Exploring Responses to Events: A Spatiotemporal Analysis of Twitter Activity

Thomas Martin

Abstract—In this paper I demonstrate an approach to identify and explore distinct events by investigating twitter activity. In particular it is discussed how it is possible to characterise geographical areas by trends in twitter activity. It is also shown that tweets related to specific event are highly localised in time, indicating that tweets are typically in response to something happening at a given moment in time. A discussion is led as to the appropriatenes of the techniques used in the study and how the findings here related to previous literature and other possible applications of the methods explored here.

Index Terms—Twitter, tweet, geotag, event-dectection, visualisation, visual analytics

1 Introduction

1.1 Motivation

Tweets and related twitter data are a particularly useful dataset in the realm of spatiotemporal data analysis. This is because they are mostly candid submissions and are often localised in both space and time. Because of this, twitter data has often been used by authors to identify and track unusual occurrences such as dramatic news events usually through a keyword analysis [1] or to discover otherwise unreported events [2]. It would be interesting, however, to investigate whether twitter data can be effectively used to discover and characterise otherwise widely observed and regular events. Such an approach may likely make it possible, with sufficient data, to characterise spatial regions in aggregate, and may provide a foundation both to identify underlying similarities between regions and to predict future behaviours.

1.2 Data

The primary dataset used in this project comes from a set of almost 120,000 tweets located primarily in the Greater London area, over a period of 48 hours between 04:00 31st December 2014 and 04:00 2nd January 2015. Importantly, this period of time covers a set of widely observed, time-localised, events such as New Year's Eve. Altogether, the dataset contains the following features for each tweet: longitude and latitude, message date username and id, message text, location, hashtags, flag to indicate whether it was a retweet (although no tweets in the raw dataset are retweets).

Tweets contained in the dataset cover a range of message types, including messages between users and shared media. The vast majority of tweets are in English and are person generated.

It is important to note that this set of twitter data was provided by City, University of London for the purpose of this project, so I cannot comment on how it was originally obtained.

1.3 Research Questions

Overall this project aims to see if it's possible not only to identify major events from twitter data, but whether geographical areas and time intervals can be characterised by their twitter usage. Breaking the project into specific questions,

- How does twitter activity vary across both space and time for this dataset?
- Can specific events be identified through twitter usage and how?
- What were the most significant events reported and how?
- Can London boroughs be grouped by their twitter activity and how?

2 Tasks and Approach

The initial dataset required some pre-processing to get it into a suitable state. As this project will only consider tweets originating from Greater London and the City of London, I removed tweets from outside this area. For the purposes of using a template for the choropleth map, I also had to match the labelled location for each tweet to an appropriate London borough.

Initially, I wanted to investigate the overall trends in both space and time to direct the project. This was a primarily visual approach, in order to perform exploratory visual analysis. I anticipated that there will be specific localised events that will be easily identifiable e.g. New Year's Day celebrations. In order to do this, I created a box map of tweets by borough to show the spatial distribution of tweets, and a bar chart of tweets by hour over the entire time period. Box maps are particularly useful in this task as it can help identify true outliers of a spatially distributed quantity. For peaks in twitter activity in either space or time, I did further research for the corresponding events. This research helped indicate the kinds of keywords to check for in subsequent analysis.

To add to this stage of the analysis I also wanted to find how keywords varied in usage both in space and time. This computational task should complement the previous visual tasks above. In both dimensions, I identified keywords by their inverse document frequency (IDF) weighting, which is lower the more frequent the term is. IDF is an effective way of determining the relative frequency of a given term out of a larger corpus, as it considers the number of times a keyword appears in relation to the overall corpus [2]. The "corpus" here is the total number of tweets for a given borough or time interval. In order to get a greater sense of the specific content of tweets, the keywords reported in this paper are stemmed, noun phrases.

From here, the project could have gone in one of two main directions to identify hotspots: using a density-based clustering approach or using a choropleth-based approach. I chose the latter for two main two main reasons. Firstly, given the volume of tweets and lack of spatial variation, a density-based approach would require additional filtering. I was reluctant to do this as such filtering is usually subjective and may wrongly ignore certain tweets [2]. Secondly, using a choropleth map, and in particular grouping by an established

administrative boundary, in this case by borough, helped to inform analysis and discussion of the results.

