Urban Flash Flooding in England Project Documentation

Helen Jackson, ClimateNode

March 20, 2023

1 Introduction

1.1 Overview

There have been many severe surface water flooding events in England in recent years. Some have badly disrupted transport systems, damaged schools and hospitals, and even posed a danger to life. However, there is a lack of comprehensive, easily accessible and up-to-date historic data on surface water flooding. If this were available we would have a much fuller picture of the extent of the problem and its recent evolution.

Local newspaper reports often have very detailed street-by-street accounts of flood events, illustrated with photos and videos. These descriptions are by no means geographically exhaustive, and are usually able to provide only visual or qualitative information on depths and velocities; but they can at least provide an understanding of what occurred with some degree of spatial detail and fill in some of the historic data gap.

The problem is, of course, that this information is distributed amongst thousands of different documents on hundreds of websites. To turn it into geospatial data, it must be systematically extracted, geocoded, stored and manipulated. Natural language processing (NLP) can help with the first of these tasks. NLP is the branch of artificial intelligence which relates to giving computers the ability to interpret and replicate human language, either as voice data or text.

This project has used NLP techniques to detect summertime urban flash flood events from newspaper articles, and extract the names of streets, buildings and other places affected. It has then used this information to generate maps of these events. It analysed approximately 17,400 articles about flooding (2,400 of which were relevant to summertime urban flash flood events) from around 300 newspaper websites. Maps were generated for 56 key dates for summertime urban flash flooding since 2010.

The project was funded by Subak.

1.2 Scope

The dates for the 56 maps were selected according to the following criteria:

- the date fell between May and September (inclusive)
- flash flooding due to heavy rainfall occurred in at least one town or city on that date, and there is sufficient information about individual locations to map this in detail

The project title refers to 'urban flash flooding' rather than 'surface water flooding'. While the primary motivation is to compile data on surface water flooding, these events obviously do occur at the same time as fluvial flooding, and/or in some locations the type of flooding may not be clear from articles.

A map for any given date presents *all* rainfall-related[1] flood impacts identified for that date from articles, including impacts in towns/cities for which there is relatively little detail, as well as impacts in rural areas. This is to avoid making overly-rigid and possibly erroneous distinctions about what is and isn't an urban area and losing information about the spatial extent of impacts of potential interest to users. Therefore, while rural areas are not explicitly included within the project title, many of the maps do contain information about rural areas. For example, the major rural flash flood event which took place in the Yorkshire Dales on 30th July 2019 is included on the map for that date because it occurred on the same day as flash flooding in the West Midlands conurbation.

1.3 Outputs

The project outputs are:

- A set of downloadable interactive maps of flooding impacts, one for each of the 56 dates selected (`UFFE_map_YYYYmmdd.html`)
- A geoJSON file containing the data for each map (`UFFE_YYYYmmdd.json`)
- A single geoJSON file containing the data for all maps (`UFFE_all.json`)
- A csv file listing each date identified by the project as having had a surface water flood event occur on it, including dates which were not mapped. Urls to relevant articles are provided for all dates (`UFFE_events.csv`)

[1] As opposed to flooding from burst water mains and indoor pipes.

2 Methodology

The first task for the project was to compile a body of articles potentially about flooding in England. Automated searches were conducted on newspaper websites for potentially flood-related articles using simple text search. Each article was subsequently analysed during the NLP stage to determine whether it was indeed likely about flooding (as opposed to metaphorical flooding).

The analysis used two particular NLP techniques:

- named entity recognition (NER): this technique is commonly used to identify place names within texts. The analysis used a modified 'off the shelf' NER model which was trained to recognise flooding and heavy rain as 'hazard' entities in addition to standard entities such as places. The presence of more than one hazard entity was taken to indicate an article likely about flooding.

- text classification: this technique assigns categories to entire texts. This was used to classify articles by flooding sub-topic (urban flash flooding, winter storm, flood alerts, etc.). *Provisional* urban flash flood events were identified by using the dates of articles placed within the 'urban flash flood' category (with dates later refined, see below).

The urls of relevant articles were then stored in a database along with the details of any location and asset entities they mentioned.

Coordinates were derived for locations from OpenStreetMap data via the Nominatim API (geocoding). These coordinates were viewed as a first guess and subject to quality control (see below).

