McNeill et al. *EPJ Data Science* (2017) 6:24
DOI 10.1140/epjds/s13688-017-0120-x

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Estimating local commuting patterns from geolocated Twitter data

Graham McNeill, Jonathan Bright* and Scott A Hale

*Correspondence:
jonathan.bright@oii.ox.ac.uk
Oxford Internet Institute, University of Oxford, Oxford, United Kingdom

Abstract

The emergence of large stores of transactional data generated by increasing use of digital devices presents a huge opportunity for policymakers to improve their knowledge of the local environment and thus make more informed and better decisions. A research frontier is hence emerging which involves exploring the type of measures that can be drawn from data stores such as mobile phone logs, Internet searches and contributions to social media platforms and the extent to which these measures are accurate reflections of the wider population. This paper contributes to this research frontier, by exploring the extent to which local commuting patterns can be estimated from data drawn from Twitter. It makes three contributions in particular. First, it shows that heuristics applied to geolocated Twitter data offer a good proxy for local commuting patterns; one which outperforms the current best method for estimating these patterns (the radiation model). This finding is of particular significance because we make use of relatively coarse geolocation data (at the city level) and use simple heuristics based on frequency counts. Second, it investigates sources of error in the proxy measure, showing that the model performs better on short trips with higher volumes of commuters; it also looks at demographic biases but finds that, surprisingly, measurements are not significantly affected by the fact that the demographic makeup of Twitter users differs significantly from the population as a whole. Finally, it looks at potential ways of going beyond simple frequency heuristics by incorporating temporal information into models.

Keywords: transport; mobility; social media; geolocation

1 Introduction

Population movement is a key issue for contemporary policymakers, who need to optimise transport infrastructures and services which are under ever increasing pressure. However these policymakers often operate in an information scarce environment: existing measurement instruments such as transport surveys (which may involve stopping people as they cross the border or drive through a specific road) are time consuming and expensive as well as being a source of frustration for the local population. Hence data collection is infrequent and decisions are often based on incomplete, out-of-date estimates.

The emergence of social media as a potential window on population movement offers a huge opportunity in this regard [1–3]. Data from social media platforms is often available to policymakers at relatively low cost, sometimes even for free. It can also be sourced without creating disruption, as the data is generated as a byproduct of interaction with

the social media platform. Furthermore, it is available in huge quantities on an ongoing basis, meaning that real time changes can be tracked. Hence the potential of social media as a secondary data source is considerable.

The availability of social media data has made it a rich area for academic research on the extent to which it can offer census like indicators in a whole variety of areas such as unemployment [4], consumer behaviour [5], stock market movements [6], health outcomes [7, 8] and elections [9, 10]. Some initial work has started to emerge in the area of population movement in particular. For example Liu *et al.* [11] have investigated the correlation between population and Twitter data in Australia, finding that at large scales population levels could be estimated from the prevelance of geolocated Twitter content. Hawelka *et al.* [12] meanwhile, looked at international population mobility, tracing the number of international tourists, again based on geolocated Twitter data; whilst Beiró *et al.* took a similar approach to domestic commuting patterns [13]. Finally, a variety of studies have successfully modelled intra-city movement patterns using Twitter and Foursquare check-in data [14–16]. However, while these initial results are promising, much more remains to be done in terms of understanding the extent to which social media can be used systematically as an accurate indicator of population movement.

The aim of this article is to build on this existing literature by examining the extent to which local commuting patterns in the United Kingdom can be inferred from data sourced from Twitter. We make three contributions in particular. First, we show that heuristics applied to geolocated Twitter data offer a good proxy for local commuting patterns; one which outperforms the major existing method for estimating these patterns (the radiation model). This finding is of particular significance because we make use of relatively coarse geolocation data (at the city level) and use simple heuristics based on frequency counts. Second, we investigate sources of error in our proxy measure, showing that the model performs better on short trips with higher volumes of commuters; we also look at demographic biases, and find that, surprisingly, our measurements are not significantly affected by the fact that the demographic makeup of Twitter users differs significantly from the population as a whole. Finally, we look at potential ways of going beyond our simple heuristics by incorporating temporal information into our models.

The rest of the article is structured in the following way. Section 2 sets out the methods and data, explaining the means of estimating commuting flow from Twitter and how these estimates are validated against a ground truth census dataset. It also outlines the radiation model of commuting which we make use of as a benchmark. Section 3 presents the results of the model, comparing our Twitter estimates to the benchmark and also exploring sources of bias and error in the estimation. Section 4, finally, explores temporal based extensions to the model.

2 Methods and data

In this section we will describe our approach to modelling commuter flow with Twitter data. We will also describe the census data which we use as a ground truth for evaluating the accuracy of the Twitter model, and the radiation model which we use as a benchmark.

Twitter data is based on messages, known as 'tweets', that people who are members of the social networking site send when making use of the service. Only a small percentage of these tweets are 'geotagged', by which we mean that they come with metainformation containing the location from which the tweet was sent [17] (geotagging often occurs when

people send tweets from their mobile phone [18]). Although historic studies of geotagged tweets have made use of exact co-ordinate data, currently the majority of geotagged data produced by the social media network is relatively coarse, accurate only to the city or municipality level.[a]

Geolocated tweets indicate, of course, where a user currently is, rather than anything about any journey they may make. However, it seems reasonable to assume that, whilst making use of the social network, many people may tweet from both their home and work locations. Hence a pattern of geotagged tweets, over a period of time, ought to contain information about patterns of commuting. Of course, there will be a certain amount of noise in the data as people will tweet from other locations: on their way to/from work, from restaurants, on holiday, etc. Furthermore, not all Twitter users will have a job, nor will all jobs require regular commuting. One of the central questions in this article is to observe the extent to which commuting patterns can be inferred in spite of this noise.

