provide insight into consumer flows, it does not offer a means of systematically quantifying retail centre catchments due to the biases in the data.

That said, it is hoped that these analyses provide some insight into using social media data as a proxy for the digital engagement of consumers within the UK, in addition to exploring the applications and limitations of applying these data in a retail centre context. Despite limitations, an important advantage of these data are that they are a free and powerful tool that can be applied to any retail centre. It has also been demonstrated here that there is a significant amount of retail centre content that could be further utilised for consumer insights. Future directions, although beyond the scope of this paper, could focus on the analysis of the content of general and retailer Tweets across different retail centres, which may add a dimension of understanding as to the motivations behind behaviours and journeys. Furthermore, the data have the capability to be applied on an international scale, which could have useful applications for understanding the digital footprints of consumers and promoting centre resilience across the wider retail landscape.

Acknowledgements

This research was funded by the Economic and Social Research Council's Consumer Data Research Centre (Grant code ES/L011840/1). Thanks are given to the Local Data Company for providing the retail centre data.

Appendix A. Spam words used to identify spam Tweets and clean the database.

	Spam trigger phrases
1	Earn $ from home
2	Employability skills
3	eCommerce
4	Social media marketing
5	Free info
6	Free investment
7	Income from home
8	Limited time offer
9	No obligation
10	Online marketing
11	Fixed income

Appendix B. Examples of major retailer Twitter usage variations, used to create the 'retailer mentions' subset of Tweets.

No.	Store	Query usage
1	Abercrombie & Fitch	Abercrombie Abercrombie & fitch Abercrombie and fitch Abercrombie&fitch Abercrombieandfitch
2	Cath Kidston	cath kidston cath_kidston cathkidston cathkidston
3	Claire's Accessories	clairesstores
4	Harvey Nichols	Harvey Nichols harvey nicks harvey nics Harveynichols harveynicks harveynics
5	Marks & Spencer	M&S Marks & sparks Marks & spencer Marks & spencers Marks and sparks Marks and spencer Marks and spencers marksandsparks Marksandsparks Marksandspencer Marksandspencers

Appendix C. Top and bottom 10 retail centres for Twitter activity.

Panel A The top 10 UK retail centres for Twitter activity.

1	Manchester	54,250
2	Edinburgh	32,594
3	Liverpool	28,176
4	Glasgow	27,667
5	Leeds	26,919
6	Newcastle Upon Tyne	22,369
7	Cardiff	20,577
8	Brighton and Hove	18,802
9	Sheffield	17,431
10	Nottingham	17,102

Panel B The bottom 10 UK retail centres for Twitter activity.

1	Downham Market	0
2	Hunstanton	0
3	Kings Lynn	4
4	North Seaton Industrial Estate	14
5	Ottery St. Mary	15
6	Headcorn	17
7	Ramsey	17
8	Ilminster	18
9	Eccleshall	23
10	Southlands Road, Bromley	24

References

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

Adnan, M., Lansley, G., & Longley, P. A. (2013). *A geodemographic analysis of the ethnicity and identity of Twitter users in Greater London* (pp. 1–6)*Proceedings of the 21st conference on GIS research UK (GISRUK)*.

Birkin, M., Clarke, G., & Clarke, M. (2010). Refining and operationalizing entropy-maximizing models for business applications. *Geographical Analysis*, *42*, 422–445.

Bithell, J. F. (1991). Estimation of relative risk functions. *Statistics in Medicine*, *10*, 1745–1751.

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.

Brandwatch (2015). *Brandwatch Report/Retail Report: An analysis of retail brands through the lens of social media.* ([Online] Available from: https://www.brandwatch.com/wp-content/uploads/2015/01/Brandwatch-Retail-Report.pdf. [Accessed: August 2015]).

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

Calenge, C. (2011). *Home range estimation in R: The adehabitatHR package. [Online] Office national de la classe et de la faune sauvage, Saint Benoist, Auffargis, France.* (Available at: ftp://mi.mirror.garr.it/pub/1/cran/web/packages/adehabitatHR/vignettes/adehabitatHR.pdf. [Accessed: 13th May 2015]).

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. J. (2011). *Exploring millions of footprints in location sharing services* (pp. 1–8)Proceedings of the fifth international conference on weblogs and social media, Barcelona.

Davies, T. M., & Hazelton, M. L. (2010). Adaptive kernel estimation of spatial relative risk. *Statistics in Medicine*, *29*, 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*, 1–14.

Dolega, L., Pavlis, M., & Singleton, A. (2016). Estimating attractiveness, hierarchy and catchment area extents for a national set of retail centre agglomerations. *Journal of Retailing and Consumer Services*, *28*, 78–90.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*, 260–271.

IRUK (2016). *Internet retailing UK Top 500.* ([Online]. Available from: http://internetretailing.net/iruk/. Accessed: May 2016).

Kelsall, J. E., & Diggle, P. J. (1995b). Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine*, *14*, 2335–2342.

Lansley, G., & Longley, P. A. (2016). The geography of twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96.

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*, 61–77.

Longley, P. A., & Adnan, M. (2016). Geo-temporal twitter demographics. *International Journal of Geographical Information Science*, *30*(2), 369–389.

Mckinsey & Company (2011). *Retail coops: Staying competitive in a changing world.* ([Online]. Available from: http://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/strategy/mckinsey%20on%20cooperatives/pdfs/mck_on_cooperatives-retail_coops_staying_competitive_in_a_changing_world.ashx. [Accessed May 2016]).

Mequoda (2015). *Subject Line Spam Trigger words.* ([Online]. Available from: http://www.mequoda.com/articles/audience-development/subject-line-spam-trigger-words. [Accessed: October 2015]).

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (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* (pp. *9*).

Nielsen Media Research (2014). 80% of UK users access Twitter via their mobile. *[Online] Official Twitter blog, February 2014* (Available at: https://blog.twitter.com/en-gb/2014/80-of-uk-users-access-twitter-via-their-mobile. Accessed: July 2015).

Pew Research Center (2015). *Demographics of key social networking platforms.* ([Online]. Available from: http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/ [Accessed: October 2015]).

Retail Research (2016). *Online retailing: Britain, Europe, US and Canada.* ([Online]. Available from: http://www.retailresearch.org/onlineretailing.php Accessed: August 2016).

Silverman, B. W. (1986). *Density estimation for statistics and data analysis (Vol. 26)*. CRC press.

Terrell, G. R. (1990). The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, *85*, 470–477.

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]).

Twitter (2015). *Twitter usage/company facts.* ([Online]. Available from: https://about.twitter.com/company. [Accessed: September 2015]).

Van Winkle, W. (1975). Comparison of several probabilistic home-range models. *The Journal of Wildlife Management*, 118–123.

Wang, A. H. (2010). *Don't follow me: Spam detection in twitter. In Security and Cryptography (SECRYPT)* (pp. 1–10)*Proceedings of the 2010 International Conference* (IEEE).

Weltevreden, J. W. (2007). Substitution or complementarity? How the Internet changes city centre shopping. *Journal of Retailing and Consumer Services*, *14*, 192–207.

Wrigley, N., Lambiri, D., Astbury, G., Dolega, L., Hart, C., Reeves, C., ... Wood, S. M. (2015). British high streets: From crisis to recovery? *A Comprehensive Review of the Evidence* ([Online]. Available from: http://eprints.soton.ac.uk/375492/. Accessed: September 2015).