Office for National Statistics

GSS Methodology Series No 41

Using geolocated Twitter traces to infer residence and mobility

Nigel Swier, Bence Komarniczky
and Ben Clapperton

October 2015

Official Statistics

ONS Official Statistics are produced to the high professional standards set out in the Code of Practice for Official Statistics.

About us

The Office for National Statistics

The Office for National Statistics (ONS) is the executive office of the UK Statistics Authority, a non-ministerial department which reports directly to Parliament. ONS is the UK government's single largest statistical producer. It compiles information about the UK's society and economy, and provides the evidence-base for policy and decision-making, the allocation of resources, and public accountability. The Director-General of ONS reports directly to the National Statistician who is the Authority's Chief Executive and the Head of the Government Statistical Service.

The Government Statistical Service

The Government Statistical Service (GSS) is a network of professional statisticians and their staff operating both within the Office for National Statistics and across more than 30 other government departments and agencies.

Contacts

This publication

For information about the content of this publication, contact
Susan Williams
Tel: 01329 444641
Email: nigel.swier@ons.gsi.gov.uk

Other customer enquiries

ONS Customer Contact Centre
Tel: 0845 601 3034
International: +44 (0)845 601 3034
Minicom: 01633 815044
Email: info@ons.gsi.gov.uk
Fax: 01633 652747
Post: Room 1.101, Government Buildings,
Cardiff Road, Newport, South Wales NP10 8XG
www.ons.gov.uk

Media enquiries

Tel: 0845 604 1858
Email: press.office@ons.gsi.gov.uk

Copyright and reproduction

Table of Contents

1. Executive Summary:

The ONS Big Data Project was set up in January 2014 to investigate the potential of big data and to understand what it means for official statistics. The potential of geolocated activity traces from Twitter to provide insights into population and mobility was one of four initial pilots[1] designed to understand the practical issues with working with new big data sources and technologies and to identify statistical benefits.

Twitter is a popular micro-blogging platform where users post short messages, or "tweets", with a limit of 140 characters. Large volumes of these messages and accompanying metadata may contain a range of insights. Users tweeting from a smartphone or other device providing location services may choose to provide a precise GPS location. These are referred to as geolocated tweets. Although less than 2 per cent of tweets are geolocated, the volumes of data are still considerable with hundreds of thousands of such tweets being sent every day within Great Britain.

The aim of this pilot is to establish whether it is possible to use geolocated activity traces from Twitter to infer a user's residence and analyse mobility patterns. These data could provide new insights into the population and how different groups move around the country. They could also be used to help better understand and validate official population estimates. It may even be possible to produce estimates for different population bases that cannot be produced from existing data sources.

Data was collected on all geolocated tweets sent within Great Britain over a seven month period (1 April to 31 October 2014). This involved collecting data through a combination of real-time collection through the Twitter API and procurement of a bulk point in-time extract. Methods were developed for clustering individual geolocated activity traces and combining these clusters with additional address data to provide information about the location of clusters. Although individual level data are required to derive these clusters, this analysis is solely concerned with aggregate patterns. A number of analyses were then made of the resulting clusters, including aggregate comparisons with 2011 Census data.

[1] The other pilots were: Web scraping for consumer price statistics, electricity smart meter data for modelling occupancy, and mobile phone data for commuting patterns

Key highlights:

User activity traces can be clustered to infer de facto residence	It is feasible to cluster geolocated activity traces from Twitter to infer a user's location of residence providing there is sufficient user data. DBSCAN[2] is a highly suitable algorithm for this purpose. AddressBase[3] can be used to find the nearest address point and to classify the clusters by type (i.e. residential, commercial, or other). The residential cluster with the highest number of tweets (referred to as the dominant residential cluster) is assumed to be the location of usual residence. There were about 340,000 Twitter users for whom there were sufficient data to infer a location of residence for a period of at least one month. The resulting penetration rates of users by local authority found the highest rates in Central London, and in urban areas of the North West, Wales and Scotland. The lowest rates tended to be in peripheral rural areas, although there were also low rates in North London and urban areas in the West Midlands.
Analysis of monthly net flows shows a strong pattern of mobility between local authorities that follows the cycle of the academic year	The activity traces for each user can be broken down into months and then dominant residential clusters can be identified for each month. When the dominant cluster from one month to the next is in a different local authority, this can be inferred as a mobility flow between local authorities. When these net flows for each local authority are compared with the proportion of students in the population (based on 2011 Census data) there is a distinct signal that follows the cycle of the academic year. For example, in June there is a net flow out of student areas coinciding with the end of studies. Then in September and October, there is a net inflow back into these areas. This pattern cannot be detected from existing sources and so could be used as a supplementary source of intelligence on the movement of student populations.

[2] DBSCAN - Density based spatial clustering algorithm with noise (Ester et al, 1996)
[3] AddressBase is the definitive source of address information within Great Britain.

Geolocated Twitter data is uneven and patchy across the user base	Half of all geolocated tweets were made by just 4 per cent of users while 17 per cent of users only sent one tweet. Thus, the volume of geolocated Twitter activity is very uneven across the user base. Only 46 per cent of all users had sufficient detail to infer a location of residence. In addition, the median time span between a user's first and last tweet was 47 days. This suggests that many users go through a phase of sending geolocated tweets but do not continue doing so. Thus, Twitter may have limited value for monitoring longitudinal change over periods of more than a few months.
Social media data is unstable and may be affected by unexpected technological and behavioural changes	A 25 per cent drop in the volume of geolocated tweets during September 2014 coincided with the release of the iOS8 iPhone operating system. This included changes to the management of privacy and location settings. This illustrates how the collection of data from social media can be impacted by a combination of technological change (including those of third parties) and the behavioural response of users. This has implications for time series analysis and illustrates why caution is needed when using social media data to inform decision-making.
Analysis in this pilot is based on un-weighted counts and new methods would be needed to produce robust estimates.	Although these analyses can provide new insights into population and mobility, they are based on un-weighted counts and are not estimates. This is an important consideration as Twitter users are not representative of the general population. One possibility for producing estimates could be to infer socio-demographic characteristics of Twitter users and then calibrate to other sources, such as the mid-year population estimates. Another approach might be to use a benchmarking survey to measure rates of Twitter usage across the population. These avenues of research are already being taken forward by ONS.

Moving this from research into operations would require procurement of Twitter data	The pilot started by collecting data through the Twitter public streaming API. Although it is straightforward to collect the target data using this approach, collecting data at this scale falls outside Twitter's terms and conditions. Thus, if this research were to be made operational, then the data would need to be purchased.
There are important ethical considerations to be made when using these data	Twitter is designed to be public facing and in addition users must agree to certain conditions about how their data (including optionally provided location data) are used. Nonetheless, there are ethical considerations to make when processing social media data, especially when dealing with precision location data. The value of this research for statistical purposes is in understanding of aggregate rather than individual patterns and so privacy rights have been respected.
There are further insights that could be gleaned from these data	This pilot has focused heavily on data collection, processing methods and use of technology. The analysis on derived clusters had been cursory and there a number of other aspects that could be explored. Avenues for further exploration include: • Analysis at lower level geographies • Incorporation of time of day into cluster formation • Investigation into whether tweet content could help identify different types of user (e.g. international tourists) and validate anchor point classifications. For example, tweets at residential locations may have different features from those at work locations.

2. Introduction

The aim of this pilot is to explore the feasibility of using geolocated activity traces from Twitter to gain new insights into residence and mobility patterns.

The original intention was to focus on internal migration as this is a key component of change for sub-national population estimates in the United Kingdom. Internal migration in this context refers to moves between local authorities in England and Wales as well as moves to and from the rest of the UK (i.e. Northern Ireland and Scotland) (ONS, 2012). The main source of internal migration is the GP patient register. When people move and re-register with a doctor within England and Wales, this change of address is recorded within the patient register system. ONS receives an annual snapshot of the patient register and any changes with the previous year are used as a proxy for an internal migration move.

A well documented issue with the GP patient register is that students generally, and young men in particular, are less likely to re-register with a GP when they change address compared with the general population. In 2014, 9 per cent of the population were between the ages of 18 and 24, yet this group make up almost a quarter of all Twitter users (eMarketer, 2015). Therefore the premise is that geolocated Twitter data could be particularly useful for gaining insight into student age migration.

Geolocated tweets are generated when a user allows their location coordinates to be shared, through a smartphone, or other device with location services. Although less than 2 per cent of tweets are geolocated, the sheer volume of global Twitter traffic generates considerable volumes of such data. As a consequence, hundreds of thousands of geolocated tweets are made each day within Great Britain. These data represent activity traces that could be used as the basis for identifying patterns of movement over time. This pilot targeted all geolocated tweets sent within Great Britain over a seven month period (1 April to 31 October 2014) totalling 81.4 million tweets.

The basic methodological approach involves clustering the data pertaining to each user to identify frequently visited locations and then infer residence based on these patterns. As this pilot progressed, it became clear that any longitudinal analysis of change comparable to the current internal migration estimates methodology (i.e. annual change on a mid-year basis) was impractical. This was partly because of time and cost constraints in obtaining sufficient data. However, it also became clear that the majority of Twitter users who send geolocated tweets stop within a few months of starting (See Section 5.6). Therefore, the number of users who have geolocated activity traces spanning twelve months or more is too small to enable useful insights to be gleaned.

Despite these limitations, this pilot provides clear evidence that geolocated Twitter data can be used to detect mobility patterns over shorter time periods. In particular, it is possible to

detect a distinct pattern of mobility in and out of student areas that follows the cycle of the academic year.

The following section describes current research on geolocated Twitter data and elaborates on the key research areas within this pilot.

3. Literature Review and Discussion

Some National Statistics Institutes (NSIs) have already started investigations into the potential of Twitter to support official statistics. Research by the Netherlands Central Bureau for Statistics (CBS) suggest that Twitter could help with official statistics as it contains large volumes of information on a broad range of topics. However, it is recognised that selection bias remains a major methodological obstacle (Daas et al, 2012). In Mexico, the National Statistics Office (INEGI) has investigated how Twitter could be used to gain insights on a range of topics from domestic tourism, border mobility and well-being (INEGI, 2015). These early investigations suggest that exploring the potential of social media to support the production of statistics is a valid field of inquiry for NSIs.

There is a strand of research focusing on geolocated Twitter data to analyse movement of people across national borders. Hawelka et al (2013) found that these data can identify distinctive seasonal patterns of international travel for residents from different countries. Zagheni et al (2013) used Twitter to estimate out-migration rates from OECD countries and propose a difference-in-differences method to tackle the issue of selection bias. Blanford et al (2015) explore the use of geolocated tweets within Kenya to investigate regional connections and cross-border movements.

Other research has focused more on mobility within national borders. Brogueria et al (2015) use geolocated tweets to identify spatial and temporal variations in different regions of Portugal. This study concluded that Twitter can be used as an indicator for when the population is higher or lower during different times of the year. A number of studies have applied the concept of the radius of gyration to geolocated Twitter data (e.g. Jurdak et al, 2014; Yan et al, 2013). This is a geometrical approach to modelling usual activity space from a set of activity traces. This approach has also been used in the analysis of mobile phone data (e.g. Xiao-Yong et al, 2013).

The concept of radius of gyration is useful for analysis that is not concerned with geographic boundaries. However, it is less useful for demographic accounting approaches that are used in the production of official population statistics. These are underpinned by administrative and statistical sub-national geographies with defined boundaries and population definitions. In the UK, population statistics are generally based on the concept of *usual residence*, which is defined as the UK address where people spend the majority of time (ONS, 2009a).

If the objective is to use activity traces to identify location of residence by local authority, we need to consider that a user's usual activity space will often cross local authority boundaries. In 2009, an estimated 41 per cent of UK workers lived and worked in a different local authority (ONS, 2011). The use of Twitter data to support sub-national demographic accounting is an area that remains unexplored and is therefore a key focus of this research.

Another commonly used population definition is the *de facto* population, which is the population present at the time of enumeration (United Nations, 2008). This is relevant in the context of using geolocated activity traces, since this records a user's location at a specific point in time. This distinction between the usually resident and de facto population is particularly important for certain populations. For example, students are counted as being usually resident at their term time address, but may spend part of the year living at their parent's home address. We can expect these issues to be a complicating factor in assigning location of usual residence. However, it may be very useful for producing de facto population measures. This could offer new opportunities to produce indicators based on a range of definitions, including time of day, day of week, or seasonal population changes, as well as international tourism.

Turner and Malleson (2012) observe that geolocated tweets for individual users are often concentrated in a small number of locations. These are thought to be anchor points that are part of routine patterns of movement, which are assumed to include a home address. It is suggested that analysis of tweet content in relation to different anchor points might provide some insight into the nature of these locations.

This pilot leans heavily on this concept of anchor points and proposes a methodology for systematically identifying them. This approach involves using DBSCAN[4] (Density Based Spatial Clustering Algorithm with Noise). A number of research projects have already applied DBSCAN to geolocated Twitter data. Wayant et al (2014) use DBSCAN as a way of summarising and studying the spatiotemporal patterns of event driven activity on Twitter. Bawa-Cavia (2010) uses DBSCAN with Twitter data to identify fragmented areas of high density social activity within cities. Steiger (2014) investigates the potential of DBSCAN as a method of automatic feature extraction from Twitter data. For example a DBSCAN algorithm applied to the coordinates of geolocated tweets referencing "Oxford Street" produces a vector that corresponds to Open StreetMap[5] data.

However, using DBSCAN to identify residential anchor points from Twitter data seems a very relevant application, but is reamins unexplored. This together with exploring a framework for analysing Twitter within a demographic accounting framework provides the main focus for this pilot.

