From Big Noise to Big Data: towards the verification of large datasets for understanding regional retail flows

Robin Lovelace, Philip Cross, Martin Clarke and Mark Birkin

Abstract

There has been much excitement amongst quantitative geographers about newly available datasets, characterised by high volume, velocity and variety. This phenomenon is often labelled as 'Big Data' and has contributed to methodological and empirical advances, particularly in the areas of visualisation and analysis of social networks. However, a fourth v — veracity (or lack thereof) — has been conspicuously lacking from the literature. This paper sets out to test the potential for verifying large datasets. It does this by cross-comparing three unrelated estimates of retail flows — human movements from home locations to shopping centres — derived from the following geo-coded sources: 1) a major mobile telephone service provider; 2) a commercial consumer survey; and 3) geotagged Twitter messages.

Three spatial interaction models also provided estimates of flow: constrained and unconstrained versions of the 'gravity model' and the recently developed 'radiation model'. We found positive relationships between all data-based and theoretical sources of estimated retail flows. Based on the analysis, the mobile telephone data fitted the modelled flows and consumer survey data closely, while flows obtained directly from the Twitter data diverged from other sources. The research highlights the importance of verification in flow data derived from new sources and demonstrates methods for achieving this.

Introduction

Much has been written about 'Big Data': definitions, ethics and the methodological challenges the phenomenon poses (Boyd and Crawford 2012; K. Davis 2012; Madden 2012). There has also been speculation that it could revolutionise Regional Science, Geography and related fields (Arribas-Bel 2014; Kitchin 2013). Many new datasets are geographically referenced: smart ticketing systems monitor transport flows at known locations; mobile phone logs are linked to specific masts. Yet the spatial element of the Big Data revolution has yet to be fully articulated with reference to empirical foundations related to data quality, notwithstanding excellent theory and position papers on the subject (see Rae and Singleton 2015).

Amongst the excitement surrounding Big Data overall, there has been little time to pause for thought about the kinds of application are most suited. Less still has been written about where big datasets are most (in)appropriate for different purposes (Crampton et al. 2013). With the growing volume of data available, there has been a tendency to proceed uncritically with the analysis. This has resulted in many beautiful visualisations and new insights (King, Pan, and Roberts 2014) but little in the way of systematic evaluation of the quality of large datasets. Whilst veracity in Big Data has been discussed previously,[1] the associated concepts and methods are largely undefined in the peer-reviewed academic literature, particularly in relation to spatial data. This paper sets out to fill this gap by articulating the concept in the context of the estimation of retail flows — movement of people for shopping.

Continual questioning and testing for the veracity of data and models is central to the scientific method in quantitative geography: "the geographer-scientist must continually test his theories and models, and he must not expect certainty" (Wilson 1969, 227). We argue that, in the era of Big Data, this commitment to testing should not be forgotten. Indeed, we argue that veracity becomes even more important, due to the increased potential to arrive at erroneous conclusions based on the sheer volume of data. As Taleb (2012, 430) provocatively puts it, Big Data brings "cherry-picking to an industrial level". Seeing past the sensationalist

[1] See http://www.ibmbigdatahub.com/blog/measuring-business-value-big-data

hyperbole of Taleb's statement, the wider point is that new datasets bring about new risks as well as the much trumpeted opportunities. Without rigorous tests, the credibility of new datasets risks being undermined by faulty conclusions from unverified sources.

In this context, we demonstrate methods for the verification of Big Data, based on an empirical foundation: three large, unrelated datasets on movement patterns within Yorkshire, UK. The aim is to demonstrate and implement methodologies for verifying Big Data to begin to evaluate the strengths and weaknesses of diverse sources for geographical modelling. To achieve this aim we pursued the following objectives:

- Convert the three disparate datasets into a single format: an origin-destination 'flow matrix' representing the amount of flow (in terms of human movement) from 120 residential zones to 79 known retail centres of varying sizes.
- Develop and implement models of expected flow, using the classic gravity model and the more recently published radiation model (Simini et al. 2012).
- Make comparisons between the retail flows which result from each of these 6 sources — three empirical and three theoretical — in terms of model-data fit.
- Discuss the implications of this research for understanding the quality of different datasets and for processes of verification in research involving Big Data.

The lack of cross-comparison and validation in the emerging field of Big Data for geographical modelling is problematic because large datasets are inherently diverse. For example, a huge dataset comprising social media scraped intermittently from the internet will have very different attributes than a smaller but more systematically collected dataset on travel habits from a market research survey. Yet each dataset may be designated as 'big'. This paper discusses this issue with reference to three large yet unrelated datasets collected from Yorkshire, UK: location data from a major mobile phone operator; a year's worth of geotagged tweets; survey data on demographics and shopping provided by the consultancy Acxiom.

Each of these sources of information could also be defined as 'big' in some way. Yet, without a framework to assess their relative merits, it is difficult to assess the applications for which each dataset would be most and least suited. There is a clear need for new sources of information to inform research on spatial behaviour. From a retail analysis perspective, for example, datasets to inform and calibrate models have been patchy. Whilst developments in modelling have been impressive — see Birkin, Clarke, and Clarke (2010) for a recent review — the data to support such advances has been elusive.

Until recently, the vast majority of data on retail flows (origin-destination matrices) have only been available from two sources, each with its own limitations:

- Retail loyalty card data: this source is limited to a specific retailer, providing only a partial summary of consumer behaviour from a potentially non-representative sample.
- Market research data: this source (exemplified by a dataset provided by Acxiom in this research) tends to record solely a person's main retail destination. As we shall see, this is a major limitation because most consumers shop at several locations.

Defining Big Data

This paper reflects on an often-cited definition of Big Data as newly available information that is high in volume, velocity and variety (Laney 2001).[2] To this list we advocate the addition of a fourth v: veracity, or lack thereof. Although such definitions are frequently mentioned in talks on the subject, criteria defining whether or not a dataset is 'Big' have seldom been explored in detail. Rather than focussing on normative definitions of Big Data and strict criteria for inclusion, we base our definition on how the term is used in practice. Big Data is, in practice, an umbrella term capturing unconventional datasets that are difficult to

[2] This understanding suggests that Big Data should more correctly be called 'V data'. It would seem difficult to dislodge the currently fashionable term, but the three (now four) v's are certainly worth mentioning.

analyse using established methods. Often this difficulty relates to size but the form, format and complexity are equally important.

The four v's provide a basis for considering the relative merits of large datasets. The consequences of each attribute (or its absence) for the types of application for which 'Big' datasets are suited is rarely discussed. We thus start from the premise that each of the aforementioned attributes of Big Data can provide both advantages and disadvantages to the researcher. One 'big' dataset may be completely different from the next in terms of its relative merits.

Whereas Laney (op. cit.) is primarily concerned with data management for e-commerce, we consider here the advantages and disadvantages of volume, velocity, variety and degree of veracity in the context of contemporary spatial modelling. Each has important consequences for the applicability of the datasets for the real world.

The *volume* of data available is clearly related to its spatio-temporal coverage (in terms of the frequency of data and its geographic density). On the other hand, highly voluminous data may be of relatively low quality and contain much noise. Separating the signal from the noise in a highly voluminous dataset may be a difficult and time consuming process, meaning that volume is not always a benefit.

High *velocity* implies that the data is collected in real time and is highly dynamic in that it has high temporal (second-by-second) resolution. This is advantageous for applications exploring behaviour that is highly dynamic. Yet velocity has drawbacks: the processes observed may only be ephemeral and not represent overall patterns. An example would be someone who tweets only when they visit new locations, providing an unrepresentative sample of travel behaviour.[3] Real time monitoring systems to process output from the 'fire hose' requires more computational work than a static dataset (Driscoll and Walker 2014). Processing large datasets 'live' has overheads and may only be necessary for informing real-time decision making applications, such as disaster response. Therefore static datasets are used in this study.[4]

Variety can be used to refer to the diversity of information within a single 'big' dataset (intra-dataset variety). It can also refer to the diversity of datasets that fall within the Big Data umbrella (inter-dataset variety). For a single dataset to be diverse implies that it may be able to provide insight into many processes within a single phenomenon, adding nuance to our models. Yet variety can also be associated with a lack of focus: some information about many aspects of a system may not be as useful as highly concentrated information about one specific topic. An example of intra-dataset 'variety' is provided by loyalty card data, which contains a range of information from one purchase per year through customers whose transactions are recorded 5 times or more a week.

Veracity refers to the degree of truthfulness associated with a dataset. We argue that it is **lack of** veracity that is most associated with Big Data, compared with more conventional official datasets used in geographical research. Hence veracity (and the related term of verification), will be used here as an expression of the desire to establish broad parameters of data reliability. Large datasets sometimes lack the quality assurances associated with smaller datasets taken from official sources. Some authors and communities prefer the use of validation for the assessment of model results against real world patterns, reserving verification to refer to rigour in the design and implementation of software codes. In the current context, verification is a better expression of the desire to establish broad parameters of data reliability. The question of whether the sheer volume of 'big' datasets can compensate for the lack of verification has yet to be resolved and is clearly context-dependent.

Veracity, or lack thereof, is probably the least discussed of the v's. To what extent is Twitter data representative of a given population? How do we know that the data provider has not pre-filtered commercial datasets before making it available to researchers? These types of question, we argue, are neglected in the field and deserve robust answers.

