Studying the Syrian Civil War with GDELT

David Masad
Caerus Analytics

New Scientist magazine demonstrated GDELT's potential by using it to analyze the ongoing Syrian civil war[1], and showed that the volume of violent events has been waning. Jay Ulfelder[2] contrasted New Scientist's GDELT analysis with other data to argue that the apparent decline in violence is largely driven by media fatigue rather than actual change in the intensity of the fighting.

We attempted to replicate the New Scientist analysis, and perform more rigorous validation of GDELT results against Syria Tracker's crowdsourced death reports and UN refugee data. We identify data quality issues across all three datasets. However, we also identify strong correlation between GDELT violent events and Syria Tracker death tolls at the national level, and weaker correlations at the level of the separate governates. We also identify lagged correlation between GDELT events and refugee registration. We analyze the network formed by actor dyads and experiment with automatically discovering factional affiliations to actors and events.

Data

- The full version of GDELT, through 05/27/2013. Consists of events extracted from media sources, consolidated to at most one event of any type between two specific actors at a particular geolocation per day[3].
- Syria Tracker's monthly counts of individuals reported killed by governate, March 2011 - May 2013[4].
- UNHCR counts of Registered Syrian Refugees, overall as well as separately for Turkey, Lebanon, Jordan, Iraq and Egypt[5].