[4] See Section 6.2 for a detailed technical explanation
[5] http://www.openstreetmap.org

4. Research Outline

The research described in this report falls into three broad areas:

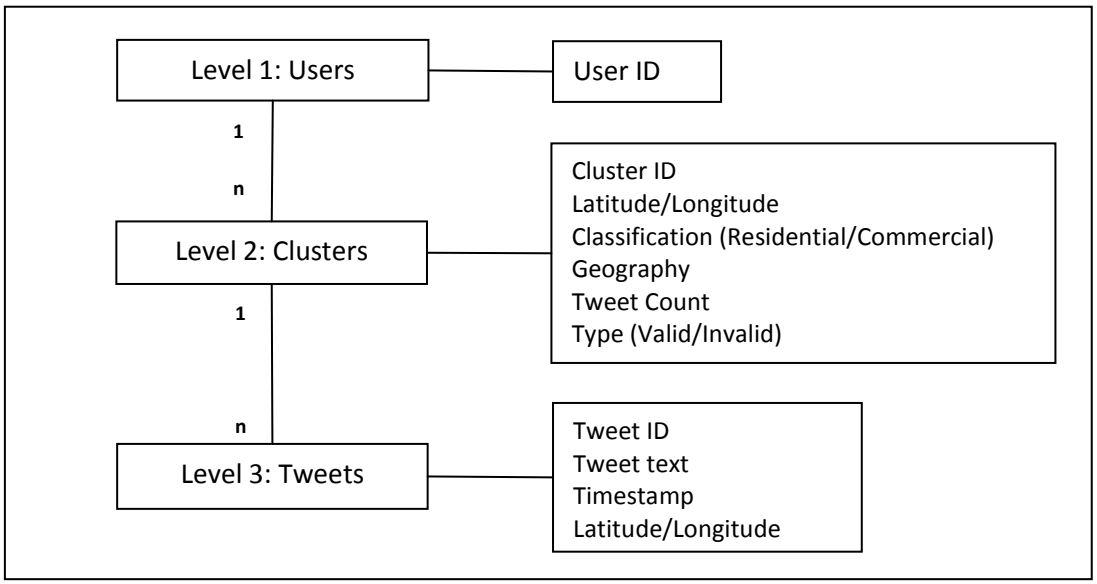
i) Data Collection: This describes the activities involved in collecting the geolocated tweets and exploratory analysis of the raw data.

ii) Methodology: This describes the process of clustering user tweets to identify anchor points and classifying these by type (i.e. residential, commercial and other).

iii) Analysis of Anchor Points: This covers the analysis of anchor points by user including the penetration rates by local authority and inferences about mobility over time.

The data used in this pilot has been organised into a hierarchical relational data model (Figure 1). This provides an outline of what the data looks like when it is collected and then how the data is transformed to derive clusters, which form the primary unit of analysis.

Figure 1: Cluster-based Data Model for Geolocated activity traces



The top level (Level 1) represents the 'user' entity[6]. This represents every individual user in the data and is defined explicitly through each unique user_id, which already exists within data. The bottom level (Level 3) represents the 'tweets' entity. This is also already defined within the data through each unique tweet_id. The middle level (Level 2) is the "cluster" entity. This is a derived entity and is the product of a DBSCAN clustering algorithm. This entity organises the tweets of each user that relate to the same location. Each cluster is

[6] Although it is assumed that 'a user' corresponds to 'a person', this may not always be the case. A person could have more than one user account or could share one account with another person.

assigned a unique identifier, which is the concatenation of the user_id and an ascending number based on the number of clusters for each user.

Clusters containing at least three points are flagged as 'valid' and are assumed to represent an anchor point for a particular user. This could be a home address, work, a friend's house, a cafe, or any other location from where they tweet on a regular basis. Information about the address location from AddressBase is used to classify the nature of these locations (residential, commercial or other) using a nearest neighbour method.

Single isolated tweets or clusters of just two tweets are classified as 'invalid' clusters. Only valid clusters (defined as clusters or three points or more) are selected for analysis of residential mobility. Each valid cluster has a derived set of coordinates based on the weighted centroid, a classification, and a count value for the number of tweets. Clusters are also assigned a range of statistical and administrative geographies. Every user in the final dataset will have at least one cluster, even if it is an invalid cluster consisting of a single tweet. In contrast, highly active Twitter users may have multiple clusters.

This data model enables the entire data set to be stored in a framework that supports the analysis of residence and mobility but could also support other types of analysis in future.

The remainder of the report follows the logical structure presented at the beginning of this section. Section 5 covers data collection, Section 6 describes the method for deriving and classifying clusters, while Section 7 presents and discusses the results of the analyses of residential mobility. Section 8 provides the conclusion and outlines avenues for future research.

5. Data Collection

5.1 Overview

The target data for this pilot was all geolocated tweets made within Great Britain between 1 April and 31 October 2014 (i.e. seven months). The precise composition of this source data set is complex to describe as it comprises data obtained via two different methods:

- Data collected via the Public Twitter Stream (from 11 April to 14 August 2015)
- Data purchased from GNIP (1 April to 10 April 2015and 15 August to 31 October 2015)

This section explains in detail each of these approaches, why a combination of approaches was taken, and the differences between them.

5.2 Twitter API

The Public Stream of the Twitter API[7] enables tweets to be collected in real-time based on user defined criteria. These include key words in the tweet text, or geolocated tweets within a defined set of coordinates. The main advantage of this collection approach is that data is free of charge. The main limitation is that the maximum number of tweets that are available at any one time is set at 1 per cent of the current Twitter feed. If the proportion of tweets matching the selection criteria exceeds this limit then only a sample of tweets are available.

Only a small proportion of the global Twitter feed has precise geolocation data. In 2013, this was around 1.6 per cent (Leetaru, 2013). The proportion of global Twitter traffic based in the UK is about 5.6 per cent[8]. Therefore, the target number of tweets at any one time would, on average be less than 0.1 per cent of the global Twitter feed. This is well under the 1 per cent rate limit and would be only rarely exceeded. Conversations with other projects involved in using Twitter's Public Stream have raised questions about whether all the target data would have been available for collection. However, we have been able to compare a sample of data collected by this method with data we have purchased (based on all available tweets) and this confirms that the Twitter API provided the expected volumes of target data.

A Twitter application was developed in Python[9] and deployed in the ONS innovation lab[10]. The selection criteria involved a set of bounding rectangles covering the British Isles. Each tweet captured via the API contains useful information such as tweet_id, user_id, user name, timestamp, location information and tweet text, but also less useful information (e.g. various urls). Thus, while there is a limit of 140 characters of each tweet, the full payload of associated metadata provided through the Twitter API is much larger, typically amounting to several kilobytes per tweet.

The Twitter API was the sole source of data generated by the pilot from the early experiments until 15 August 2014 when this approach to data collection was halted.

5.3 Point-in-time data purchase

A decision to stop collecting data through the Twitter API was due to advice from Twitter that the application was in breach of the Twitter Developer Rules[11]. In June 2014, the Beyond 2011 Privacy Advisory Group[12] reviewed all four ONS big data pilots and questioned

[7] https://dev.twitter.com/streaming/overview
[8] Source: Statistica.com: http://www.statista.com/chart/1642/regional-breakdown-of-twitter-users/
[9] Note all the code used in this pilot is available on Github: https://github.com/niczky12/ONS_Twitter
[10] http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/ons-working-paper-series/mwp1-ons-innovation-laboratories.pdf
[11] https://dev.twitter.com/overview/terms/agreement-and-policy
[12] See www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/privacy-impact-assessment.pdf (p.18)

whether the Twitter pilot was operating within the relevant developer rules. The pilot team could not initially establish with certainty whether the application was operating within the rules. However, correspondence with Twitter UK established that it was not.

Although no definitive reason was given, the main issue appeared to be around the scale of the data collection operation. The Twitter Developer Rules states:

"If your application will need more than 1 million user tokens, you must contact us about your Twitter API access, as you may be subject to additional terms." (Section 1c)

To ensure compliance the application developed by the API was halted. The pilot was advised to contact GNIP (a reseller of data, now owned by Twitter) to discuss requirements with a view to purchasing the required data.

GNIP advised that as the data supply agreement covers the use of Twitter data, this could also be applied to cover data already collected by ONS through the API. Thus, this data was combined with purchased data to minimise project costs. Additional data was procured covering the period 15 August to 31 October 2014 and from 1 April to 10 April 2014 to give seven full months of data.

The collection of data through the Twitter API is a sometimes thorny subject. In conversations with other organisations collecting data through the Twitter API, there is generally an awareness of the Twitter Developer Rules, but these are not always followed. For example, in one organisation, a risk-based assessment was made that the worst case would be that Twitter would simply block access to their API key. In any case, there is certainly little evidence of Twitter enforcing their developer rules. These rules emphasise principles of courtesy and "being a good partner" rather than enforcement and sanctions. Thus, although the pilot could have probably continued to collect data through the Twitter API without material consequences, this would be inconsistent with the ONS aim of being a good partner organisation.

In conclusion, any large scale use of Twitter data, including any future extension of this work, would require commercial arrangements to acquire data. Based on the experience of this pilot, this would be a small fraction of the cost of running a similarly sized survey. Although there are clearly major issues around representativeness of data, there may be a business case for procuring Twitter data, providing it offers sufficient benefit.

5.4 Differences between Twitter API and point-in-time data purchase

There are a number of differences between the data collected through the Twitter API and the point-in-time data purchased through GNIP.

Geographic coverage:

The Twitter API data was defined by a series of bounding rectangles covering the British Isles, while the specification of the GNIP data was defined by tweets with a "GB" country code. The country code is a derived field based on the GPS coordinates provided by the user, but this is not an available option for selecting tweets from the Twitter API. The main reason for selecting GB tweets (as opposed to the British Isles) was to align the data collection with the coverage of AddressBase[13], which was used later to classify clusters by address type.

Missing data:

There is missing data in the Twitter API data as the result of outages in the ONS innovation lab environment. These included both planned outages (e.g. moving of IT equipment) and unplanned (broadband router failure). There are no such time gaps in the data purchased from GNIP.

Streaming vs point-in-time extraction:

The most profound difference stems from how the data are collected. The Twitter API data was collected in real-time, while the GNIP data is a point in time bulk extract. These different collection approaches result in differences in the tweets that are included in each sample.

Tweets are public information by default and can be viewed on-line, extracted via the Twitter API, or purchased as a bulk extract. However, users may opt to protect their account so that only approved followers can view their tweets. These data are not made available for analysis purposes. A user with an open account may chose to delete an individual tweet after having posted it. A user with an open account may also decide to delete their entire account. Twitter and resellers of Twitter data aim to respect user privacy by ensuring that any point-in-time extracts exclude tweets from protected and deleted accounts as well as individual deleted tweets. Thus, the data purchased from GNIP excludes all such tweets. In contrast, if a user has an open Twitter account and then subsequently protects it, or if any individual tweet is posted and then deleted, any tweets captured before these actions will remain in any data set collected in real-time data.

These differences will result in inconsistencies between the Twitter API and the GNIP extract. Data from the 10 April and the 15 August (where there was some overlap between the sources) revealed a small number of tweets in the Twitter API data that were not in the GNIP data. Analysis of a small sample found that all were associated with protected

[13] This is AddressBase Premium, the most comprehensive AddressBase product: https://www.ordnancesurvey.co.uk/business-and-government/products/addressbase-premium.html

accounts. It was therefore decided to remove all tweets for these users across the entire data set. The main reason was to respect the privacy of those who decided to protect their accounts during the period of the study. However, this was also sensible from a methodological perspective as it minimised the differences in the composition of each source.

These differences could not be eliminated entirely because users who have retrospectively protected their accounts but did not send a geolocated tweet during the two days of overlap could not be identified. Something that would help researchers respect the privacy of Twitter users in these circumstances would be if Twitter maintained a list of all protected and deleted account user identifiers. This could then be used to filter out these accounts prior to analysis and would help researchers respect user privacy.

5.5 Merged data set

The total number of records acquired through both collection methods was over 106 million. The full payload of each tweet is made available in JSON (Javascript Object Notation), a lightweight data interchange format. A key feature of JSON is data is this format does not require a schema. The structure of the data is defined within each record as a set of key-value pairs and lists. The obvious choice for storing this type of data is a NoSQL document database. MongoDB[14] was chosen as it is widely used, well documented and open source.

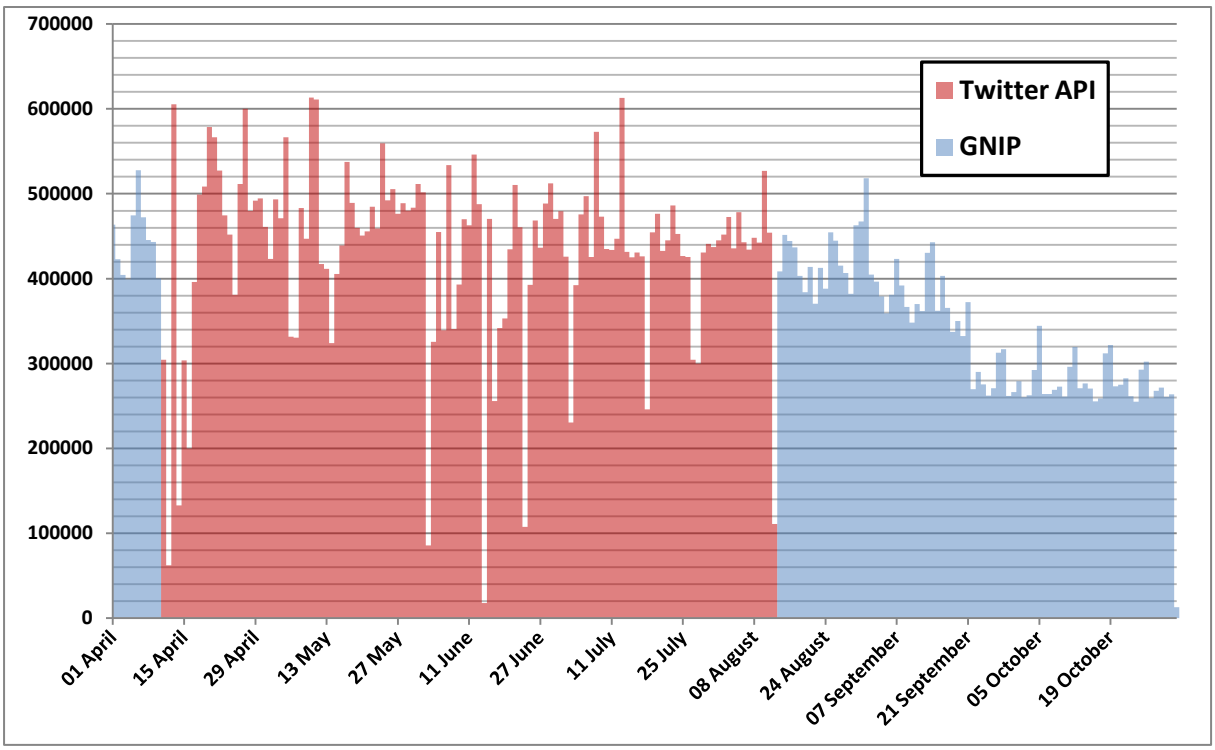
A number of processes were run to create a final clean version of the tweet-level data set for further processing. These included:

- Removal of Twitter robots. These are automated Twitter accounts that post high volumes of tweets, but do not represent the activities of a real person.
- Removal of non-GB tweets from the Twitter API data (mainly those from the Republic of Ireland)
- Removal of geolocated GB tweets without GPS precision location (e.g. sent from a desktop computer).
- Removal of a very small number of GB labelled tweets with precision coordinates that could not be assigned British Map Grid coordinates. It is assumed these had been assigned GB country codes by Twitter in error.
- Removal of duplicate tweets from the time periods on 10 April and 15 August when there were overlaps between the Twitter API and GNIP data.
- Removal of all tweets from the Twitter API relating to users that were not in the GNIP data where these two sources overlapped (as discussed in section 5.4).