In order to find another grouping of boroughs by their characteristic twitter activity I first had to find a representation of twitter activity by borough, which meant finding the normalised number of tweets sent in hourly intervals by borough. Given this new feature set, I performed k-means clustering, optimising against the silhouette score, to determine the optimal number of clusters. I then repeated the previous steps, but in this case finding the normalised number of hourly tweets by cluster. This will indicate general trends in behaviour and allow for comparison between these clusters. A complementary choropleth visualisation helped illustrate the spatial distribution of these clusters.

Finally, I also wanted to discover how key events are reported. Combining the steps above, for events identified in time e.g. New Year's Eve celebrations, I found the set of most frequently occurring terms according to their IDF weighting. For these terms, I tracked their usage by cluster for the entire time interval and grouped by cluster. Grouping by cluster I believed would help to add detail to the twitter activity. For instance, clusters may have a peak in activity at a different time, but the subject matter may be the same. Peaks were easily identified by plotting the number of related tweets on a line graph.

3 Analytical Steps

Initial investigation into the spatial variation in the number of tweets yielded a box map and histograms. From Fig. 1, it is clear there is a high degree of variability of tweets by borough, although no outliers were found. There is no simple characterisation of activity by location, Inner London boroughs generally demonstrate higher twitter activity than Outer London boroughs, with the exception of City of London. (See Table 1 for Inner and Outer London boroughs.) Both the borough with the smallest number of tweets, City of London (1,062), and the borough the greatest number of tweets, Lambeth (4,817) are Inner London boroughs.

Fig. 2 plots the distribution of total tweets by borough. This shows a somewhat bimodal distribution indicating two more or less distinct groupings of boroughs by overall usage: boroughs of high usage, and boroughs of low usage.

Table 1: Inner and Outer London Boroughs

Inner London Boroughs	Outer London Boroughs
Camden	Barking and Dagenham
Greenwich	Barnet
Hackney	Bexley
Hammersmith and Fulham	Brent
Islington	Bromley
Kensington and Chelsea	Croydon
Lambeth	Ealing
Lewisham	Enfield
Southwark	Haringey
Tower Hamlets	Harrow
Wandsworth	Havering
Westminster	Hillingdon
Camden	Hounslow
Greenwich	Kingston upon Thames
	Merton
	Newham
	Redbridge
	Richmond upon Thames
	Sutton
	Waltham Forest

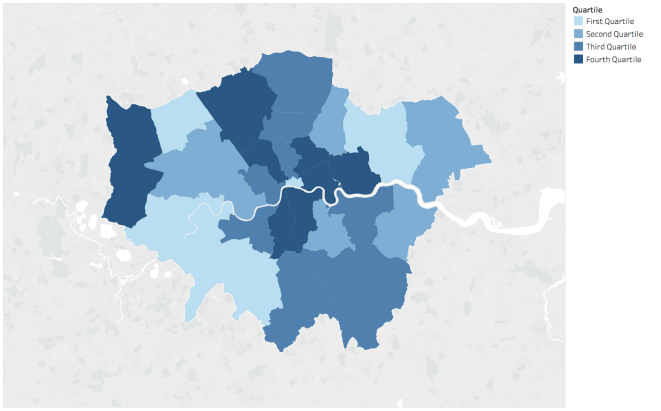

Fig. 1. Box map of total number of tweets sent during time period under consideration.

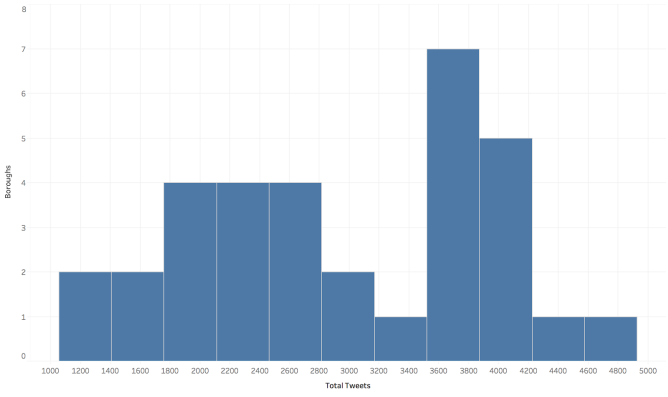

Fig. 2. Histogram of the distribution of total tweets by borough.