[3] Data collected via the fitness app Strava, which records users' position over time using the GPS receiver contained in most modern smartphones, is another good example of this time bias. Routes are more likely to be recorded during exceptional times rather than the more frequent everyday patterns due to the competitive nature of the app.

[4] Note there is often a link between temporal velocity and spatial specificity. Frequent data collection can often be allocated to more specific points on the map (as with geotagged tweets) although this is not always the case (as with raw mobile telephone data).

The four v's are useful for identifying which datasets are most suitable for different applications *a priori*. For example, a high volume of data should normally be advantageous, but this could be mitigated by greater velocity or additional variety in a less voluminous dataset. Ultimately, however, *empirical* evidence must be used as the 'gold standard' of datasets' suitability for different applications. In, practice the generation of a complete and truthful origin-destination matrix for shopping in the Leeds region is not currently achievable. This explains the focus on cross-validation of the data sources and model outputs in this paper. Moreover the experience of using diverse data to explore the same question is itself revealing, illustrating ways to cross-validate seemingly incommensurable datasets. The research question used to assess each dataset is: how well can each dataset be used to model flow to retail centres?

The methods adopted here demonstrate the benefits of cross-validating large datasets as a step towards the verification of Big Data and highlight some methods towards achieving this aim. The processes of geographical aggregation and filtering demonstrated in this paper illustrate that, with the appropriate analysis, large datasets that seem very different in the first instance can be made commensurable. The wider point is that no single large dataset should be considered to be 'true', meaning that cross-validation of data sources and model outputs, so often lacking in Big Data research, is a critical stage in the process.

A more theoretical contribution of this paper is a critical assessment of Big Data with an emphasis on veracity which is under-explored in the literature. We demonstrate that, although diverse, it is possible to perform quality checks on large datasets through cross-validation. We use the empirical results to argue that it is debatable whether sources that meet the other three criteria of Big Data, but not the 'veracity test', should be referred to as Big Data at all. 'Big Noise' may be a more appropriate label before cleaning, where the real value of large datasets is added. This understanding can help researchers identify which datasets are most in need of verification and cleaning and which can be used in a near-raw state. The distinction between Big Data and Big Noise also provides an indication of the amount of work needed on a particular dataset. However, even after processes of cleaning, filtering and aggregation (outlined in the Methods section), Big Data cannot be guaranteed to have veracity. As illustrated in the Results, it may only be with the cross-comparison of Big Data that the degree of veracity or otherwise can be ascertained.

Data

For some retail organisations, customer interaction data has been relatively easy to acquire. For example, in the automotive industry, vehicle registration has allowed individual sales to be tagged to unique customer and dealer locations (Birkin et al. 1996). Other sectors have now begun to assemble valuable intelligence about customers from loyalty card schemes (Humby, Hunt, and Phillips 2008). While loyalty cards are themselves an emerging Big Data source, they have their disadvantages. These include coverage (relating to a single retailer), sample size (not all customers are cardholders), and availability. Loyalty cards are therefore problematic as a Big Data source for *academic* research on local economic activity.

The classic source of intelligence for the retail industry has been market research.[5] Traditionally, the resulting datasets are based on small yet representative stratified random samples. This approach is still popular for monitoring political and social attitudes and behaviour and, with the addition of geodemographic profiling, it is possible to infer social and economic characteristics of respondents (V.-W. Mitchell and McGoldrick 1994). Yet with this approach temporal movement patterns cannot be inferred and the same issues that surround any door-to-door sampling apply. Retailers within a sector may be willing to syndicate larger samples yet the outputs are often only available at a coarse spatial resolution inaccessible and expensive to acquire for third party organisations.

Data on origins and destinations consisted of 79 shopping areas served by 120 residential zones. After processing, each input dataset was aggregated into the same format: a 'flow matrix' consisting of 120 rows, representing residential origins, by 79 columns, representing shopping destinations. The three input datasets were (as summarised in Table 1, below):

[5] This can be seen in the growth of dedicated market research corporations such as MORI and Verdict.

A dataset of geotagged Twitter messages ('tweets') was collected using the Twitter Streaming Application Programming Interface (API). The messages were collected during 445 days between 2011-06-22 and 2012-09-09. The full dataset consisted of 992,423 tweets originating in West and North Yorkshire. Each record in the dataset represents one tweet, including time-stamp, user id, message text and coordinates of the location where the tweet originated. Because our 'listener' was set-up with a geographical filter, only geotagged tweets were collected. Twitter data have been used for a range of purposes in academia, including economic forecasting (Bollen, Mao, and Zeng 2011), analysis of political sentiments (Bermingham and Smeaton 2011), and tracking disease (Signorini, Segre, and Polgreen 2011). Inspired by such research, this paper seeks to explore Twitter data as an alternative to more established sources of shopping behaviour with an emphasis on Twitter's veracity.

A travel flow matrix derived from mobile telephone records was provided by a major telephone operator. The telephone data were collected between 1st February and 28th March 2014. Although the mobile service provider did not reveal the precise number of users/mobile devices, the data suggest this number is ~750,000. (see Table 1). The data represents home location and frequency of trips between major retail centres across Yorkshire. The dataset is not 'geolocated' in the sense that the Twitter data is (with one set of coordinates per connection). Instead, mobile telephone data provides information about the mast or Wi-Fi network to which each phone is connected at any point in time. The spatial resolution is thus not high but sufficient to infer movement behaviours to retail centres: the spatial resolution of mobile phone mast data depends on the density of masts in the local area. In urban areas (where most retail centres are located) the spatial resolution is comparatively high. In rural areas, by contrast, the nearest mast may be several miles away and spatial resolution is comparatively low.

An individual-level **geolocated survey dataset on shopping habits** was provided by the consultancy Acxiom who do surveys across the UK, collecting ~1 million records yearly. The data has a fine spatial resolution (full postcode) and many attributes of interest for market research. Specifically, the question on "Where do you shop most often for non-food goods like clothes, shoes, jewellery etc.?" provides insight into the centre most associated with each area. A limitation of this dataset is that it only identified the most frequently visited destination, unlike the Twitter and mobile phone data, which provides multiple destinations for each person.

Each of these datasets has advantages and drawbacks due to the nature of the quantitative and qualitative information they contain and how they are collected, in terms of penetration and geographical and temporal resolution. These are described in some detail below. It was the task of the analysis phase to identify what these were and their consequences for the utility of each dataset for modelling spatial behaviour.

Table 1: Summary of the input datasets. Note 'N. flows > 0' refers to the number of 'flowlines', O-D pairs, for which a positive flow was recorded, of a possible 9600 records and 'N. observations' refers to the number of flows after filtering.

Data source	Timeframe	N. observations	N. users	N. flows > 0
Twitter	2011-06-22 - 2012-09-09	15730	8740	1250
Mobile telephone records	2014-02-01 - 2014-03-28	751580	?	5445
Consumer survey	2011	21552	21552	1021

Penetration and content

An important consideration related to sample bias is 'penetration' — in this context the uptake of a particular data collection technology within the population (Hilbert 2013). Penetration is analogous to the 'response rate' in surveys and refers to the proportion of the population who are users (or frequent users) of the data-collecting system. The penetration of mobile phones, for example, is substantially higher than the equivalent metric (sample size) for the Acxiom survey data. This was despite the fact that only one mobile telephone network (of several operating in the area) was used, illustrating the sheer volume of this data source. Twitter data has the lowest penetration of the sample population of the three datasets because a) many people do not use Twitter and b) only a few percent of tweets are geotagged. Exploring this issue a little further will help shed light on the findings, presented in the Results section.

Mobile telephone penetration in the UK has reached 93% (Ofcom 2012), and in Yorkshire and Humberside, a region incorporating the study area, a survey by YouGov suggested nearly 99% of adults own a mobile phone. In addition, the survey found that 90% of UK residents aged between 18 and 24, and 95% of those aged 55 or over own a mobile phone (YouGov 2011). By contrast, the uptake of Twitter is relatively low at only a fifth of the UK population (EMarketer 2014).

The age dependency of Twitter use is stronger: 24.7% of 18-24 year-olds regularly tweet. This percentage falls to 4.5% for UK residents between 55 and 64 years of age, and to 2.1% of those over 65. Based on these statistics, mobile phone datasets are more representative of the UK population when compared to Twitter data. The activity of a mobile phone owner at any time is not reported, whereas the activity of a Twitter user can sometimes be inferred from the publicly available content of the tweet. The Acxiom dataset has an intermediate coverage between mobile phone and Twitter data.

Geographical resolution

Although 80% of UK Twitter users are active on a mobile device, only 1-2% of all tweets are geotagged. We used only geotagged tweets with exact coordinates, indicating location by a GPS or Wi-Fi connection. An additional complication is that Twitter users who enable geotagging may not be representative of Twitter users overall, let alone all people (L. Mitchell et al. 2013). Further, people who send messages very rarely or continuously can bias the results of social media data. This was resolved by excluding individuals who had sent fewer than 20 or more than 10,000 tweets during the data collection period.

