Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Great Britain's spatial twitter activity related to 'fracking'

P. Bartie [a,*], A. Varley [b], J. Dickie [b], D. Evensen [c], P. Devine-Wright [d], S. Ryder [e], L. Whitmarsh [f], C. Foad [g]

[a] Department of Computer Science, Heriot-Watt University, Edinburgh, Scotland EH14 4AS, United Kingdom
[b] Biological and Environmental Sciences, University of Stirling, Stirling FK9 4LA, Scotland
[c] Politics and International Relations, University of Edinburgh, Chrystal Macmillan Building, 15a George Square, Edinburgh EH8 9LD, Scotland
[d] Geography, University of Exeter, Amory Building, Rennes Drive, Exeter, England EX4 4RJ, United Kingdom
[e] Department of Sociology & Anthropology, Utah State University, 0730 Old Main Hill, Logan UT 84322-0730, USA
[f] Department of Psychology, 10 West, University of Bath, Claverton Down, Bath, England BA2 7AY, United Kingdom
[g] Department of Psychology, University of Reading, Harry Pitt Building, Earley Gate, Whiteknights Road, Reading, England RG6 6AL, United Kingdom

ARTICLE INFO

Keywords:
χ-Squared expectation surface
Social media
Geolocated tweets
Fracking

ABSTRACT

Fracking has proven to be a contentious issue in Great Britain, receiving wide press coverage from the initial sale of exploration and development licences, to the current moratorium. This research tracks the public activity online related to this 'fracking' journey by analysing over 317 million geolocated tweets from 2015 to 2020, mapping their location to compare the spatial distribution against the shale gas exploration sites. To spatially normalise the results for population density a χ-squared expectation surface was generated revealing higher than expected levels of interest near the previously active fracking site of Preston New Road and licenced extraction blocks in Lancashire. The data granularity allows for peaks of activity to be identified and topics analysed at higher temporal and spatial resolution than previously possible with more traditional surveys. The paper demonstrates the use of χ-squared expectation surfaces for normalising geotweets and the value of social media spatial-temporal analysis for monitoring local involvement in environmental issues, and for monitoring the changing level of interest across different regions in reaction to political decisions.

1. Introduction

Energy security is high on the policy agenda for governments (Watson et al., 2018), brought even more into the public focus since the invasion of Ukraine in 2022. In the last decade the USA, Canada, China, and Argentina have successfully increased their fossil fuel resources by extracting shale gas using hydraulic fracturing, also known as 'fracking', whereby chemicals and sand are pumped into the porous shale to force trapped natural gas into drilled wells (Speight, 2013). The UK set about copying this strategy and issued Petroleum Exploration and Development Licences (PEDLs) to control the potential extraction sites (Cotton, 2017). However shale gas extraction is contentious as it involves vertical and horizontal drilling that can cause small earth tremors from the delivery of chemicals to force out the gas, and can lead to pollution from flowback water, as well as industrialise the countryside.

After sales of PEDLs to companies such as Aurora Energy Resources, Ineos, Third Energy, and Cuadrilla, the only UK site to progress to active drilling was Cuadrilla's Preston New Road site, in Lancashire. The Welsh and Scottish governments imposed moratoriums on fracking in 2015, with Scotland effectively 'banning' it in 2017 after public consultation (Watterson & Dinan, 2018). This Scottish precautionary stance (Stephan, 2017) was mirrored in 2019 in England, where a moratorium on hydraulic fracturing was introduced based on a report from the Oil and Gas Authority, and all drilling ceased.

Social acceptance of a new technology is an important factor in determining energy policies and its adoption, which can be a result of public discourse (Boudet, 2021). While in some countries, such as Turkey, shale gas has been socially accepted for the energy security it brings (Kânoğlu-Özkan & Soytaş, 2022), the public response in Great Britain has been more hostile leading to a 'moratorium' by the UK government in 2019 (Devine-Wright et al., 2021).

Public opinion can be gathered through surveys or interviews but these are elicited by researchers and only access limited sections of the population, in contract social media posts provide insights on public opinion that are spontaneously and publicly expressed. Moreover, they allow for fine-grained geospatial analysis of how responses vary by

* Corresponding author.
E-mail address: phil.bartie@hw.ac.uk (P. Bartie).

https://doi.org/10.1016/j.compenvurbsys.2023.101978
Received 4 August 2022; Received in revised form 8 March 2023; Accepted 27 April 2023

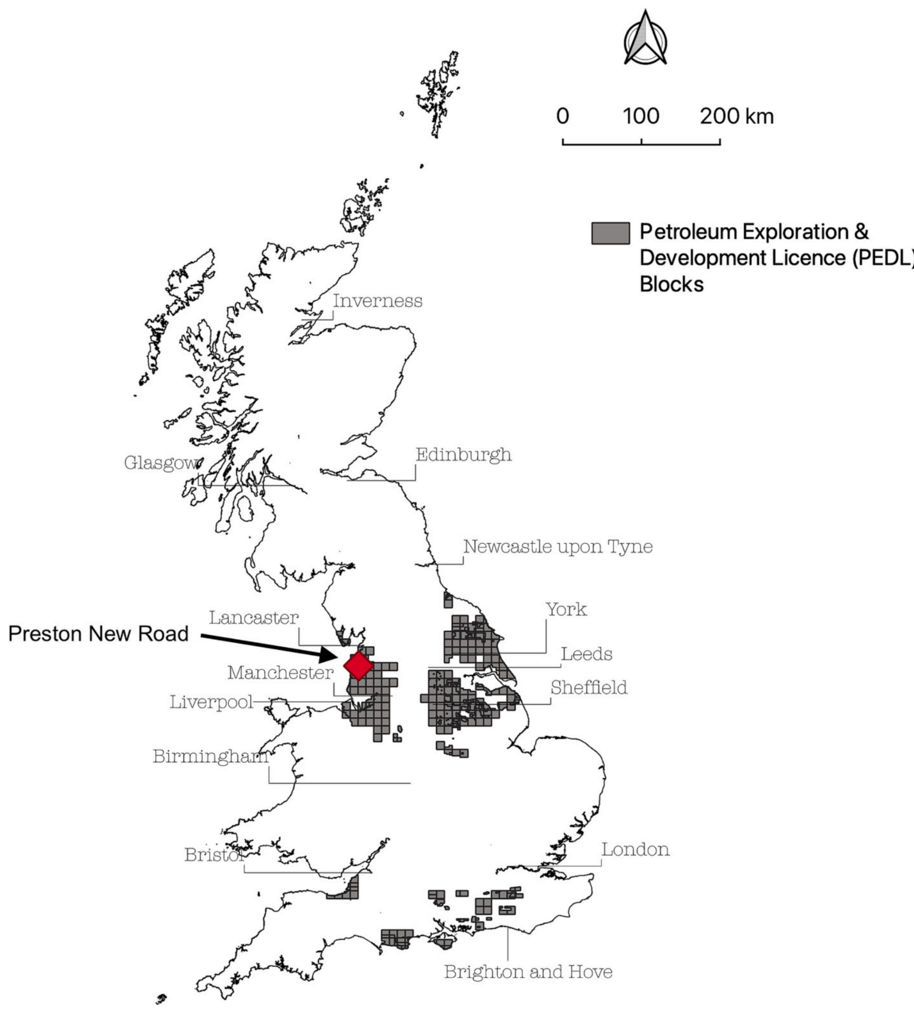


Fig. 1. Preston New Road - the only fracking site to have been active in Great Britain.

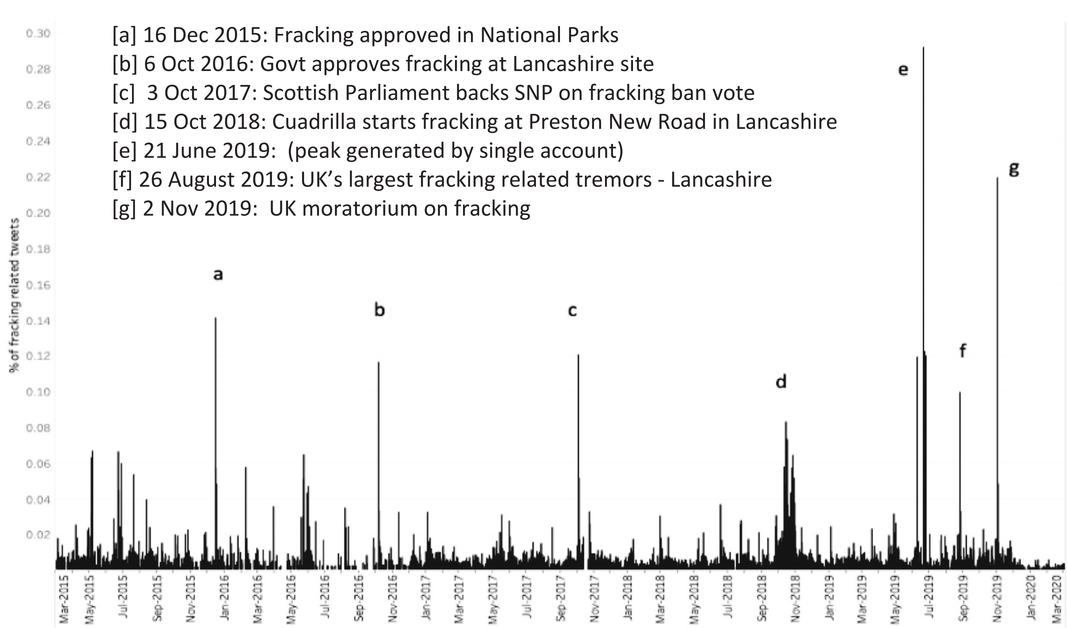


Fig. 2. Tweet Activity mentioning Fracking as a Proportion of all geotweet content in UK from 2015 to 2020.