Fig. 3 shows the number of tweets created binned by hour. Given the tweets only cover a 48-hour period, it hard to generalise, however we see a distinct ebb and flow of twitter activity: rising during the day and dropping during the night. We also notice a number of periods of peak twitter activity, the top 5 are shown in Table 2.

To provide more context around these peak times, I researched some coincident events.

- New Year's Eve celebrations 00:00 1st January 2015
- Premier League football matches 15:30 - 19:00 1st January 2015
- EastEnders TV Special 20:30 - 21:35 1st January 2015

However, at this point in the study, it is unclear whether these events are reported in tweets.

Delving into the spatial and temporal nature of the tweets a bit more, I found the IDF of keywords by both borough and hourly interval. For most boroughs the number one keyword was either "new", in the case of Kensington and Chelsea and Tower Hamlets, and "year" for the others with one exception: the top keyword for "Barking and Dagenham" was "rt" (short for "retweet"). The remaining top keywords were all related to New Year's. Performing the same exercise for hourly interval, we find that the top keywords were mostly related to New Year's, but from the 2nd and 3rd most frequent keywords we find a mixture of subjects, however these are still exceptional. The top keyword for 8-time intervals was in fact "thi", a stemmed version of the pronoun "this" but taking this alone

Table 2: Peak hourly intervals by number of tweets

Hourly Interval	Number of Tweets
00:00 1st January 2015	4,447
18:00 1st January 2015	4,055
21:00 1st January 2015	4,024
19:00 1st January 2015	4,018
16:00 1st January 2015	3,676

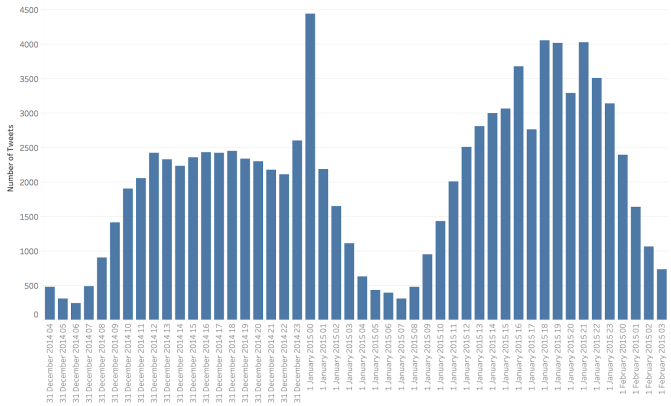

Fig. 3. Bar graph of tweets binned by hour of creation.

does not help identify what these time intervals focussed on - rather it might be just an indication that New Year's tweets were proportionally less frequent in these times. Table 3 lists some of these IDFs.

In the next step I wanted to determine if there was a richer relationship between boroughs, based not only on the total number of tweets, but the relative number of tweets sent in a given time interval. As we've seen above, particular times of day can be linked to specific events in the time frame considered, so it may well be the case that different usage of tweets relate to underlying differences between boroughs. As detailed in the previous section this required creating a set of vectors of the number of tweets sent per hourly interval, normalised across a given borough. Before performing clustering, I performed principal component analysis (PCA) to find a reduced set of features, in this case reducing the dimensionality of feature vectors from 48 to just 2. There are number of reasons for this approach. Firstly, it is much easier to provide illustrations for a two-dimensional dataset. Secondly, as k-means determines clusters using a distance metric, this tends to perform poorly in high-dimensional spaces. Fig. 4. illustrates the most successful clustering results determined by the silhouette score out of a range of cluster values. Note the cluster labels used in Fig. 4 are not consistent with labels used elsewhere in this paper.

