Office for National Statistics

Global Database of Events, Language and Tone (GDELT) data quality note

Assessing the quality of unofficial data sources in the context of 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 Data quality note addresses some of the more general quality questions relevant to the use of unofficial data sources in the context of disaster reporting.

The other pages are:

- a main Article, which presents the findings of the investigation and discusses GDELT data potential, limitations and quality, and outlines suggestions for additional research and different applications
- 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

2 . Data quality

Suitability of using a non-official data source

This data will most likely not be considered a source of robust data. It could potentially provide estimates of the scale of an event of interest, or add information in combination with other data sources.

There are alternative official UK data sources that could be used instead of the Global Database of Events, Language and Tone (GDELT) data to report on numbers of deaths, but not numbers of missing persons and directly affected persons by disasters. There are official mortality statistics that could be used to calculate numbers of deaths attributed to disasters. However, this data is often not available until two years after an event.

The global disaster database EM-DAT provides estimates of lost lives and number of people affected, as well as economic losses associated with a disaster (see Appendix for more details). However, the criteria for a disaster to be listed on EM-DAT mean that smaller disasters are not included and EM-DAT's quality assurance process means that figures are often missing where data quality standards are not high enough.

With respect to economic loss figures, official sources generally only provide detailed loss analyses for large or severe catastrophes.

The main value of the GDELT data lies in:

- its timeliness (updated every 15 minutes)
- its coverage across the globe
- its coverage of small-scale events

As such, GDELT might contribute to:

- more timely estimates of a disaster's magnitude, which can be updated once official figures are available
- estimates of disaster consequences in geographical areas with limited or no official data on mortality
- information about small-scale events that might only be reported in local news outlets

Research context

The GDELT project is set up as not-for-profit, but is supported by Google. The initiator of the GDELT project is Kalev Leetaru, an entrepreneur and Media Fellow at the RealClearFoundation, and a Senior Fellow at the George Washington University Center for Cyber and Homeland Security. He is a former Yahoo! Fellow in Residence of International Values, Communications Technology and the Global Internet at Georgetown University's Edmund A. Walsh School of Foreign Service and was a 2015 to 2016 Google Developer Expert for Google Cloud Platform.

Western media, and in particular US media, seems to be overrepresented in GDELT, and a Western reporting perspective potentially introduces bias to the overall way of event reports.

Quality assurance

It is unclear whether any specific quality assurance mechanisms are built into the GDELT data collection and processing process. GDELT should still be considered an experimental project. Only one of the algorithms that extract and compile the data is described in detail, making it impossible to reconstruct the process of data extraction from individual news articles. While the GDELT's Global Knowledge Graph (GKG) data selection process is theoretically clear (that is, extracting all data within a set of categories from a news article), the process for the GDELT events database involves aggregation of individual news articles into events, adding a further layer of unexplained complexity.

Comparisons with other data sources would be useful. Initial comparisons with EM-Dat data on UK disasters in December 2015 showed consistency with respect to the presence of events and the locations of these events. Figures on mortality or people affected could not be extracted from GDELT yet, so comparisons with EM-DAT figures are outstanding.

Relevance

The data is collected at the very high frequency of every 15 minutes, allowing for very timely data. Coverage is broad, including many British news outlets as well as smaller local papers. While the coverage is extensive, there is a bias towards Western media, with US media being particularly overrepresented.

The data that are of interest to the Sustainable Development Goals (SDG) target indicators of interest is not explicitly collected and as such cannot easily be extracted from GDELT.

Accuracy

There are significant concerns about the data quality of GDELT.

Firstly, there are general quality issues regarding figures reported in the news media which GDELT draws upon. On the one hand, news reporting around disasters can be biased (for example, to mobilise more emergency aid) and on the other hand, figures often evolve over time as the disaster conditions change.

Secondly, there are concerns related to GDELT's algorithms, most of which have been described in the Section 4 of the main article.

Data availability

Data collection started relatively recently (version 2 started its extensive collection in February 2015) and data collection and aggregation methods might change. However, handling of changes so far suggests that changes will not be implemented at short notice. As part of the one and only major change that GDELT has undergone so far in 2015, it was ensured that the old way of data collection was continued for over a year to allow data users to transition to the new version.

We are aware that a new version of GDELT will be made available early in 2020.

GDELT references in the academic literature

Whether and how GDELT is used in academia could be an additional indicator for the quality of the database. Few academic articles appear to use or discuss GDELT data and several of these publications are of low quality. The limited use of GDELT in the academic literature might be related to the data source being relatively new, but more likely related to issues with data quality or difficulty to extract meaningful information.

There are a few articles by a team of researchers at the Qatar Computing Research Institute, but these are mostly exploratory conference papers and not published in peer-reviewed journals. Where the authors compare GDELT with other datasets it is a high-level comparison and based on a small subset of events. Nonetheless, one of their interesting findings should be noted: they find that a lot of sources were not covered by GDELT in 2015 and that, based on a 10-day sample, the top 10 ranked list of news sources in GDELT and Event Registry (a start-up offering global media monitoring of 35,000 sources) did not overlap at all. This is largely because the top Event Registry's sources are very niche. Of their top five sources, one is about news from Burkina Faso, one is a small, very local German newspaper and another one a website providing information about health insurances in Germany.

Other literature deals with very specific aspects of GDELT only, such as the paper by Hammond & Weidmann on geolocation methods that 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".

Notes for: Data quality

Kwak, H., & An, J. (2016). Two tales of the world: Comparison of widely used world news datasets GDELT and EventRegistry. In Tenth International AAAI Conference on Web and Social Media.

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