Table 1
Key Events related to fracking in Great Britain.

Date	Event	Source
28 Jan 2015	Wales and Scotland moratorium on fracking	https://www.gov.scot/policies/oil-and-gas/unconventional-oil-and-gas/
16 Dec 2015	Announcement of the 14th Onshore Oil and Gas Licensing Round; Fracking approved in National Parks	https://www.theguardian.com/environment/2015/dec/16/fracking-under-national-parks-approved-by-mps-amid-acrimony https://www.gov.uk/government/news/new-onshore-oil-and-gas-licences-offered
June 2015	Lancashire council refused Cuadrilla planning permission to frack at Preston New Road, Little Plumpton, Roseacre Wood	https://www.crowdfunder.co.uk/the-fracking-threat-to-lancashire-is-growing-daily
6 Oct 2016	Govt. approves fracking at Preston New Road	https://www.crowdfunder.co.uk/the-fracking-threat-to-lancashire-is-growing-daily
Mid-2017	Cuadrilla construction of PNR site begins; drilling begins mid 2017	https://cuadrillaresources.uk/our-sites/preston-new-road/
Oct 2017	Scottish fracking 'ban' vote in parliament after public consultation	https://www.theguardian.com/uk-news/2017/oct/03/scottish-government-bans-fracking-scotland-paul-wheelhouse
4 Jan 2018	Aurora leafleted residents of Great Altcar about their scoping request to Lancashire County Council for Altcar Moss site	https://www.crowdfunder.co.uk/the-fracking-threat-to-lancashire-is-growing-daily
15 Oct 2018	Cuadrilla starts fracking at Preston New Road, Lancashire	https://cuadrillaresources.uk/our-sites/preston-new-road/
March 2019	Cuadrilla move specialist equipment on site to prepare 2 wells for further hydraulic fracturing	https://cuadrillaresources.uk/our-sites/preston-new-road/
26 Aug 2019	GB's largest fracking related tremors, in Lancashire	https://www.bbc.co.uk/news/uk-england-lancashire-49471321
2 Nov 2019	GB Moratorium on fracking	https://www.gov.uk/government/news/government-ends-support-for-fracking

location, including proximity to development sites, and how they evolve over time.

To track the evolving shale gas story through space and time in Great Britain this research uses Twitter data from the 5 year period between 2015 and 2020, with geospatial location attributes showing the Tweet origin to be within Great Britain. The research demonstrates how χ-squared expectation surfaces can be applied to Twitter data to identify areas with higher and lower levels of online related activity for a specified topic than would be expected. Gaining a better understanding of public engagement with policy-relevant topics, such as shale gas extraction, will help politicians and governments to respond to the multiplicity of issues at regional and national scales. The methods developed in this research could be applied to other environmental topics (e.g. air pollution, nuclear power, carbon capture and storage) to visualise and quantify the spatial and temporal patterns of public engagement.

This paper first summarises the background to hydraulic fracturing in Great Britain, before highlighting key research in social media analysis. This is followed by details of the methods used to process the tweet messages to detect any spatial patterns, before presenting the results and conclusions.

2. Background

Hydrocarbon based energy sources, including oil, gas, and coal, have driven the developed world's machines for decades. As traditional reserves are depleted and technology advances alternative resources can become more economically viable, such as shale gas. Shale gas is considered an 'unconventional hydrocarbon', which involves the hydraulic fracturing of rock, popularly known as 'fracking'. The process involves drilling horizontal and vertical wells into which chemicals and sand are pumped to force out the trapped natural gas from the porous shale (Speight, 2013).

2.1. International reception to fracking

Internationally fracking has had a mixed reception with differing 'impact geographies' (Haggerty, Kroepsch, Walsh, Smith, & Bowen, 2018) which result from the perceived environmental risks versus benefits, experience, regional context, and political factors. For example in Spain the environmental threat received most attention from local politicians and local newspapers, leading to negative attitudes in the public (Mercado, Alvarez, & Herranz, 2014). In the USA the proximity to active sites has been linked to changes in attitude with a 'goldilocks zone' in which survey respondents near active sites (e.g. < 115 km) have shown a more positive support for fracking than those further away (e.g. 115-305 km) (Zanocco, Boudet, Clarke, & Howe, 2019). The reasons for the YIMBY ('Yes, In My Back Yard') responses are attributed to increased business and employment opportunities resulting from the operations, but may be a result of the survey timings after those opposed to fracking have already left the region (Zanocco, Boudet, Clarke, Stedman, & Evensen, 2020).

2.2. Hydraulic fracturing in Great Britain

Success with the technique overseas, notably in the USA, drove the UK government to offer companies the opportunity to purchase inland Petroleum Exploration and Development Licences (PEDL) to explore the shale gas potential for zoned regions of the UK. It was unclear at the time of offering licences if these regions would be economically viable, and companies such as Ineos, Cuadrilla, and Third Energy were effectively buying permission to test the resource potential in the UK. To date only exploration wells have been drilled at a site in Preston New Road by Cuadrilla (Fig. 1), and there has not been any commercial production of shale gas in the UK.

The British public have actively debated the 'fracking' topic with opposition often citing the environmental damage and earth tremors (i.e. induced seismic activity) caused by the process as reasons to 'ban' it, while others consider it will offer economic benefits (e.g. jobs, energy security). There is still considerable public ambivalence about shale gas, but the potential risks are better known than the benefits (Whitmarsh et al., 2015), and it appears from a review of 10 years of public surveys that level of public support is decreasing (Ryder, Devine-Wright, & Evensen, 2020). As a result of public pressure the UK government has introduced moratoriums (i.e. temporary bans) at various times.

Currently all countries of the UK have an effective moratorium on shale gas extraction, with the devolved governments of Scotland, Wales, and England having each taken decisions at different times. Scotland carried out a public consultation in 2017 which resulted in over 60 thousand responses, with an overwhelming 86% calling for a permanent ban on fracking in Scotland (Government, 2017). The English 2019 moratorium had public support but the timing may have been linked to a political agenda, that of an approaching general election (Devine-Wright et al., 2021).

Discussions on shale gas extraction have been complex and involved many different political, industry, environmental and activist groups. The following table summarises the major events relating to shale gas extract in Great Britain since 2015.

a)

b)

c)

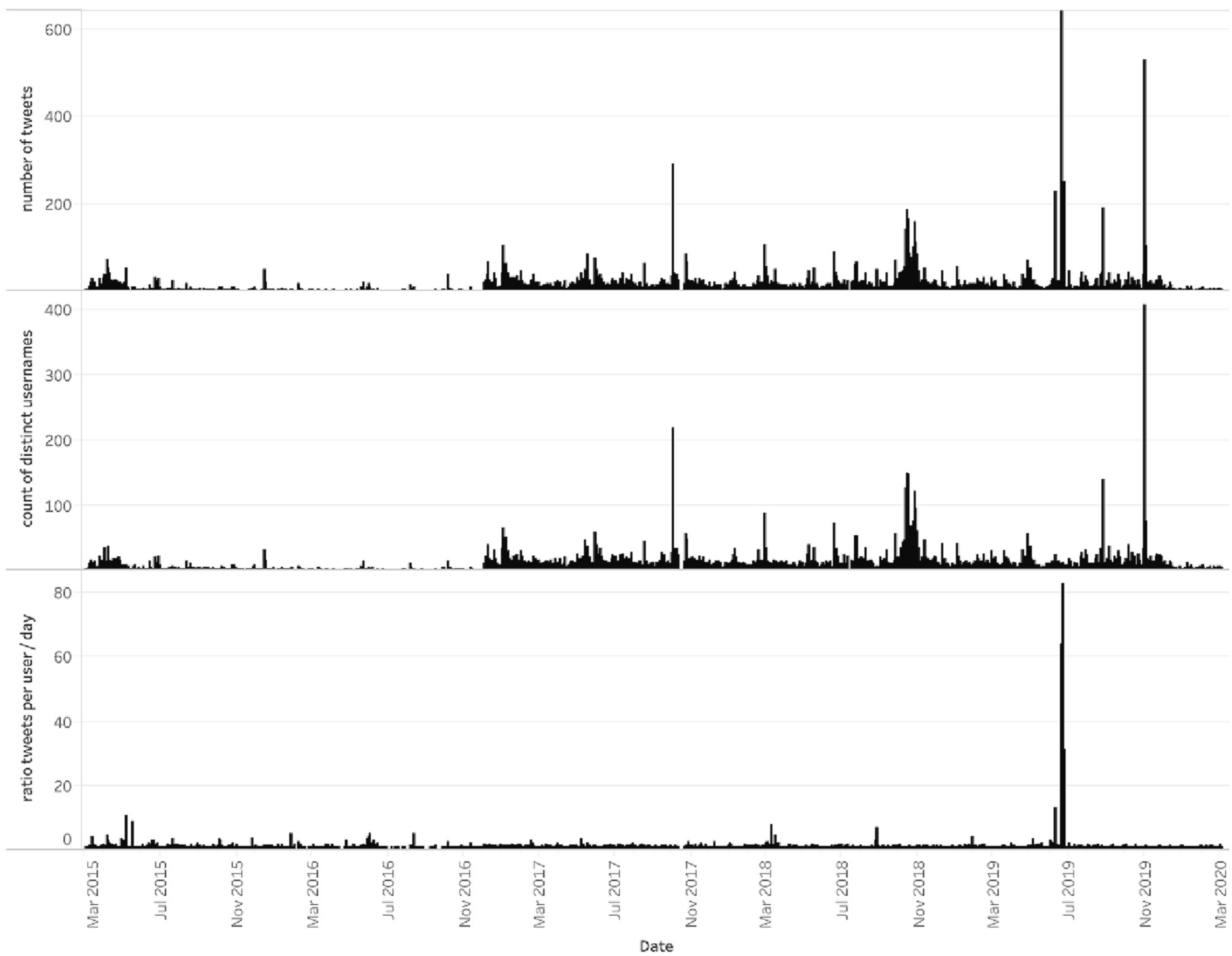


Fig. 3. (a) Tweets per day (b) Distinct Users Tweeting per day (c) Ratio of Tweets to Users per day.