Going from geolocated tweets to commuting patterns requires us to choose a heuristic for deciding which location is a 'home' location and which location is a 'work' location for each user, based on a pattern of geolocated tweets which may come from a variety of areas. There is a growing literature on the best way of inferring these locations from a pattern of digital trace data such as tweets or mobile phone calls [19–22]. We make use of arguably the simplest of these: a frequency count (although in Section 4 we experiment with ways of improving on this simple method by making use of temporal information, it is helpful to first observe the amount of signal that can be extracted from the simplest possible heuristics applied to the data). Hence, the area that a user most frequently tweets from is assumed to be their home location; the second most frequent is assumed to be their work location (users which have a tie for home and work location are discarded). All other locations are assumed to be areas visited which are unrelated to either living or working. To account for users who live and work in the same area, we use a threshold λ: if more than $\lambda\%$ of tweets are sent from the same area, we assume the user both lives and works in that area. In our results section, we experiment with different values of λ to test the sensitivity of the model to this threshold.

Having assigned users to home and work locations, we can construct a commuter flow matrix $\mathbf{T}$:

$$\mathbf{T}_{ij} = tw_{ij}, \tag{1}$$

where tw_{ij} is the number of users which have their home in location i and work in location j.

We take as our ground truth dataset commuting data from the 2011 UK census.[b] The census gives data on commuting volumes within and between 'local authorities', which are administrative regions within the UK (of which there are 378 in total[c]). By comparing estimates from our Twitter model to actual numbers derived from the census, we can assess the accuracy of commuting predictions derived from Twitter. The census also provides information on commuting volumes across different demographic groups, which allow us to assess demographic variation in the accuracy of our results.

We also want to observe how the accuracy of estimates generated by the Twitter model compares to the accuracy of existing methods of commuting flow estimation. There are, currently, two main estimation methods which are used within the literature: the widely

used gravity model (*e.g.* [14, 23, 24]), and the more recent radiation model [25]. In this paper, we opt to use the radiation model as a benchmark, as it has been shown to outperform the gravity model. It is also well-suited to our purposes since it is parameter-free and only requires basic information about each area. Hence it offers a reasonable comparison to our context, where the aim is to infer commuting patterns with only a minimal amount of observational data.

The standard radiation model estimates the commuter flow matrix $\mathbf{T}$ using:

$$\begin{aligned}\mathbf{T}_{ij} &= c_i \,\text{Prob}(\text{work} = j|\text{home} = i)\\ &= c_i \frac{n_i n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})},\end{aligned}$$

where:

- $\mathbf{T}_{ij}$ is the ijth entry of $\mathbf{T}$, the number of commuters who live in area i and work in area j;
- c_i is the total number of outward commuters who live in area i;
- n_i is the population in area i and n_j is the population in area j;
- s_{ij} is the population within a circle centered at area i and with a radius equal to the distance between areas i and j (the populations of areas i and j are not included).

The model assumes that the number of outward or external commuters from an area is proportional to its population. Hence, $c_i = C(n_i/N)$ where C is the total number of outward commuters in the population and N is the total population (given by $N = \sum_i n_i$). If C is unknown, the model can estimate Prob(work = j|home = i), but not absolute commuter numbers. It is worth noting that the radiation model does not offer a prediction for internal commuting, *i.e.* the number of people who live and work in the same area, a point to which we will return below.

Yang *et al.* [26] have introduced a 1-parameter variant of the radiation model, which has been shown to outperform the parameter-free version. We hence also include this model in our analysis. The estimated flow for the 1-parameter model is given by:

$$\mathbf{T}_{ij} = c_i \frac{[(a_{ij} + n_j)^\alpha - a_{ij}^\alpha](n_i^\alpha + 1)}{(a_{ij}^\alpha + 1)[(a_{ij} + n_j)^\alpha + 1]}, \tag{2}$$

where $a_{ij} = n_i + s_{ij}$. Yang *et al.* construct the parameter α so that it varies with the average size of regions for which commuting is being estimated. In particular, they calculate α using:

$$\alpha = \left(\frac{l}{36\ [\text{km}]}\right)^{1.33}, \tag{3}$$

where l is the mean 'length' of regions under consideration, defined as the square root of their area (for our particular case of UK local authorities, l is equal to 19 km).

We will now give details of the Twitter data collected for the study. Our data covers a one year period from June 1, 2015 to May 31, 2016. This time period is not ideal, of course, as we are comparing patterns of Twitter data to the census, which took place in 2011. Nevertheless, it is the best data available for addressing the question. Using the filter stream of the Twitter API with public, 'spritzer' level access, we collected all geotagged

tweets from within a bounding box around the British Isles.[d] It is worth remarking that the choice of a full year period is significant. We expect (and indeed we found) that user engagement with the platform is bursty [27], meaning that a large time window is required to build up a consistent pattern of tweets for one user. However, using a year long period means that we are capturing certain types of bias in our data: for example, we may pick up occasional long distance movements, such as students moving between their homes and places of study (as found by [22]), which shorter time spans might avoid.

We logged rate limiting messages from the Twitter Streaming API and found that few tweets were omitted per day due to rate limiting. We experienced no rate limiting at all for 176 days and only slight rate limiting on other days (median 4.5 tweets lost on days with rate limiting messages). Power interruptions and network connectivity resulted in additional data loss, but there is no indication of any systematic bias from these interruptions. During our time window, 1,980,600 individual users sent a geolocated tweet which fell within our bounding box.

The distribution of user activity is, as might be expected, heavy tailed, with a majority of users relatively inactive. We expect that users who tweet more frequently will give a more accurate signal about their home and work locations, hence we decided to impose a number of filters on the dataset to only include users of the platform who had a relatively high level of engagement. We make use of three filters in particular. First, we discarded users who had less than 5 tweets, as a kind of minimum threshold for extracting any kind of signal from the pattern of user engagement. Second, we discarded users who did not have either two tweets in two separate local authorities, or greater than the λ threshold of tweets within one local authority (meaning it could be assigned as both a home and work location): again, this was done to set a minimum threshold for extracting signal from a pattern of user engagement. Finally, we discarded users whose first and last tweets in their detected locations were less than 30 days apart (to try and eliminate, for example, people who only sent a short burst of tweets from a holiday destination). Application of these filters resulted in a large number of low-intensity users being discarded. After these steps the exact size of the dataset varied with the λ threshold, from just over 560,000 for $\lambda = 0.70$ to just over 380,000 for $\lambda = 0.95$. We discuss the potential impact of this filtering more in our results section below.