Conversely, a mobile phone is connected permanently to a radio mast, and the service provider knows the location of that mast. Due to variable mast density, the precise location of a mobile phone can be hard to obtain, and there is potential overlap where some areas may be covered by adjacent masts. In addition, a mobile phone may not connect to its nearest radio mast if the capacity of that mast is already used (Nanni et al. 2014). Nevertheless, for the purposes of this study, the location of a mobile phone inferred from these data is sufficient to track movement between the suburbs of a city. It should be noted that the billing address of the owner of a mobile phone is assumed to be their home address. Clearly, this could be misleading.[6] An additional issue for mobile phone location data derived from mast connections is that behaviour must be derived from the connection times at certain locations, rather than precise coordinates at a specific time. If a mobile phone is turned off, loses signal or connects sporadically to different masts, this may bias the data in unknown ways.

Temporal resolution

The time at which Twitter users are most active varies throughout the day, with the majority of tweets in the UK being sent in late morning and early afternoon, i.e., over lunchtime (360i report, July 2013). In contrast, since a mobile phone is almost always connected to a radio mast, its location can be estimated most of the time.

This study uses network data for the period between 1st February and 28th March 2014 obtained from a major mobile phone service provider. The data have been queried, pre-processed, anonymised, and extrapolated by the service provider prior to this analysis. A disclosure control process, using stochastic rounding, has also been applied to further anonymise the data.

The Acxiom dataset has the lowest temporal resolution, with only one snapshot typically provided to clients per year. For the purposes of this study, we use a snapshot year of Acxiom data from 2010, the most recently available dataset. Because the annual survey is ongoing it would be possible to obtain more up-to-date records, pending commercial restrictions. Because the Acxiom data is not continuous (i.e. it has low 'velocity',

[6] For example, business mobiles may be registered to a user's place of work rather than their home; phones owned by parents but operated by children who move away from home may also cause problems.

as defined above), one could question whether it should be classified as 'Big Data' at all. We refrain from entering this debate, instead acknowledging the contested nature of the term and emphasising its utility as a 'catch all' phrase from referring to new, unusual and non-official datasets. The continuous nature of the mobile phone and Twitter datasets, combined with the ongoing production of the survey dataset, means that there is potential for future work to assess the rate of growth and capacity to represent dynamic processes in each dataset.

Method

The first challenge we faced was the sheer diversity of information available in each of the three datasets. To tackle this problem, a single origin-destination matrix was used to ensure the resulting flows could be directly compared. The most scattered dataset spatially was the UK's geolocated tweets, which were first reduced in number with a spatial filter and then aggregated up to the level of postcode districts. Similar techniques were employed to bring the mobile phone and Acxiom datasets into a suitable form.

Flows from mobile phone data

The data are generated whenever a mobile device is connected to a radio mast. The service provider knows the location of the mast to which each device is connected and the period the device is attached to that mast, referred to here as the 'dwell time'. An 'event' is triggered every time a device moves between masts or Wi-Fi networks, a telephone call or text is sent or received, or a connection to a Wi-Fi network is made, generating a chronology of attachments (Nanni et al. 2014). This leads to a detailed picture of the locations visited by the owner of a device and an understanding of their 'activity space'. The activity which is undertaken by the owner of the device at any point in time is unknown.

The first stage of the analysis involved the identification of retail centres in the study area. The centres were identified using a commercial product called RetailVision which was supplied by the retail consultancy GMAP. Information was acquired for 79 locations in the following postal areas: BD (Bradford), HX (Halifax), HG (Harrogate), HD (Huddersfield), LS (Leeds), WF (Wakefield), and YO (York). The number of retail outlets for each of the identified centres has been used as a measure of its 'attractiveness'. In the next step 'visits' were defined as connections to towers in the retail centre made between 7am and 5pm with a dwell time of between 30 minutes and five hours. These rules target shoppers but may also pick-up on non-shopping activity such as part-time work. The 'dominant centre' for each person was that which was most visited in the period of data collection, although flows to other centres were also registered. Where two or more centres were visited equally, the centre with the greatest number of outlets was chosen.

For all 'shoppers' living within the 120 postal districts of the study area, a dominant retail centre has been identified. This results in a 120 row by 79 column matrix with each cell representing the number of shoppers in each postal district for whom a particular centre is their most dominant.

The matrix is modified to provide an indication of the number of customers in each centre. This is obtained by dividing the cells on each row by the sum of that row, and multiplying the result by the number of households within the postal district, acquired from the 2011 Census household estimates for postcodes in England. A retail flow can therefore be interpreted as the number of households in a residential area who are customers of a retail centre.

Flows from Twitter data

The general methodology used for the Twitter data is reported in Lovelace et al. (2014). Individual tweets were georeferenced and defined as `SpatialPointsDataFrame` objects in R. The dominant location during 'home hours' was assigned as the 'home' of each user and these were then linked (via user id) to 'shopping tweets' (defined below) to infer travel behaviour. Visits to multiple destinations by the same user were recorded.

Various modifications were made to the approach, to maximise the probability that destinations were related to the topic under investigation: shopping. Different definitions were tested and what is presented is the result of what was in fact one of the simplest definitions of a 'shopping tweet': any message sent outside of a user's home area, close to a shopping centre (within a 500m radius — different buffer sizes were tested, as illustrated in Fig. 1), during prime shopping times (10:00 and 17:00 on Saturdays). This crude spatio-temporal filter was favoured over more sophisticated techniques of tracking 'dwell time' (described above for mobile phone data, but unsuitable here due to the asynchronous nature of tweets), keywords related to shopping (an unreliable proxy of behaviour) and frequency of tweets, with goal of simplicity and reproducibility in mind. The method for identifying 'home tweets' relied on repeat messages from the same place during non-work hours (Lovelace et al. 2014).

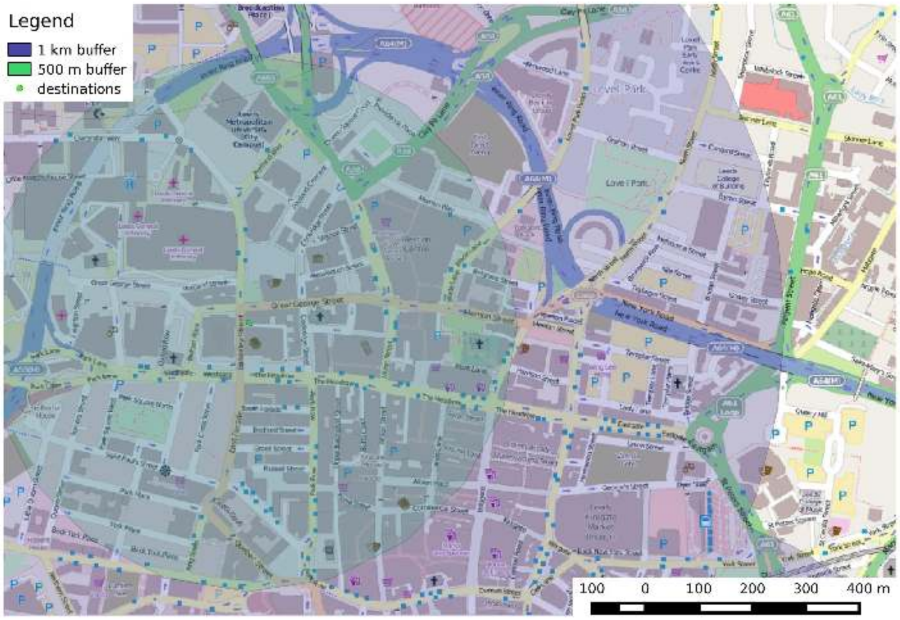


Figure 1: Illustrative 1000 and 500 metre buffers around Leeds shopping centre, as defined by Retail Vision data.

The results of this vital pre-processing phase — the importance of which should not be overlooked by data analysts (Wickham 2014) — are shown in Fig. 2 for the Twitter dataset. The figure shows a substantial concentration of tweets in the metropolitan area of Leeds and Bradford at the south-western edge of the study region. There are secondary clusters around the city of York at the centre, and seaside towns such as Scarborough on the east coast.

