Chappell P, Tse YK, Zhang M, Moore S.
Using GPS Geo-tagged Social Media Data and Geodemographics to Investigate Social Differences: A Twitter Pilot Study.

Copyright:
This is the authors’ accepted manuscript of an article that was published in its final definitive form by Sage, 2018.

DOI link to article:
https://doi.org/10.1177/1360780417724065

Date deposited:
19/01/2018

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License
Using GPS geo-tagged social media data and geodemographics to investigate social differences: A Twitter pilot study

Paul Chappell, Ying Kei (Mike) Tse, Minhao Zhang, Susan Moore

*Keywords:* Digital Methods, Twitter, Social media, Tweet analysis, Twitter analysis, Socio-spatial analysis, GPS, Coming crisis, Geodemographics, Social Class

**Abstract**

This paper outlines a new method for investigating social position through geo-tagged Twitter data, specifically through the application of the geodemographic classification system Mosaic. The method involves the identification of a given tweeter’s likely location of residence from the ‘geotag’ attached to their tweet. Using this high resolution geographic information, each individual tweet is then attributed a geodemographic classification. This paper shows that the specific application of geodemographics for discerning between different types of tweeters is problematic in some ways, but that the general process of classifying tweeters according to their position in geographical space is viable and represents a powerful new method for discerning the social position of tweeters. Further research is required in this area, as there is great potential in employing the mobile GPS data appended to digital by-product data to explore the intersections between geographical space and social position.

**Introduction**

There has been a recent move in the social sciences towards the use of naturally occurring ‘by-product’ data as a resource for analysing the social world (Beer and Taylor 2013; Rogers 2009; Savage and Burrows 2007; Tse et al. 2017), both in the sense of understanding web-based culture (e.g. Burnap et al. 2015; Dunbar et al. 2015) and the broader social world (e.g. Gleason 2013; Lotan et al. 2011). Of all the different
forms of these types of data, one of the most promising for sociological research are social media / social network data, which represent a rich resource that sociologists can call upon in order to understand more about the social (Beer and Burrows 2007; Savage and Burrows 2009). However, one of the problems with employing social media data to understand users’ perspectives, attitudes and/or opinions is that it is hard to gauge the demographic characteristics of tweeters, so comparisons between different groups of people are hard to make.

While there have been some recent advances in the area of estimating demographics, such as gender and social class, based upon information provided by social media users (Burger et al. 2011; Sloan et al. 2013; Sloan et al. 2015), the practice of deriving demographic information from social media by–product data is still underdeveloped. In this paper, we aim to show how high resolution spatial data, in the form of GPS coordinates attached to some social media posts, can plausibly be used to estimate the social position of tweeters. We show this through the analysis of Twitter data, using a process whereby GPS coordinates are converted to postcodes, each of which are then converted to geodemographic classifications, which provide information about the probable social position of the tweeter. We hope this paper will provide a first step in thinking through how we can better understand the ways in which different types of people tweet. It also has potential broader implications for research, because GPS data is not solely appended to tweets, or indeed to social media data, but is increasingly collected as a matter of course on mobile devices for a variety of purposes.

Therefore, the main aim of this paper is to outline the potential applicability of this particular innovative socio–spatial method for exploring social differences using GPS data. However, as this is an innovative method that employs by–product data in the form of tweets and accompanying geo–data, alongside geodemographic classifications, the research reported here is also interesting because we have employed a variety of different forms of data and analysis that have been developed and are commonly used in the private sector.
These data could therefore be described as forms of ‘commercial sociology’ (Burrows and Gane 2006). As such, the article has a second aim: to provide an illustration of the ways in which the innovative, commercially produced forms of data and analysis identified by Savage and Burrows (2007) as indicative of a ‘coming crisis of empirical sociology’, can potentially be imported into the academic sphere to investigate issues of sociological importance. With this second aim in mind, we outline the ‘nuts and bolts’ of our exploratory methodological process, from collecting the relevant tweets, to the derivation of socio-spatial classifications, and discuss the pros and cons of our method, as well as the pros and cons of employing geodemographic groups as measures of social position.

**Twitter and (Geo) Demographics**

**Twitter as a data source for investigating social position**

The microblogging service Twitter was launched in 2006 and has proved to be a valuable resource for social scientists working in a variety of different fields. Tweets are a good resource for researchers because Twitter has a huge user base of around 310 million active users (Statista 2016b), the majority of whom make their tweets public. Additionally, tweets are easy to collect (although this is becoming less true as Twitter restricts the availability of free tweets through its API – see Felt 2016), and easy to store. Furthermore, users are constantly commenting on topical events and on developing social and cultural phenomena. This has led to researchers in a variety of disciplines realizing the potential utility of tweets for describing and explaining the social world. To give some examples, academic studies from a variety of disciplines have examined political protests such as the Arab Spring protests (Lotan et al. 2011) and the Occupy movement (Gleason 2013), as well as exploring consumer attitudes towards a variety of different issues (Jansen et al. 2009; Tse et al. 2017).