In order to assign living and working locations to users, we first assigned each geolocated tweet to a local authority area. This assignment was achieved in one of two ways. When exact coordinates were included with the tweet, assignment was simple, as any pair of co-ordinates will fall within only one local authority. If co-ordinates are not included, what Twitter includes instead is a bounding box around a given place or region of geographic interest (for example, a city, a county or even a country); Twitter also includes information about the type of bounding box.[e] If the type of place is defined as a 'city' we use the centroid of the bounding-box as our point for geolocation, on the assumption that the majority of cities do not cross local authority boundaries (it is worth noting that a 'city' in this context also refers to an area of London). In total, we were able to assign 87% of tweets to a local authority, or 122 million tweets in total. Geolocated tweets which could not be assigned are those where the area of geolocation was too high to meaningfully assign to a local authority (for example, tweets can be geolocated to 'United Kingdom' or 'East England').

Of course, we do expect this geographical assignment process to contain some error within it. Users may assign any place name to a tweet: they are not required to assign

the 'correct' name. Furthermore, some bounding boxes may cross local authority boundaries, making the centroid an unreliable means of distinguishing location. Nevertheless, we expect the process to be broadly accurate. This is something supported by an observed strong correlation ($r = 0.78$) between geolocated tweets and the population of each local authority (a finding which also offers further confirmation for the results from [11]).

3 Modelling commuting with Twitter

We will now move on to discuss the results of the study. We will begin by looking at how predictions from our Twitter model compare to the ground truth census data; and also whether the accuracy of Twitter based predictions can improve on predictions from the radiation model. Following [26] and [24], we use the 'common part of commuters' [CPC] score based on the Sørensen index [28] to assess the accuracy of commuting flow estimates. A CPC score essentially compares the similarity of two matrices, $\mathbf{L}$ and $\widetilde{\mathbf{L}}$, and is given by the following equation:

$$\mathrm{CPC}(\mathbf{L}, \widetilde{\mathbf{L}}) = \frac{2\sum_{i,j=1}^{K} \min(\mathbf{L}_{ij}, \widetilde{\mathbf{L}}_{ij})}{\sum_{i,j=1}^{K} [\mathbf{L} + \widetilde{\mathbf{L}}]_{ij}}, \tag{4}$$

where K is the number of rows in the matrix (in our case the number of local authority areas). CPC scores lie in $[0, 1]$ with 1 indicating perfect agreement, *i.e.* $\mathbf{L} = \widetilde{\mathbf{L}}$. We also calculated all our results using Cosine distance and the Pearson product-moment correlation, to check whether our findings were sensitive to the metric used. The conclusions from these other two measures were essentially identical, and hence have not been reported here.

As described above, estimates from the radiation model are typically normalised, with row i tranformed to sum to $c_i = C(n_i/N)$, where C is the overall volume of commuting, n_i is the population of the local authority represented by row i, and N is the overall population. Here, we perform this normalisation on both the estimates from the radiation model and the Twitter model, to make results generated comparable. C is calculated by summing the commuting matrix from the census; we also make use of local authority population figures from the census to calculate n_i and N.

Table 1 shows the CPC values for the Twitter-based estimates, with values of λ varying between 0.70 and 0.95, and both variants of the radiation model (using equation (3) to estimate the parameter). As mentioned above, the radiation model does not offer an estimate for internal commuting (*i.e.* the number of people who live and work in the same area), whereas the Twitter model does. To properly compare the two approaches, we hence also produced a version of the Twitter model which considers only external commuting: *i.e.* the diagonal entries of the flow matrix are set to zero. All estimates have bootstrapped 95% confidence intervals generated from 10,000 bootstrap samples.

Three findings are evident from the table. First, the CPC scores from the Twitter models are high, reaching above 0.7 for higher values of λ and above 0.8 if internal commuting is included. Hence the proxy is quite good in absolute terms. Second, the Twitter model outperforms both versions of the radiation model for all values of λ considered (and, as the confidence intervals of the radiation models do not overlap with any of the Twitter models, these differences could be considered statistically significant). The results hence show that Twitter data can offer a good measurement of local commuting patterns that

Table 1 CPC scores for comparisons of the Twitter model and the radiation models to commuting data from the census. Brackets contain bootstrapped 95% confidence intervals

	CPC scores	
	All commuting	**External commuting**
Twitter Models		
$\lambda = 0.70$	0.70 (0.68-0.71)	0.67 (0.66-0.68)
$\lambda = 0.75$	0.73 (0.71-0.74)	0.68 (0.67-0.69)
$\lambda = 0.80$	0.76 (0.75-0.77)	0.69 (0.68-0.70)
$\lambda = 0.85$	0.79 (0.78-0.81)	0.70 (0.69-0.71)
$\lambda = 0.90$	0.83 (0.82-0.83)	0.71 (0.70-0.71)
$\lambda = 0.95$	0.82 (0.81-0.82)	0.71 (0.70-0.72)
Radiation Models		
Standard	N/A	0.57 (0.55-0.58)
1-parameter	N/A	0.62 (0.61-0.63)

improves on existing freely available models. This is especially significant if we consider both the simplicity of the location assignment technique, and the relatively coarse grained nature of the data (based on relatively wide area bounding boxes). Finally, higher values of λ are associated with higher CPC scores. This might indicate that, typically, users make much more use of the social media network when they are at home rather than at work: hence even a small amount of tweets in another area might indicate a pattern of external commuting. However, it is worth noting that this may also be related to the fact that, as described above, more users are discarded at higher levels of λ.

