Response Memo

Dear Erik, dear Will,

Thanks for your comments on the first version of our paper, and the decision to let us resubmit a revised version. In this memo, we respond to the comments raised by the Associate Editor (AE). We hope that you find this version to be an improvement over the previous one, and look forward to hearing from you. Please note that due to the issues currently surrounding GDELT, we have prepared a separate letter to the editors and the reviewers. We would be grateful if this letter could be distributed alongside the manuscript.

Sincerely

Jesse Hammond  
Nils B. Weidmann

*General comments*

We acknowledge that the previous version was not sufficiently precise when it came to the contribution we wanted to make. We have revised the ms. to make this clearer. The paper is now framed as an assessment of the distinctive feature of the “new” generation of machine-coded event datasets: geo-localization. The paper shows that the significant discrepancy of GDELT with other human-coded datasets is largely due to the fact that it seems to be placing events in the wrong locations. If we ignore location and collapse the data into time-series, the datasets actually correlate relatively well. We believe that this is an important message to scholars using machine-coded event data for geospatial analysis, but at the same time suggest an important dimension for future improvements of machine-coding techniques.

*Main comments from the AE*

* The AE felt that the ms insufficiently specifies “what we already know about content analysis”. We chose brevity in reviewing the existing literature due to the RaP word limit of 4,000 words, not due to “poor scholarly practice” as the AE suggests. We revised pp. 3-4 of the paper to reflect better our narrow focus on the new generation of geolocalized datasets, and their potential problems.
* The AE rightly points out that there are many potential reasons for the lack of disagreement between the datasets. In response to this, we have expanded our discussion of coding to briefly describe how GDELT, ACLED and GED turn sources into event records on pages 5-6. We also run a parallel set of analyses (Appendix D) where we subset the ACLED/GED data to only include sources also used by GDELT. Our findings are substantively identical.
* The AE asks for an explicit discussion of the coding rules both in the Introduction and the body of the paper. Due to the low word count, we don’t see how we can possibly fit this into the main paper, and thus left this part in the appendix.
* The AE pushes us to further examine the reasons for disagreement between the different datasets. For the results we present in the revised version of the paper, it becomes clear that one of the main reasons is differences in geolocalization. So, we cannot speak to all potential reasons for disagreement, but are able to identify one of the main ones, which has implications for the type of analysis to be performed with machine-coded dataset. We believe that in a short paper, there is not much room to go beyond this.
* The AE failed to find a definition of a “conflict case”. This definition can be found on p. XX, the coding rules are detailed in the appendix.
* The AE points out that ACLED and GED are not “true records” of conflict cases. We use these two datasets since they constitute the current standard in the geospatial analysis of violence. We believe that it is reasonable to assess how GDELT compares against this standard, even if this standard is not perfectly accurate itself.
* The AE asks us to expand on the finding that there is spatial variation when it comes to the fit between the datasets. In response to this point, we have revised the paper to focus more on an assessment of the spatial dimension (see general comment above).

Other issues

* The AE wonders about our use of the term “reliability”. This was misleading, and has been rephrased.
* The AE complains that we present machine coding as something new. This impression is not correct. On pp. XX, we explicitly describe the “old generation” of machine-coded datasets, and highlight what the “new generation” adds to this.
* The AE asks us to cite sources for the expectation of “urban bias”. This is now done on p. XX.