Office for National Statistics

Exploration of the Global Database of Events, Language and Tone (GDELT), with specific application to disaster reporting

This article summarises an investigation into the potential, within statistics, of using data available in the Global Database of Events, Language and Tone (GDELT) with specific application to disaster reporting.

Contact:
Susan Williams
susan.williams@ons.gov.uk
+44 (0)20 7592 8637

Release date:
9 January 2020

Next release:
To be announced

Table of contents

1 . Other pages in this release

This release is split into an article and two accompanying notes.

This article contains the main findings of the investigation, and the strengths and potential limitations of using the Global Database of Events Language and Tone (GDELT) data within our application of disaster reporting.

The two accompanying notes are:

- an Appendix, which provides technical details for using GDELT data, including an overview of data access options, relevant databases, main variables, and examples of inaccuracies discovered in the data that should be considered when using GDELT
- a Data quality note, which addresses some of the more general quality questions relevant to the use of unofficial data sources in the context of disaster reporting

2 . Introduction

The United Nations' Sustainable Development Goals (SDGs) are an initiative to support global sustainable development. United Nations (UN) countries need to report on progress towards each of the SDGs via a set of defined indicators. The Office for National Statistics (ONS) is responsible for identifying and putting in place mechanisms to source and publish indicator data for the UK.

Many of the UK indicators are readily available from official data but, at the time of this research, there was a data gap around the economic and social impact of disasters. More specifically, quantitative data were needed on the number of deaths, missing persons and directly affected persons, as well as the direct economic loss attributed to disasters.

This detail of information is required by the following SDG indicators:

- SDG Indicator 1.5.1: number of deaths, missing persons and directly affected persons attributed to disasters per 100,000 population.
- SDG Indicator 1.5.2: direct economic loss attributed to disasters in relation to global gross domestic product (GDP).

For an overview of current reporting progress, see the SDG UK data website.

While various maintained databases exist recording the global impacts of disasters, the UK does not usually feature within them as the social and economic impacts of UK-based disasters are deemed too low to be included. However, UK reporting for these SDG indicators has proposed to extend the definition of a disaster to that of a hazard, such as flooding or storms, and these do occur regularly in the UK.

The Global Database of Events, Language and Tone (GDELT) project publishes data on broadcast, print and web news articles. The GDELT data contain information automatically extracted from online news media around the world. Supported by Google Jigsaw, the project uses advanced machine learning techniques and text analysis of these news articles to identify various items such as events, themes or topics, number of mentions, or tone of text.

The main benefits of the data within the production of statistics are its timeliness, geographical reach, level of spatial reporting and detail. The data are also freely accessible and actively maintained.

This research explores the data available in GDELT with specific application to SDG disaster reporting. Its objectives are:

- to learn how to access and handle the GDELT data
- to better understand the variables within GDELT data
- to better understand the quality of GDELT data
- to develop and evaluate approaches to identify “disasters” through GDELT data
- to provide proposals for further research.

3 . Things you need to know

What is GDELT?

The Global Database for Events, Language and Tone (GDELT) project aims to extract as much information as possible from news articles reported around the world. GDELT captures details such as who is mentioned, what is talked about and where things have happened, together with metrics such as how many times an article is published and when.

GDELT also derives information by using the text in articles to categorise them and map their content into themes or topics. The taxonomy used within GDELT is based on the Conflict and Mediation Event Observations (CAMEO) framework, although there are additional categories as CAMEO is not well suited to cover all types of event. These additional categories make it is possible to track impacts of events like industrial accidents, natural disasters and disease epidemics.

Other derived information includes a sentiment score based on an article’s text, as well as the identification of single events by grouping articles that are referring to the same event.

A collection of databases is produced resulting from scanning news outlets across the globe every 15 minutes. The latest version of GDELT, which we used in our research, starts with data collection from February 2015 onwards, although we are aware that a new version is due out early 2020. An earlier version reaches back to 1979.

GDELT databases

The GDELT project has a massive scope, with data organised in different “collections”, which allow different views of partially overlapping aspects of the news landscape. This research focused on GDELT’s Global Knowledge Graph (GKG) that records every news article alongside all the people, organisations, themes, locations and numbers mentioned in the article. Table 1 illustrates the subset of variables used in our research within the GKG database.

Table 1: Example news article entry extracted from the GDELT's Global Knowledge Graph (GKG) database: subset of fields used in our analysis

V2SOURCECOMMONNAME	V2COUNTS	V1THEMES	V1LOCATIONS	V1PERSONS	V1ORGANIZATIONS
www.leighjournal.co.uk/news	AFFECT#60000000#People#United Kingdom; etc	TAX_FNCACT_MAN; KILL; CRISISLEX_T03_DEAD; etc.	4#Caernarfon, Gwynedd, United Kingdom;	adrian holme; etc.	cumbria police; co antrim; etc.
V1.5TONE	V2GCAM	V2.1ALLNAMES	V2.1AMOUNTS	V1.5TONE	
4.60921, 2.00400, 0.0, etc.	wc:969, c1.3:3, c12.1:56, etc.	Storm Desmond, 35;Environment, 47; etc.	130,flood warnings, 144; etc.	4.60921, 2.00400, 0.0, etc.	

Source: GDELT project

The variables in Table 1 are briefly explained as:

V2SOURCECOMMONNAME: the Uniform Resource Locator (URL) for the top-level domain of an article.

V2COUNTS: the list of counts found in the article. Each count entry will have the following form:

- Count Type, which is most commonly "Affect", "Arrest", "Kidnap", "Kill", "Protest", "Seize" or "Wound", as specified in CAMEO.
- Count, which is the actual count being reported; in our example, 60 million.
- Object Type, which identifies what the number relates to; in our example, people.
- Location, which is GDELT's derivation of the place to which the article mainly refers.

V1THEMES: lists all the themes (XLS, 27KB) that GDELT has mapped the article's text into. Again, these themes are mainly based on CAMEO. Our example only shows a subset of the themes for this article.

V1LOCATIONS: lists all locations mentioned in the article, and is prefixed with a location type; in our example "4" indicates the location as a WORLDCITY.

V1PERSONS: lists all person names found in the text.

V1ORGANIZATIONS: lists all company and organisation names found in the text.

V1.5TONE: a set of six core emotional dimensions found in the text of the article, each delimited by a comma. The main three are:

- Tone, which is an average tone of the document as a whole; the score ranges from negative 100 to positive 100
- Positive Score, representing the percentage of words that were found to be positive
- Negative Score, representing the percentage of words that were found to be negative

GCAM: the Global Content Analysis Measures (GCAM) system is a rich source of information which provides the number of words that are found in a wide collection of data dictionaries. In our example, “wc:969” tells us there are 969 words in the article, “c1.3:3” tells us there are three words in the dictionary coded as c1.3, and “c12.1:56” tells us there are 56 words associated to the dictionary coded as c12.1.