While analysis of Twitter is now widespread, demographic–based analysis of tweets remains relatively rare because the appropriate demographic information about tweeters is not always available.
Fortunately, there is a developing literature on identifying the demographic characteristics of tweeters. Burger et al (2011) and Sloan et al (2013) have shown the potential for identifying gender based upon Twitter user profile data and there have been successful applications of this type of analysis. For example, Tse et al (2017) have used this type of analysis to show gender differences in responses to the Volkswagen emissions scandal. Sloan et al. (2015) have also developed a method for estimating the occupational social class position of tweeters based on the occupation provided in their user profiles, by employing ONS SOC2010 classifications of each occupation. Although Sloan et al.’s study was broadly successful in classifying tweeters according to social class, they identified a number of problems with the method, including particular issues with identifying people from NS–SEC groups 2, 6, and 7, and a relatively high level of misclassification without human validation of their automated coding. For these reasons, it is reasonable to suggest that other methods of estimating social position are worth exploring. In this study we suggest a method based upon deriving information about social position from geographical position, through the use of geo-tagged tweets, specifically through the geodemographic classification system Mosaic.

**Geodemographics**

Geodemographic classification schemes are commercial systems of classification that classify households as a certain type according to postcodes. They are mainly employed by professional marketers to understand consumer preferences in order to better target the products and services they are selling. Working on the principle that ‘birds of a feather flock together’ (Burrows and Gane 2006), geodemographic schemes are constructed through a process in which a variety of data sources (e.g. survey data, commercial transactional data, census data) that describe the characteristics of people living in certain residential areas are combined and then clustered in order to derive a system of classification which describes the likely characteristics of people (in terms of demographics, lifestyle, employment) living within different types of geographical areas or
neighbourhoods (Harris et al. 2005). The end product is a number of ‘segments’ that every household in the UK can be placed within, through reference solely to their postcode. The two main geodemographic schemas within the UK are CACI ACORN and Experian Mosaic. They are similar in terms of the number and type of segments that they assign postcodes to. The ACORN classification system consists of 6 main categories, with 59 sub-categories whereas Mosaic’s classification scheme is comprised of 15 different groups, each of which contain between 3 and 9 different types, for a total of 66 sub-categories. We employ Mosaic here because it has been most discussed in the relevant academic literature, although either would have been acceptable because the exact details of the processes followed are commercially sensitive and therefore somewhat ‘black box’ to academic researchers anyway.

Each Mosaic classification describes a ‘type’ of consumer, who is likely to lead a certain ‘type’ of lifestyle. As Burrows and Gane (2006) note, the characteristics given to each classification can be seen as examples of ‘ideal types’ – in that while the characteristics of the people who live in areas categorized in certain ways will only very rarely, if ever, match the exact characteristics of the description, the list of characteristics will still describe the characteristics that each person / household within each area are relatively likely to have. Table 1 shows some examples of the groups alongside descriptions of their characteristics. These groups are only a small sample of the whole scheme – further information about Mosaic groups is available from Experian (see e.g. Experian 2014). These particular groups have been chosen for inclusion in this table because they act as good exemplars of how geodemographic classifications combine geographical features, such as the urbanity or rurality of a place, with social and lifestyle / cultural factors, and because they are the groups most discussed in this paper.

Table 1. Subset of Mosaic geodemographic groups and associated characteristics

<table>
<thead>
<tr>
<th>Mosaic Group</th>
<th>Example ‘types’</th>
<th>Geographical characteristics</th>
<th>Economic characteristics</th>
<th>Cultural characteristics</th>
<th>Typical age</th>
</tr>
</thead>
<tbody>
<tr>
<td>(sub-groups)</td>
<td>cs</td>
<td>profile</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>----</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prestige Positions</td>
<td>Premium Fortunes, Alpha Families</td>
<td>Tend to live in the suburbs. Live in large 4+ bedroom houses that they tend to own outright</td>
<td>Well educated. Work in senior and managerial positions, or are professionals.</td>
<td>Use the internet relatively frequently, take expensive holidays frequently, are attracted to premium brands</td>
<td>56 – 75</td>
</tr>
<tr>
<td>Country living</td>
<td>Wealthy landowners, Rural Vogue</td>
<td>Likely to live in rural areas. Homeowners.</td>
<td>Relatively likely to be self-employed. Tend to work in agriculture. Homeowners</td>
<td>Tend to own pets. High car ownership. Low engagement with latest technology</td>
<td>Working age</td>
</tr>
<tr>
<td>Senior Security</td>
<td>Bungalow Haven, Classic Grandparents</td>
<td>Have lived in same residence for a long time. Typically 3 bedroom houses and bungalows. Homeowners.</td>
<td>Retired. Have equity in their homes. Worked in variety of occupations before retirement</td>
<td>Replace consumer items only when necessary. Unlikely to invest in digital technology</td>
<td>75+</td>
</tr>
<tr>
<td>Municipal Challenge</td>
<td>High Rise Residents, Streetwise Singles, Low Income Workers</td>
<td>Urban locations. In council flats / rented accommodatio n. Often multi-storey / high rise or terraces. Low home ownership</td>
<td>High unemploymen t. Work tends to be in low level service jobs. Often receive benefits. On some of the lowest incomes.</td>
<td>Watch lots of television. Shop locally rather than online. Relatively high proportion don't own a car</td>
<td>Working age</td>
</tr>
</tbody>
</table>
Geodemographic systems of classification have been the subject of much discussion among sociologists in recent years, partly in relation to the role that such commercial systems of classification can play in the perseverance of social inequality (see Burrows and Gane 2006; Crang and Graham 2007; S. Graham 2005; Uprichard et al. 2009), but also in relation to Savage and Burrows’ (Burrows and Savage 2014; Savage and Burrows 2007, 2009) thesis of a ‘coming crisis of empirical sociology’. The ‘coming crisis’ hypothesis posits that while the traditional tools employed by sociologists (e.g. sample surveys and interviews, analysed using inferential statistics and qualitative data analysis methods) are still the dominant methods of data collection and analysis within empirical sociology, there is a whole other world in the commercial sector. Here, traditional methods are employed, but also combined with other newer, often digital, methods, in order to explore the social world. Geodemographic schemes have been suggested as excellent exemplars of this type of ‘commercial sociology’ (Burrows and Gane 2006) because they can be thought of as analogues for social class.