There then followed three post-processing tasks based on human interpretation. Articles relating to each provisional flash flood event were examined to determine the date of the event (which, of course, need not be the same as the publication date of articles). When linking flooding impacts to a date, the intention is to attribute the impact to the rainfall which occurred on that date. For example, if a flood impact is discovered on a Monday but due to rainfall which happened on Sunday evening, the impact is linked to the Sunday. Where this is ambiguous because of overnight rainfall, this is indicated via a quality flag. Where there is uncertainty about the date of an impact in an individual location, a judgement is made about the most likely date and a quality flag raised (see Section 3.5). Some articles describe more than one event; every effort has been made to ensure information linked to a date is relevant to that date rather than any other mentioned in the article.

Next, for each urban flash flood event, articles relating to each individual location/asset entity associated with the event were rapidly examined to determine what could be stated about what had occurred there via short, standardised

annotations (presented in popups on maps). This may include depth information, whether internal flooding occurred, etc. Repeatedly used annotations are referred to as 'tags' and an explanation for each of these is described in Section 3.4. In some source articles, users may need to click through a photo gallery or watch a video to find a reference to the relevant location.

Finally, there was a quality control check to ensure that geocoding results for each point of interest corresponded to the location described in relevant articles, and derive correct coordinates where not (e.g., to ensure 'Church Lane' is the one in the town mentioned, rather than one of the hundreds of other Church Lanes in England).

The analysis used newspaper websites represented by the UK newspaper industry's collecting society, NLA Media Access, and an appropriate webscraping licence was obtained.

This analysis is not exhaustive. It has not included every possible text source which could describe English urban flash flood events, for example Local Authority Section 19 flood investigation reports. Furthermore, only articles which are currently online were used. There are likely many relevant articles which are no longer available online, e.g. because a newspaper has been taken over, and unfortunately it was not possible to include these during this project. The volume of relevant articles with sufficient detail available online declines significantly prior to around the middle of the last decade. This data has been compiled in the expectation that it may complement other datasets and flood chronologies, such as the British Chronology of Flash Floods. Users are welcome to extend it, subject to the licence conditions[2], and are also welcome to contact ClimateNode with collaboration proposals.

3 Interpreting the maps and datasets

3.1 Dates

See the previous section for comments on how to interpret event dates.

Note that in the `UFFE_all.json` file the date of impact is provided for each location; impacts which occurred at the same location on different dates are listed separately.

[2] As OpenStreetMap data has been used to derive coordinates, Urban Flash Flooding in England project data is available under the same conditions. Please see OpenStreetMap's Copyright and Licence page for more details.

3.2 Coordinates

Marker coordinates should be treated with caution and generally do not represent true coordinates, which it is often impossible or too time-consuming to determine from articles. Popups present annotations linked to the geographical entity or approximate location mentioned in the header, not necessarily the coordinates of the marker.

Asterisks in popups are used to indicate locations where information in source articles has been used to narrow down where the impact occurred:

** indicates 'true coordinates': references to junctions, buildings etc., or images, have been used to identify where the flooding occurred fairly precisely

* indicates 'approximate coordinates': information in the article has allowed an approximate location to be identified, for example a section of a street between two reference points

(in the geoJSON files these are indicated via the `true_coords` and `approx_coords` properties.)

3.3 geoJSON properties

Each geoJSON file contains a standard geoJSON FeatureCollection comprising of Points. The properties of each Point defined by the project are as follows:

`name:` Name of the point of interest

`popup_title:` Self-explanatory; this is usually the same as `name`, but in a small number of cases, the popup title may differ from the POI's name (for example, where there is ambiguity over the location)

`date:` Date of relevant flood event/map (`UFFE_all.json` only).

`osm_id:` OpenStreetMap ID

`osm_ele_type:` OpenStreetMap element type (see `https://wiki.openstreetmap.org/wiki/Elements`).(Note that the OpenStreetMap ID and element type in combination give a unique identifier for the map feature, which is why is why this property is provided).

`osm_type:` OpenStreetMap map feature type (see `https://wiki.openstreetmap.org/wiki/Map_features`.)