In addition to this general measurement of the accuracy of the Twitter model, it is also worth exploring some of the sources of error in the predictions it generates. These errors are interesting from a scientific point of view, but they are also of policy relevance. For instance, if Twitter data offers better estimates for certain demographic groups, certain areas, or certain types of trip, then the people or places for whom the estimates are better may be favoured if Twitter data is used in policy decisions (for example, transport infrastructures might be unknowingly adapted more toward the needs of those who make more use of Twitter).

We explore sources of error in the model in a variety of ways below: in each case we make use of the best performing model identified, which was the Twitter model with internal commuting included and with $\lambda = 0.90$. The first source of error we investigated was the volume of commuting between local authority pairs: we might expect local-authority pairs that share lots of commuters to be estimated better than pairs that have just a few commuters, as the signal will be stronger and hence less affected by noise in the data. Figure 1 investigates this possibility, by showing a heatmap of the estimates for the Twitter-based approach against the census data (estimates for the one parameter radiation model are also included for the purposes of comparison).

The figure shows, first of all, that the majority of data represents local authorities pairs with low levels of commuting between them (more than 75% of local authorities have 10 or less people commuting between them). This is to be expected: most outward commuting from one local authority goes to just a few geographically proximate local authorities. These types of local authority pairs are also where both the Twitter and radiation models are most unreliable: there is in fact little observable correlation between Twitter estimates and census results for local authority pairs with less than 10 commuters going between them. However, it is also worth highlighting that, whilst this represents a significant amount of data in terms of local authority pairs, when considered in terms of total

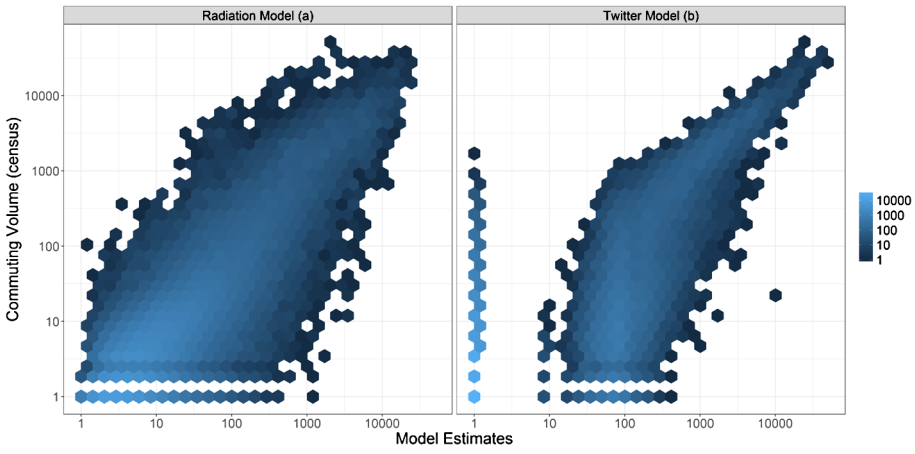

Figure 1 Estimated commuter flow. (a) 1-parameter radiation model. **(b)** Twitter model ($\lambda = 0.90$). Axes are log transformed, with all data points incremented by 1 before transformation.

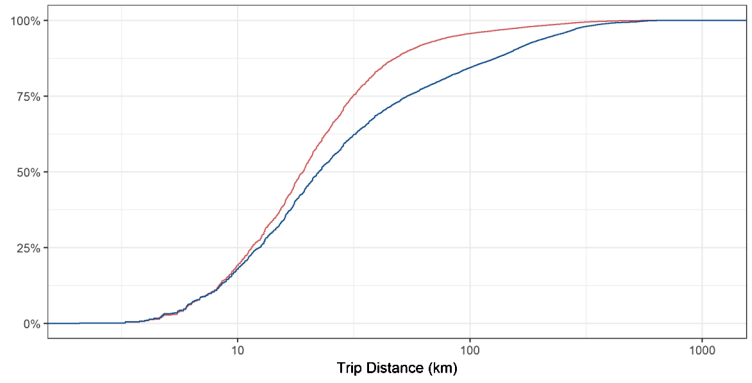

Figure 2 Empirical cumulative distribution of commuting trip distances. Red: census data. Blue: Twitter model.

volume of commuters it is much less significant: less than 25% of total commuters travel on these low intensity routes.

A further way of investigating sources of error in the model is to look at the distribution of commuting trip distances. We might expect our poorly predicted low commuting volumes to occur disproportionately for local authority pairs that are far away from each other (who are of course likely to have fewer commuters). Figure 2 shows the empirical cumulative distribution of commuting distances (with distances themselves based on local authority centroids) for the census and for our Twitter model. The main difference between the distribution of trips on Twitter and the ground truth census data is that Twitter predicts more long range trips: more than 15% of the trips predicted in the Twitter model are of more than 100 km in length, compared with only 4% of trips in the ground truth census data (a Kolmogorov-Smirnov test confirms the statistical significance of the difference between the distributions in both normal and log scale). We speculate that the inaccuracy of the Twitter model in this respect stems from the fact that we are making use of one year of Twitter data, and hence may observe long term mobility patterns even whilst trying to estimate short ones: this is likely to inflate the number of long trips we es-