[14] www.mongodb.com

Together these steps reduced the total volume of tweets to 81.4 million. The distribution of these tweets (Figure 2) shows the final merged data set showing the volume of tweets by time period.

Figure 2: Daily Volumes of Geolocated Tweets by Source (Great Britain 1 April 2014 to 31 October 2014)

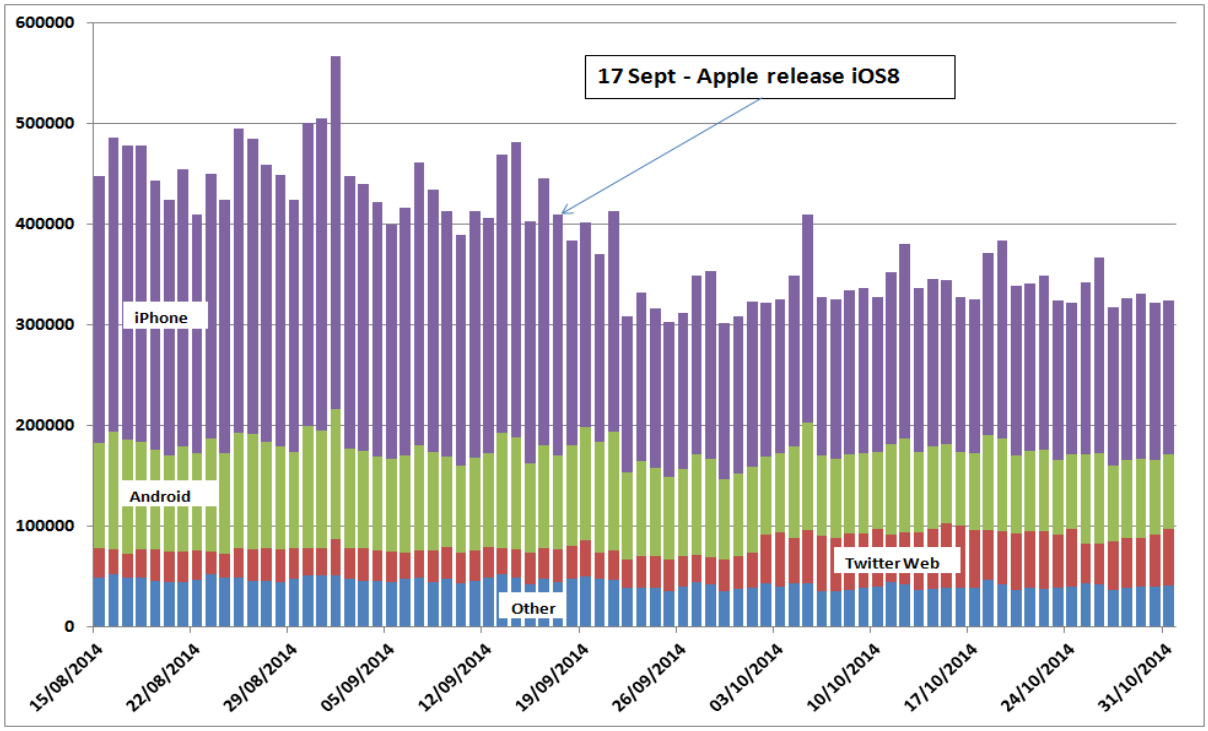


This clearly shows missing days corresponding to outages during the Twitter API collection. Apart from where there is missing data there is a regular spiky pattern, which is mostly explained by increased activity during weekends. There is also around a 25 per cent drop in daily volumes during the second half of September.

Investigations into the reason for this decline in volumes during September identified a link with the release of the iPhone iOS8 operating system. This included changes to how privacy and location are managed[15]. An analysis of tweets by device type from the GNIP data shows that this decline is indeed almost entirely explained by a decline in volumes from iPhone devices (Figure 3). This suggests that many iPhone users took the opportunity to exert greater control over their location settings which subsequently impacted the overall volume of geolocated tweets.

[15] https://support.apple.com/en-us/HT203033

Figure 3: Daily Volumes of Geolocated Tweets by Device (Great Britain, 15 August 2014 to 31 October 2014)



This raises a fundamental question around the use of social media for analysis and policy formation. It is already well established that Twitter data has issues around selection bias but this combination of technology and social change is yet another confounding factor. This drop occurred within the space of a week without any warning. Indeed, to date the impact of the release of the iOS8 on geolocated Twitter volumes seems to have gone entirely undocumented.

Not only did the volume of tweets drop, but the change resulted in a lower proportion of tweets coming from iPhone users. This is problematic because iPhone users tend to have distinct characteristics. On average, they are older than Android users and have higher socio-economic status (Nanji, 2013). Thus, certain patterns in the data will be an artefact of this interaction between changing technology and human behaviour rather than any real underlying change.

Lack of control over the data source is a well documented disadvantage of administrative data for statistical purposes (UNECE, 2011). With administrative data sources there will usually be plenty of warning of any changes that might affect a statistical output allowing contingency plans to be put in place. However, the above example illustrates that some big data sources could be affected by changes with little or no warning, including by the actions of intermediate parties. Furthermore, it may not always be clear why the source has changed, or even that the source has changed at all. Thus, analysts and policy makers basing decisions on this type of data must be extremely alert to these risks.

This also raises another ethical issue around the use of these data. The fact that a large number of iPhone users chose to exert greater control over their privacy settings following the release iOS8 raises the question as to whether these users were fully aware of what was happening to their data prior to its release. These users would have provided consent for their location data to be shared through the operating system and the supporting applications. However, this does not mean that these users were fully aware of what was happening with their location data.

It is very difficult to answer this question and this poses a dilemma for research involving this type of data. One approach is to recognise these issues, mitigate them where possible (such as described in Section 5.4) and continue cautiously with research. The alternative is to stop research completely and only continue if and when these issues can be resolved.

The problem with such a risk adverse stance would be that this pilot would not gain an understanding of this type of data, its potential benefits and its limitations. For example, if this pilot had not undertaken this research, the methodological issues of the iOS8 release and its impact on geolocated Twitter volumes would have remained undocumented. Thus, research sometimes has to continue proceed against a backdrop of ethical uncertainty. This stance can be justified on the grounds that ONS research is focused on aggregated patterns and so identifiable information about individuals will never be disclosed. In the UK, the National Statistician has recently set up a Data Ethics committee, which will be able to provide guidance on this kind of research.

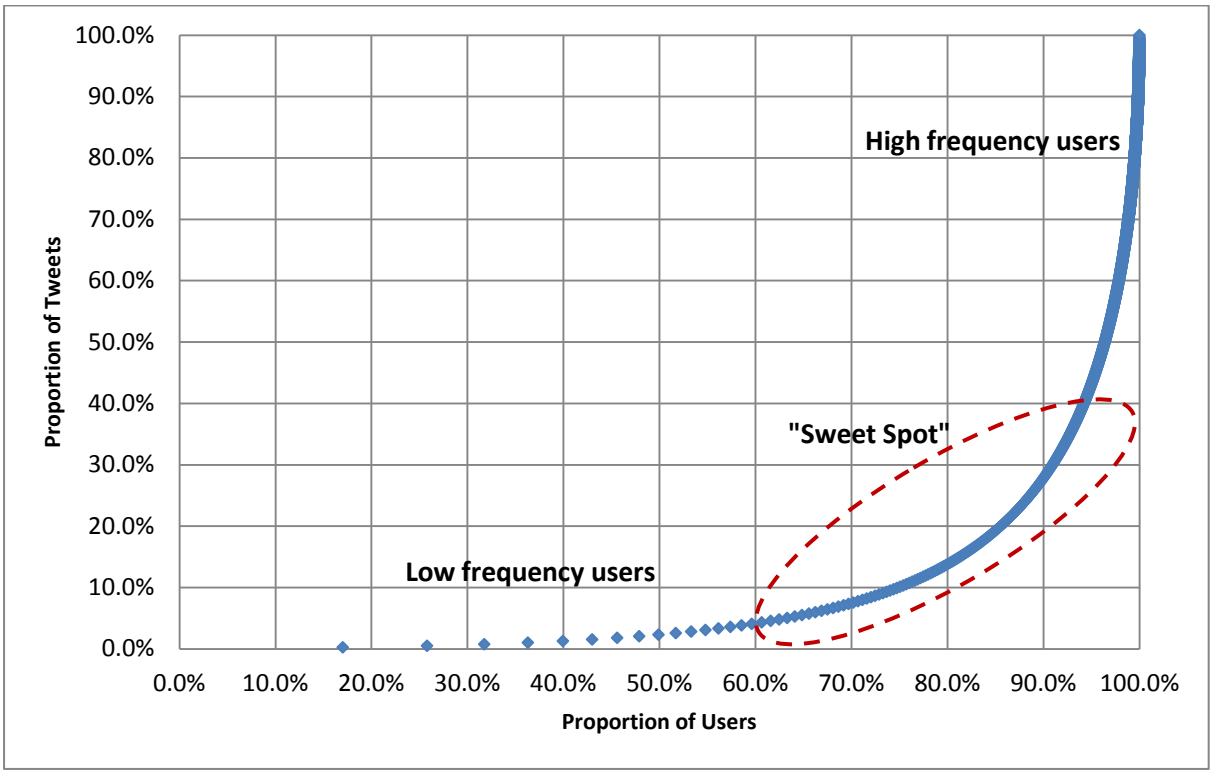
5.6 Analysis of the merged data set

This section documents preliminary analysis of the clean tweet-level data set and its key characteristics prior to clustering. This is useful in helping to understand how the methodology can be applied to the data, which is described in the following section.

Tweets by User:

A fundamental characteristic of Twitter is that some users are much more active than others (Figure 4). From the tweet-level data set, over 17 per cent of users had only one geolocated tweet over the seven month period. At the other extreme, 90 Twitter users generated more than 10,000 geolocated tweets. This means that most Twitter data is generated by a small proportion of users. More than half of all geolocated tweets were sent from just 4 per cent of Twitter accounts, while the median number of geolocated tweets by account was just 10. This is effect has been noted in other studies (e.g. Jurdak, 2014).

Figure 4: Proportion of Tweets versus Proportion of Users



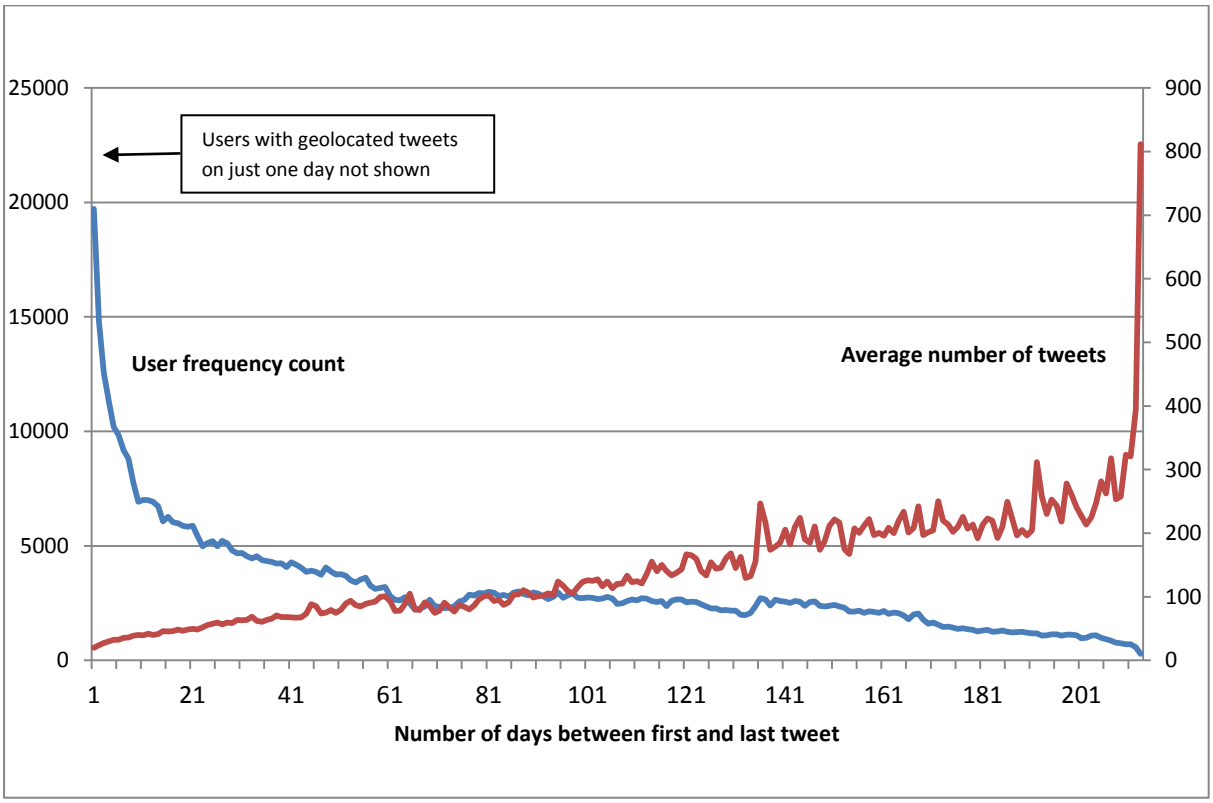
If our aim is to use geolocated activity to identify patterns of mobility then it is clearly more difficult to draw inferences for users with a small number of tweets. For those with a very large number of geolocated tweets it may be possible to build a highly detailed profile of mobility. However, there are not enough of these users to produce meaningful statistics, and so a high proportion of these tweets can be considered extraneous. There are also computational issues in dealing with users with large numbers of geolocated tweets. The most valuable data is for those users somewhere in between. This can be considered a "sweet spot" where there is sufficient data to identify patterns of mobility but without a large volume of extraneous data.