Table 3: Top keywords by IDF not related to New Year's

IDF Rank	Keyword	Time Interval
2nd lowest	eastend	21:00 1st January 2015
	london	07:00 31st December 2014
3rd lowest	chelsea	19:00 1st January 2015
	kane	18:00 1st January 2015

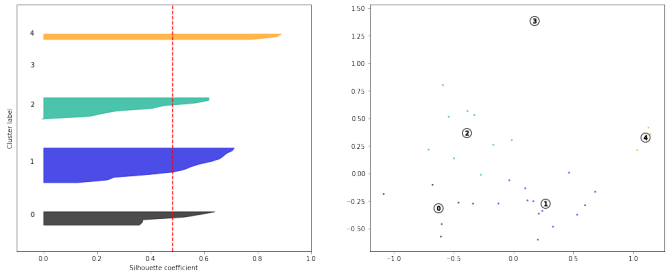

Fig. 4. For five clusters, the left-hand side shows the silhouette plot with average silhouette score indicated by dotted vertical line. The right-hand side show the plot of the actual clusters in the reduced feature space, with arbitrary dimensions.

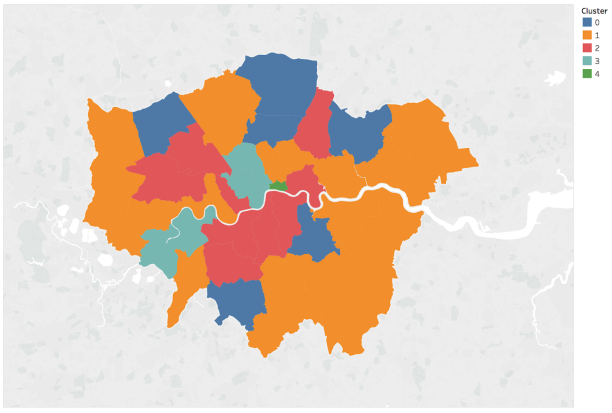

Fig. 5. Choropleth map of identified clusters.

I then matched the cluster label to the borough name to produce Fig. 5. Even more so than the box map above, it is clear that there is a high degree of spatial variation in twitter usage, for instance clusters of all types occur in Inner and Outer London boroughs. Cluster 4 is the only cluster to contain a single member, corresponding to the City of London.

Fig. 6 shows the relative number of tweets across the time period considered, grouped by cluster. Table 4 contains information from Fig. 5 for ease of reference.

- Cluster 0 had a global maximum during 19:00 New Year's Day. It had a local peak during the New Year's celebrations, but unlike other clusters this was noticeably smaller than its peak by about ~40%.
- Cluster 1 shows the peaks in relative tweets during both New Year's celebrations, and the period between 16:00 - 22:00 New Year's Day
- Clusters 2 and 3 together showed a similar pattern, with a global peak during the New Year's celebrations, and a smaller peak (~40% and ~25% smaller, respectively) during the afternoon of New Year's Day
- Cluster 4 showed peaks during 12:00 New Year's Eve, and during New Year's celebrations.

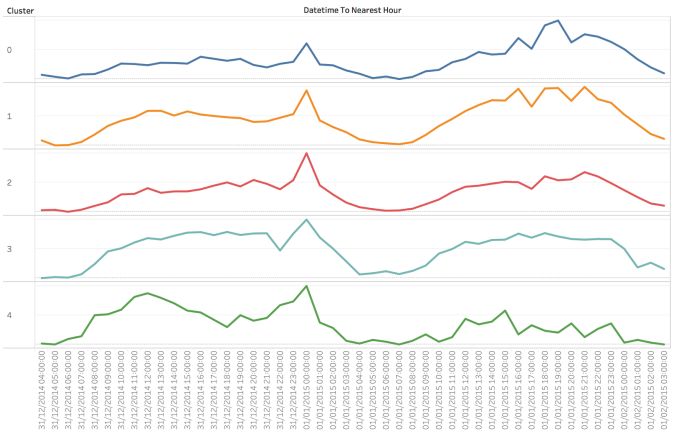

Fig. 6. Twitter activity over time period, normalized by cluster.