V2.1ALLNAMES: lists all proper names referenced in the document along with the character offsets of where they can be found in the article.

V2.1AMOUNTS: lists all precise numeric amounts found in the article, each with the following form:

- Amount, which is the actual number found
- Object, which identifies the object the number refers to
- Offset, which is the character offset of the quoted statement within the article

An overview of the different databases and versions of GDELT, as well as other example entries and a more detailed explanation of variables are included in Section 2 of the Appendix.

GDELT data access

GDELT data can be freely accessed. Access is possible in a number of ways, including:

- a full-text api that can be accessed through a browser address bar
- GoogleBigQuery
- downloading of raw files

Access options for the main databases are described in Section 3 of the Appendix.

Alternative data sources on disasters

For disaster reporting, a number of alternative data sources could also be considered. A short discussion of the main disaster databases can be found in Section 4 of the Appendix.

Analysis scope

The following analyses sometimes rely on crude assumptions and should be viewed through the lens of a “proof of concept” stage. The aim was to provide an overview of the scope of potential analyses involving GDELT data. This project only included data from February 2015 to January 2016.

4 . Research findings: exploring UK-based natural disasters in the Global Knowledge Graph 2015 data

The goals of this exploration were:

- to see if the Global Database for Events, Language and Tone (GDELT) Global Knowledge Graph (GKG) dataset offers a level of detailed information that is useful as additional information to existing disaster databases
- to explore whether disaster-related mortality data can be extracted
- to verify data accuracy

For this purpose, GDELT GKG data covering the period November 2015 to January 2016 were analysed as two big natural disasters that took place in the UK in December 2015.

The first, called Storm Desmond, brought heavy rain, flooding and severe disruption with damage to infrastructure to many parts of north west England, Yorkshire, the Scottish Borders and Northern Ireland, as well as other parts of the UK in the period around 4 to 6 December 2015.

The second was Storm Eva that brought severe flooding around 26 December 2015 and affected many areas around Greater Manchester, Lancashire and Yorkshire.

These two natural disasters had sufficient impact to be included within the international Emergency Events Database (EM-DAT) with details shown in Table 2.

Table 2: Economic and social impact of disasters
UK, December 2015

	Date	Deaths	Total people affected	Total damage
Storm Desmond	4 to 6 December 2015	3	15,600	1,200,000k USD
Flood	26 December 2015	0	48,000	1,200,000k USD

Source: EM-DAT: The Emergency Events Database - Universite Catholique de Louvain (UCL) - CRED

Identifying UK-based disasters through articles

We chose to explore whether disasters could be identified based on fluctuations in the volume of news articles referencing UK-based disasters.

To identify articles referring to UK-based stories, a number of filtering approaches were trialled before deciding to use the GKG's "V1LOCATIONS" variable. This references all locations, in the order that they are placed, in an article's text. Only articles that referenced a UK location at least twice were selected, and a secondary condition required that these locations also be in the first couple of location mentions. This signified that the UK locations are more important in the article than other non-UK locations that might be mentioned later on.

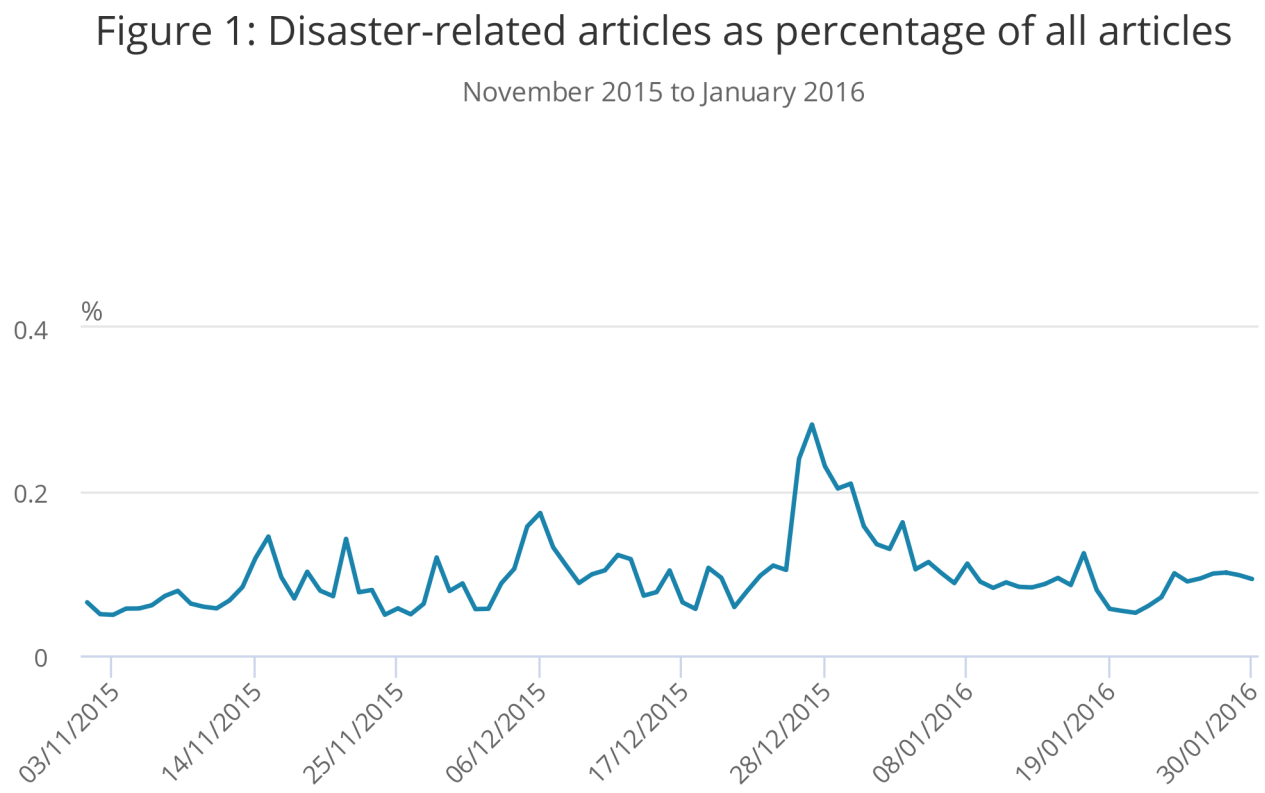
Disaster-related articles were identified using the GKG's derived "V1THEMES" variable. Although the mechanism behind GDELT's theme allocation to articles is not clear, generally it maps words or phrases found in the text to themes found in the Conflict and Mediation Event Observations (CAMEO) framework. The text in an article might be mapped into multiple themes and the GKG gives all the themes related to an article.

We made use of the "V1THEMES" variable to filter our UK-based articles down to a subset that contained at least one theme of "NATURAL_DISASTER". In a similar way to identifying UK-based articles, a secondary condition was that the theme of "NATURAL_DISASTER" should be found early in the list of themes, to indicate its prominence within the article itself.