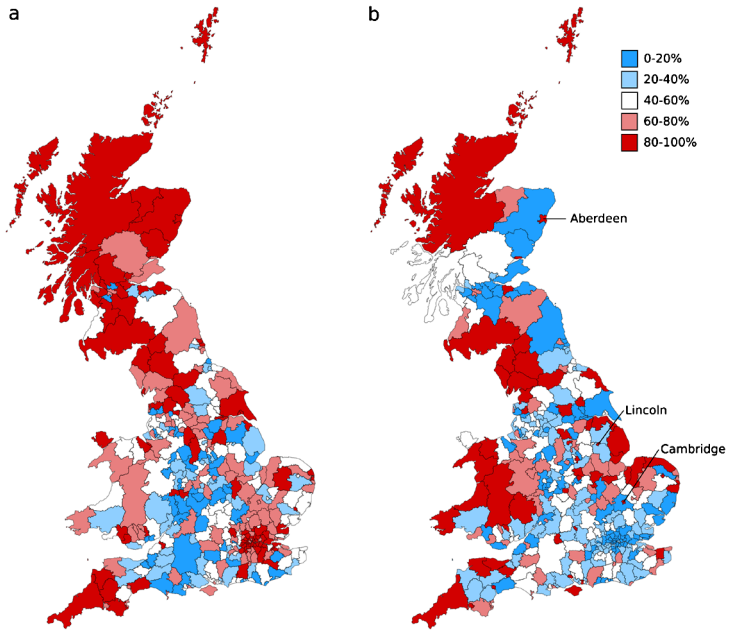

Figure 3 Absolute error in the estimated outward commuting distributions. (a) 1-parameter radiation model. **(b)** Twitter model ($\lambda = 0.90$). Color: quintiles for the combined set of errors from both approaches.

timate. This speculation is supported by [22], which has shown that Twitter data can also be used to estimate long term internal migration patterns (the study focussed particularly on students leaving their home town to go to university). There is, in other words, clearly a trade off in terms of using social media data: using a longer time period allows more data to be built up, and hence potentially a stronger signal; but it also introduces new types of bias which otherwise would not be present.

We will next look at the geographic distribution of error within the models. We might expect a variety of types of geographic variation in error: for example we might observe better predictions to be available for densely populated city areas for which there is more data. In order to investigate this, we rewrite equation (4) as:

$$\mathrm{CPC}(\mathbf{L}, \widetilde{\mathbf{L}}) = 1 - \frac{\sum_{i,j=1}^{K} |\mathbf{L}_{ij} - \widetilde{\mathbf{L}}_{ij}|}{\sum_{i,j=1}^{K} [\mathbf{L} + \widetilde{\mathbf{L}}]_{ij}}. \tag{5}$$

The numerator in equation (5) is a sum over prediction errors; so, we can look at the error associated with a given local authority by fixing i and summing over j to get the outward commuting error. Since the total number of outward commuters is different for each local authority, we first normalize each row of $\mathbf{L}$ and $\widetilde{\mathbf{L}}$ to sum to 1 and hence, consider the outward flow probability distribution for each local authority. The results are visualized in Figure 3, which contains results both from the Twitter model and the 1-parameter radiation model for comparison.

Figure 3 shows that there are some systematic patterns in the data. In the Twitter model, errors are particularly notable in medium sized cities which are surrounded by countryside (for example, the cities of Aberdeen, Lincoln and Cambridge, all with populations between 100,000 and 300,000). This highlights the fact that the model produces relatively

Table 2 CPC scores for the estimated Twitter commuting matrix ($\lambda = 0.90$) compared against census commuting volumes divided by gender, age and social class. Internal commuting is included. Brackets contain bootstrapped 95% confidence intervals. * indicates the confidence interval for the category being considered does not overlap with the interval for the full census

Commuting type	CPC score	Difference from baseline
All (baseline)	0.825 (0.815-0.835)	
Gender		
Male only	0.784 (0.773-0.795)	–0.041*
Female only	0.843 (0.835-0.852)	0.018
Age		
16-24	0.823 (0.813-0.832)	–0.002
25-34	0.760 (0.748-0.771)	–0.065*
35-49	0.816 (0.806-0.827)	–0.009
50-64	0.827 (0.817-0.836)	0.002
65-74	0.791 (0.781-0.801)	–0.034*
75+	0.770 (0.758-0.782)	–0.055*
Social Class		
AB	0.719 (0.707-0.730)	–0.106*
C1	0.804 (0.794-0.813)	–0.021*
C2	0.814 (0.804-0.824)	–0.011
DE	0.795 (0.785-0.804)	–0.030*

poor estimates of the small amounts of commuting that go out from these cities into the surrounding countryside (by contrast, the amounts of commuting into these cities is well estimated, something shown by the fact that the areas surrounding the cities are typically blue). It is worth noting that within the large metropolitan area of London, the Twitter model performs well in absolute terms, and also much better than the radiation model, which does quite poorly in this area (something also found by [29]).

A final area of potential error we considered concerned the impact of demographic factors. Twitter users are not demographically representative of the general population [30, 31], and furthermore Twitter users who geotag their tweets are not even representative of Twitter users more generally [32, 33]. We might expect this bias in the type of user included in the model to distort the results; for example, Twitter commuting estimates might be better for groups which are well represented on Twitter. In order to consider the impact of demographics on commuting predictions, we make use of further census data which describes the level of commuting between local authority pairs for a variety of different socio-demographic categories. In particular, commuting is divided up by gender, by age group, and by social class. For each of these demographic variables, we compare the performance of our predictor with the baseline performance for all types of commuting generated in Section 3 (again with $\lambda = 0.90$).

The results of this investigation are shown in Table 2. Differences can be observed in all the demographic categories we tested, approximately half of which could be considered statistically significant on the basis of whether their confidence interval overlaps with the confidence interval of the main prediction (these are marked with an asterisk). Prediction of commuting is better for females than it is for males; it is also largely better for younger age groups (less than 65) than it is for older age groups. Prediction of commuting was also better for the 'lower' social grades than the higher ones (when compared to the baseline, prediction of commuting for the AB social class was the worst of all demographics tested). However, the magnitude of the differences from the baseline in all categories is also relatively small: by and large, the difference from the baseline measure is around 0.05 CPC or

less, and in no categories was the observed CPC score lower than 0.70. From this we conclude that, even though the users of Twitter might be demographically biased, this does not hamper to a large extent our ability to infer mostly accurate commuting patterns from the data.