It is interesting to note that despite the significant body of work discussing geodemographics, there remains a gap in the relevant literature in one obvious regard. This is in the case of research that actually employs geodemographic systems empirically in order to turn them to the task of attempting to understand the social world (for a recent exception, see Burrows et al. 2016). This relative lack of empirical engagement is surprising, given the predominance geodemographics have taken in discussions around the ‘coming crisis’ and specific calls in the literature (Webber 2009) for this commercial tool to be adopted in academic analyses. In this paper, we address this gap in the literature by employing Mosaic to explore how commercially derived forms of data collection and analysis may have potential as tools for exploring the social world in an academic context.

Geodemographics and class

Before outlining the method by which we classified individual tweets, it is first necessary, given the complexity of trying to understand
geodemographic systems of classification in relation to existing sociological concepts such as social class, to examine the potential strengths and drawbacks that there may be with using geodemographics to estimate social position, as is the aim in this study. It is important to consider this issue because there are other potential options that could be employed for estimating social position through geo-tag data (for example social class measures / indices of deprivation), and it is by no means certain that geodemographics are the best method for conducting this type of work.

There are three properties of geodemographic systems that potentially make them desirable for use by sociologists interested in class and social stratification. First, and most obviously, as can be seen in Table 1: the different geodemographic groups represent different ‘types’ of people, and these different ‘types’ of people are constituted at least partly along socioeconomic lines. It is not unreasonable to say that some Mosaic groups, such as the Prestige Position group, represent elite groups, and others, such as the Municipal Challenge group, are working class groupings. It therefore would appear reasonable to see geodemographic groupings as a measure of one’s position in some form of social hierarchy, and hence of potential use to sociologists of stratification and class. Second, geodemographic classification systems offer a method of discriminating between people at a high level of granularity, if required. There are sub-groupings within each group that mean it may be technically possible to distinguish between people more effectively than is the case with, for example, the currently dominant neo-Weberian Goldthorpe classification, and the ONS NS-SEC system based upon it. Third, the geographic component to the classification fits nicely with debates around the spatialization of class. Savage et al (2005, p. 207 in Parker et al. 2007) suggest that:

“One’s residence is a crucial, possibly the crucial identifier of who you are. The sorting processes by which people choose to live in certain places and others leave is at the heart of contemporary battles over social distinction. Rather than seeing wider social identities as arising out of the field of employment it would be more promising to examine their relationship to residential location"
From such a perspective, in which social position is irreconcilably linked to spatial position, and in particular residency, the use of geodemographic classification systems can be seen as a potentially very strong route by which the spatialization of class could be investigated empirically.

However, this does not mean that geodemographic groups can unproblematically be described as examples of *social classes*. This is for two main reasons: firstly, because class is anyway a contested concept (see Crompton 2008) and secondly, even if we do take a broad ranging definition of what can count as social classes, there is still a big problem with treating geodemographic classes as a measure of social class; this is that there are other social divisions playing an important role in the demographic makeup of the geodemographic groupings.

To expand upon this first point, different conceptualizations of class operate in parallel. Marxist and Weberian definitions of class are narrow – referring to the economic dimensions of inequality through a focus on occupation – whereas geodemographic classes classify postcodes according to many other factors, including cultural tastes and practices. In this way, geodemographics can perhaps be thought of as more consistent with a Bourdieusian position (e.g. Bourdieu 2001), in which one’s social class is seen as being contingent upon the volume and composition of reserves of economic, social, and cultural capital. Members of each geodemographic group or type tend to have similar levels of economic income and wealth, and also tend to consume culture in similar ways. It could also be argued that they are likely to have similar levels of social capital because of the ‘birds of a feather flock together’ assumption inherent in the construction of geodemographic schemas – similar people tend to live in similar places, hence people in a certain geodemographic class are likely to know other people within grouping. It is therefore not unreasonable to suggest that both Bourdieusian and geodemographic classification systems are essentially based around the notion that you can ‘cluster’
similar types of people together, based upon their positions in multiple hierarchical dimensions.