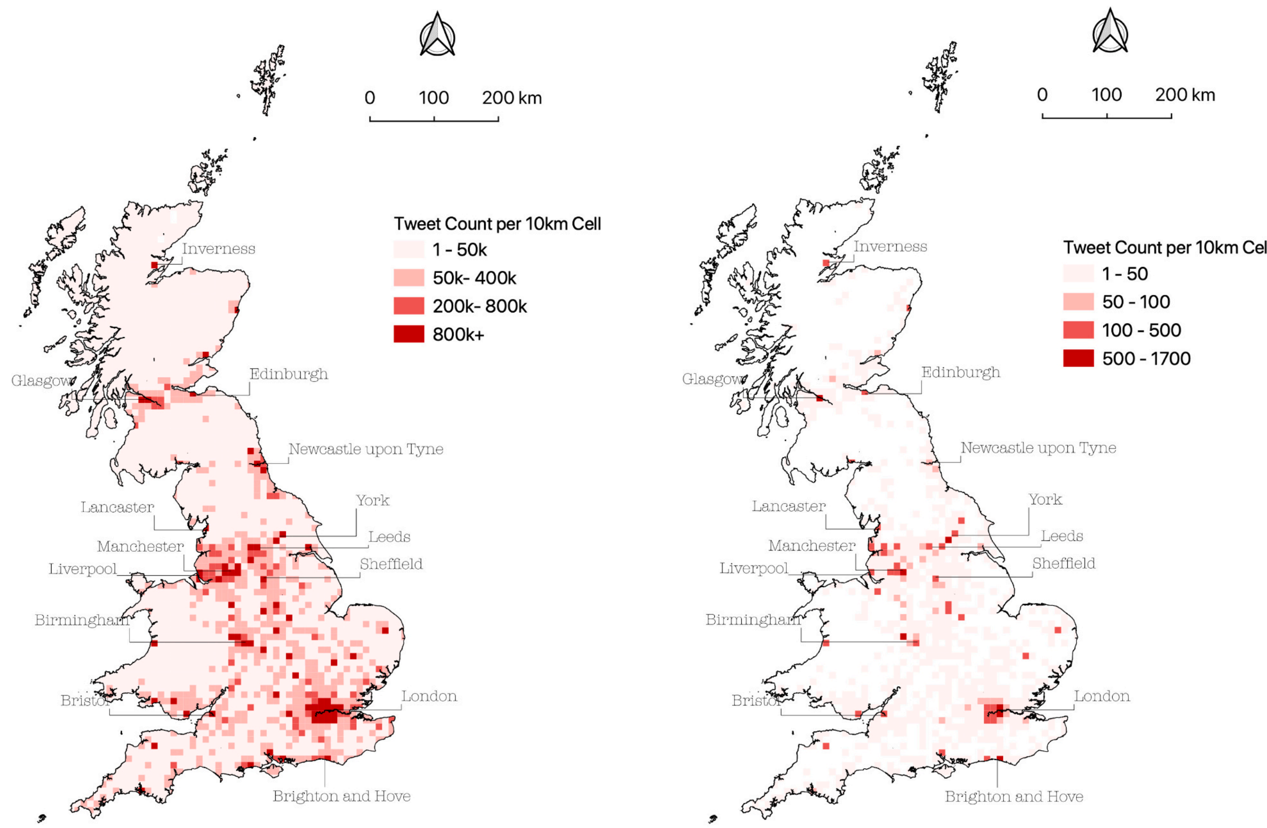
Consultations of the public attitude to shale gas tend to be at a coarse spatial scale and temporal resolution. These are based on snapshot surveys, and lack the granularity to monitor local activity relating to shale gas opposition or support. Another approach is to analyse Twitter data as it gives everyone an opportunity to voice opinion world-wide, and can offer fine grained insight into public activity levels related to a topic, including retrospectively pre-event.

2.3. Social media

Social media has seen one of the fastest adoptions of any online platform, with an expected audience of around 6 billion users by 2027 (Dixon, 2022). Twitter, a micro-blogging platform (Suh, Hong, Pirolli, & Chi, 2010), is one of the most dominant social networking platforms with over 330 million active users each month (Clement, 2019). Message length was limited to 140 characters at launch in 2006 but extended to 280 characters in Nov 2017. Accompanying each Tweet message is a timestamp, username, screenname, platform used to send the message (e.g. Twitter for iPhone, Twitter for Android, Foursquare) and occasionally (<1%) geographic location coordinates.

The location metadata option was introduced in 2009 with the intention of allowing users to filter the stream for local social media conversations, its inclusion being dependent on the user settings (Bastos, Mercea, & Baronchelli, 2018). The uptake of this feature was fairly low, but once a user had enabled location tagging their future tweets automatically included the phone's location details (i.e. GNSS coordinates). With increased awareness of privacy issues Twitter changed its location policies in 2019 (Benton, 2019), making it harder to include precise location metadata although coarser 'place' level of location tagging is still available (e.g. a point of interest, region, city), which is defined in the metadata as the corner coordinates of a bounding box. Although only a low percentage of tweets have location metadata it equates to many hundreds of thousands of geolocated messages per day in GB alone, given the high volumes of messages sent on the platform.

Social media has been extensively studied for a wide variety of topics, including trend detection (Mathioudakis & Koudas, 2010) and disease tracking (Kullar, Goff, Gauthier, & Smith, 2020), for monitoring the organisation of protest movements such as the Arab Springs revolutions in 2011 (Meraz & Papacharissi, 2013) and anti-fracking groups (Hopke, 2015). One of the reasons for its wide uptake in the academic world is Twitter's Application Programming Interface (API) which allows the automated retrieval of Tweet messages with their metadata. The API has developed over time, and currently academics are able to apply to use the APIv2 with full access to the historic archive, which dates back to the first tweet in 2006. There is also access to a live data stream which has been used for event detection, such as discovering live music concerts or potential news items (Atefeh & Khreich, 2015; Saeed, Abbasi, & Razzak, 2020; Weng & Lee, 2011). The date and timestamp metadata has allowed researchers to identify daily temporal patterns of activity, from how the scientific community react to new journal publications (Shuai, Pepe, & Bollen, 2012) to the identification of bots (Chavoshi, Hamooni, & Mueen, 2017). Through the Twitter APIv2 it is possible post-event (e.g. start of 'fracking') to collect pre-event tweets, which is important as traditional longitudinal surveys will not usually



a) b)

Fig. 4. Twitter activity from 2015 to 2020 in Great Britain (a) all geolocated tweets (b) counts related to 'fracking'.

Input

F=3 T=10	F=1 T=2
F=0 T=2	F=8 T=11

$$\chi = \frac{(obs - \exp)}{\sqrt{exp}}$$

$$\text{obs} = F\left(\frac{\sum T}{\sum F}\right)$$

$$\exp = T$$

Output

-1.19	0.06
-1.41	1.71

Fig. 5. Worked example of chi-squared surface expectation calculation (F = fracking related tweets in a cell; T = total tweets in a cell).

begin until after notable events by which time opponents may have already moved away, resulting in biased findings (Zanocco et al., 2020). Twitter also supports follower and following relationships, which can be used to derive social networks (Borge Bravo & Esteve Del Valle, 2017), although not explored in this research.

2.4. Clustering twitter data

Point data, such as geotweets, can be analysed for spatial patterns using clustering methods such as K-Means, DBScan, and Kernel Density Estimation (KDE) to reveal zones of activity. For example DBScan was used with Twitter data to detect trending events (Capdevila, Pericacho, Torres, & Cerquides, 2016), K-Means to gain a better insight into the meaning behind hashtags (Muntean, Morar, & Moldovan, 2012), and KDE to predict crime (Gerber, 2014).

K-Means is one of the most commonly used clustering algorithms, for which the desired number of classes to be found is specified (Yadav & Sharma, 2013). The algorithm is non-deterministic, meaning that the outputs can vary each time it's run for the same input data. The algorithm is not so well suited to noisy data as it attempts to classify each point, and does not allow the user to set the maximum Euclidean distance between points beyond which they are not considered from the same group. Therefore for spatial data analysis the Density-Based Spatial Clustering of Applications with Noise (DBScan) is often used. This is an unsupervised clustering technique requiring two parameters, a threshold distance (i.e. Euclidean distance) and minimum number of points (Ester, Kriegel, Sander, & Xu, 1996). The method is robust to outliers, and does not require the number of clusters to be specified in advance.