User persistence:

Another perspective related to level of usage is how persistent a user is in sending geolocated tweets over time. A simple measure of persistence is the time span in days between a user's first and last tweet. Figure 5 shows the frequency distribution of persistence by number of days (blue line) over the seven months of the study period. This declines rapidly over the first week which indicates that a high proportion of users tweet for a few days. The decline then becomes more gradual and then levels off until about 60 days. The pattern remains fairly flat until about 150 days at which point it continues a gradual decline. The median level of persistence is 47 days. Also shown is the average number of tweets for users at each of these persistence levels (red line). The average number of

geolocated tweets by user increases for different levels of persistence. This is logical since the longer a person has been active, the more tweets they are likely to make.

Figure 5: Proportion of Tweets versus Proportion of Users



These patterns suggest that many users go through a phase of sending geolocated tweets and then stop. This could be for any number of reasons from changing attitudes towards privacy, changing technology, or simply waning enthusiasm for Twitter. Thus, only a subset of users will generate enough geolocated activity to enable patterns to be detected over longer periods of time. This suggests that Twitter may be more useful for tracking longitudinal mobility patterns over periods of up to a couple of months, but may not be suitable for longer time periods (e.g. over a year). It could be that some users go through multiple phases of tweeting and/or geolocating tweets. For example, some users might only send geolocated tweets when they are go on holiday. A longer study would be required to detect these kind of patterns.

6. Anchor-Point Clustering and Classification

6.1 Overview

Having obtained a clean data set of tweets, the next step is to organise them into a framework that will support analysis of mobility patterns. The broad approach is to cluster tweets by location, identify those clusters that are in residential areas and define the residential cluster for each user with the highest number of tweets as being the most likely location of usual residence. This location is referred to as the *dominant residential cluster*. It is proposed that the dominant residential cluster can also be calculated for different time periods, for example, by month. Any changes in the dominant residential cluster across time would signal a de facto change in residence.

For statistical purposes internal migration is usually measured at a high level of geographic aggregation, that is, local authority level. Therefore, one could challenge the outline approach on the grounds that clustering to a precise location gives an unnecessary level of precision. If we are simply interested in moves between local authorities, then a much simpler method would be to aggregate the total number of tweets by local authority, define the one with the highest number of tweets as the local authority of residence and then define any changes in the dominant local authority over time as a migration move.

The main limitation with this approach is that a person's usual activity space will typically span local authorities. In 2009, an estimated 41 per cent of UK workers lived and worked in a different local authority (ONS, 2011). In addition, some Twitter users appear to be particularly active while commuting, which may also involve crossing local authority boundaries. Usual activity spaces also include activities like shopping and leisure, which may take place in a variety of locations. Therefore, a simple aggregation of tweets by local authority would often not provide an accurate indication of the local authority of usual residence. This problem will be particularly acute in areas such as London where there is a dense concentration of local authorities and user activity spaces will cross more boundaries.

In contrast, an approach which clusters tweets to specific locations, then identifies which locations are residential and finally defines the residential cluster with the highest number of tweets as being the most likely place of usual residence, should give more accurate results. This more precise approach should be particularly effective in filtering out users who tweet "on the move" between anchor points.

In summary, although clustering by location is more complex than simply aggregating by local authority, it will deliver better quality results. This approach should also provide a useful framework for other types of analyses. For example, if a user has both a dominant residential cluster and a dominant commercial cluster, then we could assume that these

represent the likely locations of both home and work and could form the basis for analysing commuting patterns.

6.2 Density-Based Spatial Clustering Algorithm with Noise (DBSCAN)

DBSCAN is a spatial clustering algorithm developed by a group of researchers at the University of Munich (Ester et al, 1996). It is based on the concept of *density-reachability* and is based on two parameters:

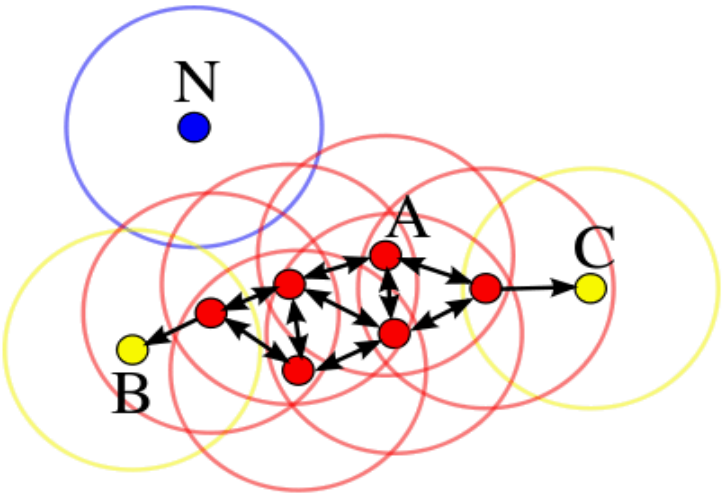
i) distance (*i*)

ii) minimum number of points (*minpts*)

The algorithm works by selecting an initial point and then searching for other points within *i*. Any points found are added to the cluster. The algorithm then keeps iterating over all points added to the cluster looking for further points to add until no further points can be found within *i*. *Minpts* is then applied to identify valid clusters. Any points in clusters with less than the required number of points (or any single unclustered points) are treated as "noise".

Core points (A) are those that satisfy both parameters, that is, the number of other points within distance (i) including itself is equal or greater to *minpts* (Figure 6). Density reachable points (B or C) are those where at least one other point is within distance *i*, but there are not enough points satisfy the condition imposed by *minpts*. Noise points (N) are those such that no other point is within distance *i*.

Figure 6: Illustration of DBSCAN Algorithm



Source: Wikipedia[16]

[16] https://en.wikipedia.org/wiki/DBSCAN

Unlike a k-means algorithm there is no need to pre-define the number of clusters. Another useful feature of DBSCAN is that it does not try to cluster all available points. This is particularly useful where a user may be sending tweets from a variety of locations, but where the interest is in identifying key anchor points, including the location of residence. Thus, infrequently visited locations can be easily filtered out. It is also important to note that the concept of density reachability means that DBSCAN algorithm can result in points in the cluster being much further apart than distance *i*. This is particularly relevant in cases where data points follow a linear pattern (e.g. a user regularly tweeting along a bus route). Thus, the distance parameter needs to be small enough to avoid clustering data showing this type of pattern.

6.3 DBSCAN Implementation

There are a number of important features about the specific implementation of DBSCAN within this pilot.

Coordinate Transformation:

The geolocated data from Twitter is provided as decimalized latitude and longitude coordinates. These values do not produce consistent clustering results between locations as they are based on a coordinate system that follows the curvature of the Earth. The problem is that the distance in metres between degrees of longitude is greatest at the equator and converges to zero approaching the North Pole. This means that a certain configuration of points in the South of England could produce different clustering results from the same configuration of points in Scotland.

To ensure consistency these coordinates were transformed into British National Grid Coordinates (BNG) using a transformation algorithm[17]. The BNG is a traverse mercator projection, which transforms the latitude and longitude coordinates on to a flat plane. The BNG system allows any location in Great Britain to be represented as a set of northing and easting coordinates from an origin point to the south west of the Isles of Scilly off the Cornwall coast. These coordinates are expressed in metres, is widely used within Great Britain and is the geographic referencing system used within AddressBase.

Treatment of density-reachable points:

This implementation makes no practical distinction between core and density-reachable points (as described in Figure 6). Both are given the same weight as we are only really concerned as to whether a group of tweets are in the same general location. The weighted centroid of all such points within a cluster is used as the generalised point location of the

[17] http://www.hannahfry.co.uk/blog/2012/02/01/converting-british-national-grid-to-latitude-and-longitude-ii

cluster. This is calculated as part of the algorithm implementation and is appended as an additional variable to each cluster.

Setting the minimum points parameter (minpts):

The minimum number of points (minpts) is set at three. In setting this parameter there is a balance to be struck between identifying good quality anchor points and having as many users with at least one anchor point. Single points and two-point clusters are assumed to be insufficient to define a meaningful anchor point. However, it is not desirable to set this parameter too high since the higher the value, the fewer valid clusters will be created. A minimum of three points was considered a pragmatic limit.

Setting the DBSCAN distance (i) parameter:

A sensitivity analysis was performed on the DBSCAN distance parameter with a series of tests comparing different values between 10 and 100 metres. This suggested 20 meters to be a suitable distance parameter. A more detailed analysis is presented in Annex A.

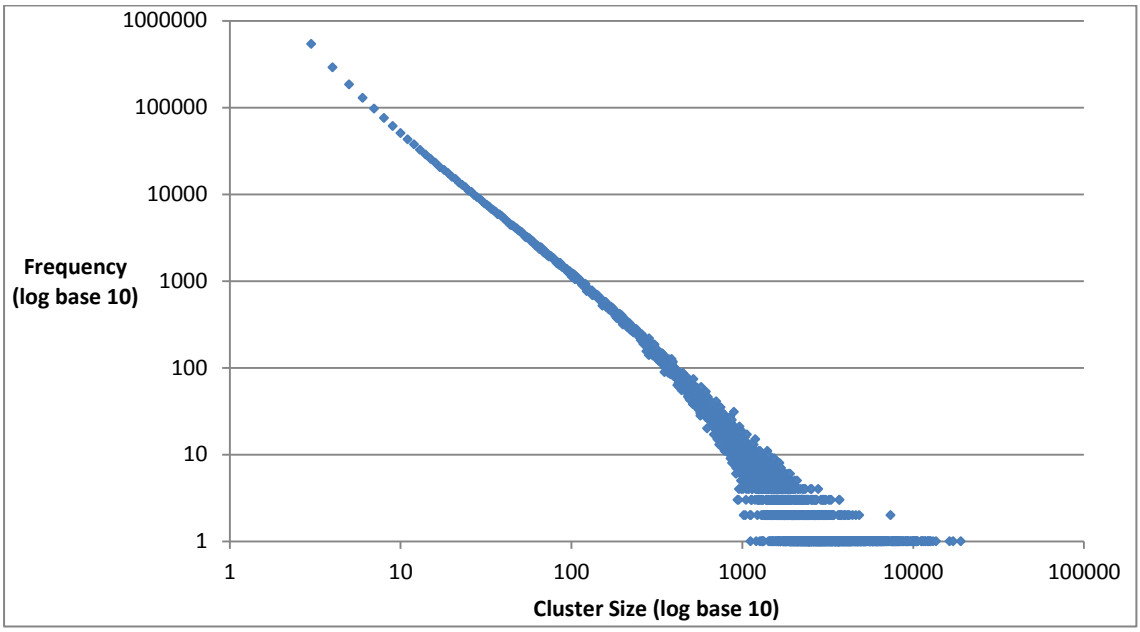
Computational efficiency:

Applying a DBSCAN algorithm to a large amount of data may be very computationally intensive because the number of calculations increases exponentially with the number of points to be clustered (Tsai & Wu, 2009). This proved to be highly relevant for this pilot where a large number of geolocated tweets are concentrated within a relatively small number of highly active users. The final version of the clustering code was able to process over 80 million data points in around 45 minutes compared with earlier versions that took over a week. Details of these steps are presented in detail in Annex B.

6.4 Clustering Results

Of the 81.4 million points in the pre-clustered data set, 67.6 million (or 83 per cent) of points formed a cluster of three points or more. The frequency distribution by cluster size is shown in Figure 7. Unsurprisingly, three point clusters (top left hand corner) are the most common with the frequency declining as cluster size increases. Some clusters are very large, with the largest containing just over 19,000 points. As the cluster size increases, the frequency values are less likely to be unique resulting in “fanning out” effect.

Figure 7: Frequency Distribution of Clusters by Size



6.5 Classification using AddressBase

The next step involved using AddressBase Premium to classify clusters by address type. AddressBase is the definitive source of address information for Great Britain and is available to Government organisations in England and Wales under the Public Sector Mapping Agreement[18]. The basic building block of AddressBase is the Unique Property Reference Number (UPRN), which operates as a persistent identifier of every Basic Land and Property Unit (BLPU) in Great Britain. Every BLPU is classified according to the BLPU Classification Schema maintained by the National Land and Property Gazetteer (NLPG)[19]. This hierarchical classification can distinguish between residential, commercial and other types of address.

A nearest neighbour method was used to identify the BLPU georeferenced address marker that is closest to each cluster centroid and then use the BLPU primary code to classify each cluster. Details of this implementation are shown in Annex C.

The results for all valid clusters are shown in Table 1. The vast majority of clusters are classified as either residential (63.9 per cent) or commercial (28.8 per cent). Around 3.4 per cent of clusters are not classified. The vast majority of unclassified tweets are from Northern Ireland, which Twitter classifies with a "GB" code even though Northern Ireland is not part of Great Britain (and therefore is not within the scope of AddressBase).