After filtering articles by UK location and a theme of "NATURAL_DISASTER", the percentage of these articles as a proportion of all UK articles followed the pattern depicted in Figure 1 during the period November 2015 to February 2016.

Figure 1: Disaster-related articles as percentage of all articles

November 2015 to January 2016



Source: GDELT project

Storm Desmond can be observed in the peak on 5 to 6 December 2015, while Storm Eva and the subsequent days of severe flooding can be seen in the peak around 27 December 2015.

The “V2SOURCECOMMONNAME” variable in the GKG contains each article’s Uniform Resource Locator (URL). Although not all article URLs provide information about the article content, they were a good source for this information with many URLs containing the article heading. Manually skimming through URLs of the article selection on peak dates suggested reasonable accuracy as most articles seemed related to Storm Desmond or the flooding in UK. A search for the terms “flood”, “storm”, “rain” and “wind” showed that the majority of URLs contained at least one of these words on the days of interest.

A specific issue in our approach to the selection of UK natural disaster-based articles, as illustrated in Figure 1, is the presence of other smaller peaks that could not directly be associated with major natural disaster events. A cursory investigation on the dates of these peaks revealed that they tend to occur at the weekends, leading us to consider if weekend reporting tends to have fuller reporting on natural disasters than midweek. This might be explored in future research.

In summary, our research showed us that filtering articles based on the theme of “NATURAL_DISASTER” and identifying specific disasters in GDELT is difficult. Identifying larger disasters based on the number of articles that have a “NATURAL_DISASTER” theme assigned to them, as in this research, seems to work reasonably well. However, smaller or less sensational events are much harder to deal with, as they might be hardly recognisable against background noise.

In respect of two main 2015 UK disasters, Storm Desmond during the period 4 to 6 December, and a major flood commencing 26 December, peaks in disaster reporting can be identified in GDELT. The relevance of extracted articles is promising but insufficient, underlining a need for better noise reduction in article selection. Exploration of further disasters or hazards of interest would be useful to draw more valid conclusions about the detectability of these events. GDELT GKG only commenced in 2015, limiting the number of disasters that the data can currently be searched for. Ideally, data for at least one more year should also be explored to check for the impact of seasonality.

Extracting numeric information from disaster articles

This section considers the two GDELT variables available in the GKG that hold numeric information extracted from news articles “V2.1AMOUNTS” and “V2.1COUNTS”.

The “V2.1AMOUNTS” variable lists all numeric data referenced in the document, together with the object it refers to. The “V2.1COUNTS” variable only lists counts that belong to specific categories found in the CAMEO framework, and adds location information to each entry based on the surrounding text. Section 3 gives more detail, and further information is included in Section 2 of the Appendix.

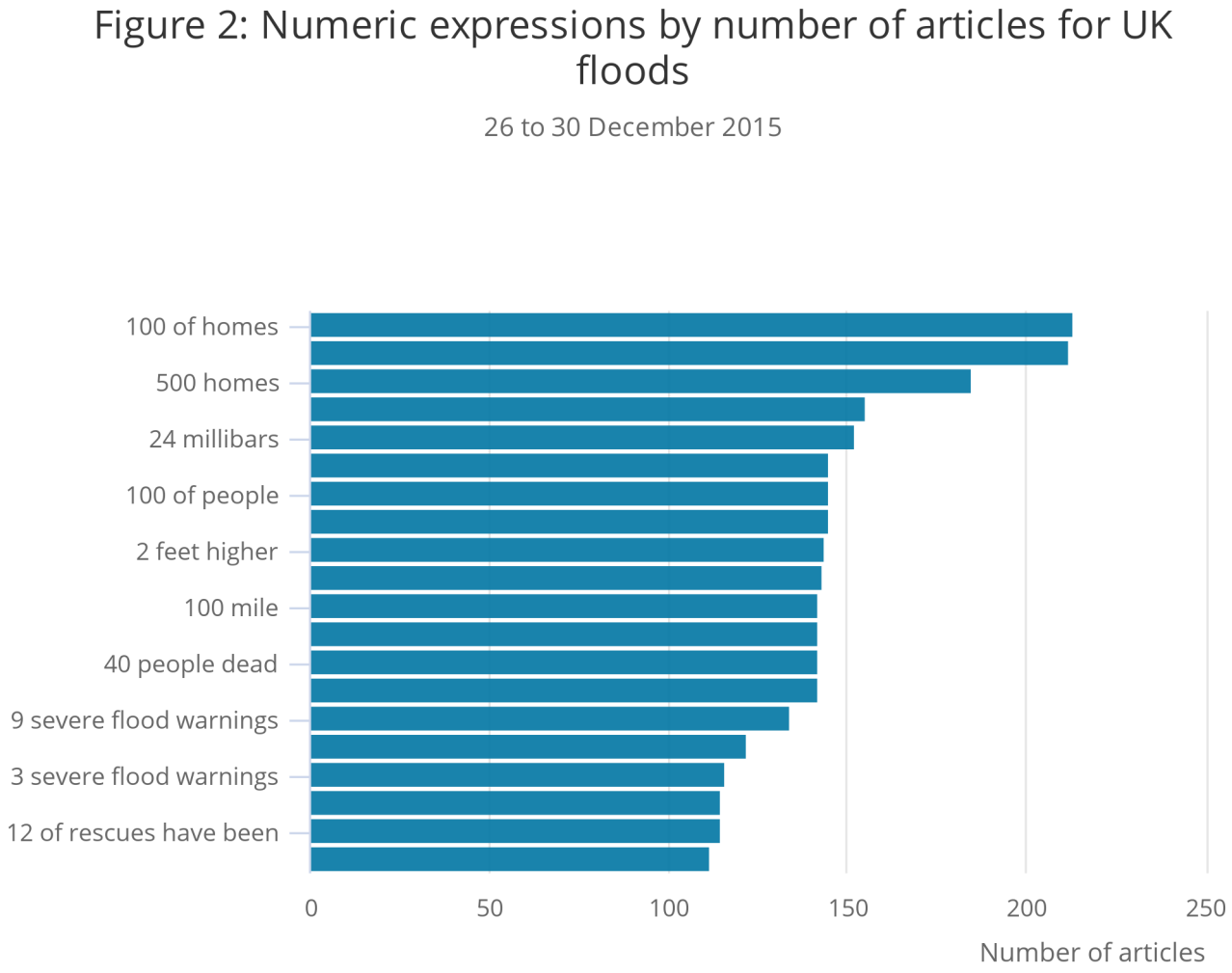
Top numeric expressions in disaster articles

Using the GKG “V2.1AMOUNTS” variable, we extracted the most common numeric expressions from our disaster articles around the two peaks in December 2015.