Clustering reveals 'hotspots' based on raw data which reflects the underlying population, in this case the distribution of Twitter users. Another approach to summarise point data is to create a lattice (e.g. 10 km by 10 km grid of cells) and count the number of events (e.g. geotweets) within each cell. This reveals the user population distribution but an additional step can be added to normalise the results for any subset of tweets related to a particular topic, such as those related to

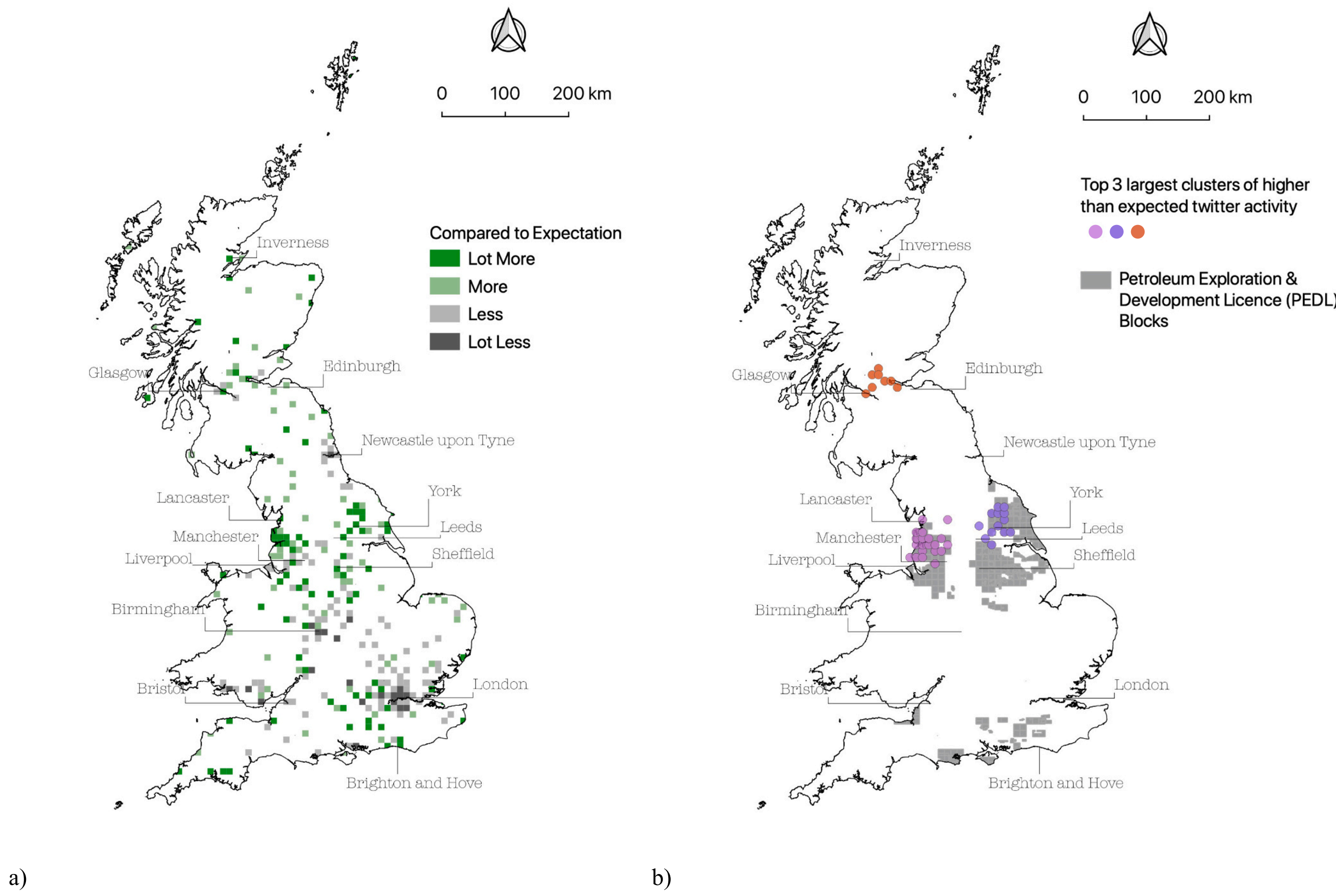


Fig. 6. χ-squared surface expectation for fracking related tweet activity - showing regions with higher and lower than expected numbers of tweets.

'fracking'. A similar approach was taken in research that used Flickr data to determine 'tranquil' regions, where a χ-squared expectation surface was calculated to visualise the regions with higher numbers of tagged images (Wartmann & Mackaness, 2020). This method takes into account the total number of tweets in the dataset as well as the ratio of tweets in each grid cell which relate to the topic. This method can be used with geo-tagged tweets to gain a better understanding of regions with higher than expected activity (i.e. online 'fracking' related activity) taking into account the background user population distribution.

2.5. Research questions

The objectives of this research are to demonstrate if geotweets are able to capture spatial and temporal trends in public engagement with 'fracking'. Two theories to be tested are that online activity is predicted by location according to theories of place identity (Hauge, 2007), in that people close to drilling sites will be more engaged online. Secondly that tweet activity will increase immediately following a relevant political or media event, as the public perceive and interpret associated risks according to the social amplification of risk theory (Kasperson et al., 1988).

3. Data collection and storage

Twitter supports a number of Application Programming Interfaces (APIs) to request subsets of data, about user accounts, historic tweets, or access to the live data stream as Tweets are sent (Streaming API). For this research tweets from the period from 3 March 2015 until 3 March 2020 (5 years) totalling 317 million geotweets were analysed for GB, of which 25,196 related to fracking based on the definition that the message contained 'frack'', 'shale gas', or 'hydraulic frac'. This filter also includes messages that contain partial matches, such as '**frack**ing', '**hydraulic frac**turing', '**hydraulic frac**tured' and so on.

While geotweets make up only a small fraction of all tweets there were still 200,000 tweets per day on average within the study region covering a wide range of subjects. Of these around 40 thousand per day were GNSS location (i.e. high spatial resolution) in mid-2015, dropping to 13 thousand per day in 2019, and totalling 61 million (19%) of the dataset. The additional geotweets are located to a coarser 'place' level which corresponds to a city, famous point-of-interest, or an area.

3.1. Database tuning

To perform the spatial and temporal analysis the tweet dataset was loaded into a PostgreSQL database, configured to maximise performance through enabling parallel workers, maximising memory usage, and moving the WAL (Write Ahead Log) to a 1 TB NVME (non-volatile memory express) drive with a write performance of 3500 MB/s. The data tables were distributed across three tablespaces each on a separate physical disk, totalling just over 1 TB of storage space including the indexes. As well as spatial and date indexes, a gin index was used with trigrams to support matching of partial Tweet messages (as below), which greatly improved search performance.

```
CREATE INDEX idx_msg ON tweets USING GIN(msg gin_trgm_ops);
```

Although generating the index takes a long time on 317 million records the benefits outweigh the costs given the dataset was being analysed as read only (i.e. SELECT SQL statements) and not updated.

This gin index is used by PostgreSQL when using ilike case

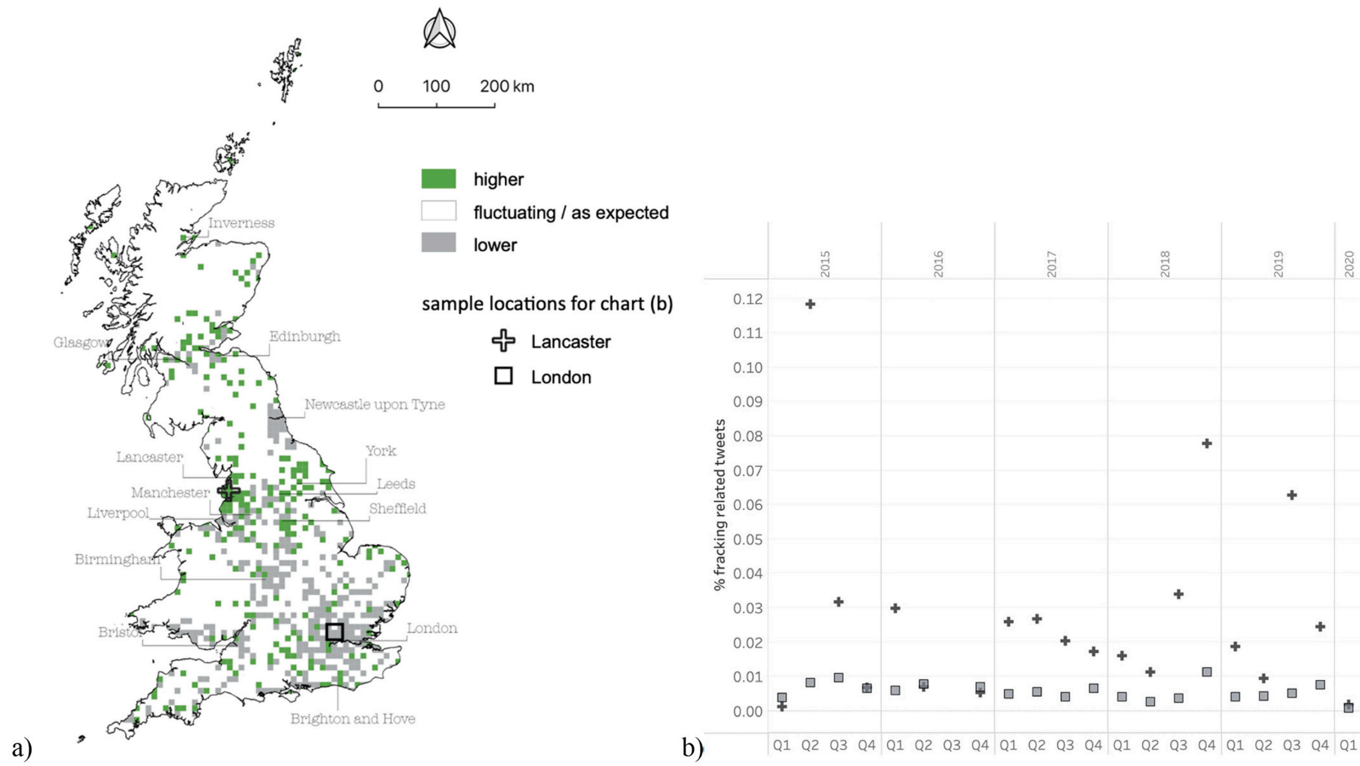


Fig. 7. Consistency of fracking related Twitter activity, showing regions that are consistently higher/lower than expected, and those which fluctuate or are as expected in years 2015–2020.