[18] https://www.ordnancesurvey.co.uk/business-and-government/public-sector/mapping-agreements/public-sector-mapping-agreement.html
[19] http://www.nlpg.org.uk/documents/LLPG_SNN_best_practice_v2.pdf

Table 1: AddressBase Primary Classification Codes

Primary Code	Primary Description	Count	%
C	Commercial	617039	28.8%
L	Land	2019	0.1%
M	Military	626	0.0%
P	Parent Shell	62318	2.9%
R	Residential	1368921	63.9%
U	Unclassified	647	0.0%
X	Dual Use	14683	0.7%
Z	Object of Interest	4285	0.2%
Not classified		72148	3.4%
Total		**2142686**	**100.0%**

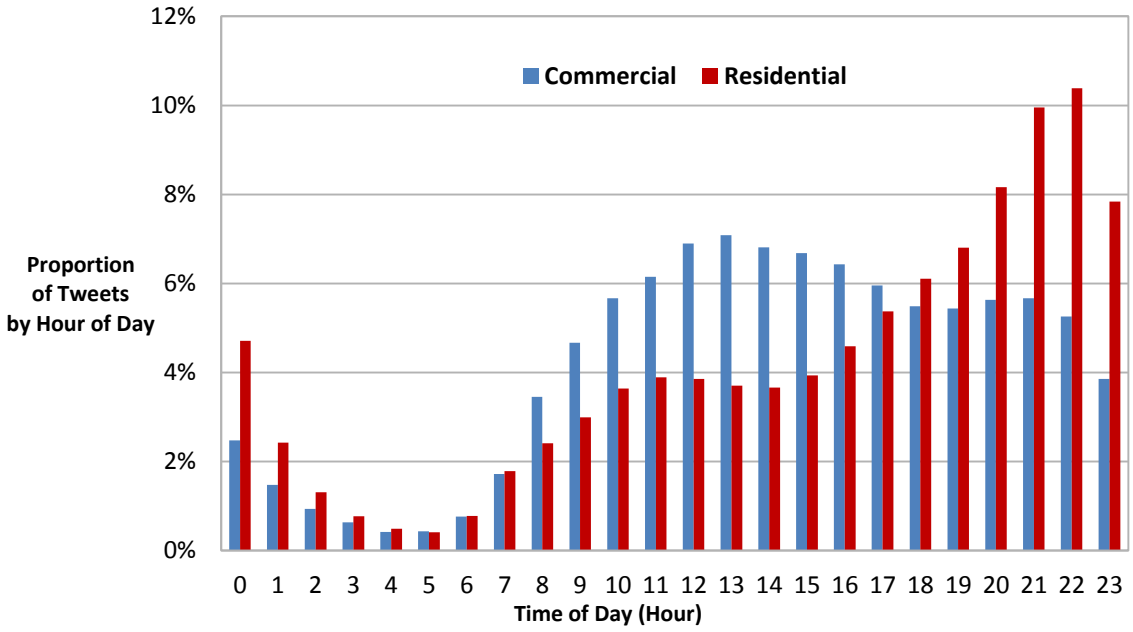
Analysis of the average number of tweets per cluster shows some major differences between residential and commercial clusters. The average size for residential clusters was 161 tweets compared with just 19 for commercial clusters. Thus, although almost 30 per cent of clusters are associated with commercial locations only about 10 per cent of all tweets are made at commercial locations.

Further analysis by type reveals that 25 per cent of commercial clusters are *Shop/Showroom*, 11 per cent were *Office/Work studio*, 7 per cent were *Public House/Bar/Nightclub* and 6 per cent were *Restaurant/Cafeteria* with the remaining 50 per cent various other types of commercial location. The low proportion of clusters at *Office/Work studio* locations seems to suggest that activity at commercial locations is associated more with leisure than with work, although of course some activity at locations associated with leisure, could be from employees working at these locations. This requires further investigation but it would seem that Twitter might have only limited application for identifying commuting patterns.

This method for classifying clusters may be subject to error in areas with a dense mix of commercial and residential activities. Areas with commercial activity at street level with residential accommodation at higher levels are particularly problematic. A more sophisticated clustering method combining location and times of day and tweet content might improve the quality of these classifications.

A comparison of the distribution of geolocated tweets by time of day shows a distinct profile for residential and commercial clusters (Figure 8). Activity for residential clusters peaks in the late evening between 21:00 and 23:00. Activity for commercial clusters is more evenly distributed throughout the day but with a slight peak at around lunchtime. This reflects a broad pattern that one would expect and confirms that this step of classifying clusters by address type does add value in terms of identifying residential based moves.

Figure 8: Distribution of geolocated tweets by address type and time of day



The final step involves deriving the dominant residential cluster for each user with a residential cluster of at least three data points. This is simply the residential cluster with the highest number of data points. This is a bold assumption and this pilot has not identified a method of validating this. It is certainly possible that this method does not always produce the right result. However, this assumption seems reasonable.

7. Analysis of Anchor Points

7.1 Total residential clusters

The number of dominant residential clusters by local authority can be divided by the population to derive an indicator of the prevalence of users who have sent geolocated tweets within each local authority. This is referred to as the *geolocated penetration rate*. There were over 563,000 dominant residential clusters produced by the clustering method defined in the previous section. This figure divided by the 2014 mid-year population estimates for Great Britain (62.7 million) thus gives a rate of 0.90%. Rates can be calculated similarly for any level of geography.

Analysis of these penetration rates is useful to understand how geolocated Twitter use varies across the country. However, the concept of the penetration rate could also form the basis of an estimation framework. The penetration rate can be thought of as being similar to a sampling fraction where the inverse provides the basis for producing survey estimates. Of course, this is much more challenging for data like Twitter, where the sample is self-selecting and unrepresentative of the population. Nevertheless, it is useful concept to bear in mind.

The rates for the top and bottom 10 local authorities in Great Britain for all dominant residential clusters is shown in Table 2.

Table 2: Top and bottom 10 penetration rates by local authority for all dominant residential clusters (Great Britain)

Rank	Local Authority	NUTS3	Population (2014 MYEs)	Dominant residential cluster count (All)	Penetration Rate (All)
1	City of London	Inner London	8,072	287	3.56%
2	Westminster	Inner London	233,292	5,800	2.49%
3	Cardiff	East Wales	354,294	5,315	1.50%
4	Camden	Inner London	234,846	3,416	1.45%
5	Kensington & Chelsea	Inner London	156,190	2,266	1.45%
6	Islington	Inner London	221,030	3,177	1.44%
7	Hackney	Inner London	263,150	3,764	1.43%
8	Southwark	Inner London	302,538	4,261	1.41%
9	Liverpool	Merseyside	473,073	6,644	1.40%
10	Hammersmith & Fulham	Inner London	178,365	2,477	1.39%
:					
371	Walsall	West Midlands	274,173	1,630	0.59%
372	North East Lincolnshire	East Yorkshire & Northern Lincolnshire	159,804	936	0.59%
373	South Holland	Lincolnshire	90,419	528	0.58%
374	Torridge	Devon	65,618	364	0.55%
375	Slough	Berks, Buckinghamshire & Oxfordshire	144,575	796	0.55%
376	Sandwell	West Midlands	316,719	1,742	0.55%
377	East Lindsey	Lincolnshire	137,623	756	0.55%
378	Orkney Islands	Highlands and Islands	21,590	115	0.53%
379	West Somerset	Dorset and Somerset	34,322	181	0.53%
380	Redcar and Cleveland	Tees Valley & Durham	135,042	706	0.52%
	Great Britain		**62,756,254**	**563,379**	**0.90%**

Eight out of the top 10 local authorities in Great Britain are Inner London boroughs. The City of London and Westminster have by far the highest penetration rates with Cardiff and Liverpool the only non-London local authorities in the top 10. The local authorities with the lowest penetration rates tend to be more spread out across the country although there are small clusters in Lincolnshire and the West Midlands.

7.2 Short-term and long-term clusters

An important factor to consider is that some dominant residential clusters may not represent users from the population of interest (i.e. those usually resident in the UK). International tourists and short-term migrants visiting the UK for less than a year are not counted as UK usual residents. Identifying these users from the data is very difficult. This is partly because the study period of seven months is not sufficient to determine their status, but also because it is difficult to distinguish between non-UK residents and the large number of UK residents who do not persist in sending geolocated tweets (see Section 5.6).

There may also be some UK residents who mostly (or only) send geolocated tweets when they are doing something outside of their usual routine, such as travelling to a different part of the country. In such cases, their place of residence would be incorrectly inferred. While it might be possible to distinguish between some of these different scenarios, for example, by looking at tweet content, this was not investigated by this pilot.

When considering Twitter traces by international visitors it is assumed that this will impact on local authority level analysis since tourism activity will tend to cluster in certain locations. However, we can also consider that data from UK residents who send geolocated tweets from their usual residential location but without doing so for more than a month are not useful in the context of measuring internal mobility patterns as there is not sufficient data to detect change over time.

If the tweets for a dominant residential cluster span a period of at least a month, then this provides a conceptual basis for filtering out some of these cases, namely, international visitors staying in the UK for less than a month, domestic tourists who do not send geolocated tweets as part of their usual activity and UK residents who geolocated tweets for less than a month. Therefore, for the purposes of investigating internal mobility patterns there is a case for focusing only on the dominant residential clusters which span a period of at least a month.

The total geolocated penetration rate was therefore split into clusters with tweets spanning time periods of 31 days or more (referred to as *long-term*) and those of 30 days or less (referred to as *short-term*). Of the total of approximately 563,000 clusters, 340,000 were long-term clusters and 233,000 were short-term. The highest short-term geolocated penetration rates are shown in Table 3.

Table 3: Top 10 Local Authorities with the highest short-term geolocated penetration rates (Great Britain)

		Dominant residential cluster penetration rates / ranks			
Local Authority	Short-term rank	**Short-term rate**	**Total rank**	**Long-term rank**	**Long-term rate**
City of London	1	1.50%	1	1	3.56%
Westminster	2	1.26%	2	2	2.49%
Kensington & Chelsea	3	0.72%	5	25	1.50%
Oxford	4	0.58%	17	62	1.45%
Camden	5	0.58%	4	6	1.45%
Southwark	6	0.56%	8	10	1.44%
Hammersmith & Fulham	7	0.56%	10	11	1.43%
Liverpool	8	0.54%	9	9	1.41%
Cardiff	9	0.54%	3	4	1.40%
Cambridge	10	0.52%	18	45	1.39%
Great Britain		**0.36%**			**0.54%**

The top 3 rankings are held by City of London, Westminster and Kensington & Chelsea. One explanation for these high short-term penetration rates are the high number of international visitors to London (17.4 million in 2014 staying at least one night[20]). These three local authorities account for less than 5 per cent of London's residential population but over 39 per cent of all accommodation bed spaces[21]. Despite these high rates, the effect of removing short-term dominant residential clusters does not affect the rankings for City of London and Westminster. However, Kensington & Chelsea drops from 5th place for all clusters to 25th place for long-term clusters.

Oxford and Cambridge hold 4th and 10th places respectively in the short-term rankings. Together with Edinburgh, these towns have the highest numbers of international visitors staying at least one night (ONS, 2014) relative to the resident population. As with Kensington & Chelsea, their rankings are affected by the removal of short-term dominant clusters with Oxford moving from 17th position (or all clusters) to 62nd (for long-term cluster only). Cambridge moves from 18th position to 45th. For remaining local authorities in the top 10, the effect of removing short-term clusters does not greatly affect the rankings.

However, there are some large rank changes in areas with lower penetration rates. Table 4 shows the largest local authority rank changes after removing short-term dominant clusters. Overall, the largest falls in rank are greater than the largest rises, with Eastbourne and Flintshire dropping over 100 places. While the pattern is not entirely clear, the 2011 Area

[20] ONS, 2014 Travel Trends: http://www.ons.gov.uk/ons/dcp171776_361237.pdf
[21] Visit England: https://www.visitengland.com/biz/resources/insights-and-statistics/research-topics/accommodation-research/accommodation-stock (Accessed on 27 July 2015)

Classification[22] provides some possible clues. Four out of the largest 10 falls in rank are for local authorities classified as *Resorts and Ports* with two classified as *Rural Scotland*. A possible explanation is the total penetration rate in these local authorities could be inflated by the activity of domestic tourists who only send geolocated tweets when they are on holiday. In contrast, six out of the top 10 rank increases are classified as either *Prosperous England,* or *English and Welsh Countryside*. The reasons for this pattern are less clear, but in any case the effect is fairly weak.

Table 4: Highest gain/loss rank changes following the removal of short-term geolocated clusters (Great Britain)

Rank	Local Authority	Area Classification Subgroup	Total Geolocated Penetration Rate		Long-term Geolocated Penetration Rate		Rank position gain/loss
			Rate	Rank	Rate	Rank	
1	Wycombe	Prosperous England	0.83%	210	0.55%	142	**68**
2	Mole Valley	Prosperous England	0.77%	265	0.50%	200	**65**
3	Pendle	Mining Heritage & Manufacturing	0.70%	328	0.46%	268	**60**
4	Malvern Hills	English & Welsh Countryside	0.83%	202	0.55%	145	**57**
5	Mid Sussex	Prosperous England	0.81%	230	0.52%	180	**50**
6	Daventry	English & Welsh Countryside	0.78%	259	0.50%	209	**50**
7	Maldon	English & Welsh Countryside	0.78%	261	0.50%	213	**48**
8	Epsom and Ewell	Prosperous England	0.90%	134	0.60%	87	**47**
9	Warwick	Prosperous England	0.88%	156	0.57%	109	**47**
10	Gedling	Mining Heritage & Manufacturing	0.76%	280	0.48%	234	**46**
371	Denbighshire	Remoter Rural	0.88%	157	0.49%	216	**-59**
372	Scarborough	Coastal and Rural	0.90%	135	0.51%	195	**-60**
373	Highland	Rural Scotland	0.98%	81	0.54%	149	**-68**
374	Wrexham	Mining Heritage & Manufacturing	0.88%	158	0.48%	229	**-71**
375	Thanet	Resorts and Ports	0.84%	196	0.45%	278	**-82**
376	Clackmannanshire	Rural Scotland	0.83%	209	0.44%	294	**-85**
377	Worthing	Resorts and Ports	0.90%	136	0.48%	230	**-94**
378	Hastings	Resorts and Ports	0.81%	224	0.42%	320	**-96**
379	Flintshire	Mining Heritage & Manufacturing	0.87%	171	0.45%	279	**-108**
380	Eastbourne	Resorts and Ports	0.96%	97	0.50%	208	**-111**

In conclusion, there is some evidence of confounding effects linked to international tourism and some similar effects for domestic tourism. However, the relationships are complex and difficult to unpick. For example, a cluster derived from a short-term visitor tweeting from a hotel should be classified as a commercial rather than a residential cluster and therefore should already be filtered out. Thus, these short-term visitor effects would be caused by the

[22] http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/index.html

subset of those staying at residential addresses. It is likely though that the vast majority of these short-term clusters are made by UK residents whose geolocated tweets span a month or less, which are not useful in terms of identifying mobility patterns. For these reasons, the remainder of this analysis focuses on long-term clusters only.