Retail flows from Acxiom data

Text string manipulation was used to match the 8 character postcode provided per respondent in the Acxiom data to the 3 or 4 character postcode district codes used as the origins in the study. Identifying the destinations where people undertake their shopping was more challenging, as respondents are provided with an open-ended question on "Where do you shop most often", hence the great diversity of answers. Still, 89% of responses were repeated at least once by other people. Typical examples of unique responses, highlighting this diversity, were: "Trafford Centre, Man" (the 'Man' referring to Manchester), "Sreat Universal" (likely to be typo where the respondent was referring to the Great Universal mail order company) and "Dont have enough". Clearly, only the first of these can be geocoded, and the latter two must be allocated *NA* values. In addition, the second most common answer is a blank, and many common responses are generic non-geographical names

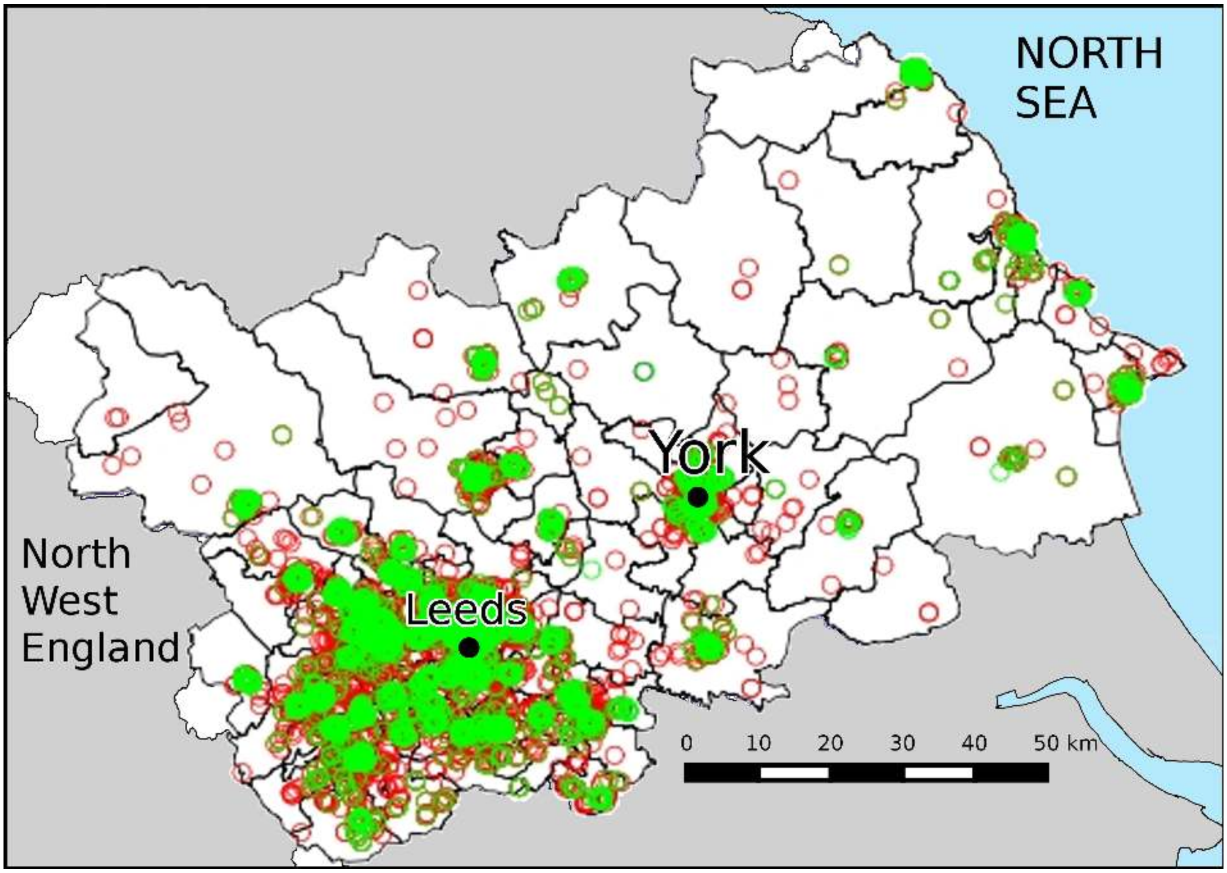


Figure 2: Illustration of the 'home tweets' (one per user, red) and 'shopping tweets' (one or more per user, green) identified from the national Twitter dataset

such as “Primark”, “Catalogue”, “In the town” and “on line”. Therefore a high proportion — roughly half of respondents — have no identifiable shopping destination.

Of the more commonly used destinations, including “Leeds”, “York” and “Huddersfield” (these are the three most common responses, each stated by more than 1000 people), the majority could clearly be linked to one of the 79 shopping destinations taken from the RetailVision data. To link each character string to a destination involves either fuzzy matching of characters, or ‘geocoding’ — linking words to coordinates — and then selecting the nearest store. Neither will produce a 100% certain match but, having tried both methods, it was found that the geocoding option was far more reliable and accurate.

Google’s geocoding API was called from within R to perform the this task, using the `geocode` command from the **ggmap** package (Kahle and Wickham 2013). A useful feature of this method is that it flags values that could not be successfully geo-coded: 76.2% of the 24,075 destinations recorded for households in the study area were successfully geocoded. A further 16.9% of the destinations were geocoded to destinations outside the study area, leaving 59.4% rows of Acxiom data with geocoded destinations inside the study area. These coordinates were allocated to the nearest shopping centre. The 50 most common destination labels (accounting for 80% of responses) were checked manually to ensure the results made sense. Overall, there were 14,292 flows represented in the resulting flow matrix from the Acxiom data.

Modelling retail flows

Various retail modelling approaches (including those which we articulate below) were deployed to determine the rate of interaction between residential zones and retail destinations. The retail flow ‘system’ was be modelled as a spatial network (a ‘flow matrix’) with the model parameters obtained from the origin-destination points described above. The model can then be used to predict the response of shoppers to proposed changes, such as the addition of a new retail centre, or to plan transport strategies. The main models employed by researchers to analyse the movement of people, goods, and communications are variations on the ‘gravity model’ and the ‘intervening opportunities’ model. All interactions were measured as person trips, although it would be possible to extend the analysis to estimate expenditure.

The unconstrained gravity model

The gravity model is based on Newton’s law of gravitational force, and states that, for the purposes of the present paper, the rate of travel flow is directly proportional to the ‘attraction’ between the origin and destination, and inversely proportionally to some function of the distance between them. It is therefore assumed in the model that shoppers aim to minimise the ‘cost’ of a journey. Clearly this will be related to the distance between the origin and destination, in addition to other factors. Distance was calculated as the Euclidean distance in metres between shopping centre centroid and the centroid of the residential postcodes based on the ‘OSGB 1936’ coordinate reference system. In addition to ‘distance decay’ (which represents how much distance is a barrier to travel via the the δ in the equation below), the gravity model used here has two other tuneable parameters. The resulting model, incorporating the tuneable parameters α, β and γ is defined as follows,

$$S_{ij} = \alpha \frac{O_i^{\beta} W_j^{\gamma}}{d_{ij}^{\delta}} \tag{1}$$

where S_{ij} the number of people who travel from origin postal district i to retail destination j; O_i is the number of people in postal district i; W_j is a measure of the attractiveness of the retail destinations j (here this is simply the number of shops in each centre); and d is the distance matrix between origins (rows represented by i) and destinations (columns, j). The parameters were tuned such that the modelled flow fits the actual data contained within the origin-destination matrix. This is achieved using unconstrained non-linear optimisation, minimising the least squares error between the flows predicted by the model and the actual values in the origin-destination matrix.

Production-constrained gravity model

Although the gravity model has been used successfully to model traffic flows and population movement, it is clear that if the values for O_i and W_j are doubled then, according to equation (1), the number of trips between them would quadruple. Yet intuition suggests they should double. In the second model the gravity law is modified such that the sum of the rows in the origin-destination matrix is equal to the number of trips originating from each origin. Further, an exponential distance decay function is introduced following Wilson (1971). This reflects the statistical derivation of this version of the model (Wilson 1967). The production-constrained gravity model is written as:

$$S_{ij} = A_i O_i W_j exp(-\beta d_{ij}) \tag{2}$$

where S_{ij}, O_i, W_j and d_{ij} are the same as in equation (1) and β is a distance decay parameter. The variable A_i,

$$A_i = \sum_j W_j exp(-\beta d_{ij}) \tag{3}$$

is a balancing factor to ensure that the following equation is true for all flows:

$$\sum_j S_{ij} = O_i \tag{4}$$

Thus the balancing factor is calculated as follows:

$$A_i = 1/\sum_j W_j exp(-\beta d_{ij}) \tag{5}$$

The distance decay parameter can be obtained by logarithmic transformation of the exponential form described above to a linear form as follows:

$$\log S_{ij} = \log O_i + \log W_j - \beta d_{ij} \tag{6}$$

Linear regression models of the gradient of the response $\log S_{ij}$ to the predictor d_{ij} can be used to find the optimal value of $-\beta$. For a more formal treatment of gravity models, the interested reader is directed towards Wilson (1971).

The radiation model

The radiation model proposed by Simini et al. (2012) derives its name from the analogy of a particle emitted from a source and absorbed by a destination location. The theory behind the model assumes there is no direct relationship between the number of trips and distance. Instead, it proposes that the number of trips over a given distance is proportional to the attractiveness of the destinations at that distance and inversely proportional to the attractiveness of the intervening opportunities. In its basic form, the model is parameter-free, and can therefore be applied in cases where there are no measurements of trip numbers, only requiring information on the distribution of populations. Although the mathematical formulation of the radiation model for trip flows is quite recent, the importance of intervening opportunities provides a clear link to established theories (Stouffer 1940).

In the original model developed by Simini et al. (2012), m_i and n_j represent populations, and the number of trips between i and j are predicted by the model. In this version of the model, these parameters are replaced by O_j and W_j, as defined above. Further, we add a 'home-field advantage' parameter for each origin, ϵ_i. This represents the increased attractiveness of retail opportunities available close to the home location, for example due to knowledge of the local area. A tuneable scaling parameter, C, is included in the model in place of the scaling factor, T_i, employed by Simini et al. (2012) in their model to scale the flows between an origin and destination by the proportion of commuters in the study area. This also removes the need for the additional parameter in the numerator as used by Simini et al. (2012). The revised model is therefore written:

$$S_{ij} = C\frac{O_i + \epsilon_i}{(O_i + \epsilon_i + P_{ij})(O_i + \epsilon_i + W_j + P_{ij})} \tag{7}$$

The matrix P_{ij} is the number of retail outlets in the circle of radius d_{ij} centred at i, excluding outlets at the source and destination. The values for ϵ_i and C are estimated using unconstrained non-linear optimisation with the objective function to minimise the square of error between estimated and observed flows.