insensitive wildcard searches, as shown below, to significantly increase search performance.

```
SELECT * FROM tweets WHERE msg ILIKE '%frack%' OR msg ILIKE
'%shale gas%' or msg ILIKE '%hydraulic fract%';
```

4. Analysis and results

This section covers the temporal, spatial and textual analysis of the tweets dataset, highlighting significant related events in the GB shale gas timeline. The likelihood of automated bulk messaging (i.e. bots) is also addressed through comparing the number of daily messages against the number of daily distinct users, based on the Twitter account used to send each tweet.

4.1. Temporal

Twitter data reveals natural cycles from people's daily and weekly routines, but by normalising the total number of fracking related tweets by the total number of tweets collected each day it is possible to identify periods of greater online activity (Fig. 2). There are a number of obvious peaks which tie in with significant newsworthy events (labelled items a-g), also identified in Table 1. Notably the peaks tend to be very short lasting, with the exception of the sustained higher activity period [d] when Cuadrilla started fracking at Preston New Road in Lancashire.

4.1.1. Distinct users

Twitter results can be impacted by very active users or 'software bots' through automated bulk sending of messages. Natural language processing has been used in an attempt to find patterns of language use and identify bots (Allen, 2003; Nadkarni, Ohno-Machado, & Chapman, 2011), but another approach is to group the tweets by date and then calculate the number of distinct users (based on Twitter account names) sending the messages.

Analysis of the geotweets shows that there were 3.5 million distinct users in the dataset of which 9.3 k distinct users mentioned 'fracking' related content. Those 9.3 k users sent an average of 2.6 tweets per account related to fracking, with the 12 most active accounts sending over 100 tweets each. The daily ratio of messages to distinct users gives a clear indication of when a few users, possibly bots, have swamped the channel with content. Fig. 3 shows a chart of these results over the 5 year period, at the top the number of tweets related to 'fracking' each day, in the middle the number of unique usernames sending related tweets per day, and at the bottom the ratio of tweets per user per day.

Around July 2019 there is an obvious change in the ratio of tweets per user per day (see Fig. 3c), which on investigation revealed a single account had been used to send many tweets. This account was identified and removed from further analysis.

4.2. Spatial

Geotweets report location as either a point coordinate sourced from the phone's GNSS (e.g. GPS), or a place tag (e.g. York) defined as a bounding box. Analysis of the place tag bounding boxes revealed that 68% were under 100km^2 (e.g. 10 km $\times$ 10 km), and 84% under 400km^2. For analysis purposes the centroid coordinate was used to represent the place regions, so that all calculations are based on point data.

The spatial distribution of all 317 million geotweets point locations were aggregated into 10 km by 10 km cells, revealing the densely populated urban centres as shown in Fig. 4(a). However only a small percentage of these relate to 'fracking', with the distribution shown in Fig. 4(b).

To understand the population's interest and online activity in the topic it's necessary to normalise the 'fracking' related data as a proportion of all Twitter activity to take account of the variations in the Twitter user population density across GB.

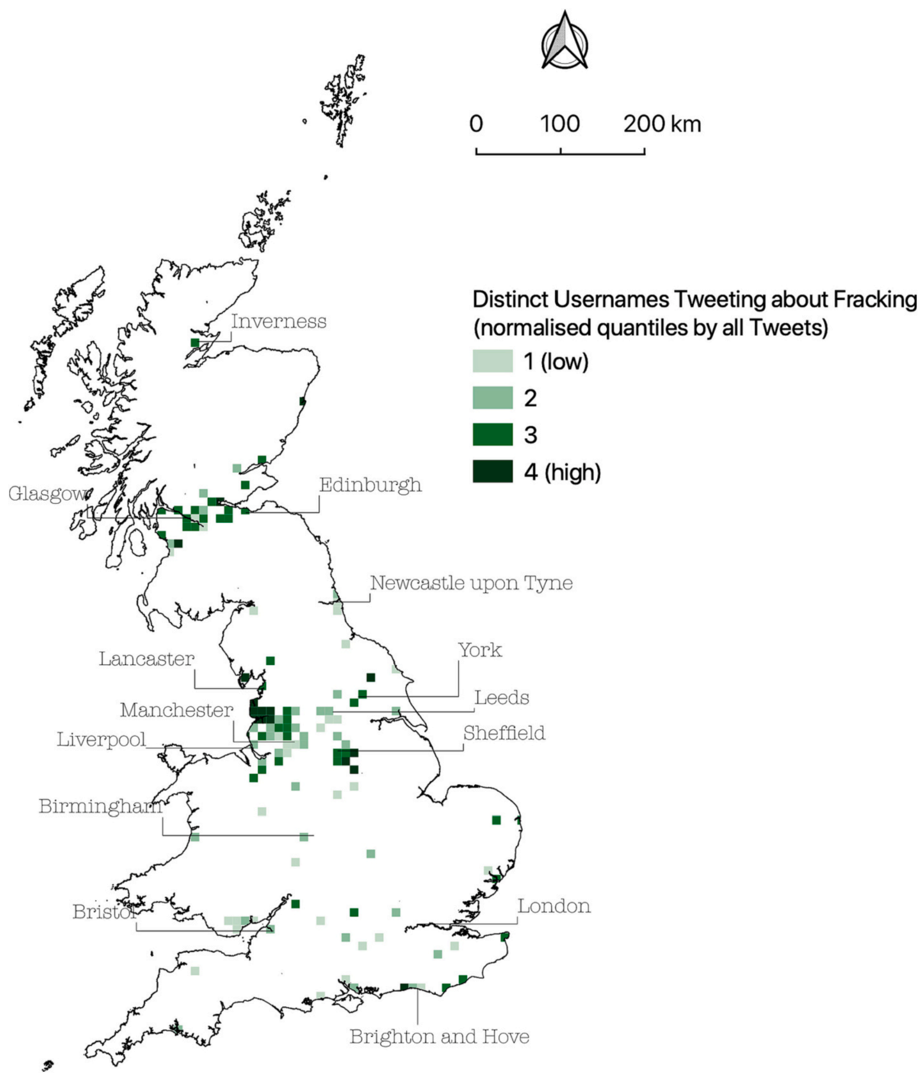


Fig. 8. Number of distinct users per grid cell over 5 years mentioning fracking keywords normalised by unique usernames from all tweets in the cell (shown as quartiles).

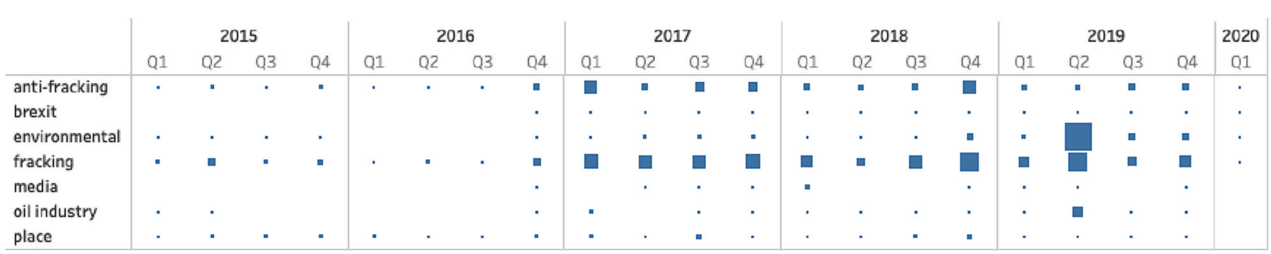


Fig. 9. Top hashtags used over the study period, grouped by topics.

4.3. *Chi-squared output*

Normalisation is required to take account of the variability of the twitter user density across GB, otherwise 'hot spots' will most likely be a result of highly populated urban regions. One approach is to take account of the Twitter population (i.e. the 317 M tweet locations) by calculating a χ-squared expectation surface. This approach was used by Wartmann and Mackaness (2020) to find regions of tranquillity based on Flickr data, as it takes into account the distribution of expected location (exp) in each grid cell and the observed number (obs) of event related records in each cell. In our case the expected is the total number of geotweets in a grid cell, and the observed number is calculated as the total number of all tweets divided by total number of fracking related tweets, multiplied by the number of fracking related tweets in the cell. A worked example for 4 grid cells is shown in Fig. 5.

For this research cells with positive values have greater activity related to 'fracking' than would be expected, while negative results indicate less activity than expected, and values around 0 showing a level of activity proportionate to expectations. Fig. 6(a) maps the results, showing greater than expected interest in fracking around Lancaster (Lancashire), and lower than expected in Birmingham, Newcastle and London.