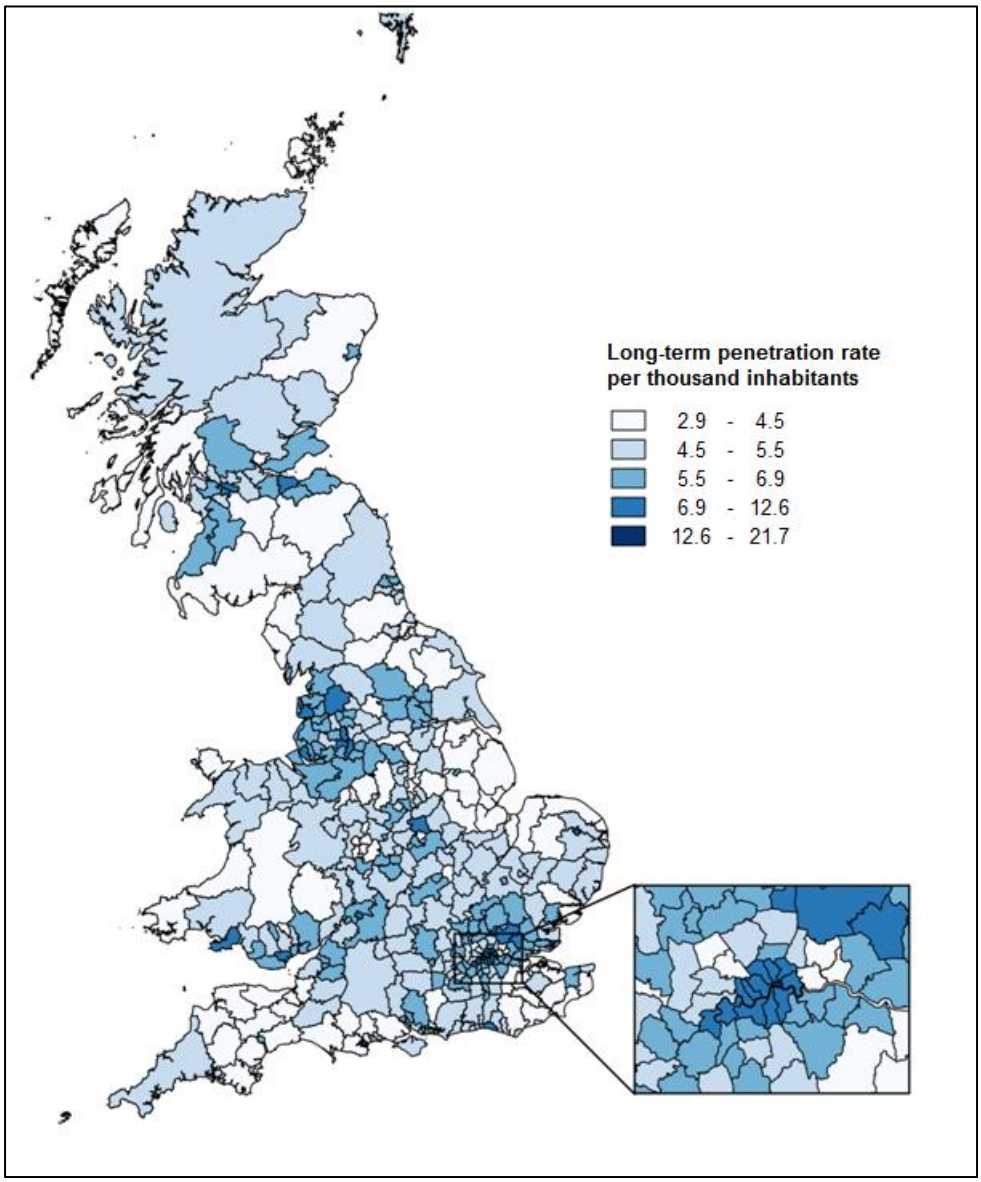
7.3 Long-term geolocated penetration rates

The long-term geolocated penetration rate for Great Britain is 0.54%. This means that there are approximately 185 UK residents for every Twitter account for which there is sufficient geolocated information to infer a location of residence over a period of at least a month. For local authorities in Great Britain the rates vary from 2.06% for the City of London to 0.29% in Slough. The rates vary by Great British region/country from 0.63% in the North West to 0.45% in the West Midlands. The long-term geolocated penetration rates by local authority for the whole of Great Britain are shown in Figure 9.

Broadly speaking urban areas have higher penetration rates than rural areas. The highest penetration rates are in Central London, Essex, and the main urban centres in the North West, Scotland and South Wales. Areas of low penetration include rural areas in South West England, Kent, Wales, Norfolk, Lincolnshire and North Yorkshire. This skew of Twitter usage towards urban areas has also been found in other studies (e.g. Mislove, 2012)

However, there are some more nuanced patterns within this broad picture. For example, when grouped by the 2011 Census Area Classification, *London Cosmopolitan Central* has a penetration rate of 0.87% whereas the rate for *Cosmopolitan North London* is less than half this rate at 0.42%. There are also pockets of low penetration rates in urbanised areas of the West Midlands including Birmingham (0.43%). Thus, while there is a general association between high penetration rates and levels of urbanisation, there are clearly other factors at play.

Figure 9: Long-term geolocated Twitter penetration rates



7.4 Comparisons with 2011 Census Data

An investigation into possible other factors was made by searching for correlations between long-term penetration rates by local authorities in England and Wales and a range of 2011 Census variables. Selected correlations[23] are shown in Table 5.

[23] Correlations are with the 2011 Census variables calculated as rates. Analysis is for England and Wales only

Table 5: Local authority level correlations between geolocated penetration rates and 2011 selected Census variables (England and Wales)

2011 Census Variable	Pearson's Correlation Coefficient
Ethnic Group: Asian/Asian British: Chinese	0.605
Adult Lifestage: Age 25 to 34: No dependent children in household	0.589
Dwelling Type: Unshared dwelling: Flat, maisonette or apartment	0.584
Marital Status: Single (never married or never registered a same-sex civil partnership)	0.578
NS-SeC: 1.2 Higher professional occupations; measures: Value	0.573
Industry: M Professional, scientific and technical activities	0.576
Highest Level of Qualification: Level 4 qualifications and above	0.507
Sex Ratio	0.390
Economic Activity (Student): Full-time students: Economically inactive	0.356
Marital Status: Married	-0.472
Age: Over 60	-0.400
Highest Level of Qualification: Level 2 qualifications	-0.509
Industry: F Construction	-0.512
Age: 10 to 17	-0.518
Dwelling Type: Unshared dwelling: Whole house or bungalow	-0.589
Social Grade: C2 Skilled manual occupations	-0.617

There is a positive correlation (R = 0.589) for the proportion of the population aged 25 to 34 and with no dependent children. In 2014, about quarter of Twitter users in the UK were in this age group[24] although they comprise less than 14 per cent of the UK population.

This analysis also suggests a link with higher educational attainment (also identified by Koetsier (2013)) and socio-economic status (also identified by Sloan et al (2015)). A positive correlation with the sex ratio suggests higher Twitter penetration rates for men, which is consistent with previous studies (e.g. Mislove et al, 2012; Zagheni et al, 2014). There are other demographic characteristics that tend to be co-correlated with these primary factors, such as marital status (e.g. younger population more likely to be single) and industry of occupation (higher socio-economic groups less likely to work in construction).

Although these findings are broadly consistent with previous research, there are obvious limitations with correlating data aggregated at the local authority level. There is likely to be other confounding factors that are masking the relationships between the geolocated penetration rates and the aggregate local authority characteristics. As such, these correlations do not provide much additional insight into spatial distribution of penetration rates. A more complete understanding might be achieved by deriving socio-demographic characteristics from the Twitter data itself.

[24] eMarketer.com http://www.emarketer.com/Article/More-than-One-Fifth-of-UK-Consumers-Use-Twitter/1010623 Accessed on 02-08-2015

Another important confounding factor may stem from inferring residence from the full seven months of data. This is particularly applicable in the case of students, especially considering that study period ran from 1 April to the 31 October. This period covered the university summer break as well as the Easter recess. For about half of this period many students would not have been present at their term time address. It is likely then than some students would have been assigned a residence to their home address rather than the term time address, which is the residential definition on which 2011 Census estimates are based

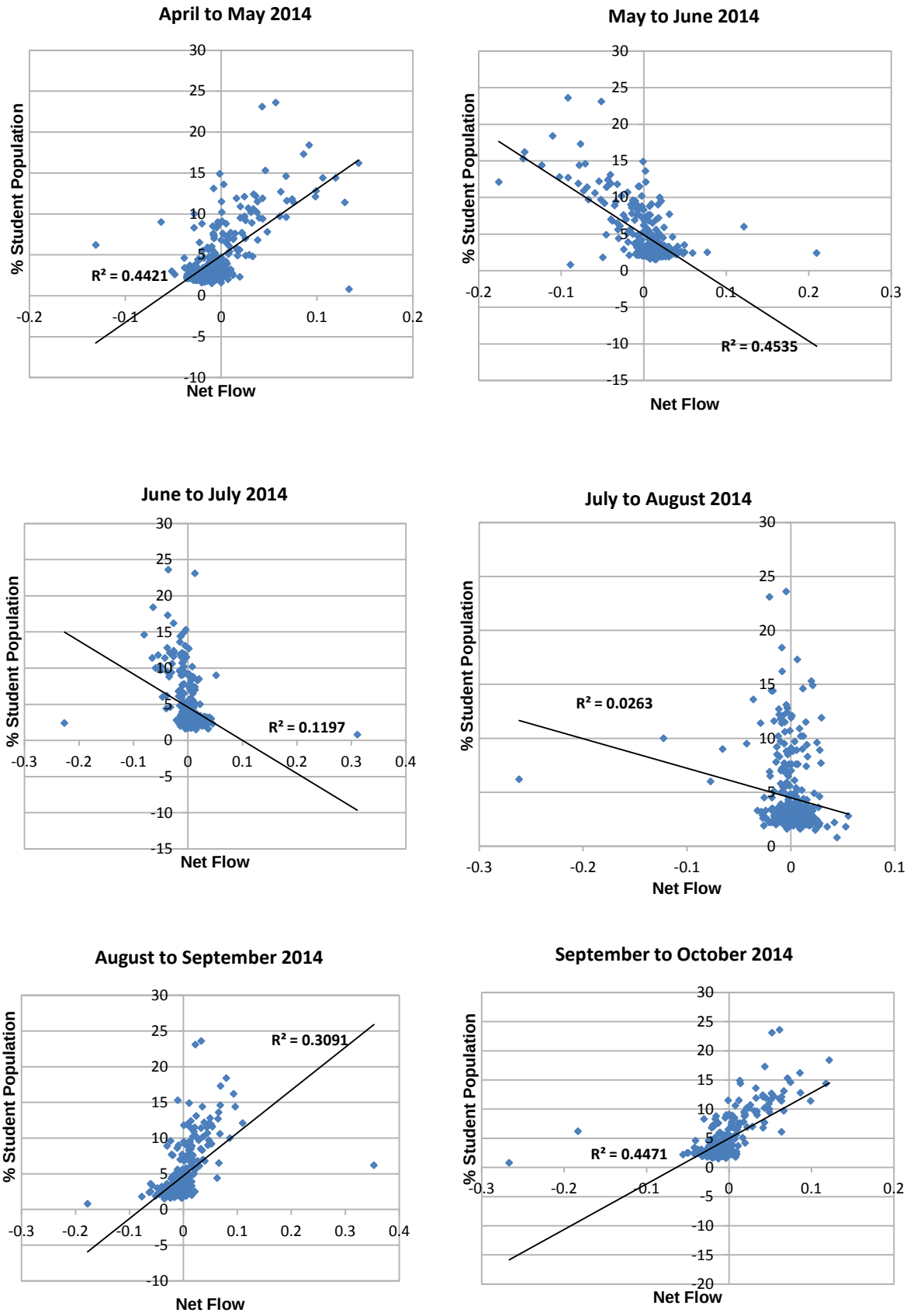
This most likely explains why the correlation between geolocated penetration rates and the proportion of full time students (R=0.356) is much weaker than what we might expect. It is interesting to note that the highest positive correlation was with Chinese ethnicity. This could be related to Chinese student immigrant populations and the fact that such populations will not have a home address within Great Britain. This is another area requiring further research.

7.5 Student Mobility

This section discusses work undertaken to tackle the issue around term-time and home address by breaking the data into months and inferring residence separately for each month. This also provides a basis for exploring net flows between local authorities for each month. These net flows are derived by producing dominant residential clusters by month and then summing the instances where the dominant cluster for a user is in a different local authority than the previous month. These changes are assumed to represent a de facto change of residence.

The monthly pattern of net flows reveals a striking series of correlations with the proportion of students over 18 and in full-time study based on the 2011 Census (Figure 10). A positive slope indicates a net flow into student areas while a negative slope indicates a net flow out of student areas. From April to May, the gradient slopes upward with a moderate correlation, representing a net flow into areas with higher student populations. This is consistent with a pattern of students returning from the Easter break to sit end of year exams. From May to June, the gradient slopes downward suggesting a net outflow, which corresponds to the end of studies.