Results

To be useful for retailers, developers and Local Authority planners, estimates of retail flow predicted by a model should match the source data. This is critical for assessing the impact of proposed changes, such as the addition of a new retail centre.[7] Model predictions for each retail centres were compared to the corresponding values obtained from the mobile network data. These results are compared below.

The origin-destination matrices obtained directly from the mobile phone data, the Twitter data, and the Acxiom data are visualised in Fig. 3. This shows the much larger size of the mobile phone dataset and the relatively sparse nature of the Twitter dataset compared with the other two sources.

To quantify the fit between the flow matrices, the coefficient of determination (R_T^2) — a measure of model fit — was used. The overall fit per matrix was calculated (Table 2), in addition to R_T^2 per origin zone, to identify the geographical distribution of model fit. R_T^2 is defined as follows (Steel and Torrie 1960):

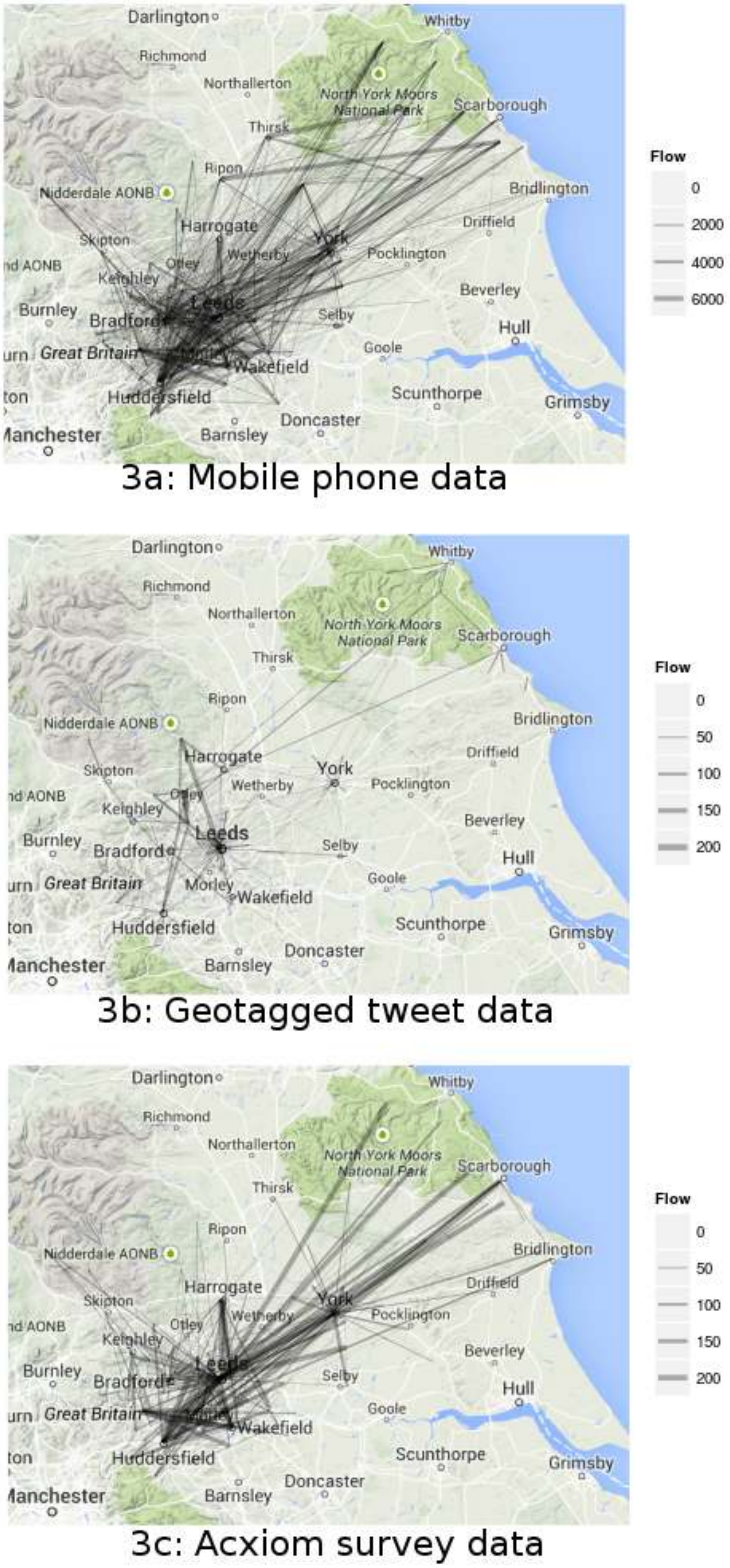
$$R_T^2 = 1 - (\sigma_e^2)/(\sigma_y^2) \tag{8}$$

where σ_e^2 is the sample variance of the residuals in the observed output and σ_y^2 is the variance of the residuals in the model. If the variance of the model residuals is low compared with the variance of the observed output, R_T^2 tends to unity, indicating the model gives a good explanation of the observed data. If the variances are similar in magnitude, R_T^2 tends to zero, indicating a poor fit. If the variance in the model residuals is large compared to the variance in the observed data, R_T^2 will be negative, indicating a very poor fit. The correlation between flows obtained directly from the phone, Twitter, and Acxiom data and the parameter-free radiation model was calculated using the coefficient of determination. Table 2 summarises the coefficient of determination for the three flow matrices obtained from the phone, Twitter, and Acxiom data and a fourth flow matrix obtained from the parameter-free radiation model.

Table 2: Coefficient of determination between the matrices of flows obtained directly from the phone, Twitter, and Acxiom data and estimated by the parameter-free radiation model.

	Mobile phone	Twitter	Acxiom	P-F Radiation
Mobile phone	1			
Twitter	0.127	1		
Acxiom	0.653	0.2	1	
P-F Radiation	0.793	0.069	0.548	1

[7] Note it is generally assumed that the population of each postal district is located at the centroid of the corresponding polygon and the models presented in this paper are no exception. Clearly, this is not an accurate reflection of the population distribution; however, for the purposes of comparing the relative performance of the modelling techniques described in the previous section, this is a reasonable approximation.



3a: Mobile phone data

3b: Geotagged tweet data

3c: Acxiom survey data

Figure 3: Flows derived from three sources: mobile phone data (top), geotagged tweets (middle) and retail survey data (bottom).

It can be seen from Table 2 that the matrix of flows obtained directly from the Twitter data do not coincide with those obtained from the mobile network data particularly well. This is partly due to the sparsity of the Twitter data, which represents only 8,500 flows compared to almost 100 times more (740,000) flows obtained directly from the mobile phone data. The flows obtained from the phone data are considered to be the 'benchmark' for the study, and for the remainder of this section the spatial variability of the fit of the flows obtained from the Twitter data, the Acxiom data, and the parameter-free radiation model to the flows obtained from the telephone data (i.e. the first column in Table 2) are discussed.

There is good agreement with the mobile network data for the area around central Bradford, the areas to the north, east, and west of York, and the area around Selby. This may be due to central Bradford having a much higher proportion of users represented in the Twitter data compared to other postal districts in the study area. The reason the other areas show good agreement to the mobile network data is less clear. In contrast, areas with a high proportion of Twitter users in the data, for example, central York and the area around Settle, have a very poor fit to the mobile network data.

Despite the relatively small sample size of the Acxiom dataset (just under double the size of the Twitter data), the resulting flows fit those predicted by the much larger mobile phone data surprisingly well. City centres, such as central York, Halifax, and Leeds, have a comparatively high proportion of respondents and show a good match to the mobile data. There is a poor fit for central Bradford, due in part to the low proportion of respondents in this area. Suburban areas of the major towns and cities in the area also show a poor fit to the mobile data. This may be due to the responses being simply "Leeds" or "Bradford" and not the name of the particular suburb in which the respondent predominantly shops. A similarly poor fit is apparent with rural areas; this is probably because the mobile data tracks users who need to travel to a retail centre for all types of purchases, in contrast to the Acxiom data which are for purely non-grocery goods.

The parameter-free radiation model provides the best fit to the mobile phone data, indeed the model fit is good for the majority of the study area, particularly around the major retail centres of Leeds, Bradford, Scarborough, York, Huddersfield, and Wakefield. The model fit for many of the rural areas is relatively poor, this may be due to the lack of 'intervening opportunities', meaning the flows are overestimated.

The matrices of flow obtained directly from the Twitter, phone, and Acxiom data described above have each been used to obtain parameters for three varieties of spatial interaction model: an unconstrained gravity model, a production-constrained gravity model, and a radiation model, using the methodology described in the previous section. Table 3 summarises comparisons between the retail flows at each of the retail centres obtained from each of nine models generated using the mobile phone, Twitter, and Acxiom data.

Table 3: Average coefficient of determination values for the flows per origin postcode and the retail destinations its inhabitants visit obtained from mobile, Twitter and Acxiom data (rows) and the models generated from these data sources (columns).