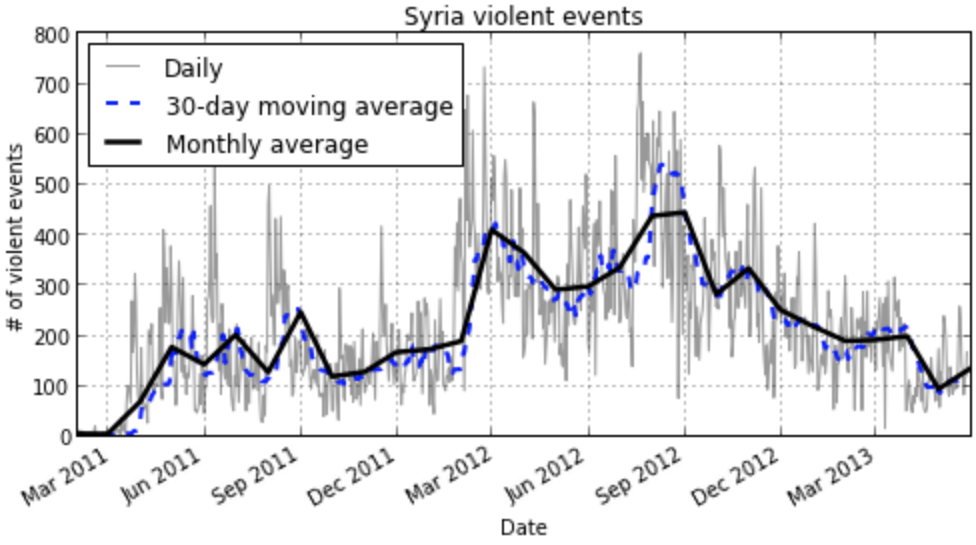
Data Subsetting

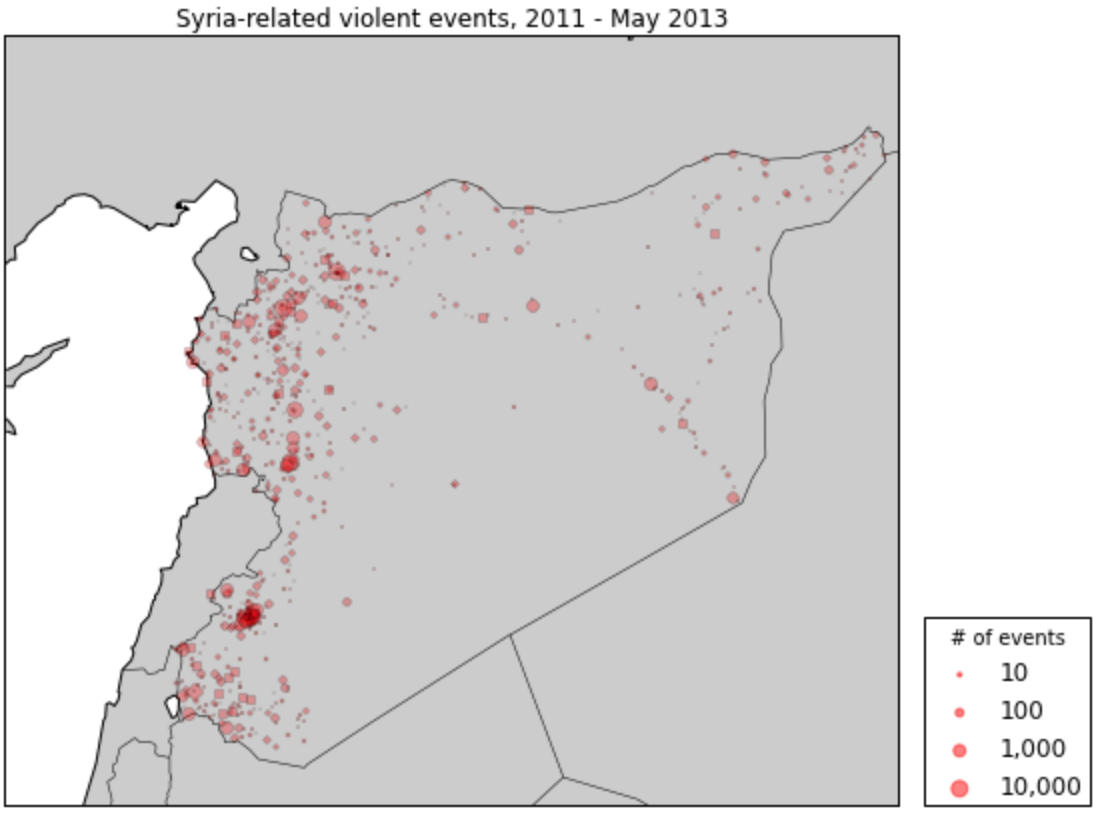
We use GDELT records from 2011-present, and attempt several methods to subset records related to the Syrian conflict. After some experimentation, we settled on selecting events where the action country geocode is Syria, and at least one of the actors has Syria as an actor country-code. We further select event coded as involving Material Conflict[6] with a Goldstein score of -8.7 or below indicates that the event involves military or nonmilitary destruction and injury[7].

GDELT events are recorded daily; however, daily counts tend to be extremely noisy. The

[1] http://syria.newscientistapps.com/
[2] http://dartthrowingchimp.wordpress.com/2013/05/16/challenges-in-measuring-violent-conflict-syria-edition/
[3] http://gdelt.utdallas.edu/
[4] https://syriatracker.crowdmap.com/reports/view/2888
[5] http://data.unhcr.org/syrianrefugees/regional.php
[6] http://eventdata.psu.edu/papers.dir/ISA.2013.GDELT.pdf
[7] http://www.chsbs.cmich.edu/fattah/courses/empirical/jgscale.htm

emerging best practice[8] is monthly aggregation; this is also the temporal scale that the Syria Tracker and UNHCR datasets are recorded at.





8

http://eventdata.psu.edu/papers.dir/Working%20with%20Event%20Data-%20A%20Guide%20to%20Aggregation%20Choices.pdf