But even though a Bourdieusian position allows more room for manoeuvre than a more traditional economically focused definition, this does not take into account the second problem with treating geodemographic classifications as social class groupings, and that is that the characteristics that make up geodemographic groups refer to more than just a position on a social hierarchy. Geodemographic groups also provide information about other social divisions, the most important of which would appear to be related to age and life course stage: people within each of the classes are likely to be of a certain age. For example, the Mosaic Senior Security group is made up of 90% of people over the age of 65 (Experian 2014). With such a high proportion of people being of a certain age, it is clear that position on a socioeconomic hierarchy is not necessarily playing a key role in structuring the makeup of the groupings, and if it is, it is doing so in an interaction with age.

This means that any empirical application of Mosaic groupings in which such groups are treated as measures of social class will at best use geodemographics as a fairly loose proxy for social class, and will at worst be theoretically incoherent and methodologically unsound. Having said this, there are still three strong arguments for exploring the potential of geodemographics within theoretically-minded empirical sociology. Firstly, as we have outlined above, there is a need for the academy to engage with ‘commercial sociology’ in more ways than just providing critical commentary and the application of resources such as Mosaic in empirical research will help to close this gap. Secondly, despite its drawbacks, many sociologists support the idea that geodemographic classifications may actually have some desirable properties for exploring the social world. If class is increasingly becoming spatialized, as has been argued by the likes of Savage et al (2005), then it may be the case that a combined measure of space and social position could provide novel insights, and analyses using systems such as Mosaic provide a potentially illuminating way of approaching certain empirical problems. Thirdly, and related to this
second point, geodemographic classifications have been shown to have impressive predictive capacity (above and beyond more traditional measures of social stratification) for understanding various patterns of cultural ‘behaviour’ (Webber 2004), and also for predicting students’ exam performances (Webber and Butler 2007). We would not suggest that predictive capacity is the be all and end all when it comes to gauging the usefulness of any analytical tool or variable, however these findings still bode well for the potential merit of using Mosaic in empirical research, as in this study.

Method

Data collection

The data employed here were originally collected for a study into the horse meat contamination scandal of 2013. In order to achieve an appropriate corpus of tweets for analysis, we followed a multistage sequential data collection process. There were three main considerations that guided the data collection process: firstly, we aimed to identify as many tweets addressing ‘Horsegate’ as possible, whilst minimising false positives; secondly, we selected only the tweets with geographic information attached; finally, we aimed to ensure that a maximum number of tweets possible were tweeted from the home of the individual. This third point was necessary because geodemographic classification systems are based upon the type of people who are likely to live in a particular residential area. The data collection process was conducted sequentially, addressing each of these three considerations in turn.

The first stage of the data collection process gathered a years’ worth of tweets referencing the horsemeat scandal and produced a broad sample. All the English language tweets from the UK from the period 15/01/2013 – 15/01/2014, that had some form of attached geotag data were purchased from Twitter archive company GNIP for US$4100. The relevant tweets were identified using 101 keyword rules. There is no scope to outline the whole list of rules here but some examples include: any tweets that included both the words ‘horse’ and ‘meat’
within a single tweet, ‘horse’ and ‘eat’ within a single tweet, ‘horse’ and ‘dinner’ within a single tweet, and so on. A full list is available from the first author on request. This first part of the data collection process led to a total of 26,737 tweets being collected.

Twitter geotag data comes in three different formats (see Moffitt 2014 for more information) so some further tweets had to be removed in the second stage of the data collection process. The format of geographic data required to conduct this particular analysis was the Activity Location data, which specifies the precise longitude and latitude coordinates of the location the tweet was sent from. This is collected in real time based upon GPS data from the mobile phone, or other GPS enabled device, sending the tweet. Only a subsection of the tweets in the sample had this particular form of geographic data attached to them, so the data was filtered and any tweets without this information were discarded from the sample, leaving a total of 22,035 tweets with this particular form of geotag data.

The third step was to select a sub-sample of the tweets that would be most likely to have been tweeted from each individual’s place of residence. This was necessary because geodemographic classifications are derived from residential postcodes. This means that many of the tweets sampled would have been sent from locations that differed from the home address of the tweeter, and if geodemographic classifications were derived based upon all of these tweets, instead of residences many of the locations would be places such as their school or work or tweets sent while travelling to and from any given destination. This would mean that the geodemographic classifications would be incorrectly applied to these locations rather than tweeters’ home addresses. While this issue cannot be completely mitigated (as we have no way of knowing whether or not someone is at their home location at any one time), we attempted to restrict the extent of this problem by selecting a further sub-sample of tweets based upon temporal factors. That is, we selected only the tweets sent between 6.30pm and 8am on weekdays, when the largest proportion of people tweeting could be presumed to be home. The spreadsheet containing each tweet was filtered by the time of day and then by day of the week.
before tweets not meeting the above time-based criteria were deleted from the sample. This process further restricted the sample to 4,990 tweets.

Each pair of long-lat coordinates were then converted to postcodes using a batch reverse geocoding tool (available at <http://www.doogal.co.uk/BatchReverseGeocoding.php>) and then corresponding Mosaic classifications were purchased for each of the postcodes from Experian. Some of the coordinates could not be matched to a postcode and some of the postcodes did not have corresponding Mosaic classifications so these cases were removed from the sample. The final sub-sample that was taken on for further analysis consisted of 4,200 cases. For each of these cases we had the body of a tweet from a unique user and a corresponding Mosaic classification. The geographic distribution of these tweets can be seen in Figure 1 (note that they are unsurprisingly clustered in urban areas) and the distribution of these tweeters across some of the different Mosaic geodemographic groups can be seen in Table 2.
Figure 1. Geographic distribution of tweets about horsemeat across the UK. N = 4200.