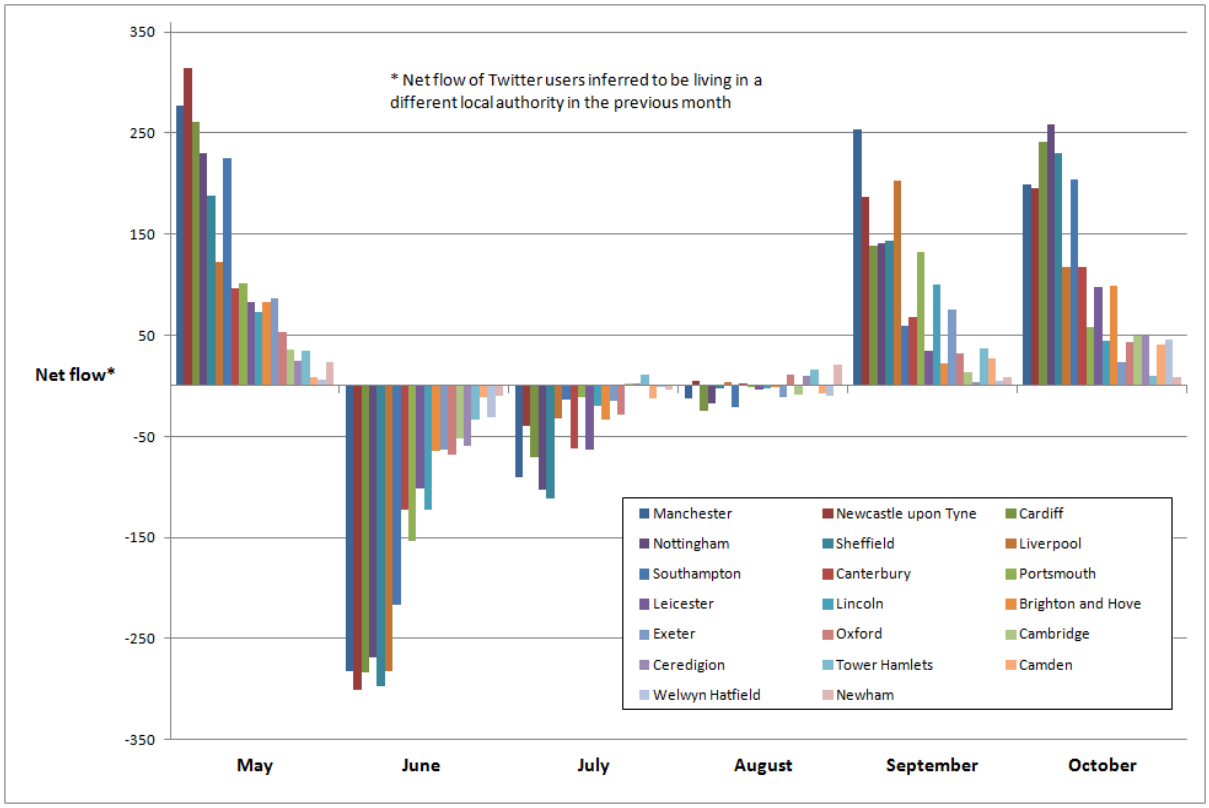
Figure 10: Scatterplots of Monthly Net Flows and % Student Population by Local Authority (England and Wales)



From June to July, and from July to August the correlations are very weak indicating lower levels of student mobility. From August to September the gradient is resumes an upward slope with a strengthening correlation between September and October. This is consistent with a pattern of students moving back into student areas at the beginning of the academic year.

Figure 11 shows the same month on month comparison of flows for the LAs with the highest proportion of students. The pattern is strongest for larger regional centres (e.g. Sheffield, Newcastle, Manchester). The pattern for Oxford and Cambridge is weaker despite having a very high proportion of students in the population. The reason for this is unclear, but one possibility is that it relates to international students who have a home address outside of the Great Britain and any geolocated Twitter activity outside of Great Britain would not have been picked up in this study. In future in may be possible to identify these students in which cases these could be removed from this analysis. The pattern of net flows is weakest in London. This may relate to a more general problem of usual activity spaces crossing many local authorities. A dense mix of different address types may also be a factor.

Figure 11: Net flows by month for the 20 local authorities in England and Wales with the highest proportion of students



These results show that it is possible to detect distinct movements of the population in student areas throughout the year. This data could help explain particular anomalies. For

example, there is anecdotal evidence that in some areas, students are removed very quickly from the GP patient register after they complete their studies, that is, before the mid-year extract is provided to ONS. This has a cumulative effect of undercounting student age populations in these areas. While these patterns of student move during different times of the year is hardly surprising, having more reliable evidence of when the timing and size of these moves could help support methods for improving population estimates.

However, there is further potential within this data that remains unexplored. For example, this pilot has not attempted to identify which Twitter users are actually students. This certainly could be feasible by using data from their user profile, their tweet content and which accounts they follow. Building a socio-demographic profile of Twitter users and their associated mobility patterns would provide more reliable evidence and so this would be an avenue for further research.

Although this pattern of student mobility is hardly surprising, it is a useful illustration that usually resident population estimates, while vitally important, do not provide an indication of short-term mobility. In reality, people are constantly on the move, whether it is students moving between a term-time and home address, commuters travelling between home and work, people congregating for large events, or people engaged in domestic and international tourism. Various de facto population bases such as daytime, weekend, and seasonal populations, may be useful for a range of policy purposes such as, transport planning, service provision, and civil defence. Twitter may be able to provide timely insight into these alternative population bases that existing sources and estimation frameworks cannot.

8. Conclusion

This pilot has demonstrated that it is feasible to use geolocated activity traces from Twitter to infer de facto residence and to identify mobility patterns. DBSCAN is well-suited for identifying the anchor points of Twitter users while AddressBase is useful for identifying which anchor points are residential. The residential anchor point with the highest number of data points, or the *dominant residential cluster*, can be assumed to be the user's de facto location of residence.

These data can then be aggregated by geography and compared directly with existing population statistics. For example, it can be used to calculate penetration rates by local authority, which in turn can be compared with Census and mid-year population estimates. Thus, it is feasible to use Twitter to support analysis for existing demographic accounting frameworks. Also, as Twitter data is continually generated, it is possible to analyse intra-year mobility patterns. Month on month change in the nu**m**ber of dominant clusters shows a distinct ebb and flow for local authorities with high student populations that follows the cycle of the academic year. Such insights cannot be detected from existing data sources.

However, there are also important limitations with Twitter data. First of all, Twitter users are not representative of the general population. They tend to be young adults with high socio-economic status. So, while this pilot has shown that it could be used to detect mobility patterns for students, it could not be used for example, to detect equivalent patterns for people over retirement age. Also, within the Twitter user base there is huge variation in the level of use both in terms of number of tweets and for how long users keep sending geolocated tweets. This means that only a subset of Twitter users provide enough information for mobility patterns to be detected.

One of the biggest problems with Twitter data is the lack of control over the data source. The release of Apple's iOS8 operating system in September 2014 has had a major impact on the volume of data collected and very likely the socio-demographic characteristics of users sending geolocated tweets. This problem is of a magnitude far beyond that encountered with traditional sources for official statistics. Thus, it is very difficult to see how Twitter, or indeed any social media data, could be incorporated directly into current frameworks for official statistics.

At the same time, we cannot ignore the fact that Twitter has the potential to provide new insights that current statistics, including thosethat could be used to improve government decision-making. One solution could be a new framework for big data and official statistics. This might include explicit recognition that some big data sources lack stability and NSIs have no control over them. Therefore, statistics based on these sources might not have the longevity that is traditionally associated with official statistics. Such a framework should also require clear advice and guidelines on the limitations of big data for informing public policy.

Future Research:

This pilot has produced a large and rich data set and there are a number of potential avenues for further analysis.

One option for further analysis would be to repeat the analysis described in Section 7, at a lower level of geography. Middle Super Output Areas (MSOAs) are a statistical geography within local authorities and contain between 5000 and 15000 people. The size of the Twitter "sample" is large enough to allow meaningful analysis at this level and offers greater granularity and units of more consistent size. Another avenue would be to repeat the penetration rates analysis based on term time clusters only (i.e. May, September, October). This would help remove the confounding effects of mobility between term-time and home address.

There may also be scope for improving the methods set out in this research. One interesting possibility would be to incorporate a time of day and/or day of week element into the clustering algorithm. This could provide further insight into the nature of different anchor points and provide a more reliable method for identifying residence. Another possibility

would be to look at methods for analysing user profiles and tweet content to identify international tourists. One might suppose their tweets would be quite different from those of the resident population. This might in itself provide new insights to support tourism policy, but could also be used to improve overall user segmentation, and thus improve the quality of analysis of the resident population.

A very important direction of future research is the development of methods to derive socio-demographic indicators of users based on information within their Twitter profiles and their corpus of tweets. At one level this is useful in terms of gaining a deeper understanding of how these characteristics relate to specific residential and mobility patterns. However, this information would also be useful in gaining an understanding of issues around selectivity and whether it might be possible to overcome these issues to develop an inferential framework for producing estimates. There is already some considerable research interest in deriving socio-demographic characteristics from Twitter (e.g. Sloan et al, 2015; Daas & Burger, 2015). The problem of selectivity is one of the biggest methodological challenges for the use big data within official statistics frameworks and so undertaking work in this area is a high priority.

References:

Backlund H, A. Hedblom, N. Neijman, 2011, Linkopings Universitet, *DBSCAN - A Density-Based Spatial Clustering of Application with Noise,* Available at: http://staffwww.itn.liu.se/~aidvi/courses/06/dm/Seminars2011/DBSCAN(4).pdf Accessed on 25-03-2014

Bawa-Cavia, 2010, "Using location-based social network data in urban analysis", Urbagram, Available at: http://urbagram.net/media/SensingTheUrban-WP.pdf Accessed on 01-08-2015

Blanford J., Z Huang, A Savelyevm A M Mac Eachren, 2015, "Geolocated Tweets. Enhancing Mobility Maps and Capturing Cross-Border Movement", PLoS One, Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4473033/ Accessed on 01-08-2015

Brogueira G., F Batista, J P Carvalho, "Using Geolocated Tweets for Characterization of Portuguese Administrative Regions", Agilie 2015 Conference, Available at: http://www.agile-online.org/Conference_Paper/cds/agile_2015/shortpapers/102/102_Paper_in_PDF.pdf Accessed on 01-08-2015

Daas P., Roos M.,Van de Ven M.,Neroni J. (2012) "Twitter as a potential source for statistics", CBS Netherlands, http://www.cbs.nl/NR/rdonlyres/04B7DD23-5443-4F98-B466-1C67AAA19527/0/201221x10pub.pdf Accessed on 13-05-2015 25-07-2015

Daas P. Burger, J. (2015), "Profiling Big Data Sources to Assess their Selectivity", NTTS 2015 Conference Proceedings, Brussels, Available at: http://www.cros-portal.eu/sites/default/files//Presentation%20S17AP1.pdf Accessed on 16-08-2015

eMarketer.com, 2015, "More than one fifth of UK consumers use Twitter", Available at: http://www.emarketer.com/Article/More-than-One-Fifth-of-UK-Consumers-Use-Twitter/1010623 Accessed on 02-08-2015

Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining http://www2.cs.uh.edu/~ceick/7363/Papers/dbscan.pdf Accessed on 27-06-2015

INEGI, 2015, "Mobility Analysis from Twitter Data", 2015 NTTS Conference Satellite Workshop, http://www1.unece.org/stat/platform/download/attachments/109252755/Big%20Data%20Satellite%20Workshop.pptx?version=1&modificationDate=1425992394314&api=v2 Accessed on 31-07-2015

Huang Q., Cao G., Wang C., 2014. From Where Do Tweets Originate? - A GIS Approach for User Location Inference. In Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN '14), Available at:

http://faculty.ce.berkeley.edu/pozdnukhov/lbsn14/camera-ready/WhereTweetsOriginate.pdf Accessed on 01-08-2015

Hawelka, B., I Sitko, E Beinat, S Sobolevsky, P Kazakopoulos and C Ratti, 2013 "Geolocated Twitter as the proxy for global mobility patterns" http://arxiv.org/abs/1311.0680 Accessed on 19-03-2014

Jurdak, R., K Zaho, J Liu, M AbouJaoude, M Cameron, D Newth, 2014, "Understanding Human Mobility form Twitter", PLoSONE, Avaliable at: http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0131469.s001 Accessed on 18-08-2015

Koetsier, J. 2013, "Only 16% of U.S. adults use Twitter, but they are young, smart and rich". Available at: http://venturebeat.com/2013/11/04/only-16-of-u-s-adults-use-twitter-but-theyre-smart-young-and-rich/ Accessed on 18-03-2014

Leetaru, K., S. Wang, A. Padmanabhan, and E. Shook. 2013. "Mapping the Global Twitter Heartbeat: The Geography of Twitter." First Monday 18 (5) Available at: http://firstmonday.org/article/view/4366/3654 Accessed on 17-08-2015

Mislove A, S Lehmann, A Yong-Yeol, J Onnela, J N Rosenquist, 2012, "Understanding the Demographics of Twitter Users", Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, Available at: http://www.ccs.neu.edu/home/amislove/publications/Twitter-ICWSM.pdf, Accessed on 02-08-2015

Nanji A, 2013, How iPhone and Andriod Ownership varies by demographic, Available at: http://www.marketingprofs.com/charts/2013/10957/how-iphone-and-android-ownership-varies-by-demographic , Accessed on 15-08-2015

Office for National Statistics (ONS), 2009, "Final Population Definitions for the 2011 Census", Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/2011-census-questionnaire-content/final-population-definitions-for-the-2011-census.pdf , Accessed on 01-08-2015

Office for National Statistics (ONS), 2011, "Commuting to Work 2011", Available at: http://www.ons.gov.uk/ons/dcp171776_227904.pdf , Accessed on 23-06-2015

Office for National Statistics (ONS), 2012, "An Improved Method for Estimating Student Migration", Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/population-and-migration/internal-migration-improved-method-of-estimating-student-migration.pdf , Accessed on 16-08-2015

Sloan L., J Morgan, W Housley, M Williams, A Edwards, P Burnap, O Rana, 2013, "Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter", Sociological Research Online, Available at: http://www.socresonline.org.uk/18/3/7.html , Accessed on 02-08-2015

Sloan L., J Morgan, P Burnap, M Williams, 2015, "Who Tweets? Deriving the Demographic Characteristics of Age, Occupation, and Social Class from Twitter Use Meta-Data", Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115545 , Accessed on 02-08-2015

Steiger E., J Lauer, T Ellersick, A Zipf, 2014, "Towards a framework for automatic geographic feature extraction from Twitter", 8th International Conference on Geographic Information Science Vienna, Available at: http://koenigstuhl.geog.uni-heidelberg.de/publications/2014/Steiger/GISCIENCE_Steiger_et_al_2014_geographicfeatureextraction.pdf , Accessed on 01-08-2015