Figure 2 shows us that the most common expressions for these articles in late December all seemed to be related to the flooding that started on 26 December.

Figure 2: Numeric expressions by number of articles for UK floods

26 to 30 December 2015

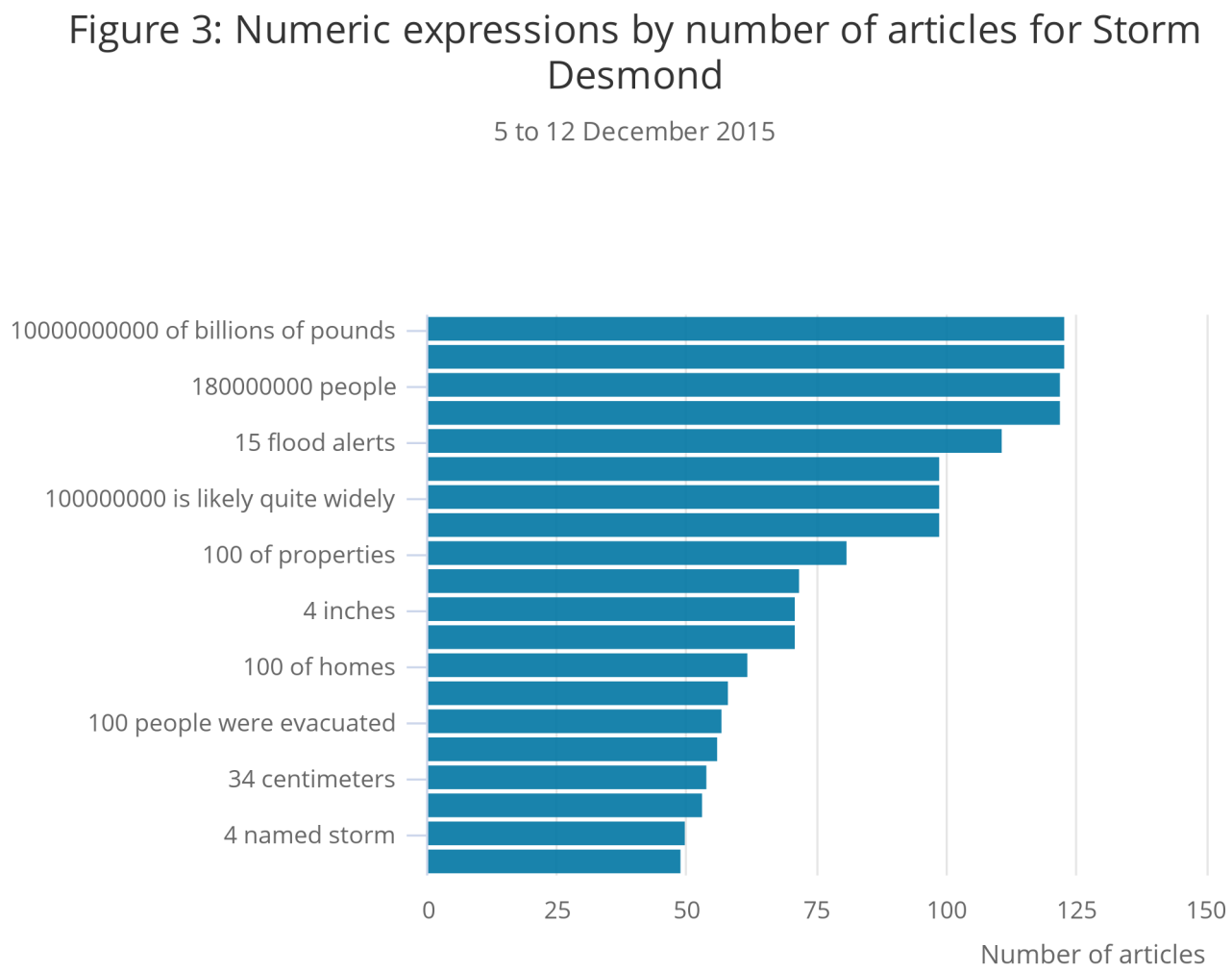


Source: GDELT project

By contrast, Figure 3 shows that the top four expressions in our articles during 5 to 12 December 2015 (the period covering Storm Desmond and its immediate aftermath) actually originate from references to the UN Climate Change Conference in Paris where billions of pounds in aid were promised to help millions of people. More sophisticated filtering to select articles referring to current UK-based disaster articles might help to improve this analysis.

Figure 3: Numeric expressions by number of articles for Storm Desmond

5 to 12 December 2015



Source: GDELT project

A more targeted analysis on numeric data in the GKG “V2.1AMOUNTS” variable in the context of mortality was also performed. Three people died as a consequence of Storm Desmond, and expressions involving mortality were extracted from our disaster articles during the period 5 to 12 December 2015.

There were only 27 articles for which consistent phrases around mortality could be extracted, and the top mortality expression in them of “300 dead” was a reference to an historical event – a flood in the 1980s that had killed 300 people. As illustrated in Figure 3, this expression on mortality did not manage to be one of the top phrases within all our disaster articles from this period.

In summary, using GDELT information to extract specific numeric data from relevant articles is challenging. However, a more sophisticated text analysis on the numeric expressions available in the GKG could improve this information. For example, it might be useful to group together expressions found in the GKG “Amounts” variable that refer to the same thing, such as treating “500 properties”, “500 homes” and “500 houses” as the same expression.