Table 2. Frequencies and proportions of selected Mosaic group membership across our final sample and across the UK population.

<table>
<thead>
<tr>
<th>Mosaic Group</th>
<th>N (Percentage of final sample)</th>
<th>Percentage of households across UK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prestige Positions</td>
<td>286 (6.8)</td>
<td>7.4</td>
</tr>
<tr>
<td>Country living</td>
<td>205 (4.9)</td>
<td>6.1</td>
</tr>
<tr>
<td>Senior Security</td>
<td>274 (6.5)</td>
<td>8.4</td>
</tr>
<tr>
<td>Municipal</td>
<td>194 (4.6)</td>
<td>6.4</td>
</tr>
</tbody>
</table>
Potential validity issues

As the process outlined above is somewhat novel, it is important to acknowledge potential weaknesses. First, we have no way of knowing for certain whether or not the tweets were actually sent from the homes of the tweeters in question. While we have tried to maximise the number of tweets sent from home residencies by concentrating on the times of the day and periods of the week when people are most likely to be at home (weekday evenings), there will undoubtedly be errors.

In order to test the extent to which this method produced valid results in this regard, we used a method consistent with the general aim of making use of available ‘commercial sociology’ resources: we selected 100 tweets at random from our sample, and then viewed the precise coordinates the tweet was sent from on Google Maps. We used Google Streetview to zoom in and view the individual houses / neighbourhoods that the tweets were sent from in order to gauge the likelihood that the relevant tweets had indeed been sent from someone’s home address. Out of the 100 tweets investigated, 90 appeared very likely to be tweeted from a residential address. Of the remaining locations, 1 was from a school, 1 from a commercial area, 1 from a hotel, 1 from a sports club, and 6 of them were considered inconclusive (for example, in an area that had bars and houses in close proximity or a building with a bar on the bottom floor but a residence above).

This test would therefore appear to indicate that the vast majority of tweets we were identifying were being sent from residential areas. While we have no way of confirming that the tweeters lived in the neighbourhoods in question or if they are instead the neighbourhoods of tweeters’ friends and/or families, it seems likely that a large proportion are the homes of the tweeters. Even if they are not, the people living in these residencies will likely have a similar social
standing to them tweeters (people tend to have friends and partners from social positions similar to themselves, as has been shown by researchers employing occupational survey data – see Prandy and Lambert 2003; Stewart et al. 1973). Therefore for the purposes of this study, this is not a huge problem. It is worth noting that most of the coordinates appeared to be extremely accurate – normally showing tweets as hailing from individual houses, even in areas surrounded by open spaces. There were a few anomalies that bring the accuracy of some of the GPS data into question; for example one person appeared to be tweeting from a field adjacent to a housing estate. It is plausible they were in the field but given the time of day (9.32pm) and time of year (February) it is likely an error. Nevertheless, the vast majority of tweets appeared to hail from residential properties, so this process certainly supports the general idea that the tweets in this sample are being sent from residential areas, and not, for example, from places of work.

The second potential weakness with this methodology is the likely problem with ecological validity. Members of each geodemographic classification are said to possess a list of (demographic, lifestyle) characteristics so one could make the assumption that the people within each classification within our sample therefore also possess these characteristics. Such a position could be seen as somewhat defensible if the people in our sample were representative of geodemographic groups, but even this would actually be a problematic position to take because there are severe problems with the representativeness of the sample (indeed we make no claims to the representativeness in this study). This is because Twitter use is not evenly spread among the population. In fact, its use is biased towards young people (Statista 2016a), and towards middle class people (Sloan et al. 2015). This means that a geodemographic group that is disproportionately composed of a certain type of person at the population level is unlikely to be represented by the same type of people, in the same proportions, within our sample. To illustrate using the example of age, over 90% of the Senior Security group are aged over 65 (Experian 2014), but a lower proportion of this group are over this age in our sample, because the younger people within the
Senior Security group are far more likely to be the people using Twitter. This problem is further complicated by the fact that Twitter users with geotagged tweets make up only 0.7% of the overall population (M. Graham et al. 2014) and because the tweets that are geotagged are not representative of the whole Twitter population (Sloan and Morgan 2015).

To explore the extent to which we need to be cautious of attributing aggregate level characteristics to individuals, we examined the profile information of tweeters in one potentially problematic Mosaic group (Senior Security), and compared the descriptions that they gave about themselves to the characteristics of this group provided by Mosaic. There are 276 tweets categorized as coming from Senior Security tweeters in the final sample. Of these tweeters, 51 suggested they had jobs or were currently students, compared to just 1 who suggested they were retired. 18 of the 276 tweeters listed their ages on their online profiles, and these ages ranged from 13 to 50, with no-one suggesting they were older than this. These findings suggest that, in the context of age at least, the characteristics of the people within the sample do not match up well with the characteristics of the population. This means that the Mosaic classifications are not going to be very useful for gauging the age of the tweeters (i.e. it would be poor practice to compare the content of Senior Security tweets to a young geodemographic group and suggest the findings showed age differences).