`UFF_type:` One of 10 map feature classes used within the project database corresponding to the map legend as follows:

`urban` corresponds to *city/town/village*: a human settlement, neighbourhood, suburb or housing estate

`subnat` corresponds to *county or district*: a subnational administrative unit such as a local authority district or civil parish

`street` corresponds to *street*: a named road with properties adjoining it

`asset` corresponds to *building*: a distinct, usually commercial, building such as a shop, office block, pub, care home etc.

`healthAsset` corresponds to *healthcare facility*: a hospital, medical practice, etc.

`school` corresponds to *school*: self-explanatory; childcare, pre-school and higher education facilities are classed as assets

`railInfra` corresponds to *rail infrastructure*: a railway station or other piece of railway infrastructure other than a railway bridge, or a point on a railway line

`railBridge` corresponds to *railway bridge*: self-explanatory; these are treated separately from other railway infrastructure as (i) roads below railway bridges are common urban flash flood hotspots; and (ii) in such cases it is the road beneath which is affected rather than rail services.

`roadInfra` corresponds to *road infrastructure*: a piece of road infrastructure such as a major junction, bridge or tunnel, or a point on a numbered road

`urbanTrans` corresponds to *urban transport*: a piece of urban transport infrastructure such as an underground or bus station, or a point on an urban public transport line

Where `null` (or on the map, a marker without an icon) the point of interest does not belong in any of these ten categories.

`true_coords`: references to junctions, buildings etc., or images, have been used to identify where the flooding occurred fairly precisely

`approx_coords`: an approximate location has been identified, for example a section of a street between two reference points

`quality_flags`: flags raising an issue over the quality of the data point (see Section 3.5)

`annotations`: Annotations are short descriptions of what can be understood from articles to have occurred in the stated location. Annotations are either 'tags' (see Section 3.4) or free form.

Note that not all map features have an OpenStreetMap equivalent, and in some cases the OpenStreetMap equivalent was not identified. osm_id, osm_ele_type and osm_type are `null` for these map features.

3.4 Tags

Tags are systematised annotations used repeatedly in different locations as a shorthand for frequently-encountered or expected types of information about flood impacts.

Annotations are intended to provide granular information and are generally applied to the lowest geographical level possible. For example, if internal flooding occurred in a neighbourhood on streets named in the article, the 'internal flooding' tag is linked to those individual streets, not the neighbourhood. If internal flooding occurred in a neighbourhood but no street names are mentioned, the 'internal flooding' tag is linked to the neighbourhood.

Transport impacts are generally described if there is information on the location of floods/landslides causing disruption. For example, if an article describes flooding causing disruption of train services between London and Crewe, that does not give precise enough information on where the flooding occurred, while flooding occurring between stations within a few miles of each other is generally included.

The full set of tags is as follows:

Basement/cellar flooding: Flooding in rooms below ground level.

Basement/cellar flooding implied: Flooding in rooms below ground level likely/implied, but unclear.

Burst river/stream/reservoir banks: Self-explanatory. See also 'River/stream flooding' below. This tag is more likely to be used where the location is near to where the water body burst its banks. Minor flooding immediately adjacent to the water body in parks, fields etc. where it is not problematic is unlikely to be tagged.

Closure: Premises closed during flood event. This may or may not be due to internal flooding, with further information possibly provided by other tags.

Closure (internal flooding presumed): Premises closed during flood event, likely caused by internal flooding, but not explicitly stated in sources.

Closure (unclear if flooded): Premises closed during flood event, less likely to be due to internal flooding, and more likely to be precautionary or due to access problems.

Damage reported: Sources mention either damage to fixtures and fittings or structural damage. This tag is provided to indicate more serious flood impacts.

Drone footage/photo(s) available: Self-explanatory.

External flooding: External flooding occurred at a property.

Fatality: Fatality due to urban flash flooding (single case identified).

Fatality (river or stream): Fatality due to a person falling in a river or stream.

Fire & Rescue Service call (flooding unconfirmed): Fire and Rescue service call which is likely due to flooding, but this it not clear in the source.

Flooding: Default tag where there is flooding but few/no further details on the type or impact.

Flooding affecting train/urban transport services: Self-explanatory. Tag is only provided with adequate geographical information on where the disruption occurred.

Flooding below: Flooding beneath a bridge.

Flooding implied: Flooding likely/implied, but unclear.

Flooding implied or anticipated: Flooding is either implied or distinct actions are being taken (e.g. householders are placing sandbags in doorways) and it is not clear whether the flooding has actually occurred.