Extracting geographical information from disaster articles

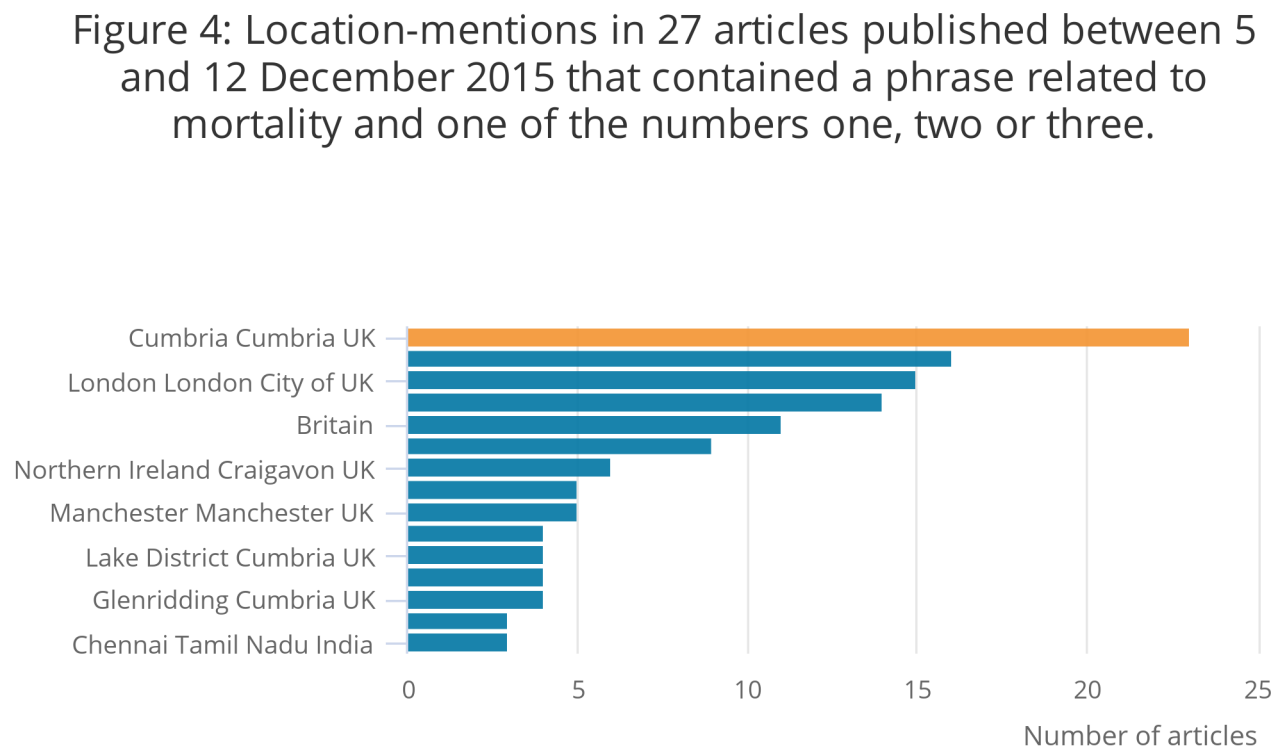
This section contains our research into the locational information found in articles and made available in GDELT's GKG database. As a specific use case, we explored the locational information around the known deaths attributed to Storm Desmond.

The GKG "V1LOCATIONS" variable captures every mention of a location in an article, and this information was used in the following investigation.

In total, there were three deaths directly attributed to Storm Desmond: one person drowned in the River Kent in Cumbria, one person was blown into the side of a bus in London and one person was a resident of Northern Ireland who was found in a river in the Republic of Ireland, close to the Irish border.

We started by considering all the articles we had found using the filtering criteria for detecting UK-based disasters above, namely that the articles should mention UK locations prominently and that they should have at least one theme of "Natural_Disaster". As Storm Desmond passed over the UK during the period 4 to 6 December 2015, we further restricted our articles to those being published during the period 5 to 12 December 2015. For our restricted subset of articles, the "V2.1AMOUNTS" variable from the GKG was analysed for articles referencing one, two or three deaths, and 27 articles were identified. Using the "V1LOCATIONS" GKG variable on these articles, the frequency of locations mentioned is shown in Figure 4.

Figure 4: Location-mentions in 27 articles published between 5 and 12 December 2015 that contained a phrase related to mortality and one of the numbers one, two or three.



Source: GDELT project

The locations of each of the three deaths attributed to Storm Desmond were mentioned across the 27 articles as highlighted in Figure 4. However, article numbers were very small and without prior knowledge of the location of the deaths they would not have been distinguishable from the other locations mentioned.

Issues with using the locations variable

We identified a number of issues that imply the “V1LOCATIONS” variable needs to be treated with caution. These include the following:

- GDELT struggles with ambiguous location names, and gets some location name classifications wrong across several articles. For instance, GDELT classifies some UK locations as Australian because the place names overlap; it classifies the UK location “Aberfeldy in Perth And Kinross” as “Kinross, New South Wales, Australia”.
- Country codes cannot be relied upon, so full country names need to be used. In the UK context, the code “UK” should not be used as a whole range of locations wrongly end up with this country code in GDELT. Examples include locations entries “1#Russia#UK#RS…” or “4#Berlin, Berlin, Germany#UK#GM16...”.
- A 2014 paper[1] on the micro-level study of political violence compared GDELT’s location extraction to two hand-coded conflict event datasets. The researchers concluded “GDELT should be used with caution for geospatial analyses at the subnational level[.]… researchers studying local conflict processes may want to wait for a more reliable geocoding method”.

GDELT also assigns geographical coordinates entries in the GKG “V1COUNTS”, which represents a broad categorisation for an article, although it is unclear how the information is derived.

As introduced in Section 3, GDELT generally maps articles into the CAMEO framework in the “V1COUNTS” variable; each article most commonly belongs to an event categorised as “Affect”, “Arrest”, “Kidnap”, “Kill”, “Protest”, Seize” or “Wound”. Articles that do not have any information in the “V1COUNTS” variable might simply not have sufficient content to indicate that they are predominately discussing themes related to the CAMEO framework.

Our research used the “V1COUNTS” variable to identify all articles categorised as “Affect” to see if the locational impacts of UK-based disasters could be broadly identified. We used our initial sample of articles, published during November 2015 to Feb 2016, prominently referencing the UK and having a theme of “NATURAL_DISASTER” as before.

A basic approach was chosen for this research. As each article can have more than one entry in the “V1COUNTS” variable, only the first entry categorised as “Affect” was selected, and its referenced UK location extracted for analysis. During this step, it became evident that the “V1COUNTS” variable has a comparatively poor coverage, as only 15% of articles had a least one entry, however the location extracted was more specific than the multiple available in “V1LOCATIONS”.