4 Extending the Twitter model

As we have highlighted, the simple model using Twitter data offers a good proxy for local commuting patterns, but not a perfect one. In this further analytical section, therefore, we want to consider ways of extending the model. Our current model is simplistic: it makes use of a mere frequency count of the geographic locations of tweets to assign individuals to a home and work location. However, Twitter data contains much more information than just geographic locations, and including this extra information might improve our predictions. In this section, we explore one particular avenue, looking at whether temporal information might offer an improvement in accuracy. Temporal information is of course something which is already made use of in a variety of existing works on trace data [19–21, 26].

Working (and hence commuting) typically follows temporal patterns. Hence it might be possible to improve the accuracy of our assignment of home and work locations by considering whether a user's tweets are being sent during the typical working day. To test this possibility, we of course first need to define the limits of the working day. In Figure 4, we explore the extent to which we can do this on the basis of our Twitter data, by looking at how patterns of tweets within our data vary over time.

We begin by looking at whether there is any variation between weekdays, where the majority of work occurs, and the weekend. Figure 4(a) and (b) show the temporal distribution of tweets on Saturday-Sunday compared to Monday-Friday for each local authority. Figure 4(c) shows the means of the weekend and weekday distributions. As we might expect there appear to be slightly different patterns of Twitter usage on weekends and weekdays, with weekday use characterised by a sharper evening peak (the period which runs approximately from 18:00-24:00) and flatter use levels during the daytime (approximately 08:00-18:00). This suggests that patterns of geolocated Twitter usage are different during the working week, hence only using data created during weekdays might improve the accuracy of the Twitter model.

We will now move on to exploring temporal fluctuations during the day. In Figure 4(d), we cluster the distributions in Figure 4(b) using K-medoids clustering ($K = 2$) with the earth-mover's distance [34], a well-used metric that computes the cost of transforming one distribution into another. It is notable that one cluster (shown in blue) is characterized by fewer tweets during the daytime and a higher peak in the evenings, whilst the other (shown in red) has comparatively more tweets during the day and less in the evening. This could suggest that the blue cluster represents local authorities with high outward commuting and the red cluster represents authorities with high inward commuting. As a means of validating this proposition, we compute the ratio of inward to outward commuters for each authority and find the geometric mean of these ratios for each cluster (we use the geometric mean here as it is more appropriate for ratio data). The means are 1.38 and 0.95: on average, inward commuting is 138% of outward commuting in the red cluster, but only 95% in the blue cluster, which seems to support our assumption.

The fact that we are able to identify typical patterns of tweets for areas of high and low inward commuting also suggests that we can use these patterns to understand more

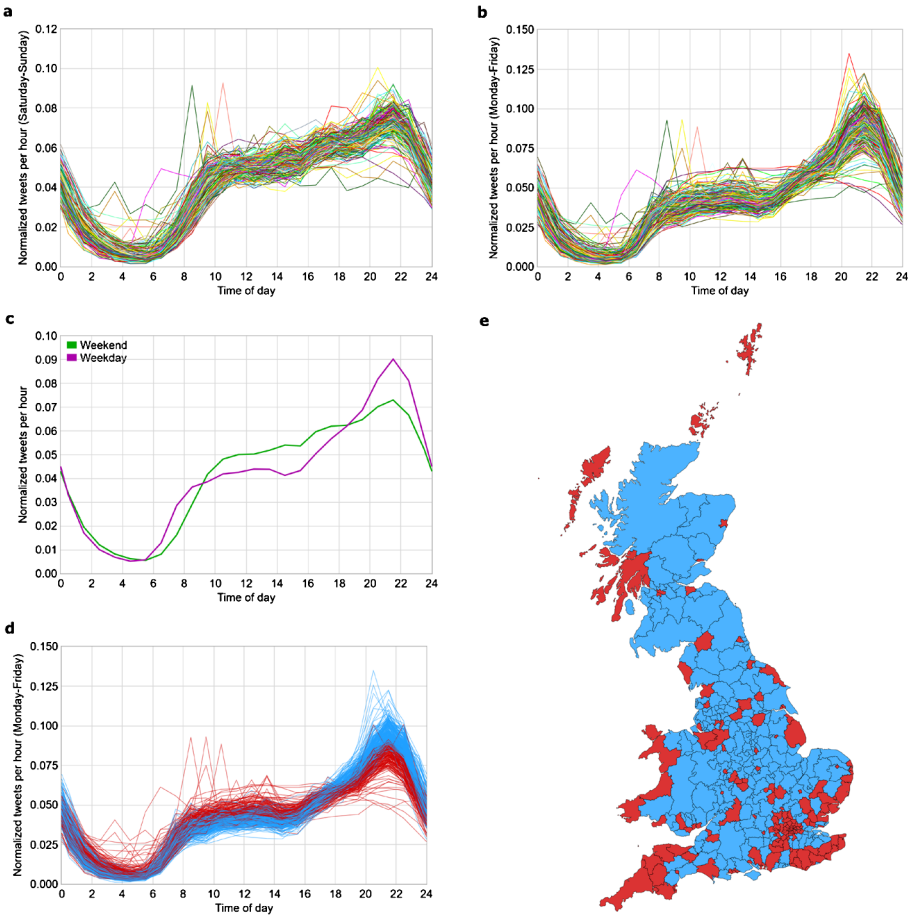

Figure 4 Temporal distribution of tweets for each local authority. (a) Saturday and Sunday. **(b)** Monday-Friday. **(c)** Mean of weekend and weekday distributions. **(d)** Clustering local authorities on Monday-Friday distributions using *K*-medioids ($K = 2$) with the earth-mover's distance. **(e)** Cluster membership of local authorities.

about the appropriate time window to use for the typical working day. In Figure 4(d), the two clusters intersect at around 9:00 and 17:00, suggesting this time period could be used as delimiting the typical working day. The clusters are also particularly clearly separated during the periods 10:00-15:00 and 20:00-23:00, which suggests that focusing on these times as a kind of 'restricted' working day may be optimal in terms of identifying work and home locations respectively. We experiment with the use of both of these approaches to defining the working day below.