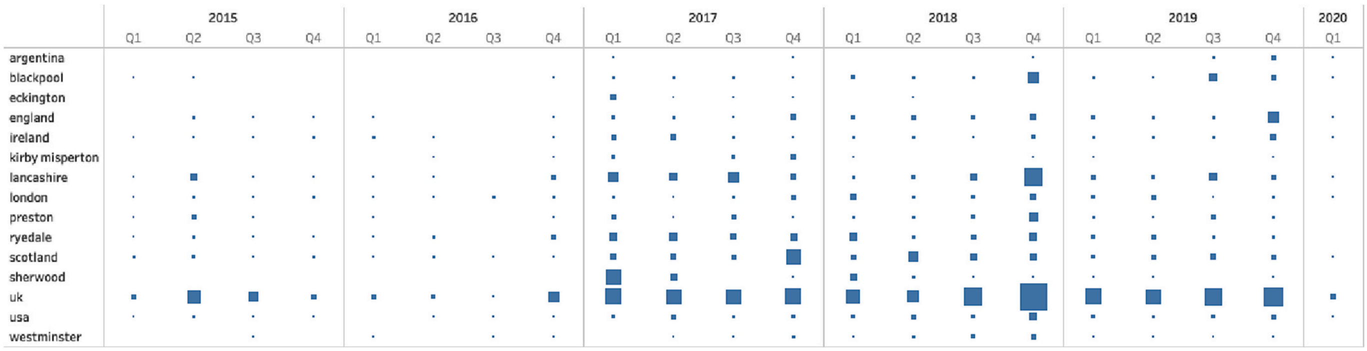


Fig. 10. Most common locations mentioned in 'fracking' related tweets from 2015 to 2020.

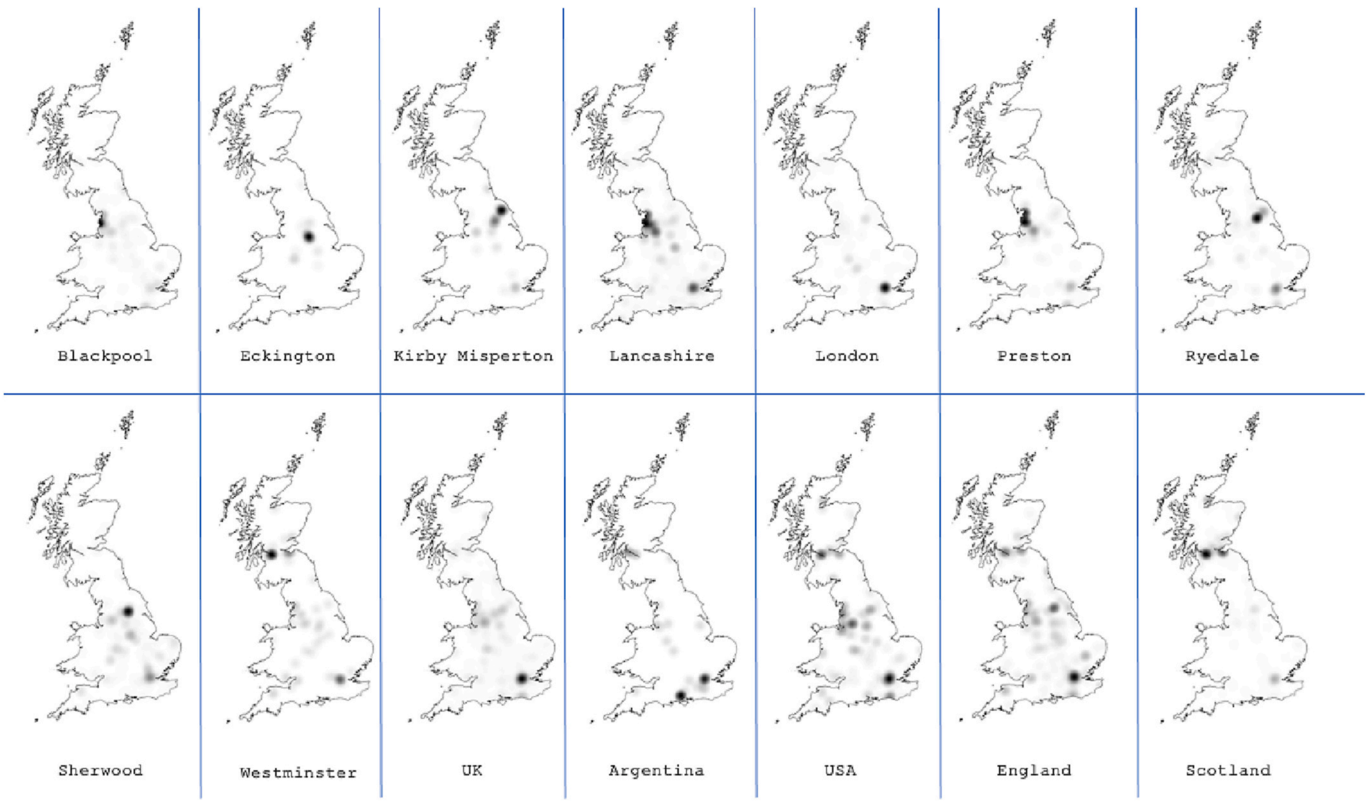


Fig. 11. Mapping the send location of 'fracking' geotweets that included a place name.

To identify the largest contiguous regions of high activity DB-SCAN was used with an epsilon of 20 km, ensuring it spanned across 10 km grid cells. The 3 largest clusters of above expected online activity are overlaid on the PEDL areas in Fig. 6(b). Notably the top 2 clusters are situated in the largest PEDL blocks, indicating a greater interest in the topic from the public in those regions. The third cluster is based in Scotland, which is not near a PEDL block, but reflects the ongoing anti-fracking debate and subsequent moratorium (see Table 1). Based on these findings it appears that the local community near the active shale gas extraction site, and the neighbouring regions, are the most active online regarding this topic. An alternative theory could be that protestors move into a region and send tweets from near the site, and this is examined in more detail in Section 4.4.

An additional consideration is that the location metadata may refer to the user's location reported by the device's coordinate at the time of sending the tweet, or the user may set a place relevant to the message content. To explore this further another χ-squared expectation surface was calculated based on the subset of geotweets which the metadata reported as 'precise', based on the user's phone's GNSS coordinates which is not easy to spoof. A correlation of 0.699 (p-value <0.0001) was found between this expectation surface and that using all geotweets, with the same pattern of higher activity around Lancashire, and a lack of activity around London. The notable differences relate to a greater coverage in the full dataset, as GNSS located tweets corresponded to just 19% of the dataset. The correlation and high significance demonstrate that the full dataset exhibits the same χ-squared expectation surface trends across the country, and gives re-assurance that the results shown in Fig. 6(a, b) relate to the location of the user sending the tweet rather than tagged news event locations.

4.4. Activity levels over time

Fig. 6 shows a summary of spatial activity across GB but does not give any indication if these have been brief or sustained online localised campaigns. Therefore a new method was developed to visualise the regions of sustained higher/lower online interest than would be expected,

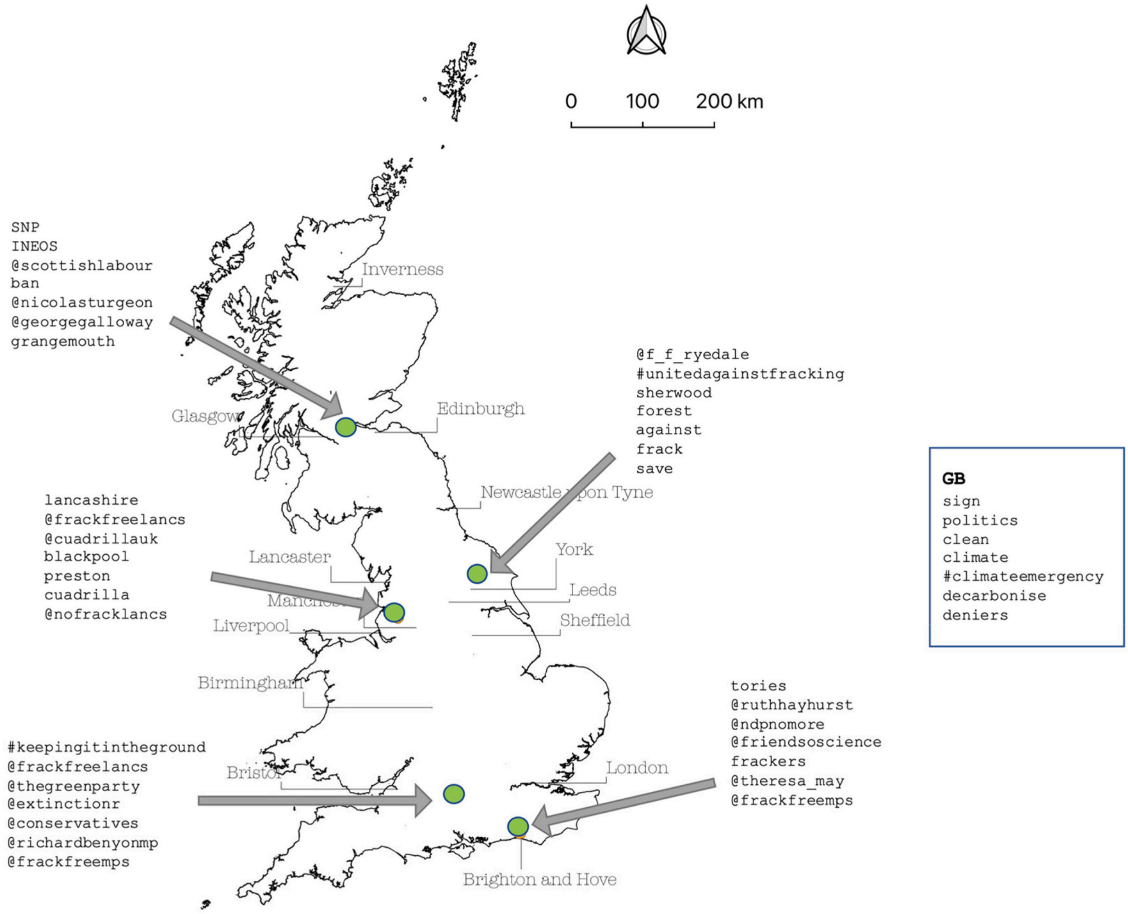


Fig. 12. The most common regional and GB wide terms related to fracking on Twitter.