	Gravity (unconstrained)	Gravity (constrained)	Radiation
Mobile phone	0.831	0.742	0.909
Twitter	0.39	0.281	0.272
Acxiom	0.853	0.608	0.84

It is clear from Table 3 that models obtained from the Twitter data are very poor in comparison to those obtained using the mobile phone and Acxiom datasets. This is simply because the Twitter data are sparse and insufficient to provide a robust and convincing fit to the model output. Again, in spite of the relatively small dataset provided by Acxiom, the models provide a good fit to the data; however closer inspection of all the predicted flows reveals that the fit is relatively poor for those areas with a low proportion of respondents.

Considering the three models obtained from the flow matrix derived from the mobile phone data (i.e., the top line in Table 3), Fig.s 4 to 6 illustrate the spatial variability in the coefficient of determination between the modelled flow matrices and the source data. Fig. 4 illustrates the coefficient of determination of the prediction generated by the unconstrained gravity model. It can be seen that this model provides a reasonable fit to the source data, with the model providing a good model fit close to the major retail centres of Leeds, Bradford, Scarborough, York, Huddersfield, and Wakefield. Away from these centres, the performance of the model deteriorates, especially at the boundary of the study area.

Fig. 5 reveals that the model fit is acceptable for the majority of the study area. However Table 3 suggests that the fit of the output of the production-constrained gravity model to the source data is poor in comparison to the unconstrained gravity model, and that the fit is less satisfactory for the more populous area around Leeds and Bradford. This is slightly surprising and may be due to the adoption of a very simple model form in equation (2). While it is likely that an enhanced specification of this model e.g. to include a different functional representation of the attractiveness or distance effects, such model optimisations were considered to be outside the scope and purpose of this paper.

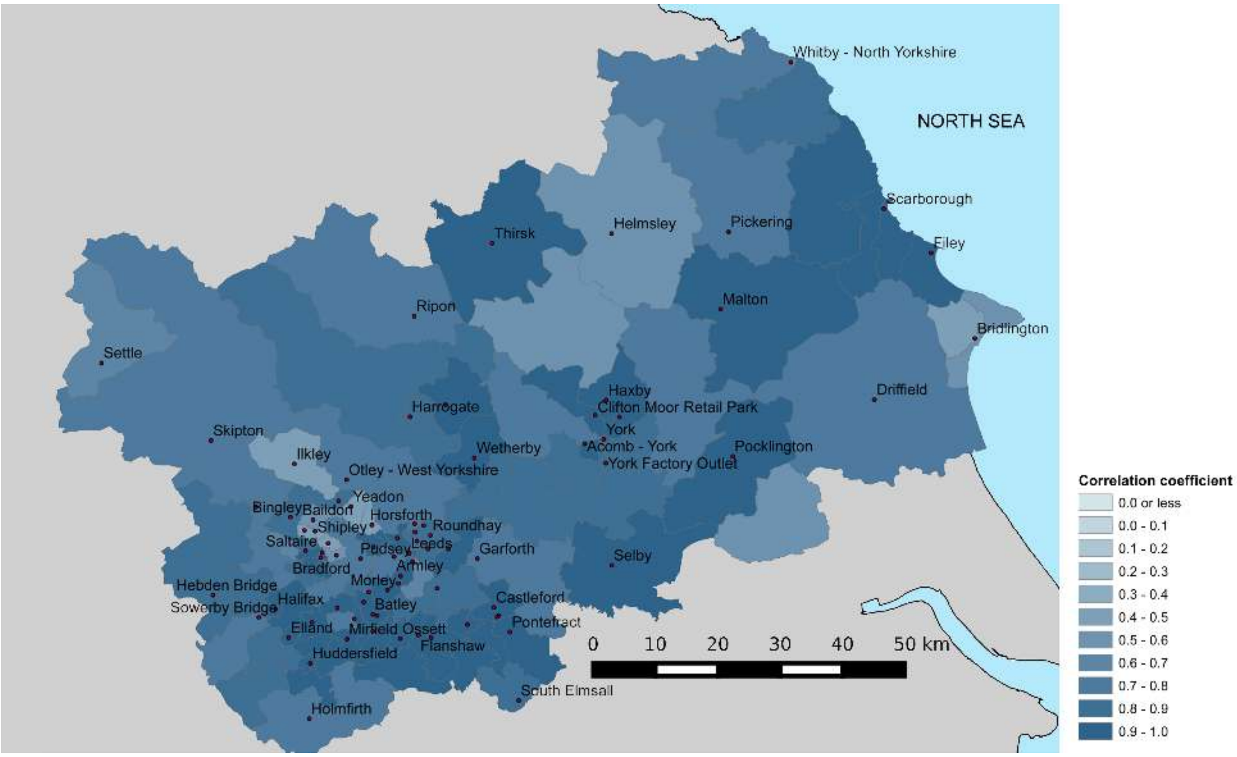


Figure 4: The coefficient of determination between flows obtained from unconstrained gravity model and flows inferred from mobile telephone data by postcode district of origin.

The radiation model provides a good overall fit to the source data, as illustrated in Fig. 6. The model also appears better able to predict retail flows at the boundary of the study area compared to the spatial interaction models. It is possible the better model fit to the source data is due to the radiation model having more tuneable parameters in the 'home advantage' form used here than the other models. This may lead to 'over-fitting' to the source data, and therefore reduce the effectiveness of the model when predicting the impact of proposed changes.

Discussion and conclusions

This paper has explored the relative utility of three large datasets for exploring retail-related travel behaviours. The work is of some empirical interest, reaffirming the long-standing idea that "the influence of a city of this category [a 'Regional Capital'] radiates over an extensive area — its region" (Dickinson 1930). The results, primarily obtained from the Acxiom and mobile telephone datasets, demonstrate the potential for new Big Data to explore long-standing issues such as those raised by Dickinson more than 80 years ago. Our work

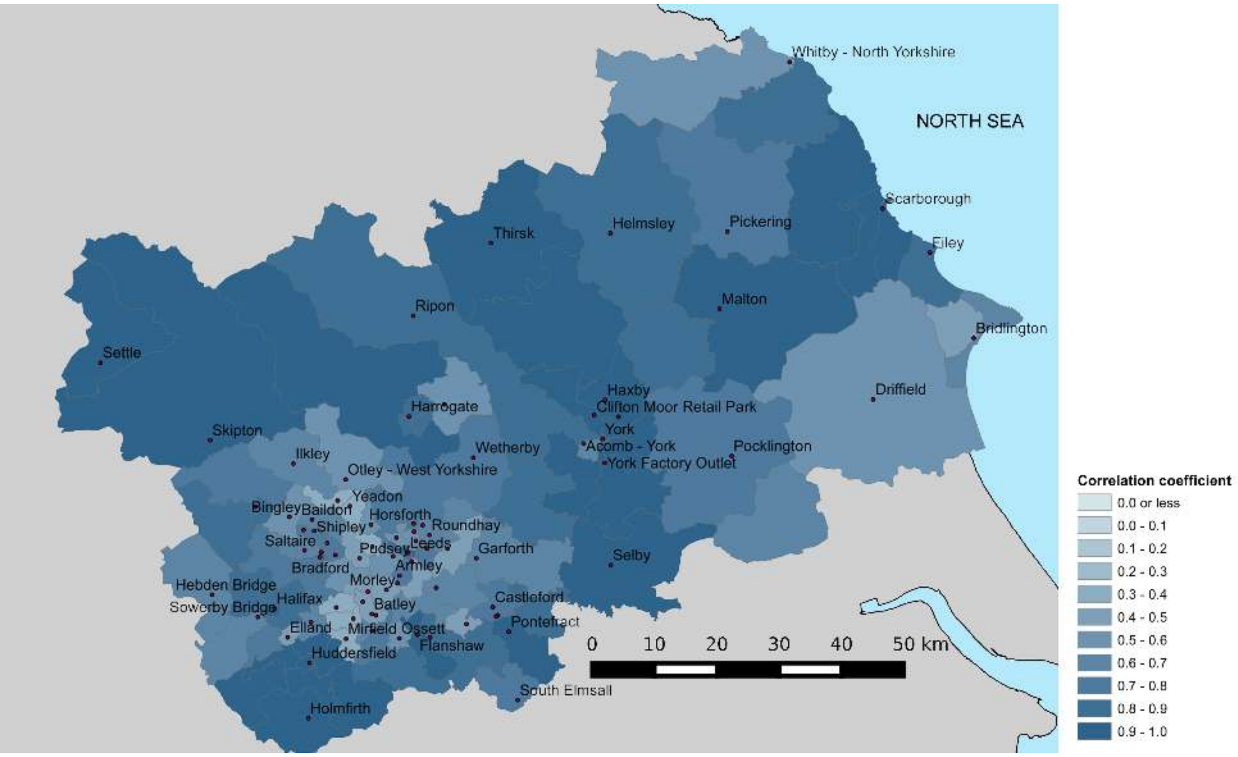


Figure 5: The coefficient of determination between flows obtained from production-constrained spatial interaction model and flows inferred from mobile telephone data by postcode district of origin.