Tsai C, C. Wu, 2009, "GF-DBSCAN A New Efficient and Effective Data Clustering Technique for Large Databases", Proceedings of the 9th WSEAS International Conference on Multimedia Systems and Signal processing", Available at: http://www.wseas.us/e-library/conferences/2009/hangzhou/MUSP/MUSP38.pdf , Accessed on 15-08-2015

Turner A, N Malleson, 2012 "Applying geographical clustering methods to analyse geolocated open micro-blog posts", Proceedings of GISRUK 2012, Available at: http://www.geos.ed.ac.uk/~gisteac/proceedingsonline/GISRUK2012/Papers/presentation-82.pdf , Accessed on 01-08-2015

UNECE, 2011, "Using Administrative and Secondary Sources for Official Statistics – A Handbook of Principles and Practices", Available at: http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf , Accessed on 15-08-2015

UNECE, 2015, Mexico (INEGI) Tweet Analysis http://www1.unece.org/stat/platform/display/BDI/Mexico+%28INEGI%29+-+Tweet+Analysis , Accessed on 28-06-2015

United Nations, 2008, "Principles and Recommendations for Population and Housing Censuses – Revision 2", Available at: http://unstats.un.org/unsd/demographic/sources/census/docs/P&R_Rev2.pdf , Accessed on 01-08-2015

Wayant N., A Crooks, A Stefanidis, A Croitoru, J Radzikowski, J Stahl, J Shine, "Spatiotemporal Clustering of Twitter Feeds for Activity Summarization", GIScience, Available at http://css.gmu.edu/andrew/pubs/giscience2012_paper_78.pdf , Accessed on 01-08-2015

Xiao-Yoing T, H Xiao-Pu, W Bing-Hong, Z Tao, 2013, "Diversity of individual mobility patterns and mergence of aggregated scaling laws", Scientific Reports, Available at: http://www.nature.com/srep/2013/130918/srep02678/full/srep02678.html Accessed on 31-07-2015

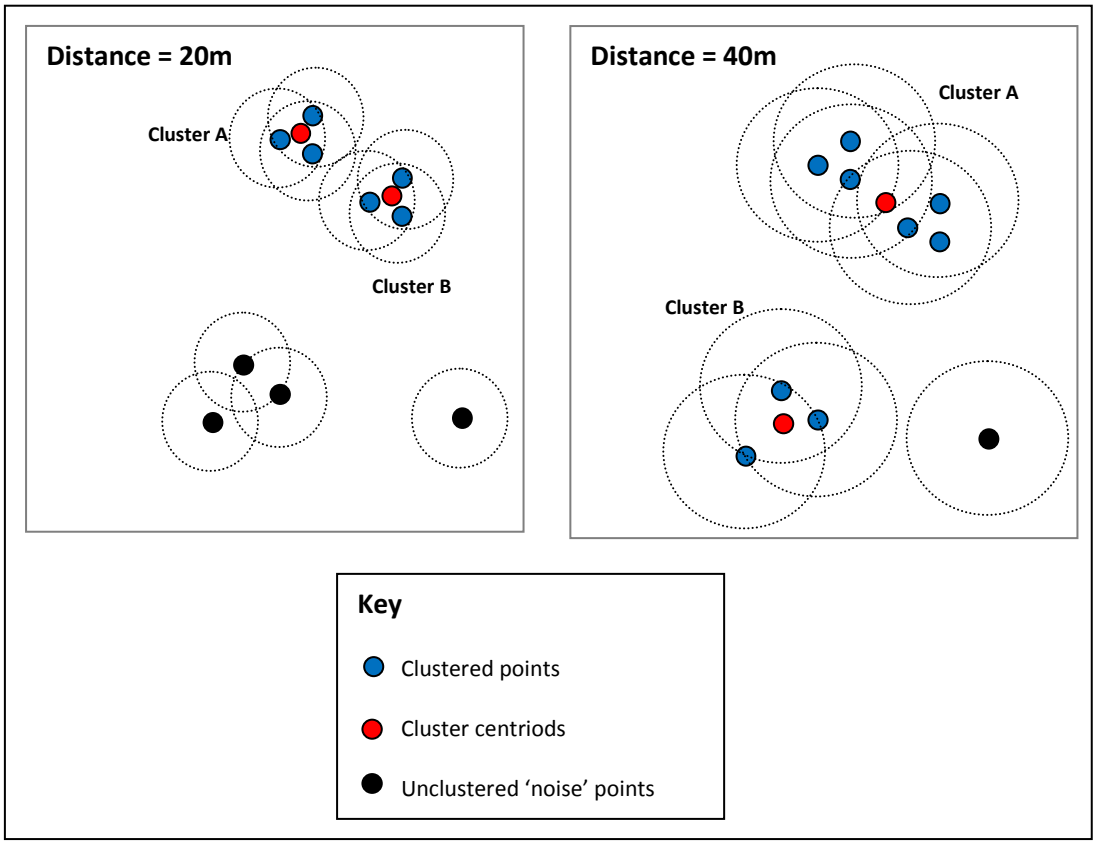
Zagheni E, V Rama Kiran, G Ingmar Weber, B State, “Inferring International and Internal Migration from Twitter Data”, Zagheni.net, Available at: http://zagheni.net/uploads/3/1/7/9/3179747/wbsc03-zagheni.pdf , Accessed on 31-07-2015

Annex A: Selecting the distance parameter for DBSCAN

GPS precision data is not fully accurate. The GPS Standard Positioning Service set out by the U.S. Government[25] offers a worst-case pseudorange accuracy of 7.8 metres with 95 per cent confidence. However, there are other factors that may affect positional accuracy such as, atmospheric effects, receiver quality and sky blockage. The latter is particularly relevant in the context of identifying geolocated Twitter activity from residential addresses, which will typically be generated inside a building. However, it is important to consider the accuracy needed to define the location of an anchor point. The aim is to be able to associate activity to an address not, for example, to a specific room within a building. Thus, ten metres was chosen as a minimum distance value.

There are some complex effects that occur as the value of the distance parameter is increased. Some points that were noise points at a lower distance parameter will come into range of two-point proto-clusters, thus increasing the number of valid clusters. In contrast, some clusters within close proximity may become absorbed into one larger cluster, thus reducing the number of clusters (Figure A1). However, the former cases outnumber the latter and so the number of anchor points produced tends to increase as the distance parameter increases.

Figure A1: Illustration of varying distance parameter on DBSCAN cluster formation



[25] http://www.gps.gov/systems/gps/performance/accuracy/#difference

Another effect is that as clusters merge or as new points are absorbed into existing clusters, the cluster centroids tend to shift. In general, increasing the distance parameter tends to increase the distance that a centroid can move. However, since the centroid is weighted by the number of points included in the cluster, the degree of movement is inversely proportional to the size of the cluster. Thus, small clusters are more sensitive to changes in distance than larger clusters.

A more profound effect is where increasing the distance parameter affects the pattern of cluster formation so that the dominant residential cluster shifts to a completely different location. The shifting location of the dominant cluster centroid represents an error related to the correct geographical identification of the cluster centroid.

This pilot did not initially set out to identify an optimised distance parameter. This investigation was undertaken after considerable effort had already been expended using a distance parameter of 20 metres. The following analysis aims to establish whether the value is sensible.

The optimal value of the distance parameter can defined as one that:

a) Maximises the number of dominant clusters

b) Minimises the centroid identification error

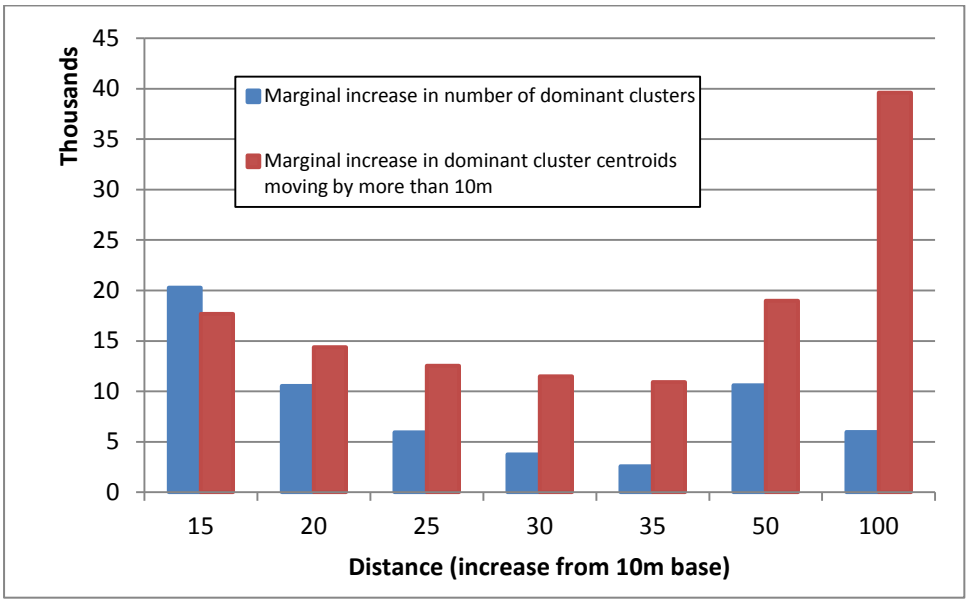
The latter is defined as being the number of instances where the dominant cluster centroid moves by more than 10 metres. This is the same value as the minimum distance parameter as it is assumed to be a typical distance separating addresses within Great Britain. Figure A2 shows the marginal effects of increasing the DBSCAN distance parameter from the minimum value of ten metres[26].

The results show that the net increase in the number of clusters formed declines as the distance parameter is increased. Increasing distance from 10 to 15 produced more additional clusters than the number of cluster centroids that moved by more than 10 metres. Moving from 15 to 20 metres produced fewer additional clusters than the number of centroids that moved more than 10 metres.

However, the percentage impact of these different values on cluster formation was quite small. Increasing the distance from 15 to 20 metres delivered an increase in dominant clusters of 1.9 per cent, while the number of dominant clusters that shifted more than 10 metres increased by 2.5 per cent. These differences are assumed to be small enough to justify setting the distance parameter at 20 metres.

[26] This analysis is based on a subset of the final data set using data collected between April and August 2014 using the Twitter API

Figure A2: Illustration of varying distance parameter on dominant cluster formation



Annex B: Improving run-time performance of DBSCAN

As discussed in the main body of the report, there were some computational challenges applying the DBSCAN algorithm to users with a large number of geolocated tweets since the number of calculations increases exponentially with the number points to be clustered. This annex documents the mains steps taken to maximise run-time performance.

1. A step was implemented prior to running DBSCAN to identify high-frequency users who had sent more than 1000 tweets with the majority of their tweets located within a 2 meter grid. These data were removed and processed separately. This step helped limit the exponential growth in the number of calculations that is inherent when clustering large number of data points.

2. Most modern computers have multi-core processors, mainly to support graphic intensive processing, such as gaming applications. Standard applications process tasks sequentially through a single processing unit. Joblib[27] is a Python library that enables processing tasks to be split and piped through all available cores of a multi-core CPU. This is a simple form of parallel computing. Joblib was incorporated into the clustering code and improved run-time performance by a factor of eight.

3. Further improvements were through very simple code changes. The Python 'set' function was used in place of the 'list' function within the part of the code that determines whether a data point has already been incorporated in the cluster. Sets and lists are functions that can be used interchangeably but their efficiency depends on the programming task. Specifically, a set is faster when identifying whether an object, such as a reference ID, is present in a set, but slower when iterating over the contents.

The final improved version of the clustering code was able to process over 80 million data points in around 45 minutes compared with earlier versions that took days.

[27] https://pythonhosted.org/joblib/

Annex C: Classification of Cluster Centroids using AddressBase

Cluster centroids are classified using AddressBase to distinguish between residential, commercial and other types of addresses. For each cluster centroid a simple nearest neighbour method is used to identify the nearest address point in AddressBase. AddressBase contains over 27 million separate address points, which could potentially involve a vast number of calculations. This Annex describes the data and methods used to reduce the number of calculations.

The AddressBase data used was a slimmed down extract containing:

- UPRN (Unqiue Property Reference Number)
- x_coordinate
- y_coordinate
- classfication_code
- postcode

These data were loaded into the R software package in the form of an indexed list. The list contained about 4000 smaller lists and for each of those smaller lists correspond to a data-frame with addresses within a 100m x 100m square.

One element of the list can be seen below:

Postcode_list[[“314900”]][[“ 132500”]]

Row	Postcode	UPRN	Easting	Northing	Classification code	...
12019997	KY4 8DX	320140660	314995	690346	RD	...
12019998	KY4 8DX	320140659	314999	690348	RD	...
12019999	KY4 8DX	320140628	314932	690322	RD	...
12020001	KY4 8DX	320140627	314937	690325	RD	...
12020003	KY4 8DX	320140663	314976	690338	RD	...
...	...	...	...	...	...	...

The names of the list (“314900”, “132500”) refer to the easting and northing coordinates of the addresses contained in this particular table.

This indexed list hashes AddressBase into an object that can be quickly and efficiently subsetted. When searching for the nearest address, the floor function is used on the x,y coordinates of the centroid location , which is then used to look up the addresses within the

nearest nine 100m x 100m squares. An average square contains only 20-30 addresses, which will involve a much smaller number of calculations.

Some densely populated areas could still have a lot of addresses contained in one square. In order to account for this, the algorithm first looks in the single closest square. Then if the nearest address point found is more than 10 meters away from the actual cluster centroid, then it will look in the adjoining eight squares as well.

Even though the method reduces the number of calculations considerably, it is still computationally intensive. To reduce processing time the search algorithm is wrapped in a single function definition, which searches for the closest address for one particular cluster. This can then be parallelised using the “doParallel” package in R. Eight Xeon processing cores can easily be assigned to the job of postcode matching and then our code is effectively searching for eight closest addresses at a time.

This parallelisation also increases the memory usage of the instance, but one can modify the number of clusters processed within a parallel process and output results more frequently. This has the effect of reducing the amount of required memory.