This analysis would appear to provide evidence that, at least where age is concerned, the use of geodemographic classification systems as proxy measures for demographic characteristics is problematic because the process involved is not accurately classifying the demographic characteristics of the individuals being researched. But could geodemographics be more useful for comparing groups based upon position on a social hierarchy? It seems reasonable to assume that, while you will always get some younger people living in and/or visiting neighbourhoods of predominantly older persons, it is less likely that you will find people from one end of a social hierarchy in a neighbourhood generally occupied by people at the opposite end of
the hierarchy. To take economic capital as an example, it seems unlikely that you will encounter many people with large amounts of inherited wealth or highly paid jobs living on council estates, or unemployed people living in expensive neighbourhoods, so there may be a case for still using geodemographics as a method for distinguishing between people at the top and at the bottom of a social hierarchy, as long as we are cautious in how results are interpreted.

In order to test this idea, we employed Google Maps and Streetview to inspect the neighbourhoods of a sub-sample of tweeters from two very different geodemographic groups (50 Prestige Position and 50 Municipal Challenge individuals each). Experian documentation (2014) suggests that Prestige Position individuals are likely to have reached senior and managerial positions in companies, have accomplished professional careers, and/or be directors of their own companies. On the other hand, the Municipal Challenge group are likely to be some of the lowest earners, and are the group with the highest proportion of people currently seeking employment. So, the comparison between the two should be one of a relatively privileged vs an underprivileged group.

Visual inspection of the neighbourhoods in which the tweets were sent (again, via Google Streetview) does indeed confirm that this is the case. Prestige Position tweeters were tweeting from clearly affluent suburban neighbourhoods in England, Scotland, Wales, and Northern Ireland. The majority (66%) were in neighbourhoods with mostly large detached houses. 14% were in areas where most of the houses could not be seen from the street due to being largely or completely concealed by hedges, trees, fences, gates, and walls. 16% had large semi-detached houses, and there was one tweet from a Holiday Home and one from a hotel. The sub-sample of 50 Municipal Challenge residences, on the other hand, were living in urban and suburban areas in council estate type terraces (54%), blocks of flats with three or four floors (32%), tower blocks (4%), as well as some small semi-detached houses (10%), again in all four countries of the UK.
The quick visual analysis described above illustrates the difference between *Prestige Position* and *Municipal Challenge* neighbourhoods, and fits in with the descriptions of these types of residencies provided by Experian (2014). This process of visually inspecting the precise location that the tweets were sent from also served to demonstrate qualitative differences between the two types of neighbourhoods. The *Prestige Position* houses were generally much larger and had bigger gardens with different types of cars outside and on the surrounding roads (notably newer cars and more expensive brands in *Prestige Position* areas); there was much less space between residencies in *Municipal Challenge* neighbourhoods, and a far greater variance in the design of houses in the *Prestige Position* homes. There were also more people on view in the *Municipal Challenge* neighbourhoods. This process of visual validation may be unconventional but it was actually very effective in showing that we are clearly looking at two very different types of residential areas, and that the real world characteristics of these residential areas appear to be in line with what would be expected according to the socioeconomic characteristics of the two relevant geodemographic types. This indicates that the method outlined here is somewhat effective at assigning appropriate geodemographic classifications to the areas that tweets were sent from, although the point remains that it is not appropriate practice to infer aggregate–level characteristics to the individuals classified as members of each geodemographic group.

**Further considerations**

It is worth reflecting at this point on the forms of analysis that make sense with such data, and the types of claims that can be made based upon these analyses. The sampling process here has been successful in selecting two very different types of tweeters, and we can clearly see that these tweeters are from very different social backgrounds, thus making any differences revealed in their tweets very interesting and important. However the black–box nature of the derivation of Mosaic classifications, combined with the research process employed here, mean that no direct comparison (in quantitative terms) can be made.
between the findings of this study and any other quantitative findings based upon survey data. People classified as Municipal Challenge / Prestige Position should not even be seen as representative of broader groups within a broader population (whether this population be all people, all Twitter users, or all geotag-enabling Twitter users), because they were not selected randomly from these populations.

The fact that we are making no claims to representativeness has knock-on effects for the forms of analysis that can be employed. It is not appropriate to use traditional inference-based statistical methods to draw conclusions about differences between class groups. Even if a form of quantitative analysis was employed, and differences between content in tweets quantified, then there would be no way that we could show that the perspectives of certain tweeters were representative of people within any broader population, and therefore no reason to employ inferential tests to indicate whether these differences were ‘statistically significant’, because such tests are based upon the assumption that a sample is representative of a broader population.