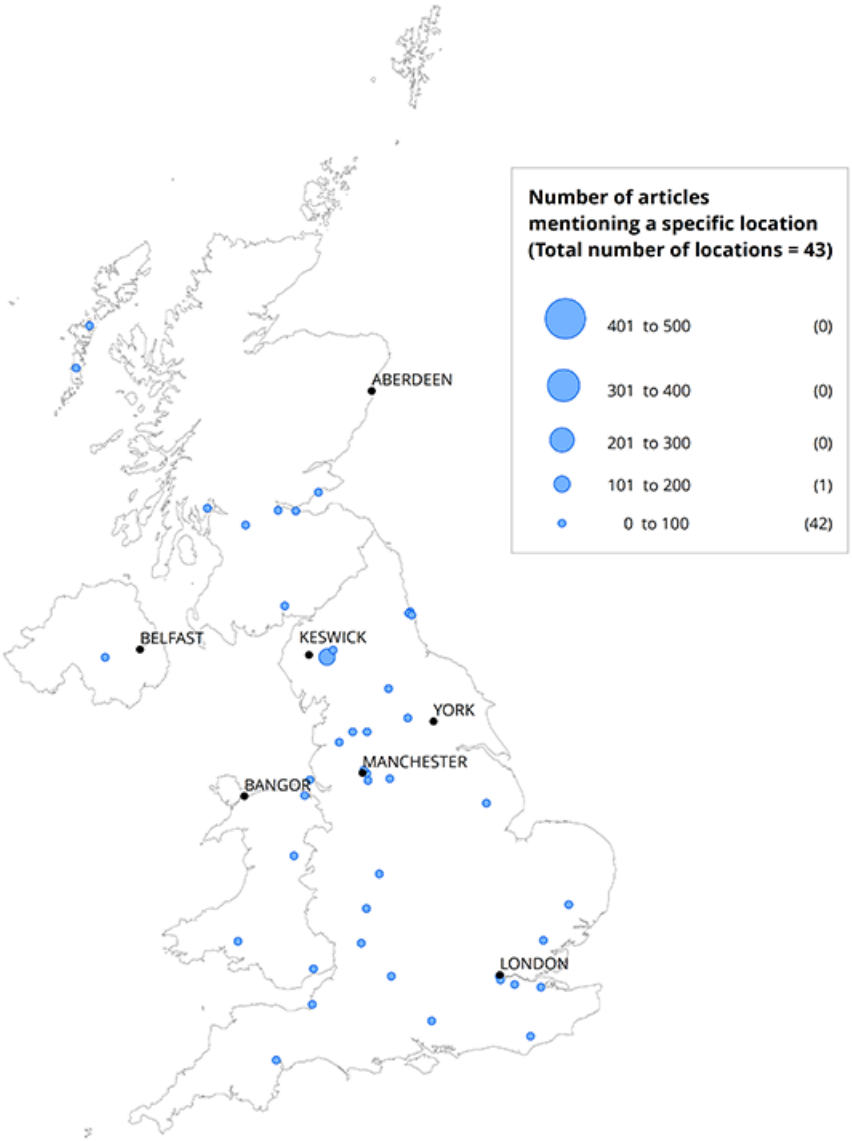
Articles where the first entry categorised as “Affect” had the UK coordinates latitude 54 and longitude negative 2 were removed as this coordinate represents GDELT’s default UK location. This pre-processing resulted in a total of 4,197 articles across the three-month period.

Figure 5 to 7 show the locations extracted from these articles, aggregated for each month of November 2015, December 2015 and January 2016. The size of the dots represents the number of articles referencing the specific location where they are situated on the maps. The maps highlight the dispersal of location across the UK as well as by time.

There were relatively few articles about natural disasters in November 2015; Cumbria had the most with over 200. In December 2015, there were noticeably more such articles across the country, but especially in the North West area including Cumbria and the cities of Manchester and Leeds. London also featured prominently. In January 2016, the number of natural disaster articles reduced, although similar areas to December 2015 are still observed.

Figure 5: Number of articles mentioning a specific location of a natural disaster

UK, November 2015,

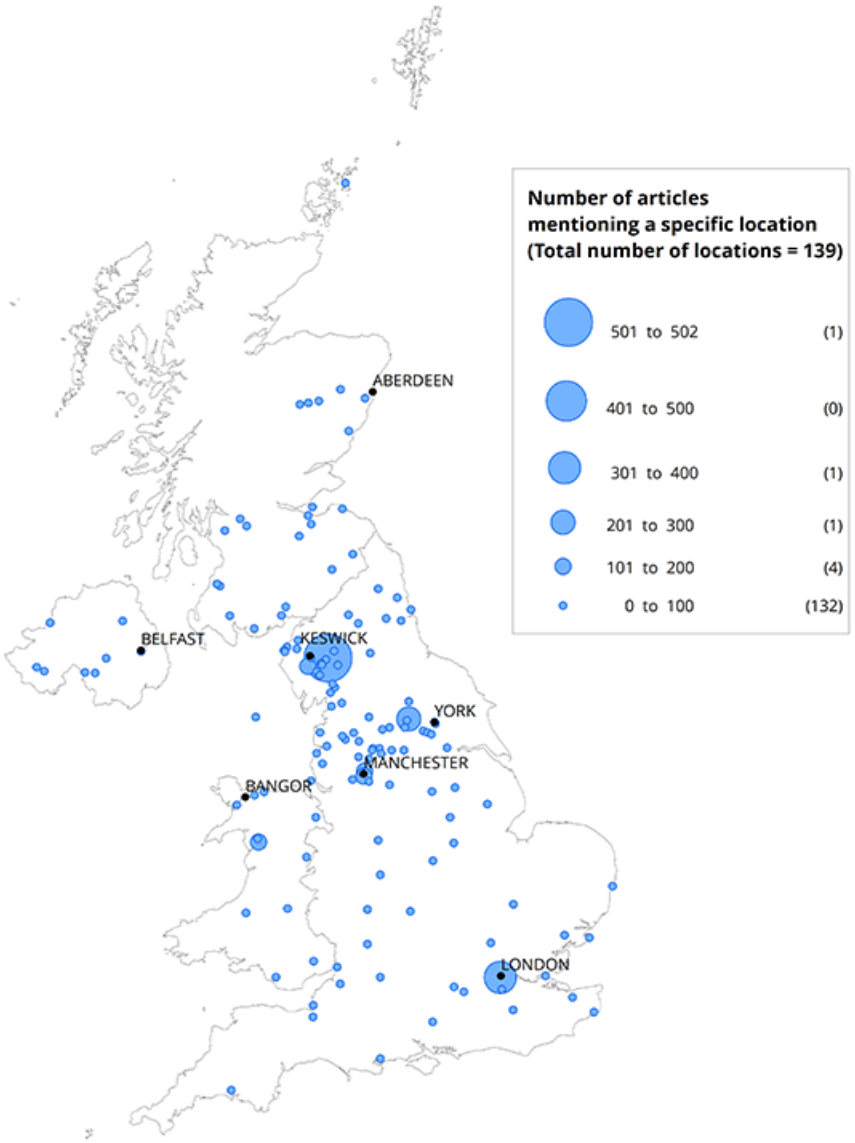


Source: GDELT Project (https://www.gdeltproject.org/)

Graphic created by ONS Geography

Figure 6: Number of articles mentioning a specific location of a natural disaster