by time-slicing the data using a moving 12 month time-window, incremented in 1 month steps. A new χ-squared expectation surface was created for each time-step, and the values for each cell (10 km by 10 km) were totalled across all of the time-slices. Resulting large positive numbers indicate a sustained higher than expected online activity for the cell, while large negative numbers reflect consistently lower than expected activity. For cells with a sum of near zero there had been about the expected level of activity online, or a great deal of fluctuation in interest.

Fig. 7a shows the mapped output, revealing regions which have had sustained interest in fracking related topics within the dataset. To illustrate the temporal variations two grid cells with extreme results were selected (Fig. 7b), Lancashire (marked on the map as +) and London (marked with □). Fracking was mentioned in tweets at these locations on 400 separate days over the 5 year study period, and for 99% of that time (396 days) the Lancashire location had a greater percentage than the London location, concurring with the map output. The chart shows that Lancashire sustained higher ratios of fracking related tweets by a large margin on many occasions.

While each peak of activity may be brief, as shown in Fig. 2, the overall sustained levels of online participation in 'fracking' related discussion are higher in the top 3 regions illustrated in Fig. 6(b). Notably Lancashire (i.e. Preston New Road's once active drilling site) shows a sustained higher than expected interest, while London and Newcastle-upon-Tyne exhibit a sustained lower than expected interest on social media. This would support the ideas of place theory with those located nearer the site having a greater attachment and sense of place, while the other more distant urban centres are comparatively less interested even though as a potential energy source it could have an impact on their future living costs and energy security.

4.5. Spatial pattern of distinct users

The previous maps show the spatial activity on Twitter related to 'fracking' based on numbers of tweets, but another measure of activity is to calculate the number of distinct user accounts discussing 'fracking' topics at the 10 km cell resolution as a ratio of the total number of distinct users within that cell. The spatial distribution is similar to that shown in Fig. 6, revealing a higher number of engaged users clustered in Scotland and around the only previously active drill site in Lancashire, while notably the city of London is proportionally less engaged. The resulting map, Fig. 8, shows the regions with the highest proportion of users mentioning 'fracking' topics over the 5 year period.

In addition to the temporal and spatial aspects of geotweets is the content of the tweet messages including hashtags, frequently used terms, and mentions of other locations.

4.6. Text analysis of tweets

Examination of the account names taking part in the debate reveal there to include environmental activists, politicians, media, and anti-fracking groups. The most basic text analysis for tweets is to calculate hashtag frequencies over time. Fig. 9 shows the results of grouping hashtags by common topic areas for the 'fracking' tweet subset with at least 50 mentions. The majority of social media messages are 'anti-fracking' related, with mentions of Brexit being connected to the fracking debate. There was an increase in the use of the fracking and environmental related hashtags particularly '#climateemergency', '#singleuseplastic' and '#waronplastic' in 2019 Q2. Hashtags which relate to places (e.g. Sherwood) occur fairly consistently, and can be examined in further detail using Natural Language Processing (NLP).

Named Entity Recognition (NER) is a part NLP which identifies place names using AI based on sentence structure, rather than comparing words against a gazetteer. The Python FLAIR library[1] was used to process the tweets, due to its ease of use and state-of-the-art NLP performance, to find place names in the message content. NER is not faultless and sometimes place names identified in some messages were missed in other tweets based on the sentence phrasing. To overcome this the tweet messages were processed using NER to identify locations mentioned, and then this list of place names was used with SQL to count all the related tweets, ensuring all occurrences of the place names were located within the dataset.

The results of the most commonly mentioned places from the 'fracking' related tweets, mentioned twenty or more times in a quarter, are shown in Fig. 10. The 'UK' and 'Lancashire' are mentioned fairly consistently throughout the study period, while other places experience peaks of interest such as Scotland in 2017 Q4 when there was a public debate and 'ban'. As well as PEDL sites (e.g. Blackpool, Lancashire, Kirby Misperton, Preston) there are political locations (e.g. Westminster) and country level mentions (e.g. Scotland, England, Ireland). The moratoriums of Scotland (2017Q4) and England (2019 Q4) result in peaks of those place terms, but despite the larger English population the peak is half the size of the Scottish one. The highest variety of place mentions was in 2018 Q4, which co-insides with the start of Cuadrilla's test drilling at Preston New Road.

Mapping the locations of the tweet message that mention these place names in the message, shows that the conversations are mainly local, that is messages mentioning 'Blackpool', 'Eckington', 'Kirby Misperton' and so on are mostly sent by people from those areas (Fig. 11). The exceptions are 'Westminster' which has a hotspot in Glasgow as well as London from UK based political discussions, and 'Argentina' which is dominated by the south coast and London. 'USA' is spread across GB's urban regions as is use of 'England', however inclusion of 'Scotland' is more localised to the Scottish cities. Fig. 11 shows the results using Kernel Density Estimation to create the heatmaps, with each map using its own relative symbology range. As a result it is not possible to draw absolute comparisons between the hotspot values of different maps (e.g. 'Argentina' relative to 'UK'), however such quantity comparisons can be made using Fig. 10.

A final analysis was carried out at five locations across GB to find the top most common terms used in messages, once 'fracking' and stop words were removed. Stop words (e.g. a, the, of) can be removed in PostgreSQL using the ts_vector function, which is part of the full text searching tools.

Fig. 12 shows the results, sorted in order at each location from most to least frequently used terms. The Scottish location references the current political party in power and its leader ('SNP', @nicolasturgeon) and rival politician ('@georgegalloway'), as well as 'grangemouth' a town with an oil refinery run by 'INEOS'. Further south near York are mentions of a local anti-fracking group (f_f_ryedale - the account for Frack Free Ryedale) and the local nature reserve of 'Sherwood Forest'. To the west in 'Lancashire' the energy company 'Cuadrilla' are most mentioned along with the site they used for test drilling ('Preston') near 'Blackpool'. The two other locations in South England mention political parties (@thegreenparty, @conservatives, tories) and politicians (@Theresa_may, @richardbenyonmp), the Extinction Rebellion global action movement (@extinctionr) and other anti-fracking and UK campaign groups (e.g. @frackfreelancs, @frackfreemps), and investigative journalist (@ruthhayhurst). In summary Scotland is mostly concerned about Scottish locations and politics, the north of England mainly mentions locations and companies involved in the fracking, and the south is more politically orientated.

The results of this research reveal those near the large PEDL regions, and especially the actively drilled site (Lancashire), were more active online than would be expected based on the Twitter population density. This is probably as they have a greater ability to identify with the places and perceive any risks as more immediate threats. The content of messages were most commonly associated with anti-fracking and environmental issues, with those nearer the activity mentioning details of locations and companies, while elsewhere comments were more generalised. There were peaks of online activity which link closely to political and environmental events, although with the exception of the start of fracking at the Lancashire site the peaks were short-lived. It was also notable that Scotland had its own cluster of highly active users, with content mentioning companies, Scottish locations, and local political figures.

5. Conclusions and future work

This research has, for the first time, investigated the ways that fracking social media discourse has evolved over time and across space in Great Britain. The study analysed 317 million tweets from 2015 to 2020 to reveal the public's online engagement with 'fracking' and 'shale gas' related topics. Activity online closely mirrored the key political and newsworthy events such as the devolved moratoriums, tremors from test drilling, and generally these associated activity peaks were short term.

Regional differences were mapped using a χ-squared expectation surface to normalise the output, to take account of the variation in Twitter user population density across GB. Regions of higher than expected online activity relating to 'fracking' coincide with the largest PEDL areas, and only active UK shale gas extraction site (Preston New Road, Lancashire). A Scottish cluster can be attributed to political events, based on examining the tweet messages (e.g. '…Scottish fracking ban…'). Those in London and the south of Great Britain were proportionally less active than would be expected taking into account the user population sizes.

The use of a novel time-sliced χ-squared expectation surfaces revealed the spatial regions where interest levels had remained either higher or lower than expected for extended periods, showing that those near Preston New Road had a prolonged heightened involvement in the topic. Mapping the top terms shows spatial variation from the more specific content at local level near the once active site, compare to more generalised political content elsewhere.

While Twitter is a social media platform used by a subset of the population it does offer some advantages over other interview and longitudinal study approaches in gaining an insight into a population's engagement with environmental topics. It offers an ability to listen in to opinions without any forewarning and possible induced topic bias, and the opportunity to check pre-event comments from after an event has happened. This is contrary to longitudinal studies which often gather opinions post-event at which time those strongly opposed to a movement may have already left the region.