By providing discrete time periods within which to assign home and work locations, the temporal approach to classification also offers a second advantage, which is that it allows us to incorporate a measure of uncertainty into the assignment of home and work locations to each user. Thus far we have made use of a 'hard' assignment method: the location with the most tweets is assigned to the user as their home location, and the one with the second most tweets is assigned as their work location, with ties discarded. But theoretically there ought to be a difference in certainty according to the distribution of tweets: for example, users who had a variety of candidate areas as potential work locations

Table 3 CPC for Twitter-based estimates using different time-based heuristics, for both hard and soft assignment techniques. Brackets contain bootstrapped 95% confidence intervals. The restricted day only considers tweets from 10:00-15:00 when calculating the work location, and tweets from 20:00-23:00 when calculating the home location

Time interval	All commuting		External commuting	
	Hard assign	Soft assign	Hard assign	Soft assign
9-5 Working Day	0.64 (0.62-0.66)	0.70 (0.68-0.71)	0.68 (0.67-0.69)	0.73 (0.72-0.74)
9-5 Working Day, Mon-Fri	0.68 (0.66-0.69)	0.72 (0.70-0.73)	0.70 (0.70-0.71)	0.74 (0.74-0.75)
Restricted Day	0.65 (0.63-0.66)	0.70 (0.68-0.71)	0.67 (0.66-0.68)	0.72 (0.71-0.73)
Restricted Day, Mon-Fri	0.70 (0.68-0.71)	0.73 (0.71-0.74)	0.70 (0.69-0.71)	0.73 (0.72-0.74)

with similar amounts of tweets ought to have less certainty in them than users who had one work area which had a very clear majority.

In order to incorporate the uncertainty in assigning a home and work location to each user we consider a method of soft assignment by creating a 'location matrix' $\mathbf{L}_u$ for each user, which is given by:

$$\mathbf{L}_u = \mathbf{h}_u \mathbf{w}_u^T,$$

where $\mathbf{h}_u$ and $\mathbf{w}_u$ are the normalized distributions of home and work tweet counts for user u respectively. Hence, we can interpret the entries of $\mathbf{L}_u$ as:

$$\begin{aligned}[\mathbf{L}_u]_{ij} &= [\mathbf{h}_u]_i [\mathbf{w}_u]_j \\ &= \text{Prob}(\text{live} = i)\, \text{Prob}(\text{work} = j) \\ &= \text{Prob}(\text{live} = i, \text{work} = j).\end{aligned}$$

To estimate Prob(live = i, work = j) for the population, we take the mean over all U users:

$$\mathbf{L} = \frac{1}{U} \sum_{u=1}^{U} \mathbf{L}_u.$$

The rows of $\mathbf{L}$ are then normalized as before.[f]

Table 3 shows the CPC scores based on the temporal heuristics discussed above. Results are provided both for the Twitter model based on all commuting and the model which is based only on external commuting. For each of these two types of model, we show results using the hard assignment method which we used in Section 3 (whereby the top two locations from which a user tweets are used as their home and work locations respectively) and the soft assignment method described above which incorporates uncertainty.

The findings from this table are mixed. There is good evidence that soft assignment, which incorporates uncertainty, is better than hard assignment. There is also good evidence that focussing on Monday-Friday only gives better estimates than looking at the whole week. By contrast, there is less evidence that the time of day makes a difference, with no significant differences between the 9-5 working day and restricted working day.

Furthermore, the best performing heuristics in Table 3 only offer a modest improvement on the estimations of external commuting developed with simpler heuristics (Table 1), and no improvement when internal commuting is considered. This is perhaps surprising: it may suggest that temporal information is not as valuable as we might have expected when estimating different location patterns. However, it may also be related to the fact that, as people appear to moderate their Twitter usage during working hours, the signal contained in temporal information is already largely captured by our simple frequency counts.

5 Conclusion

In this paper we have set out to examine the extent to which Twitter data can be used to estimate local commuting flows, thus building on the nascent literature that seeks to extract reliable population indicators from social media data. We have shown that simple heuristics applied to Twitter data offer good approximations of local commuting patterns; approximations that outperform the current benchmark for commuter flow estimation models (the radiation model). We explored the sources of error in these estimations, and found that Twitter was more reliable at estimating large commuting flows over short distances, and less reliable at estimating small amounts of long range commuting. We found some evidence of geographic and demographic biases in the data, though these biases were not severe. We also explored potential extensions to the models, but in the end found that simple frequency heuristics largely outperformed more complicated models using temporal information. In conclusion, we would argue that this paper highlights again the potential of freely created and distributed social media data for understanding more about local populations.

Funding
This project was supported by funding from InnovateUK under grant number 52277-393176 and the NERC under grant number NE/N00728X/1.

Abbreviations
OD: Origin-Destination. API: Application Programming Interface. CPC: Common Part of Commuters.

Availability of data and materials
Data supporting the publication will be made available in Oxford's institutional repository following publication.

Ethics approval and consent to participate
We confirm that this study has received ethical approval from the University of Oxford's Ethics Review board (case ref: SSH OII C1A 16 076).

Competing interests
We declare no competing interests.

Consent for publication
Not applicable.

Authors' contributions
SH collected the data for the manuscript, whilst GM and JB produced the analysis. All authors contributed equally to the drafting of the manuscript. All authors read and approved the final manuscript.

Endnotes
[a] For more details on current geotagging practices in the Twitter Streaming API see: https://twittercommunity.com/t/foursquare-location-data-in-the-api/36065
[b] https://wicid.ukdataservice.ac.uk/
[c] In the census commuting data (and throughout this article), Westminster and the City of London are treated as a single local authority, as are Cornwall and the Isles of Scilly.
[d] The coordinates used form a rectangle with a lower-left corner at −13.4139, 49.1621, and a top-right corner at 1.7690, 60.8547.