And indeed from another perspective it seems rather strange to even think of these data as a sample (representativeness or otherwise) from which one can generalize, because the data are more realistically thought of as a population than as a sample – these are all the tweets sent out at the relevant times, by people who allow their tweets to be made public with accompanying geotags, in the two areas classified as Municipal Challenge and Prestige Position. As has been pointed out by Uprichard et al (2008), new forms of digital data are often not well suited to traditional forms of statistical analysis, and this is the case here. This means that if quantitative comparisons are made then they cannot be validly accompanied by measures of statistical significance, despite the established conventions we normally work to as academic sociologists. For example, in the ongoing analysis of Horsegate using the data described here, we are employing a mixed methods data analysis strategy, using thematic analysis alongside deductive human-coded content analysis to analyse subgroups of the sample. Specifically, we are comparing Municipal Challenge tweets to Prestige position tweets. The resulting quantitative analysis of our content
analysis data is descriptive and exploratory, in line with the principles of Exploratory Data Analysis (Tukey 1977). As such, this is not a process that allows for the testing of hypotheses but it should serve to paint a picture of the differential responses to Horsegate from different geodemographic groups. Other quantitative methods could also be employed with these types of data (for example word frequency analysis or Sentiment analysis) but we are not using them in this study.

In summary, three key points can be made to summarize the results of these preliminary investigations of validity. First, it appears that the process of selecting only tweets sent on weekday evenings, on the assumption that these tweets will be sent from where people live, is a very promising method for selecting out tweets that can be used for estimating demographics based on a point in geographical space, as the vast majority of these tweets appear to come from individual properties in residential areas. Second, there are some significant weaknesses in the method in terms of the utility of the *Mosaic* classification system for estimating the demographic characteristics of tweeters. The unrepresentative nature of Twitter use, combined with the fact that *Mosaic* classifications refer to multiple demographic characteristics concurrently, means the characteristics of tweeters will not necessarily match up with their attributed Mosaic classification. This problem is particularly acute for *Mosaic* classifications (such as the *Senior Security* group) that represent clusters of people of a certain age because Twitter use is particularly unrepresentative in this regard. Third, and despite the problem identified above, there is still some limited potential for using geodemographics to explore social differences, particularly if one concentrates on clearly discerning between two very different geodemographic groups.

**Discussion**

**Key findings and areas for further research**

The first aim of this paper was to explore the potential applicability of a socio-spatial method, based upon GPS geo-tag data and
geodemographics, for estimating social position in social media posts. We would suggest that we have been largely successful in this aim. We have shown that, while it is certainly not methodologically defensible to take just any geo-tagged tweet, assign a geodemographic classification and then claim that the tweeter has the demographic and cultural characteristics associated with that classification, the use of geodemographics to identify two different types of people from opposite ends of the social spectrum in order to compare their perspectives is a defensible practice, despite a number of drawbacks.

Going forward, if researchers conducting analysis into tweets or other geo-tagged data have a sufficient sample size, it should be fairly straightforward to identify a sub-sample of data hailing from specific geographic areas that correspond to specific geodemographic classifications (we would recommend Municipal Challenge and Prestige Position groups from the Mosaic schema), and then make comparisons between these groups. Having said this, we would stop short of claiming that geodemographics are the best available resource for investigating social position in geotagged tweets. Just because we have shown that Mosaic is capable of discerning between different social groups based upon geography does not mean that such a method is going to be the most effective way of classifying people into social groups based upon the spatial position they have tweeted from.

It is equally plausible that the same GPS data that we have employed here to assign a postcode and geodemographic classification to tweeters could be used in different ways. Rather than assigning a geodemographic classification to each postcode and then comparing geodemographic groups, more conventional measures of social hierarchy could be employed – such as indices of deprivation, rates of local unemployment, NS–SEC social classes derived from Census data, and so on. While there are some advantages to the use of geodemographic classifications (such as high granularity and the integration of the spatial with the social), it is our opinion that deriving geodemographic classifications may be less useful for understanding the likely demographic characteristics of a tweeter from a particular area than deriving these demographic characteristics directly. It is
worth noting that whatever form of classification is employed, problems with ecological validity will still apply because each tweet will still be classified according to the aggregate characteristics of the area they tweeted from. Further research is required into classification using other methods, in order to identify the most useful metrics / classificatory schema for this type of research.

There are two aspects of our methodology that we would like to stress as important successes, and that we believe represent useful analytical techniques that have great potential for future research. The first of these innovations is the broad idea itself: to our knowledge there is no existing research that uses the geo-tag data from social media data to try to apply social classifications to tweeters. Given that there are ongoing difficulties in trying to assign demographics to social media users (see Sloan 2015), this particular method of employing GPS data to estimate social position has great potential. Although only a small proportion of Twitter users have the real-time geolocate option turned on, if a topic is being discussed by a large enough group of people then achieving a large sample of tweets with attached geo-data should not be a problem. Additionally, the ubiquity of mobile phones and the ever-increasing reliance on social media and other applications on these phones also means that geo-tag data will likely increasingly be collected as a matter of course, by a multitude of different commercial applications. A method similar to the one outlined here, where geo-tags are used to position people within certain social groups, or on some form of social hierarchy, will be a powerful tool that the commercial sector will probably make use of – if they are not already doing so – and we should be engaging with these new forms of data from this point onwards.

The other successful aspect of our pilot research is the time-based selection method that we have employed in order to produce a valid sample of tweets. The fact that people move around in their day-to-day lives had the potential to hamstring this method at the first hurdle, because it calls into question the extent to which geo-tag data can be used to identify an individual’s place of residence. But our analyses indicate that this is not a significant problem, as long as the
time at which tweets are sent is taken into consideration, it is feasible to employ geotag data to estimate probable places of residence. It is possible that there may be more precise ways of identifying tweets sent from residential locations, but this time-based selection method represents a very strong start, and we are confident that its application leads to the exclusion of a large proportion of the tweets sent from non-residential locations, and hence is a useful method going forward.