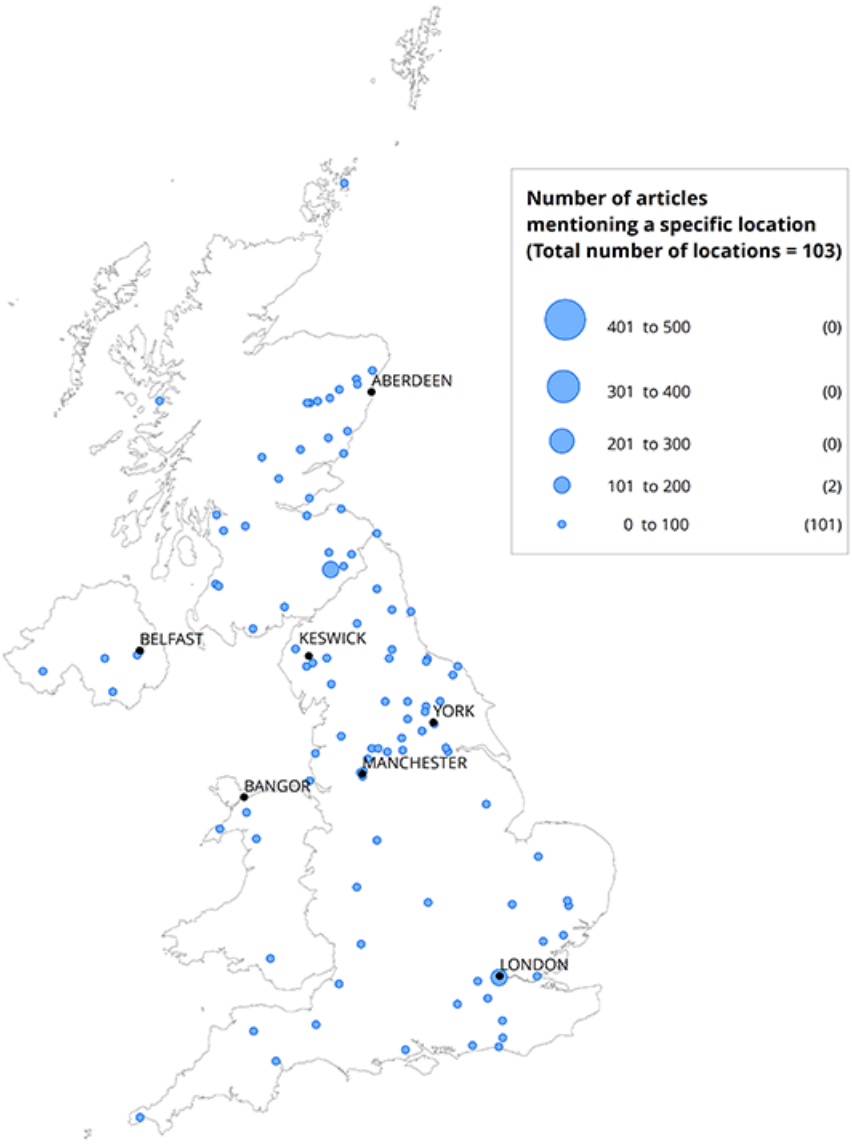
UK, December 2015



Source: GDELT Project (https://www.gdeltproject.org/)
Contains OS data © Crown copyright and database right 2019
Contains LPS Intellectual Property © Crown copyright and database right (2019). This information is licensed under the terms of the Open Government Licence (http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3).
Graphic created by ONS Geography

Figure 7: Number of articles mentioning a specific location of a natural disaster

UK, January 2016



Source: GDELT Project (https://www.gdeltproject.org/)
Contains OS data © Crown copyright and database right 2019
Contains LPS Intellectual Property © Crown copyright and database right (2019). This information is licensed under the terms of the Open Government Licence (http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3).
Graphic created by ONS Geography

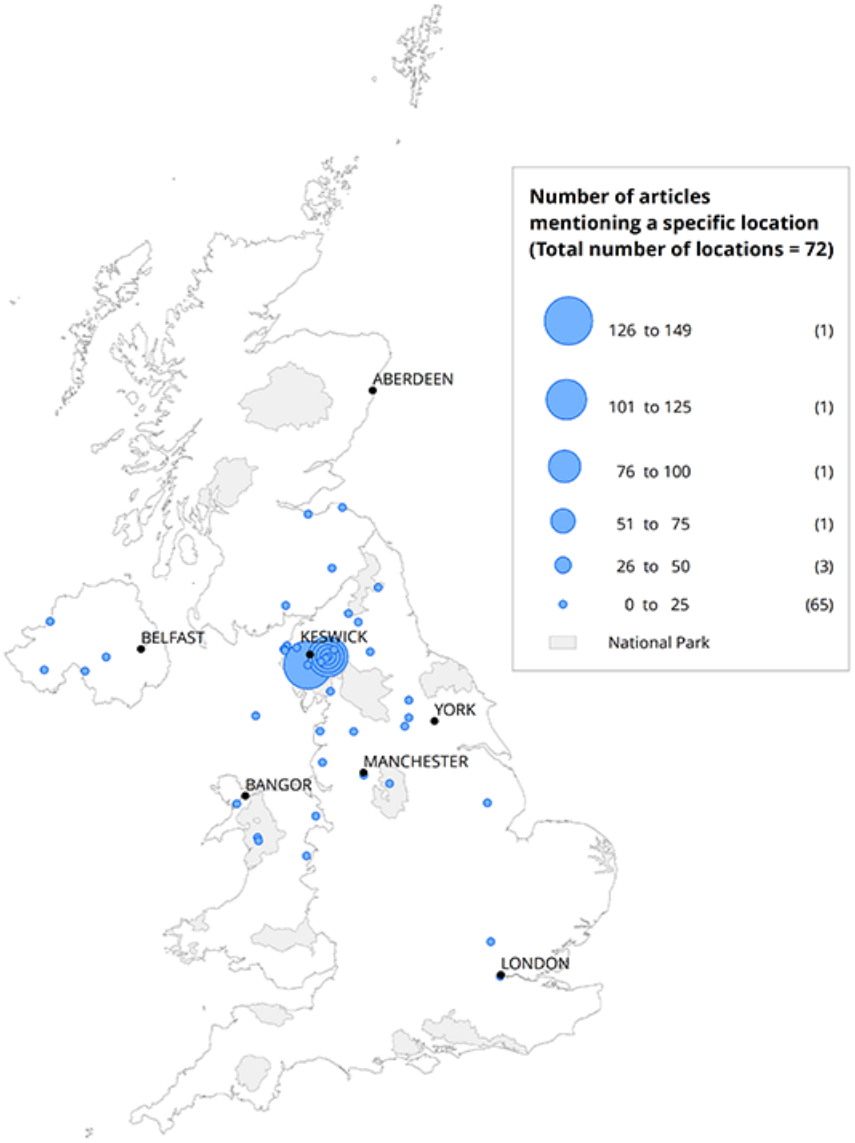
Restricting data to articles published 4 to 12 December 2015 – the time frame around Storm Desmond – revealed the map in Figure 8. Here is it evident that locations in Cumbria are most prevalent among articles discussing natural disasters in the UK. This is consistent with locations recorded in the international Emergency Events Database (EM-DAT) for Storm Desmond: Lancashire, Cumbria district, Kendal, Keswick, Appleby and Hexham.

Figure 9 similarly shows the locations referenced in natural disaster articles during 24 to 31 December 2015, the period covering the major flooding after Storm Eva. Cumbria has fewer mentions, with the main urban areas around Manchester and Leeds taking precedence. Again, this is consistent with EM-DAT records, which cited the main affected areas to include the city of York, Lancashire, Greater Manchester, Humberside, North Yorkshire, South Yorkshire and West Yorkshire.