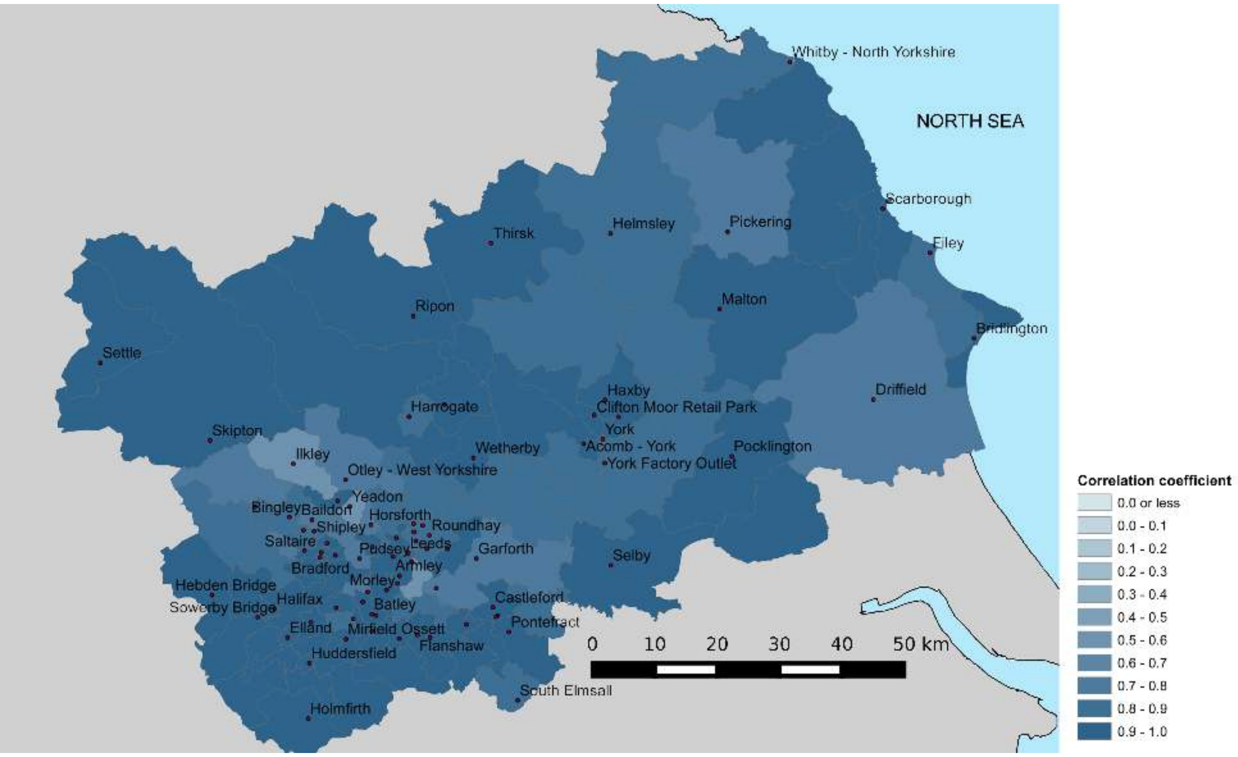


Figure 6: coefficient of determination between flows obtained from the radiation model and flows inferred from mobile telephone data by postcode district of origin.

suggests that, in the absence of reliable official datasets on retail travel (as opposed to commuter flows), emerging Big datasets can provide new insight into retail travel patterns.

However, the main contribution of this paper is methodological and theoretical. The methods will be of interest (and potentially useful) across wider fields of application including tourism, public service provision and transport planning. Others have recently noted the potential utility of mobile telephone records in tracking and perhaps controlling epidemics (The Economist 2014) and in the production of census-like migrant distributions (Tatem et al. 2014). The wider point relates to the concept of Big Noise, coined in this paper to refer to Big Data that is high *volume* but of relatively low *value* in its raw state. Even after 'data mining' processes to refine the low grade 'ore' of the Twitter data, the dataset seemed to have relatively low levels of *veracity* compared with the other datasets. These data quality judgements are rare in the emerging Big Data literature and we hope that the concepts and methods developed in this paper will lead to better assessment of data quality in future work.

Data quality is paramount in empirical work and this is particularly true for emerging data sources which lack a track record in applied research. This problem has been addressed in the present paper through cross-verification between sources and against a selection of model results. A superficial interpretation of the results suggests that the robustness of each source is related to the intensity of its coverage, hence the data *volume*. Mobile telephone data, collected the most frequently, is a promising source for retail flow data while relatively sparse and infrequent geolocated twitter messages unsurprisingly performed worst. However, volume is only one factor that should decide how (and whether) a dataset is used. There are more subtle explanations of the results. Acxiom market research data contains a very wide range of demographic attributes, as well as attitudes, lifestyles and behaviours (Thomas, Stillwell, and Gould 2014). The slow velocity and stable sample size of the surveys make the Acxiom dataset is better suited to the exploration of long-term trends than short-term shifts.

The work supports the view that variety can be a strength or weakness, depending on context. Twitter messages contain intriguing clues about activity patterns (Malleson and Birkin 2012; Lovelace et al. 2014) but cannot be assumed to represent real behaviour. For many retail applications *velocity* may be the key attribute, for identifying market trends. More sophisticated methods may be needed to capture the seasonal, weekly, and diurnal variations in behaviour patterns. But for the Acxiom and Twitter data, the flow of information is simply not sufficiently fast to pick-up on short-term shifts in shopping habits.

The empirical case study used to explore these points was deliberately limited in scope but could be expanded in future work. The year's worth of market research data could be extended to cover more recent years and the mobile telephone and Twitter data could (provided access) be updated almost continuously. However, the subset of these vast datasets analysed in this paper have been sufficient to demonstrate clear trends. A more systematic appraisal of the models could be incorporated to allow a thorough consideration of different model parameters and functional forms of distance decay. More specifically, the ongoing nature of the datasets under investigation provides an opportunity for assessing their ability to pick-up on future trends. Our preliminary hypothesis would be that the Twitter and mobile phone datasets would shine here, notwithstanding issues of changing rates of use disrupting time-series analysis.

An important finding for future researchers jumping on the Big Data "bandwagon" (Kwon, Lee, and Shin 2014) is that social media messages from Twitter messages were found to be significantly skewed in their spatial and temporal distributions. This is likely to be mirrored by bias due to a small number of high volume users accounting for the majority of Twitter-inferred flows. With this caveat in mind, we allowed multiple trips (to different destinations) per user to overcome issues of data sparsity in the Twitter data. Advanced filtering techniques and sensitivity analyses to identify and minimise such biases (D. R. Davis, Dingel, and Morales 2015) are therefore recommended areas for future research. For the models used in this paper to be more policy-relevant, they would need to estimate absolute 'flow' in absolute values, rather than the relative flow numbers presented here. Further refinements (which could be aided by the textual content of social media) would seek to quantify the expenditure resulting from different flows — the lack of monetisation of the flows is a clear limitation of this research from the perspective of private business, but not from the perspective of transport planning.

A more specific limitation of this paper relates to the scale of the case study. 'Edge effects', also known as the ‘boundary problem’ (Batty 2012), may have artificially reduced fit towards the boundary of the study area and affected the distance decay parameter. This limitation could be minimised in a nationwide study or by modelling ‘sinks’ outside the study area: it is recommended that further work focussed on developing spatial interaction models of retail flow use a wider area to train and test the models. A wider issue that this paper highlights, that is also applicable to more traditional datasets, is the context-dependence of verification. Different methods would be needed to assess the datasets’ quality for assessing user experience, or frequency of visit, for example.

Notwithstanding these limitations and drawbacks, the results of this work demonstrate the potential for emerging large datasets to be used in geographical modelling. Applications where official datasets are limited, such as retail flows, are especially well-suited to the use of Big Data. This will require new methods and concepts, some of which are explored presented in this paper. Verification in geographical research is set to become increasingly important as datasets become more diverse. There is great potential for further work on data verification of datasets as more and more Big Data sources are made available to the research community. While this paper has been primarily methodological and theoretical in its emphasis the approach, of cross-validating unrelated sources of geographically aggregated information, has great potential in applied research.

This paper also relates to the research funding landscape surrounding quantitative geography and the social sciences. In recognition of the importance of Big Data for academic research, research councils are mobilising resources to take advantage of new sources of data, from both government and commercial domains.[8] An optimistic view of such investment is that, in the long run, it will facilitate new research insights. In geographic research, for example, we have demonstrated the potential to harness new datasets to better understand retail flows. There is also great potential for using Big Data for social benefit. Understanding how mobility patterns shift over short-term time-scales and for multiple purposes could have important implications for sustainability and help answer important questions such as: How can society best transition away from fossil fuels?

For such exciting geographical Big Data research to happen, it is critical that publicly-funded researchers have access to the data. It is encouraging, therefore, to see many initiatives aiming to unlock the great potential of Big Data for academics.[9] This view seems to be shared by national statistical agencies that are considering the possibility that Big Data could eventually replace official datasets altogether. That seems to be the view of the UK’s Office for National Statistics which is looking into how synthesised datasets could replace the National Census after 2021 (Coleman 2013).