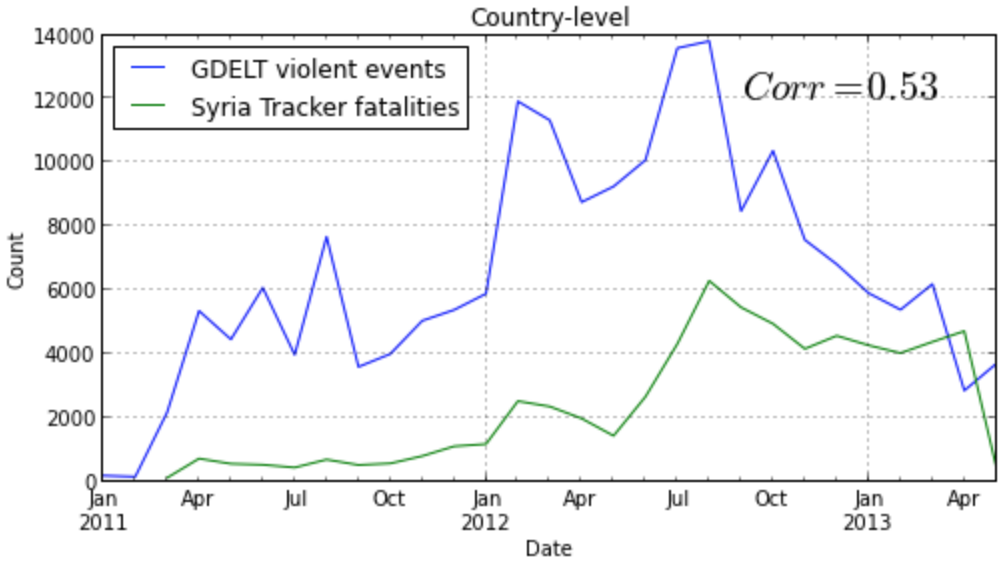
Event Time Series

GDELT is based on media reports; one way of checking its accuracy is by comparing it to other datasets about the same conflict. While these may not represent ground truth either, agreement between them suggests accuracy, while divergences may raise further questions.

Syria Tracker Reported Fatalities

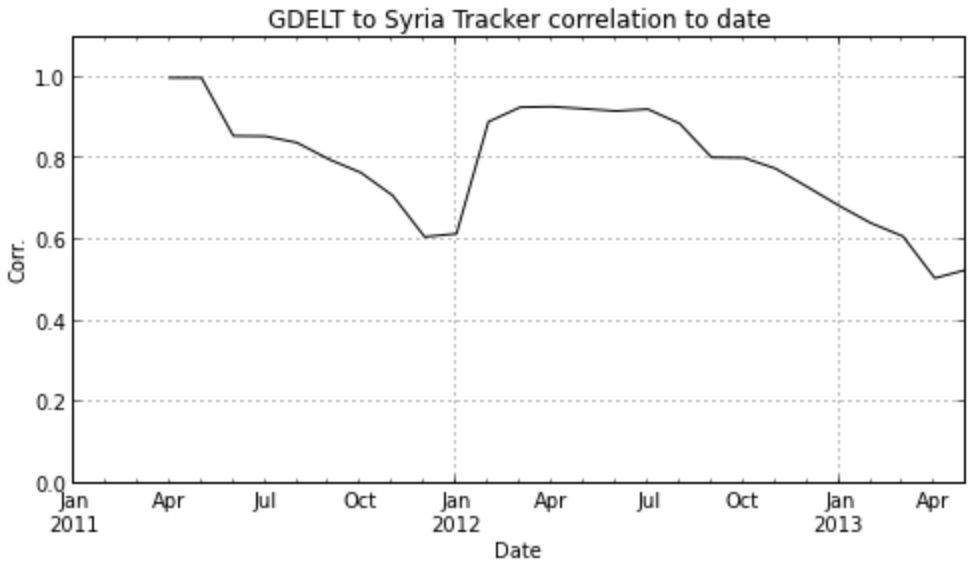
The Syria Tracker project compiles data on individuals killed in Syria from volunteered reports. While not the only such dataset, it has been widely used and was readily obtainable.

At the country level, we see a relationship between the volume of violent events and the number of reported deaths. As the graph below shows, while the number of violent events is higher than the number of reported deaths, the two series initially move together: an increase in violent events accompanies an increase in reported deaths, and vice versa.



However, we do see a falloff in the relationship between the series beginning in the summer of 2012. While the level of violence remains roughly constant as measured by fatalities reported, the number of GDELT events steadily drops. As Jay Ulfelder suggests[9], this may be evidence of media fatigue[10]: as the conflict drags on with few major developments, media interest wanes and coverage (and hence sources for GDELT to draw from) decreases

[9] http://dartthrowingchimp.wordpress.com/2013/05/16/challenges-in-measuring-violent-conflict-syria-edition/
[10] http://eventdata.psu.edu/papers.dir/EWER.pdf



.

We can see additional evidence for this hypothesis in the expanding window correlation above, showing the correlation between GDELT violent events and reported fatalities up through each date. We see a steady decline in correlation while the conflict is stable in 2011; correlation then shoots up along with violence and fatalities both in 2012, suggesting that the escalating conflict drew additional media interest. Finally, as the conflict stabilizes, the number of events steadily drops, yielding a month-over-month decline in correlation.