[e] It is worth noting that the time period considered starts after Twitter began promoting the inclusion of 'place' in tweets rather than the exact latitude-longitude coordinates. These places are represented as bounding boxes in the data. From earlier data, we observed an 80% decrease in exact-geotagged tweets in April 2015, though the the overall number of geotagged tweets (*i.e.* exact or place) remained stable.

[f] Extending the filtering from Section 3, entries of $\mathbf{h}_u$ and $\mathbf{w}_u$ that are not associated with tweets spanning longer than a 30 day period are reset to 0 before the vectors are normalized. Users that do not have at least one non-zero entry in both $\mathbf{h}_u$ and $\mathbf{w}_u$ are discarded. Depending on the precise time-windows used, 75-85% of users are discarded, leaving 287,000-496,000 users.

Publisher's Note

Received: 21 February 2017 Accepted: 14 September 2017 Published online: 02 October 2017

References

1. Mayer-Schönberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and think. John Murray, London
2. Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M (2009) Life in the network: the coming age of computational social science. Science 323(5915):721-723
3. Voigt C, Bright J (2016) The lightweight smart city and biases in repurposed big data In: The second international conference on human and social analytics (HUSO 16)
4. Llorente A, Garcia-Herranz M, Cebrian M, Moro E (2014) Social media fingerprints of unemployment. PLoS ONE 10(5):e0128692
5. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with web search. Proc Natl Acad Sci USA 107(41):17486-17490
6. Curme C, Preis T, Stanley HE, Moat HS (2014) Quantifying the semantics of search behavior before stock market moves. Proc Natl Acad Sci USA 111(32):11600-11605
7. Broniatowski DA, Paul MJ, Dredze M (2013) National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS ONE 8(12):e83672
8. Kostkova P, Szomszor M, St Louis L (2009) #swineflu: the use of Twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. ACM Trans Manag Inform Syst (TMIS) 5(2):8
9. Yasseri T, Bright J (2016) Wikipedia traffic data and electoral prediction: towards theoretically informed models. EPJ Data Sci 5:22
10. Yasseri T, Bright J (2014) Can electoral popularity be predicted using socially generated big data?. IT, Inf Technol 10.1515/itit-2014-1046
11. Liu J, Zhao K, Khan S, Cameron M, Jurdak R (2014) Multi-scale population and mobility estimation with geo-tagged tweets. In: 31st IEEE international conference on data engineering workshops, pp 83-86
12. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41(3):260-271
13. Beiró MG, Panisson A, Tizzoni M, Cattuto C (2016) Predicting human mobility through the assimilation of social media traces into mobility models. EPJ Data Sci 5(1):17
14. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. PLoS ONE 7(5):e37027
15. Liu Y, Sui Z, Kang C, Gao Y (2014) Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. PLoS ONE 9(1):e86026
16. Mcardle G, Furey E, Lawlor A, Pozdnoukhov A (2014) Using digital footprints for a city-scale traffic simulation. ACM Trans Intell Syst Technol (TIST) 5(3):41
17. Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In: Proceedings of the seventh international AAAI conference on weblogs and social media
18. Graham M, Hale SA, Gaffney D (2014) Where in the world are you? Geolocation and language identification in Twitter. Prof Geogr 66(4):568-578
19. Alexander L, Jiang S, Murga M, González M (2015) Origin-destination trips by purpose and time of day inferred from mobile phone data. In: Transportation research part C: emerging technologies, pp 240-250
20. Toole J, Colak S, Sturt B, Alexander L, Evsukoff A, González M (2015) Population bias in geotagged tweets. In: Transportation research part C: emerging technologies, pp 162-177
21. Lenormand M, Picornell M, Cantú-Ros O, Tugores A, Louail T, Herranz R, Barthelemy M, Frías-Martínez E, Ramasco J (2014) Cross-checking different sources of mobility information. PLoS ONE 9(8):e105184
22. Swier N, Komarniczky B, Clapperton B (2015) Using geolocated Twitter traces to infer residence and mobility. Office for National Statistics GSS Methodology Series, 41
23. Erlander S, Stewart NF (1990) The gravity model in transportation analysis: theory and extensions, vol 3. VSP, Utrecht
24. Lenormand M, Huet S, Gargiulo F, Deffuant G (2012) A universal model of commuting networks. PLoS ONE 7(10):e45985
25. Simini F, González MC, Maritan A, Barabási AL (2012) A universal model for mobility and migration patterns. Nature 484(7392):96-100
26. Yang Y, Herrera C, Eagle N, González MC (2014) Limits of predictability in commuting flows in the absence of data for calibration. Sci Rep 4:5662
27. Barabasi AL (2005) The origin of bursts and heavy tails in human dynamics. Nature 435(7039):207-211
28. Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skr 5:1-34

29. Masucci AP, Serras J, Johansson A, Batty M (2013) Gravity versus radiation models: on the importance of scale and heterogeneity in commuting flows. Phys Rev E, Stat Nonlinear Soft Matter Phys 88(2):022812
30. Sloan L, Morgan J, Savage M, Burrows R, Edwards A, Housley W et al (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
31. Wang S, Lo D, Jiang L (2013) An empirical study on developer interactions in StackOverflow. In: Proceedings of the 28th annual ACM symposium on applied computing - SAC'13
32. Hecht B, Stephens M (2014) A tale of cities: urban biases in volunteered geographic information. In: International AAAI conference on web and social media
33. Malik M, Lamba H, Nakos C, Pfeffer J (2015) Population bias in geotagged tweets. In: Ninth international AAAI conference on web and social media
34. Rubner Y, Tomasi C, Guibas LJ (2000) The Earth mover's distance as a metric for image retrieval. In: International journal of computer vision, pp 99-121