‘Commercial Sociology’

The second aim of this paper was to provide an illustration of the ways in which commercial tools of data and analysis could be turned to academic research. Again, the results indicate a mixed bag as far as this aim is concerned. In terms of positives, we have used commercial by-product data acquired from Twitter, and commercial data analysis methods borrowed from Experian, complemented by analytical tools provided by Google, to some success. On the other hand, there is a messiness to the data analysis that many researchers in the academy (in particular, traditional quantitative researchers) may find unsettling. We are entirely reliant on data that was not produced by us, and that we may not be able to access, or use, in its entirety, and the analytical methods we are employing are not fully understood.

Beer and Taylor (2013) point out that working with messy data is an unavoidable consequence for those who wish to work with by-product data, and that we, as social scientists, will have to adapt our methods to deal with these issues. We concur with this, and would suggest a three pronged strategy to researchers conducting this type of research. First, be willing to adopt a flexible research process. This is necessary because protocols of proper practice are not yet established, and in such situations it is not always plausible to know what will work before the actual empirics of a research project are underway. For example, in this study, we had to develop a time-based system for selecting tweets, in order to collect a sample more likely to represent places of residence, rather than transitory locations, or places of education and work. Second, attempt to validate findings as research is conducted. This will have the dual benefit of avoiding artefactual
conclusions, and help researchers aiming to do similar types of research in future, by providing some idea of the types of practice that may be reasonable, and those that are not. In the current study, we conducted ad hoc validity analyses by visually inspecting the geographical locations in our sample, in order to check if the classifications provided by Experian tallied with the residencies in our sample. Third, be cautious in the claims that are made. If the data and methods employed are incomplete, messy, and/or exploratory, then it makes sense that findings should not be overstated. In our case, we identified that it was important that the tools and language of inferential statistical analysis were not ported across indiscriminately to this domain of research, as the sampling process did not allow the assumptions of their use to be met.

**Conclusion**

This paper represents a tentative first step in the development of a methodology for using social media geo-tag data to identify people from different social groups, based upon their location in geographical space. We have shown that deriving measures of social position from geo-location data is a potentially powerful technique, and we have developed some key aspects of a methodological process that we believe can form the basis for future socio-spatial analysis of social media data. In particular, our innovation of using a time-based selection method to identify a sample of people more likely to be in their place of residence when they tweeted was particularly successful. We have also shown that while geodemographic schemes have some potential for classifying tweeters according to their location, they have significant drawbacks with unrepresentativeness and ecological validity, and are likely to be inferior to other methods. We would encourage researchers to engage with data and methods such as those outlined here because while this type of research is messy, remains under-researched, and requires flexibility on behalf of the researcher, it still represents a powerful tool that can help academic social scientists explore the intersections between inequality and space in meaningful ways. The relatively recent development of portable GPS technology within mobile devices, and the by-product data that results
from these devices’ use represents an opportunity for social researchers, and provides a good example of the way in which the ‘coming crisis’ in social research should be seen as an opportunity.

Acknowledgements

This research was supported by BA/Leverhulme Small Research Grants SG141713. Thanks to Mark Taylor for his insightful comments on an earlier draft of this paper.

References


Crang, M and Graham, S (2007), 'Sentient cities ambient intelligence and the politics of urban space', Information, Communication & Society, 10 (6), 789-817.

Crompton, Rosemary (2008), Class and stratification (Polity).


Harris, R, Sleight, P, and Webber, R (2005), Geodemographics, GIS and neighbourhood targeting (7: John Wiley and Sons).
Jansen, B, et al. (2009), 'Twitter power: Tweets as electronic word of mouth', *Journal of the American society for information science and technology*, 60 (11), 2169-88.


Parker, S, Uprichard, E, and Burrows, R (2007), 'Class places and place classes geodemographics and the spatialization of class', *Information, Communication & Society*, 10 (6), 902-21.

Prandy, K and Lambert, P (2003), 'Marriage, social distance and the social space: an alternative derivation and validation of the Cambridge Scale', *Sociology*, 37 (3), 397-411.


--- (2009), 'Some further reflections on the coming crisis of empirical sociology', *Sociology*, 43 (4), 762-72.


Sloan, L and Morgan, J (2015), 'Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter', *PloS one*, 10 (11), e0142209.

Sloan, L, et al. (2015), 'Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data', *PloS one*, 10 (3), e0115545.

Sloan, L, et al. (2013), 'Knowing the tweeters: Deriving sociologically relevant demographics from Twitter', *Sociological research online*, 18 (3), 7.


Tukey, John W (1977), 'Exploratory data analysis'.


Webber, R (2004), 'The relative power of geodemographics vis a vis person and household level demographic variables as discriminators of consumer behaviour'.


Webber, R and Butler, T (2007), 'Classifying pupils by where they live: how well does this predict variations in their GCSE results?', *Urban Studies*, 44 (7), 1229-53.