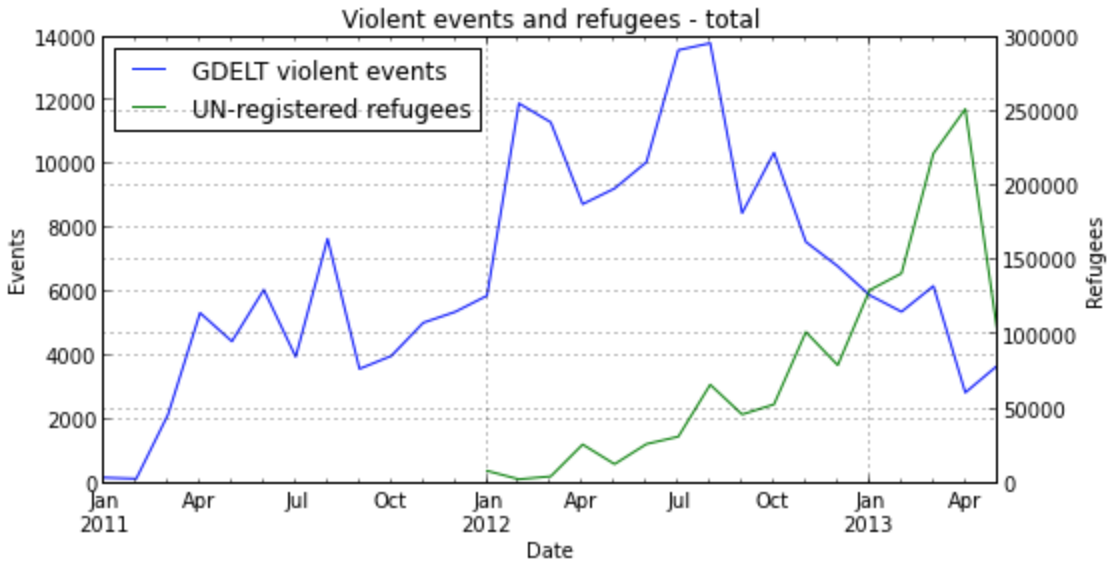
Syria Tracker reports death counts at the level of the governates and of lower-level administrative units. We associate GDELT events with governates as well, by identifying the governate within which each geocoded event falls. We then conducted a similar correlation analysis for each, comparing the volume of reported violent events and reported deaths for each month (see Appendix for full visualizations).

As we can see, the governate-level correlations are lower than at the country level, and vary widely from one another. The correlation does not appear to be tied to the volume of either events or deaths. We note that Damascus appears to have a disproportionate number of events associated with it, potentially leading to its relatively low correlation; this may be a product of many events (e.g. involving the Syrian government, or simply reported from Damascus) being improperly coded there. Other localized processes and data issues may account for the lower correlation in several other governates, and may require further analysis.

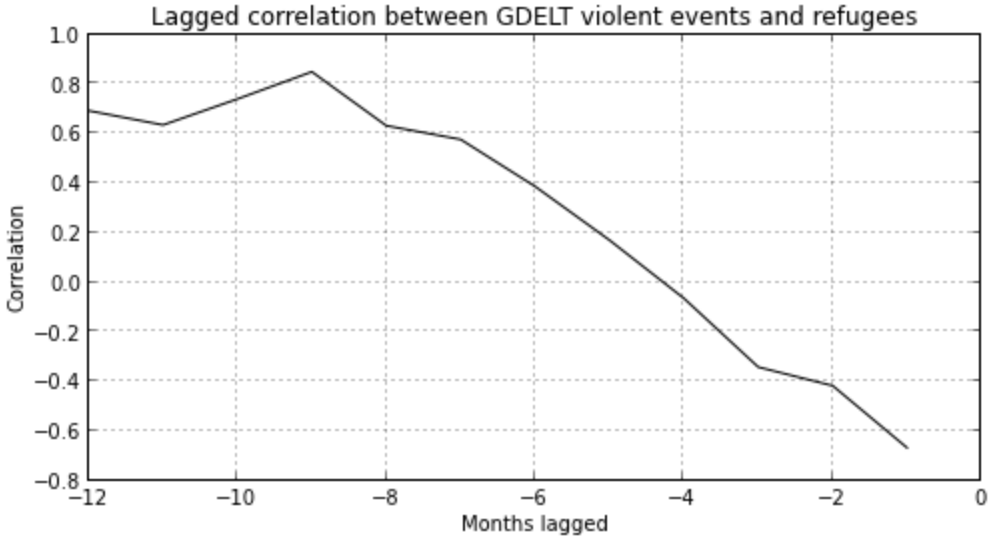
Governate	Total GDELT violent events	Total reported deaths	Time-series correlation
City Damascus	61169	5264	0.530087
Aleppo	26941	10006	0.716308
Homs	19544	10833	0.718755
Idleb	16739	7678	0.43966
Damascus Province	7914	15171	0.363328
Al Qunaytirah	6788	160	0.212933
Hama	5472	4814	-0.085631
Dara	4745	5481	-0.169697
Lattakia	3635	799	0.675003
Dayr Az Zor	3336	3836	0.585805
Tartous	2402	339	0.801152
Raqqa	1347	929	0.770112
Hassakeh	830	479	0.017108
As Suweida	411	45	0.395619