Flooding in the vicinity: Flooding has occurred either at or near the specified location.

Flooding or other storm-related incident (unclear): Described impact could be due to flooding or another storm-related impact, such as hail or lightning.

Hail: Self-explanatory.

Internal flooding: Water has entered a property.

Internal flooding implied but ambiguous: Internal flooding likely/implied, but unclear.

Internal flooding in the vicinity: Internal flooding has occurred either at or near the specified location.

Internal flooding narrowly avoided: Sources describe householders narrowly avoiding internal flooding, either through preventative actions or because rainfall was not quite sufficient to cause it.

Internal flooding via ceiling/roof and/or ceiling collapse: Rainfall has been intense enough to cause a ceiling collapse and/or significant ingress through the rood/ceiling. Minor leaks are not tagged.

Landslip: Intense rainfall has caused a landslip.

Landslip affecting train/urban transport services: Self-explanatory. Tag is only provided with adequate geographical information on where the disruption occurred.

Lane/slip road closure: Road flooding causes one or more lanes of a multi lane road (or a slip road at a major junction) to be closed without total closure of the road.

Manhole cover displacement: Water pressure from within the the sewer system has caused a manhole cover to be displaced.

Marked downhill flow: Sources describe significant quantities of water flowing downhill.

Minor internal flooding: Internal flooding where water has not entered the property in significant quantities and can be cleaned up easily, for example, where a shop does not need to close to deal with flooding.

Outbuilding flooding: Flooding in a garage, shed, or other outbuilding.

Perilous situation described: Those affected or eye-witnesses describe a situation which is particularly dangerous, frightening or life-threatening.

Photo(s) available: Self-explanatory. Users may need to click through photo galleries to find the relevant photo.

Photo(s) available (external): As above, but making clear in cases of internal flooding that the photos are external.

Property flooding: A property has been flooded, but it is not clear whether this refers to internal, external or outbuilding flooding.

River/stream flooding: Self-explanatory. See also 'Burst river/stream/reservoir banks' above. This tag is more likely to be used further from where the relevant water body burst its banks.

Road closure: A road has been closed during a flood event. This may or may not be due to flooding, with further information possibly provided by other tags.

Road closures: As above, but used for a district, town or village where the specific road closures are not known.

Road damage: The road surface or an item of road infrastructure such as a bridge has been damaged during a flood event.

Road flooding in vicinity: Road flooding has occurred either at or near the specified location.

Road flooding/surface water: Sources describe or show images of flooding affecting the surfaces of streets/roads but likely not properties. This tag is also used for road infrastructure such as underpasses, tunnels and bridges.

Road flooding/surface water implied: Road flooding likely/implied, but unclear.

Sewer overflow into street/building: Sources describe or show images of overflows from either foul or surface water sewers into public places or buildings. This includes flooding up through toilets.

Street/location(s) unspecified: Sources describe flooding in a town or village without further geographical detail.

Structure collapse: A structure such as a wall or bridge collapses during a flood event.

Subway flooding: A subway/pedestrian underpass under a road or at a railway station is flooded.

Vehicle(s) floating in floodwater: Flooding is so severe that vehicles are described as floating in floodwater.

Vehicle(s) stuck in water: Sources describe or show images of vehicles which are clearly incapacitated due to the depth of floodwater.

Video available: Self-explanatory.

Video available (external): As above, but making clear in cases of internal flooding that the video is external.

3.5 Quality flags

Quality flags are presented between square brackets [] in map popups.

ambiguity over flooding impact: sources seem to indicate that there is some kind of flooding impact but this is unclear, either due to mistakes in sentence construction or some other source of ambiguity.

ambiguity over street/location: the specified map feature may not be the one indicated in the source, for example, due to misspelling, a possible reporter's error, or difficulty distinguishing between two places with the same name.

date uncertain or ambiguous: the described impact may have occurred on the date indicated, but this cannot be determined with certainty

flooding may not be due to rainfall: flooding may be due to a burst water main occurring at the same time as the flood event

overnight, following: source indicates the flooding impact resulted from overnight rainfall; the rainfall leading to the flooding impact may have occurred on the date following the one indicated

overnight, previous: source indicates the flooding impact resulted from overnight rainfall; the rainfall leading to the flooding impact may have occurred on the date prior to the one indicated