Table 4: Boroughs by cluster

Cluster	Boroughs	Boroughs by Cluster
0	Enfield, Haringey, Harrow, Lewisham, Redbridge, Sutton	6
1	Barking and Dagenham, Barnet, Bexley, Bromley, Croydon, Greenwich, Hackney, Hammersmith and Fulham, Havering, Hillingdon, Hounslow, Islington, Kingston upon Thames, Newham	14
2	Brent, Ealing, Kensington and Chelsea, Lambeth, Merton, Southwark, Tower Hamlets, Waltham Forest, Wandsworth	9
3	Camden, Richmond upon Thames, City of Westminster	3
4	City of London	1

Table 5: Top keywords related to New Year's celebrations

Keywords	IDF
year	1.78
new	1.88
new year	1.92
happi	1.93
happi new	2.02

Given this variation in twitter activity, it is likely this leads to variations in topics being tweeted about, for instance cluster 0 may focus more on football or TV rather than New Year's celebrations.

It was previously found that the top keywords for all hourly intervals generally referred to the New Year, it would be interesting to determine whether "New Year's" tweets were localised in time or not, especially around midnight on New Year's Eve. To decide whether this was the case, I found the top keywords by IDF used during the hour of 00:00 1st January 2015, shown in table 5. Keywords were identified from n-grams selected from tweets, for n between 1-3. This range of n-grams allows discovery of individual words as well as common expressions e.g. "Happy New Year!" while allowing for some variation in the specific arrangement. Given these keywords, I decided to look for all tweets containing the keyword tokens "happi", "new", or "year". To make it easier to reason about the twitter activity demonstrated I did this on a per cluster basis. In Fig. 7, and the remaining two figures, cluster 4 was removed due to the small number of tweets.

I chose to focus on raw count rather than relative proportion due to the potential problem of false negatives, discussed below. Fig. 7 is a good indication that New Year's tweets, at least as defined previously, are generally created at the time of the New Year celebrations. This is the case, even for clusters whose peak activity occurred at a different time. It is hard to give a precise metric for the success of this approach; however, the line graphs indicate quite visually the time-localised nature of related tweets. This discussion is continued in section 5 below.

This pattern is more striking still when considering less well observed events, in this case a number of football matches and an EastEnders TV special. In the former case, I manually identified a set of relevant keywords from those with lowest IDF. Following the same process, I also found tokens for the EastEnder TV special, however this was also required identifying tokens corresponding to major plot points out of the most frequently occurring terms by IDF weight [4]. In all cases, tweets containing identified keywords were focussed around the time that these events occurred. In the case of cluster 0, which demonstrated global peak activity during the afternoon of New Year's Day, we can see from Fig. 9 that this was related to activity related to the football matches. In all cases, peak twitter was related to events as they occurred.

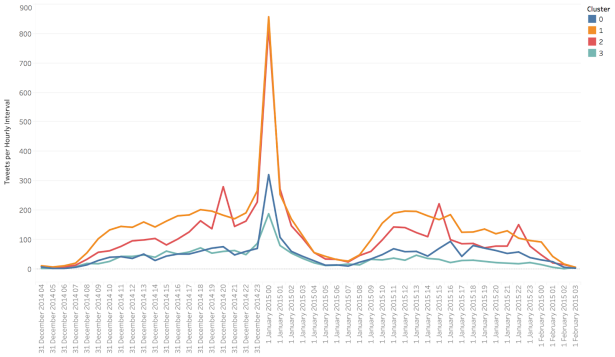

Fig. 7. Number of tweets related to New Year's celebrations over entire time period by cluster.

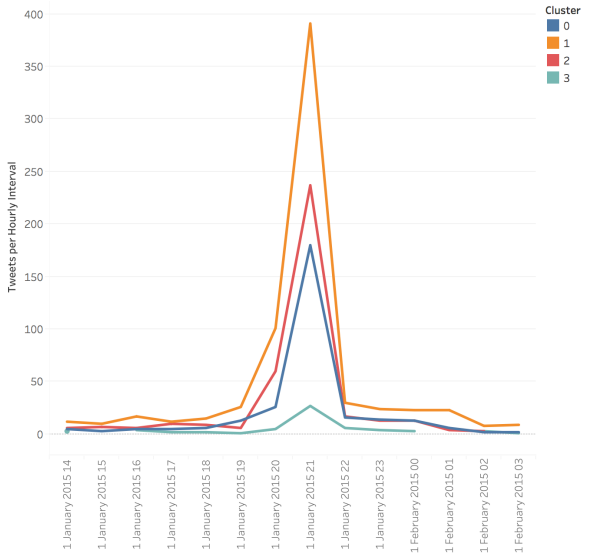

Fig. 8. Number of tweets related to EastEnders TV special, truncated to time period immediately before and after event.