The spatial, temporal, and textual dimensions of this research combine to reveal that the people nearest 'fracking' sites were the most active online for sustained periods, with peaks during the bigger environmental and political events. Not only were the number of tweets greater than expected in those regions, but also the number of distinct users taking part was the highest in the country giving some indication of the connection to 'place' felt by that community. In contrast London has demonstrated a sustained lower level of online activity, and disconnection from the topic. Many of the texts about fracking mentioned other environmental issues such as 'war on plastic' and 'climate emergency', as well as the companies involved in drilling. The most commonly used terms had local connections, and places mentioned in the tweets showed a common theme of the PEDL regions and political centres.

The combination of analytical approaches in this paper offers a meaningful way of understanding public engagement with a policy-relevant topic that would be expected to be geospatial influenced. Such insights are useful for how the national and devolved governments,

[1] https://github.com/flairNLP/flair

and individual MPs focussed on their constituencies to understand the multiplicity of issues and respond to issues.

Future work could apply these methods to other contexts (e.g. USA), or combine this social media based approach with other methods such as local case study research, and longitudinal surveys. There is also scope to track anonymised individual actors in the Twitter dataset to plot their journeys spatially and to monitor engagement frequencies and attrition rates. Furthermore connections between the various online parties could be examined through network analysis of Twitter account follower relationship, to find out more about the existence and importance of news echo-chambers.

Author contributions statement

PB, AV, and JD conducted the data analysis; All authors contributed to writing the article and interpreting the results and implications of the findings; PB, JD, DE, LW, PDW collaborated on the application for the funding secured for this research.

Acknowledgements

This research has been undertaken as part of the 'The Attitudes to Shale GaS in Space and Time' (ASSIST) project funded through ESRC and NERC (Grant Ref: NE/R017727/1). The authors also thank Twitter Inc. for supporting research through their developer programme.

References

Allen, J. F. (2003). Natural language processing. In A. Ralston, E. D. Reilly, & D. Hemmendinger (Eds.), *Encyclopedia of computer science* (pp. 1218–1222). John Wiley and Sons Ltd.

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence, 31*(1), 132–164.

Bastos, M., Mercea, D., & Baronchelli, A. (2018). The geographic embedding of online echo chambers: Evidence from the Brexit campaign. *PLoS One, 13*(11), 1–16. https://doi.org/10.1371/journal.pone.0206841

Benton, J. (2019). Twitter is removing precise-location tagging on tweets. Retrieved May 1, 2022, from https://www.niemanlab.org/2019/06/twitter-is-turning-off-location-data-on-tweets-a-small-win-for-privacy-but-a-small-loss-for-journalists-and-researchers/.

Borge Bravo, R., & Esteve Del Valle, M. (2017). Opinion leadership in parliamentary twitter networks: A matter of layers of interaction? *Journal of Information Technology & Politics*. https://doi.org/10.1080/19331681.2017.1337602

Boudet, H. (2021). Getting closer. *Nature. Energy, 6*(10), 945–946.

Capdevila, J., Pericacho, G., Torres, J., & Cerquides, J. (2016). Scaling dbscan-like algorithms for event detection systems in twitter. In *International conference on algorithms and architectures for parallel processing* (pp. 356–373). Springer.

Chavoshi, N., Hamooni, H., & Mueen, A. (2017). Temporal patterns in bot activities. In *Proceedings of the 26th international conference on world wide web companion* (pp. 1601–1606).

Clement, J. (2019). *Twitter: Number of active users 2010–2019 | Statista*.

Cotton, M. (2017). Fair fracking? Ethics and environmental justice in United Kingdom shale gas policy and planning. *Local Environment*. https://doi.org/10.1080/13549839.2016.1186613

Devine-Wright, P., Ryder, S., Dickie, J., Evensen, D., Varley, A., Whitmarsh, L., & Bartie, P. (2021). Induced seismicity or political ploy?: Using a novel mix of methods to identify multiple publics and track responses over time to shale gas policy change. *Energy Research and Social Science*. https://doi.org/10.1016/j.erss.2021.102247

Dixon, S. (2022). Number of global social network users 2018–2027. Retrieved June 28, 2022, from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *, 96. kdd* (pp. 226–231).

Gerber, M. S. (2014). Predicting crime using twitter and kernel density estimation. *Decision Support Systems, 61*, 115–125.

Government. (2017). Unconventional oil and gas consultation: analysis of responses. Retrieved September 3, 2021, from https://www.gov.scot/publications/talking-fracking-consultation-unconventional-oil-gas-analysis-responses/pages/2/.

Haggerty, J. H., Kroepsch, A. C., Walsh, K. B., Smith, K. K., & Bowen, D. W. (2018). Geographies of impact and the impacts of geography: Unconventional oil and gas in the American west. *The Extractive Industries and Society, 5*(4), 619–633.

Hauge, Å. L. (2007). Identity and place: A critical comparison of three identity theories. *Architectural Science Review, 50*(1), 44–51.

Hopke, J. E. (2015). Hashtagging politics: Transnational anti-fracking movement twitter practices. *Social Media and Society.*. https://doi.org/10.1177/2056305115605521

Kânoğlu-Özkan, D. G., & Soytaş, U. (2022). The social acceptance of shale gas development: Evidence from Turkey. *Energy, 239*, Article 122150. https://doi.org/10.1016/J.ENERGY.2021.122150

Kasperson, R. E., Renn, O., Slovic, P., Brown, H. S., Emel, J., Goble, R., … Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk Analysis, 8*(2), 177–187.

Kullar, R., Goff, D. A., Gauthier, T. P., & Smith, T. C. (2020). To tweet or not to tweet—A review of the viral power of twitter for infectious diseases. *Current Infectious Disease Reports, 22*(6). https://doi.org/10.1007/s11908-020-00723-0

Mathioudakis, M., & Koudas, N. (2010). Twittermonitor: Trend detection over the twitter stream. In *SIGMOD '10 proceedings of the 2010 ACM SIGMOD international conference on management of data* (pp. 1155–1158). ACM. https://doi.org/10.1145/1807167.1807306.

Meraz, S., & Papacharissi, Z. (2013). Networked gatekeeping and networked framing on #Egypt. *International Journal of Press/Politics.*. https://doi.org/10.1177/1940161212474472

Mercado, T., Alvarez, A., & Herranz, J. M. (2014). The fracking debate in the media: The role of citizen platforms as sources of information. *ESSACHESS–journal for. Communication Studies, 7*(1 (13)), 45–62.

Muntean, C. I., Morar, G. A., & Moldovan, D. (2012). Exploring the meaning behind twitter hashtags through clustering. In *International conference on business information systems* (pp. 231–242). Springer.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association, 18*(5), 544–551.

Ryder, S. S., Devine-Wright, P., & Evensen, D. (2020). *Briefing: Public perceptions of shale gas exploration in the UK. A summary of research (2012−2020)*. University of Exeter. February.

Saeed, Z., Abbasi, R. A., & Razzak, I. (2020). Evesense: What can you sense from twitter? *Advances in Information Retrieval, 12036*, 491.

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PLoS One, 7*(11), Article e47523.

Speight, J. G. (2013). *Shale Gas Production Processes. Shale gas production processes*. https://doi.org/10.1016/C2012-0-00596-0

Stephan, H. R. (2017). The discursive politics of unconventional gas in Scotland: Drifting towards precaution? *Energy Research and Social Science*. https://doi.org/10.1016/j.erss.2016.09.006

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. In *Proceedings - SocialCom 2010: 2nd IEEE international conference on social computing, PASSAT 2010: 2nd IEEE international conference on privacy, security, risk and trust* (pp. 177–184). IEEE. https://doi.org/10.1109/SocialCom.2010.33.

Wartmann, F. M., & Mackaness, W. A. (2020). *Describing and mapping where people experience tranquillity. An exploration based on interviews and Flickr photographs.* Landscape Research. https://doi.org/10.1080/01426397.2020.1749250

Watson, J., Ketsopoulou, I., Dodds, P., Chaudry, M., Tindemans, S., Woolf, M., & Strbac, G. (2018). The security of UK energy futures. In *UKERC research Report, (March)*.

Watterson, A., & Dinan, W. (2018). Public health and unconventional oil and gas extraction including fracking: Global lessons from a Scottish government review. *International Journal of Environmental Research and Public Health, 15*(4), 675.

Weng, J., & Lee, B.-S. (2011). Event detection in twitter. In *, 5. Proceedings of the International AAAI Conference on Web and Social Media*.

Whitmarsh, L., Nash, N., Upham, P., Lloyd, A., Verdon, J. P., & Kendall, J.-M. (2015). UK public perceptions of shale gas hydraulic fracturing: The role of audience, message and contextual factors on risk perceptions and policy support. *Applied Energy, 160*, 419–430.

Yadav, J., & Sharma, M. (2013). A review of K-mean algorithm. *International Journal of Engineering Trends and Technology (IJETT), 4*(7), 2972–2976.

Zanocco, C., Boudet, H., Clarke, C. E., & Howe, P. D. (2019). Spatial discontinuities in support for hydraulic fracturing: searching for a "Goldilocks Zone". *Society & Natural Resources, 32*(9), 1065–1072.

Zanocco, C., Boudet, H., Clarke, C. E., Stedman, R., & Evensen, D. (2020). NIMBY, YIMBY, or something else? Geographies of public perceptions of shale gas development in the Marcellus shale. *Environmental Research Letters, 15*(7), Article 074039. https://doi.org/10.1088/1748-9326/ab7d01