UN Refugee Data

Another data source is the number of refugees registered by the United Nations. Intuitively, an increase in violence seems likely to lead to an increase in the flow of refugees from Syria into neighboring countries. UNHCR provides data on registered refugees in total, as well as by country (though these reports do not seem to match; see Data Issues below), generally monthly or bimonthly. Unfortunately, the UN data available only begins in January of 2012, and thus does not overlap with the the first year of our data.



However, there are two lags involved between violence and refugees. The first is simply that it takes time for individuals to uproot and travel to a neighboring country; even if civilians leave their homes immediately upon a violent event as captured in GDELT, they will not arrive at the border instantly, much less register with the UN. Indeed, UN registration itself takes time, as evidenced by the UN's own report on the large volume of individuals awaiting registration (currently nearly 230,000) . To account for this, we lag the refugee time-series, correlating registered refugees with violence reported a set number of months previously. Happily, this provides much better overlap between the two time series as well.



As we can see, the correlation quickly increases as we begin to lag the refugee registration data,reaching levels comparable to what we observed for the fatalities series. Of course, different refugees will take different amounts of time to arrive at the various locations, and the time to become registered will vary as well. Additional work may attempt to identify

relationships between violence in particular areas and flows of refugees the nearby borders, and would take into account a better understanding of the actual UN registration process as well.

Data Quality

As we have already discussed, GDELT appears to exhibit evidence of media fatigue. There is also possible urban bias, introduced both due to the location of reporters as well as miscodings based on datelines and mentions of major cities (e.g. “several hours from Aleppo” being coded to Aleppo). The overcounting of events for Damascus may also be evidence of this. Actor coding appears to be less reliable than the event coding; additional analysis has shown that many events simply have ‘SYR’ as the actor; context suggests that this can variously refer to both government- and rebel-affiliated actions. We are developing tools and practices to disambiguate these issues, but they are still nascent.

The Syria Tracker project is crowdsourced, and thus much of its data cannot be independently verified. It’s own opposition affiliation means that it may undercount fatalities among pro-regime forces and supporters. Additionally, it is likely a source for many media accounts[11] that may in turn be GDELT inputs, creating potential for circularity. However, as we have seen, the divergence between the two indicates that this is unlikely to be a major issue.

The data provided by UNHCR only dates back to December, 2011, leaving us without a crucial year. The data records only registered refugees, though the website makes clear that tens of thousands of individuals are awaiting official registration. Additionally, there are separate datasets for the total number of registered refugees and for those in Turkey, Lebanon, Jordan, Iraq and Egypt. These country-level counts are recorded at different intervals, and do not consistently sum to the reported total. The cause of this discrepancy is not clear.

Possible Further Work

We have only scratched the surface of what can be done using the GDELT dataset in relation to Syria. Future work must go in two directions: additional data validation, and actual analysis and inference.

Data Validation

- Correlate GDELT with additional data sources (e.g. Violations Documentation Center[12]).
- More formal model of media fatigue, attempt to adjust event counts to account for it.
- Gain a better understanding of actor- and geo-coding and adjust for that.
    - The daily updates include a source URL for the originating media report; this can be used to help understand the coding system.

[11] http://www.humanitariantracker.org/#!news/c1j2a
[12] http://www.vdc-sy.info/index.php/en/about

- Experiment with additional faction detection methods.

Analysis

- Test spatio-temporal forecasting techniques for violent events
- Look for leading indicators and early-warning signs of major escalations
- Analyze violence in relation to regional demographics, key geographic features (e.g. major roads) and qualitative insights (e.g. legitimacy of local actors) and attempt to find or confirm relationships.
- Civilian displacement modeling: understand which events trigger civilians to evacuate and become IDPs/refugees, with an eye toward eventual forecasting and policy preparation.

Appendix: Governate-Level Time Series