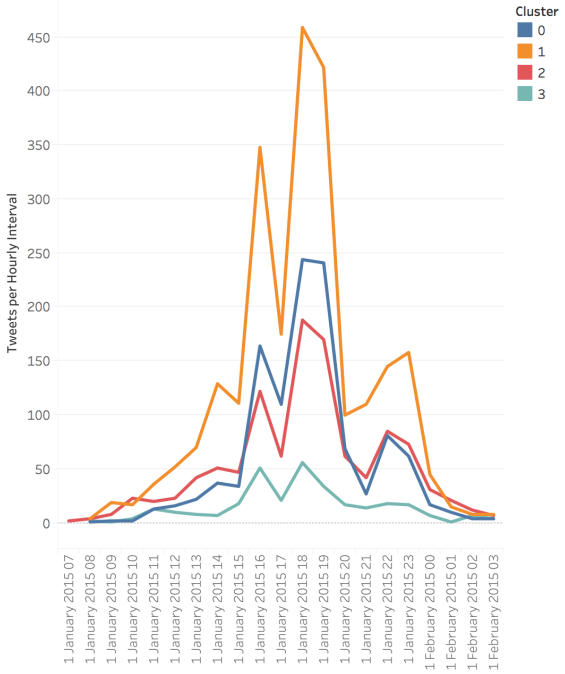

Fig. 9. Number of tweets related to football matches, truncated to time period immediately before and after events.

4 Findings

Within the geographical area and time period considered, there was a great deal of spatial and temporal variation in twitter activity. It was found that there was a bimodal distribution in the number of tweets sent over the entire time period, indicating two distinct groupings of boroughs with high twitter activity, and boroughs of low twitter activity. However, there is no straightforward explanation for this spatial distribution, as boroughs of high and low activity were scattered throughout the Greater London area. Twitter activity varied greatly by hourly interval, with five noticeably peak periods, which were found to correspond to widely observed events. These events were likely New Year's celebrations, various Premier League football matches, and the EastEnders TV special.

Events well observed by twitter activity were localised in time but not in space. Top words, by IDF, were more or less identical across all boroughs: For most boroughs, the top token was either "new" or "year". Temporal events were identifiable both by spikes in hourly tweets and by IDF scores. That is to say, observed spikes in the total number of tweets for a given hourly interval were due to a large number of tweets related to that event, rather than an overall increase in twitter activity. For instance, during the football match between Chelsea and Tottenham, the top tokens included "kane" and "chelsea", both with lower IDF weightings than "new year". This was also reflected at the cluster level. Given the time-localised nature of tweets by keyword, it is strongly indicative that, in general, relevant tweets are posted in response to specific events. However, as noted above there was a lack of rigor in establishing the whether all relevant tweets were being correctly identified or not. Also, this study cannot conclusively say whether one cluster discussed a particular topic more than another or not. I don't believe this entirely discounts the findings here, as the time-localised nature of related tweets is still valid.

It was also found that boroughs could be grouped by their overall twitter activity during this time period, which related to the relative number of tweets sent during particular times of day. For most clusters, the defining time intervals were during New Year's Eve celebrations and during the afternoon and evening of New Year's Day. For example, Cluster 0 demonstrated a peak in twitter activity during the evening of New Year's Day, almost twice as many sent during New Year's Eve celebration, whereas cluster 2 exhibited the opposite behaviour. The exception was cluster 4, corresponding to the City of London, which demonstrated much greater twitter activity overall on New Year's Eve than on New Year's Day.

5 Critical Reflection

5.1 Implications

In the context of tracking and identifying events through data mining of twitter activity, this study has shown the viability of identifying major events localised in time through twitter activity. This is perhaps more surprising in this case due to an event like New Year's, which features heavily in all tweets for the time period considered. A simple combination of calculating IDF weightings and manually finding appropriate keywords was highly effective at identifying related twitter activity.