New methods and greater experience are needed before such an ambition becomes reality. As pointed out with respect to the physical sciences, “researchers need skills in both science and computing — a combination that is still all too rare” (Mattmann 2013). This problem is even more acute in the social sciences and there is a need for training around the specific skill sets needed to deal with the deluge of new datasets and software packages available in the 21st century (R. Harris et al. 2014). On the data side, new sources should *not* be assumed to be better than existing sources simply because they are bigger, although its large size does seem to contribute to relatively good data-model fit found for mobile telephone data. The work presented in this paper outlines a small step in this direction, highlighting the importance of cross validation.

References

Arribas-Bel, Daniel. 2014. “Accidental, open and everywhere: Emerging data sources for the understanding of cities.” *Applied Geography* 49. Elsevier: 45–53.

Batty, Michael. 2012. “A generic framework for computational spatial modelling.” In *Agent-Based Models of Geographical Systems*, 19–50. Springer.

[8] In the social sciences, this can be seen in a few multi-million pound investment programmes, in ‘Big Data’, including the Australian Urban Research Infrastructure Network (https://aurin.org.au/) UK’s Big Data Network (http://www.esrc.ac.uk/research/major-investments) and the Big Data Research Initiative in the USA.

[9] See, for example, the Consumer Data Research Centre (cdrc.ac.uk/projects/), which aims to make commercial data more accessible to academics for social benefit.

Bermingham, Adam, and Alan F Smeaton. 2011. “On Using Twitter to Monitor Political Sentiment and Predict Election Results.” *Psychology*, 2–10.

Birkin, Mark, G P Clarke, Martin Clarke, and AG Wilson. 1996. *Intelligent GIS: location decisions and strategic planning*. London: Geoinformation Group.

Birkin, Mark, Graham Clarke, and Martin Clarke. 2010. “Refining and operationalizing entropy-maximizing models for business applications.” *Geographical Analysis* 42: 422–45. doi:10.1111/j.1538-4632.2010.00801.x.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. “Twitter mood predicts the stock market.” *Journal of Computational Science* 2 (1): 1–8. doi:10.1016/j.jocs.2010.12.007.

Boyd, Danah, and Kate Crawford. 2012. “Critical Questions for Big Data.” *Information, Communication & Society* 15 (5): 662–79. doi:10.1080/1369118X.2012.678878.

Coleman, David. 2013. “The Twilight of the Census.” *Population and Development Review* 38. Blackwell Publishing Ltd: 334–51. doi:10.1111/j.1728-4457.2013.00568.x.

Crampton, Jeremy W., Mark Graham, Ate Poorthuis, Taylor Shelton, Monica Stephens, and Matthew Zook. 2013. “Beyond the geotag: situating ’big data’and leveraging the potential of the geoweb.” *SSRN Electronic Journal*, 1–29. doi:10.2139/ssrn.2253918.

Davis, Donald R., Jonathan I. Dingel, and Eduardo Morales. 2015. “Spatial and Social Friction in the City: Evidence from Yelp.” *Mimeo*.

Davis, Kord. 2012. *Ethics of Big Data: Balancing Risk and Innovation*. O’Reilly Media.

Dickinson, Robert E. 1930. “The regional functions and zones of influence of Leeds and Bradford.” *Geography*. JSTOR, 548–57.

Driscoll, Kevin, and Shawn Walker. 2014. “Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data.” *International Journal of Communication* 8: 20.

EMarketer. 2014. “More than One-Fifth of UK Consumers Use Twitter.” *EMarketer.com*. http://www.emarketer.com/Article/More-than-One-Fifth-of-UK-Consumers-Use-Twitter/1010623/2.

Harris, Richard, Tate Nicholas, Catherine Souch, Alex Singleton, Scott Orford, Chris Keylock, Claire Jarvis, and Chris Brunsdon. 2014. “Geographers Count: A Report on Quantitative Methods in Geography.” *Enhancing Learning in the Social Sciences* 6 (2): 43–58. doi:10.11120/elss.2014.00035.

Hilbert, Martin. 2013. “Big data for development: From information-to knowledge societies.” *Available at SSRN 2205145*.

Humby, Clive, Terry Hunt, and Tim Phillips. 2008. *Scoring points: How Tesco continues to win customer loyalty*. Kogan Page Publishers.

Kahle, D, and Hadley Wickham. 2013. “ggmap: Spatial Visualization with ggplot2.” *The R Journal* 5: 144–61. http://stat405.had.co.nz/ggmap.pdf.

King, G, J Pan, and ME Roberts. 2014. “Reverse-engineering censorship in China: Randomized experimentation and participant observation.” *Science* 345 (6199): 1–10. http://www.sciencemag.org/content/345/6199/1251722.short.

Kitchin, R. 2013. “Big data and human geography: Opportunities, challenges and risks.” *Dialogues in Human Geography* 3 (3): 262–67. doi:10.1177/2043820613513388.

Kwon, Ohbyung, Namyeon Lee, and Bongsik Shin. 2014. “Data quality management, data usage experience and acquisition intention of big data analytics.” *International Journal of Information Management* 34 (3). Elsevier: 387–94.

Laney, Doug. 2001. “3D data management: Controlling data volume, velocity and variety.” *META Group Research Note* 6.

Lovelace, Robin, Nick Malleson, Kirk Harland, and Mark Birkin. 2014. “Geotagged tweets to inform a spatial interaction model: a case study of museums.” *Arxiv Working Paper*.

Madden, Sam. 2012. "From Databases to Big Data." *Internet Computing, IEEE* 16 (3): 4–6. doi:10.1109/MIC.2012.50.

Malleson, Nick, and Mark Birkin. 2012. "Estimating Individual Behaviour from Massive Social Data for An Urban Agent-Based Model." In *Modeling Social Phenomena in Spatial Context*, edited by Andreas Koch and Peter Mandl. Salzburg: Universitat Salzburg.

Mattmann, Chris A. 2013. "Computing: A vision for data science." *Nature* 493 (7433). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 473–75. http://dx.doi.org/10.1038/493473a.

Mitchell, Lewis, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. 2013. "The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place." *PLoS ONE* 8 (5). Public Library of Science: e64417. http://dx.doi.org/10.1371/%2Fjournal.pone.0064417.

Mitchell, V-W, and Peter J McGoldrick. 1994. "The role of geodemographics in segmenting and targeting consumer markets: a Delphi study." *European Journal of Marketing* 28 (5). MCB UP Ltd: 54–72.

Nanni, Mirco, Roberto Trasarti, Barbara Furletti, Lorenzo Gabrielli, Peter Van Der Mede, Joost De Bruijn, Erik De Romph, and Gerard Bruil. 2014. "Transportation Planning Based on GSM Traces: A Case Study on Ivory Coast." In *Citizen in Sensor Networks*, 15–25. Springer.

Ofcom. 2012. "Facts & Figures." http://media.ofcom.org.uk/facts/.

Rae, Alasdair, and Alex Singleton. 2015. "Putting big data in its place: a Regional Studies and Regional Science perspective." *Regional Studies, Regional Science* 2 (1): 1–5. doi:10.1080/21681376.2014.990678.

Signorini, Alessio, Alberto Maria Segre, and Philip M. Polgreen. 2011. "The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic." *PLoS ONE* 6 (5). doi:10.1371/journal.pone.0019467.

Simini, Filippo, Marta C González, Amos Maritan, and Albert-László Barabási. 2012. "A universal model for mobility and migration patterns." *Nature*, February, 8–12. doi:10.1038/nature10856.

Steel, RGD, and JH Torrie. 1960. *Principles and procedures of statistics.* New York: McGraw-Hill.

Stouffer, Samuel A. 1940. "Intervening opportunities: a theory relating mobility and distance." *American Sociological Review* 5 (6). JSTOR: 845–67.

Taleb, Nassim Nicholas. 2012. *Antifragile: things that gain from disorder.* Random House LLC.

Tatem, Andrew J., Zhuojie Huang, Clothilde Narib, Udayan Kumar, Deepika Kandula, Deepa K. Pindolia, David L. Smith, et al. 2014. "Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning." *Malaria Journal* 13 (1): 52. doi:10.1186/1475-2875-13-52.

The Economist. 2014. "Waiting on hold." *The Economist*, October. http://www.economist.com/news/science-and-technology/21627557-mobile-phone-records-would-help-combat-ebola-epidemic-getting-look.

Thomas, Michael, John Stillwell, and Myles Gould. 2014. "Exploring and Validating a Commercial Lifestyle Survey for its use in the Analysis of Population Migration." *Applied Spatial Analysis and Policy* 7 (1): 71–95. doi:10.1007/s12061-013-9096-5.

Wickham, Hadley. 2014. "Tidy data." *The Journal of Statistical Software* 14 (5). http://www.jstatsoft.org/v59/i10 http://vita.had.co.nz/papers/tidy-data.html.

Wilson, AG. 1967. "A statistical theory of spatial distribution models." *Transportation Research* 1 (3). Pergamon: 253–69.

———. 1969. "The Use of Analogies in Geography." *Geographical Analysis* 1 (3). Wiley Online Library: 225–33.

———. 1971. "A family of spatial interaction models, and associated developments." *Environment and Planning* 3 (January): 1–32. http://www.environment-and-planning.com/epa/fulltext/a03/a030001.pdf.

YouGov. 2011. “Blackbelt Survey Results.” YouGov.com. http://cdn.yougov.com/cumulus/_uploads/document/295n1rd92f/YG-Archive-Whiteoaks-250313-mobile-phone-data.pdf.