London is also notable in Figure 9 even though the flooding did not have a great impact there. This needs further investigation, specifically, whether London is mistakenly referenced for articles reporting on the flood in Lancashire and Yorkshire, or if other natural disasters were occurring there that got no mention in EM-DAT as their economic and social impact was too small.

Figure 8: Number of articles mentioning a specific location of a natural disaster

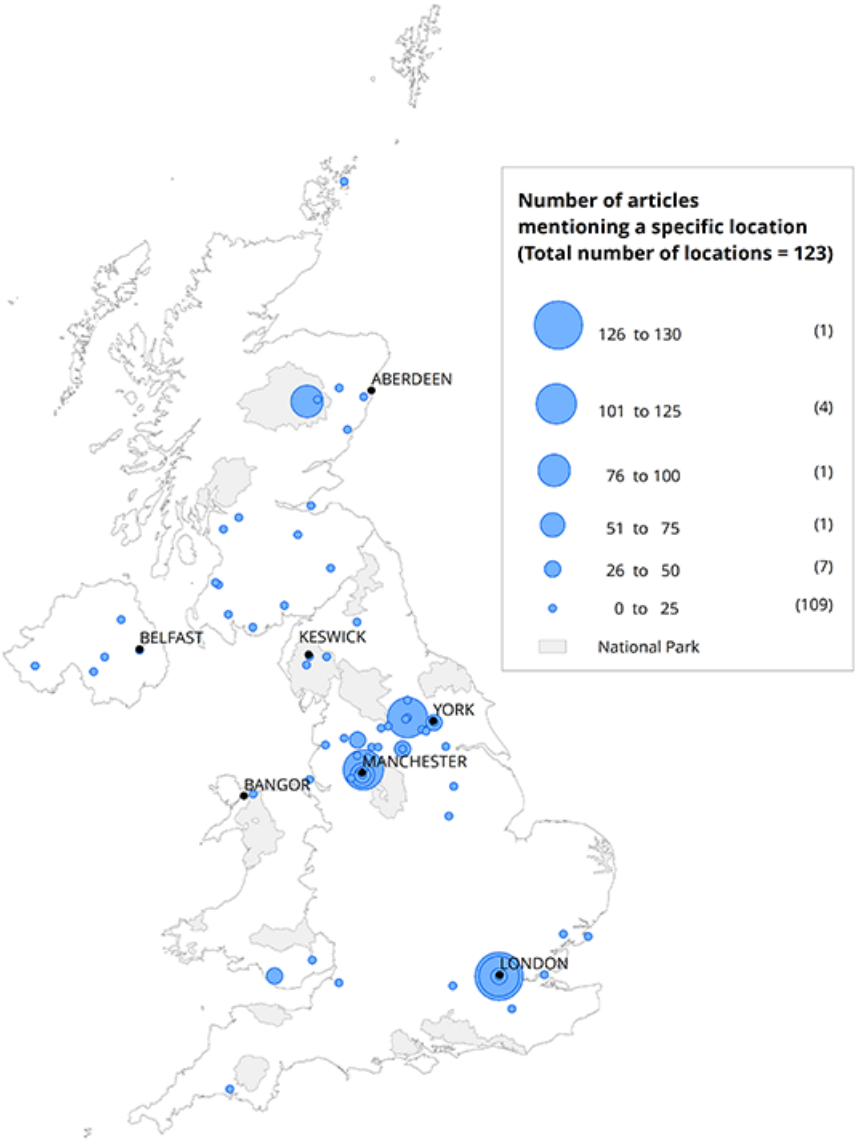
UK, 4 to 12 December 2015



Source: GDELT Project (https://www.gdeltproject.org/)
Contains OS data © Crown copyright and database right 2019
Contains LPS Intellectual Property © Crown copyright and database right (2019). This information is licensed under the terms of the Open Government Licence (http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3).
Graphic created by ONS Geography

Figure 9: Number of articles mentioning a specific location of a natural disaster

UK, 24 to 31 December 2015



Source: GDELT Project (https://www.gdeltproject.org/)
Contains OS data © Crown copyright and database right 2019
Contains LPS Intellectual Property © Crown copyright and database right (2019). This information is licensed under the terms of the Open Government Licence (http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3).
Graphic created by ONS Geography

Summarising this investigation, the GDELT variable of “V1COUNTS” might be used to give a very timely indicative picture of the areas most affected by natural disasters. However, the more localised information contained in the “V1LOCATIONS” variable is harder to extract reliably, mainly because of an inability to accurately identify the context that each location is referring to, but also as the result of issues with the location coding within GDELT’s algorithms.

Notes for: Research findings: exploring UK-based natural disasters in the Global Knowledge Graph 2015 data

1. Hammond, J., & Weidmann, N. B. (2014). Using machine-coded event data for the micro-level study of political violence. Research & Politics.

5 . Strengths and limitations

This article has summarised an investigation into the potential of using the public data held in the Global Database of Events, Language and Tone (GDELT) to inform on disaster reporting, as required by specific indicators for the United Nations Sustainable Development Goals (SDGs). Disaster-themed articles associated with the UK were identified by trialling a number of different filtering techniques.

This provided some evidence that two disasters of 2015, Storm Desmond and a major flood, could be timely observed in the GDELT data.

However, limitations included:

- GDELT's algorithms are not transparent so it is not certain how well they work at categorising information in articles. Nor is it certain how complete a coverage of media outlets is monitored.
- GDELT's algorithms map an article to many different themes including that of "NATURAL-DISASTER". Using these themes, it is difficult to identify those articles that are wholly or predominantly referring to a natural disaster.
- All locations mentioned in an article are made available in GDELT. However, there is difficulty in understanding the context that each location is referring to. This makes the identification of articles on UK-based disasters unreliable, although the filtering methods used in our research did provide a high proportion of relevant articles to our case study disasters.
- Numeric information on the reporting of "deaths" in our articles was unreliable as the text associated with these numbers is not sufficient to determine if these are deaths associated directly with the disaster of interest; the semantics of the actual phrase within the article cannot be inferred directly from the GDELT processed data.

A quality assessment was also conducted using the findings we observed throughout our investigations. For more information, see the accompanying Data quality note.

6 . Further research

Global Database of Events, Language and Tone (GDELT) data are vast, and further research will benefit greatly from downloading and integrating the data into a research database to enable more complex analysis. For this exploratory project, analyses frequently had to be restricted to a small subset of variables or to very limited timeframes to be able to process data in R.

Another suggestion for exploration is to develop an approach to dimensionality reduction when using GDELT's themes for filtering out irrelevant articles. This could significantly decrease the noise in article selections.

There are also potential use cases of GDELT data outside the context of reporting on the Sustainable Development Goals (SDGs). Promising applications are in the context of using news sentiment to feed into indicators of the economy, or by combining data with Twitter data.