This study has also demonstrated that it is possible to characterise larger groupings, in this case London boroughs by twitter activity. Other studies considered have typically focused on the burst of activity in tweets specific to certain events [1, 3]. However, in this study I have demonstrated it may be possible to also characterise aggregate groups in relation to their twitter activity, even in the context of universally observed events such as New Year's celebrations. A possible extension ofthis approach may be to detect patterns in behaviour and sentiment around regular and divisive social events such as elections. This could be used in combination with other demographic data to understand underlying similarities between geographical regions.

5.2 Discussion

I believe the use of choropleth maps in particular, were entirely appropriate for this study, especially as one of the key aims was to characterise pre-established administrative regions such as boroughs. However, considering the lack of spatial variation in keywords at least at the borough level, a density-based clustering technique may have aided in providing richer analysis. For instance, it is likely that tweets relate to local landmarks as the authors of [2] found. Overall the combination of mapping and graphing techniques proved extremely effective at answering the questions set out at the beginning of the project, especially in indicating the time-localised nature of tweets related to specific event. However, there is a lack of rigour in the computational techniques used, as it was difficult to verify evaluation metrics for the approaches used e.g. to determine the temporal spectrum of tweets. As discussed in [2], precision and recall could be used to evaluate the effectiveness of these tweets but would require manual labelling of tweets as being relevant or not. Due to lack of time, this approach was not possible.

There are a number of noteworthy limitations of this study worth discussing. First and foremost, the dataset is quite limited in scope and overall size, in particular it is limited to a single 48-hour time interval. This makes it very difficult to infer long term behaviour and justify reasons for cluster groupings. For instance, the boroughs of Tower Hamlets and Kensington and Chelsea were grouped into the same cluster, however given demographic and geographical differences, it is hard to speculate specific reasons for this similar behaviour. From a different angle, given that New Year's Day is a bank holiday, and City of London has a small residential population, it could be the case that the twitter activity found in this study does not represent typical usage. On a different note, it could be argued that twitter data is not the most appropriate means of identifying

underlying demographic similarities between spatial regions. This is because twitter activity is highly skewed by the typical user of the service, which tend to be a younger, more technologically abled person. It could also be argued that twitter data does not necessarily indicate typical usage in a given region as it is not clear if a given tweet was sent by a resident of that region - this would require cross-referencing another dataset.

The approach used to determine the temporal spectrum of tweets on a particular topic is likely to have led to a lot of false negatives i.e. related tweets ignored due to limited size of keyword corpus. For instance, for cluster 0 there are 1,845 tweets sent between 18:00-20:00 on New Year's Day, but only 485 were flagged in this approach as being relevant. Manually checking through the dataset indicated that these tweets were discussing events identified as occurring in the time interval. A different approach, used for example in [1], used latent Dirichlet allocation to mine twitter data to identify topics and related tweets. This is useful in relating multiple keywords to a specific topic, effectively automating the approach taking in this study.

5.3 Applications

Given the fairly generic approach to discovering events as they occur in time, it is entirely possible that this is applicable to a broader context and especially in the case of less well observed social events, similar to the approach taken by authors in [1].

Considering that twitter activity was found to characterise arbitrary groupings, it is entirely possible that such groupings can exist outside of a strictly spatial context. For instance, one may be able to identify or enrich demographic information based not only on the content of tweets, but also general twitter activity: working-age adults may typically tweet more so outside of working hours.

REFERENCES

[1] T. Cheng and T. Wicks, Event Detection using Twitter: A Spatio-Temporal Approach Tracing the Spatial-Temporal Evolution of Events Based on Social Media Data. *PLoS ONE*, 2014

[2] T. Sugitani, M. Shirakawa, T. Hara and S. Nishio, Detecting Local Events by Analyzing Spatiotemporal Locality of Tweets. *International Conference on Advanced Information Networking and Applications Workshops*, pp. 191 - 196, 2013.

[3] X. Zhou and C. Xu, Tracing the Spatial-Temporal Evolution of Events Based on Social Media Data. *International Journal of Geo-Information*, pp. 88 - 102, 2017.

[4] Metro, *Who will die in EastEnders this New Year's Day? The 8 prime candidates…*, https://metro.co.uk/2014/12/31/who-will-die-in-eastenders-this-new-years-day-the-8-prime-candidates-4994616/, last accessed 21st December 2018.