Reproducibility review: "Tracking Hurricane Dorian in GDELT and Twitter"

This report is part of the reproducibility review at the AGILE conference.

For more information see https://reproducible-agile.github.io/.

This document is published on OSF at https://osf.io/xs5yr/.

To cite this report use

> *Ostermann, F. O., and Nüst, D. (2020, July 12). Reproducibility review of: Tracking Hurricane Dorian in GDELT and Twitter. https://doi.org/10.17605/OSF.IO/XS5YR*

Reviewed paper

> *Owuor, Innocensia, Hochmair, Hartwig and Cvetojevic, Sreten: Tracking Hurricane Dorian in GDELT and Twitter. AGILE GiScience Ser., 1, 19. https://doi.org/10.5194/agile-giss-1-19-2020, 2020.*

Code repository: https://github.com/InnocensiaO/Tracking-Hurricane-Dorian-in-GDELT-and-Twitter

Summary

The authors do a commendable job in providing all code and all data that can be provided (given platform terms of service). The reproduction was initially made more difficult by the absence of a documentation that explains what the scripts are doing, and in which order they are to be run. While the paper's boxplots were successfully reproduced, the maps cannot be reproduced with the materials provided. Overall, the reproduction was thus partially successful.

Reproducibility reviewer notes

The materials on GitHub have an MIT license.

Data

Twitter data

The data collection and preprocessing is not reproducible, because the exact query is not given (study area?) and bot removal was conducted via external API. However, with all used Twitter IDs provided, I was able to hydrate 90% (with Hydrator v0.3) of the input data. This shows how useful and important it is to provide at least the Tweet IDs.

GDELT data

The query to recreate it is given, but it might cost a fee to access the data. If I understood correctly, the data might be downloaded if a new account is created, but then again preprocessing steps are missing.

The relevant GDELT event data is provided.

Hurricane tracks

This is linked to NOAA images, but data collection and preprocessing not reproducible. The relevant Hurricane data is provided.

Processing

ArcGIS Pro and RStudio were used to explore the data sets and scripts. The GitHub repository contains several data sets (tables, shapefiles) and R scripts, however, their purpose or lineage was not documented at the time of the reproduction and needed to be inferred through exploration and experimentation. The order in which the files have to be run was unclear. This has been addressed by additional documentation in the repository now, but we could not redo the review due to time constraints.

The R script needed adjustments for paths. They also contain git merge artefacts, e.g., "<<<<<<< HEAD", which needed to be removed before they can be run. The script Tweets_GDeltCountiesCorrelation.R has a missing library load (spatstat) and once added, still created an error in line 21 "Error in square(TweetTotal_GdeltTotal) : is.numeric(r) is not TRUE"

After manually fixing the paths to ones working on my system, I could source the script files

`Boxplots.r`

`Gdelt_DorianMedian.r`

`Gdelt_WilcoxMedianDifferences.r`

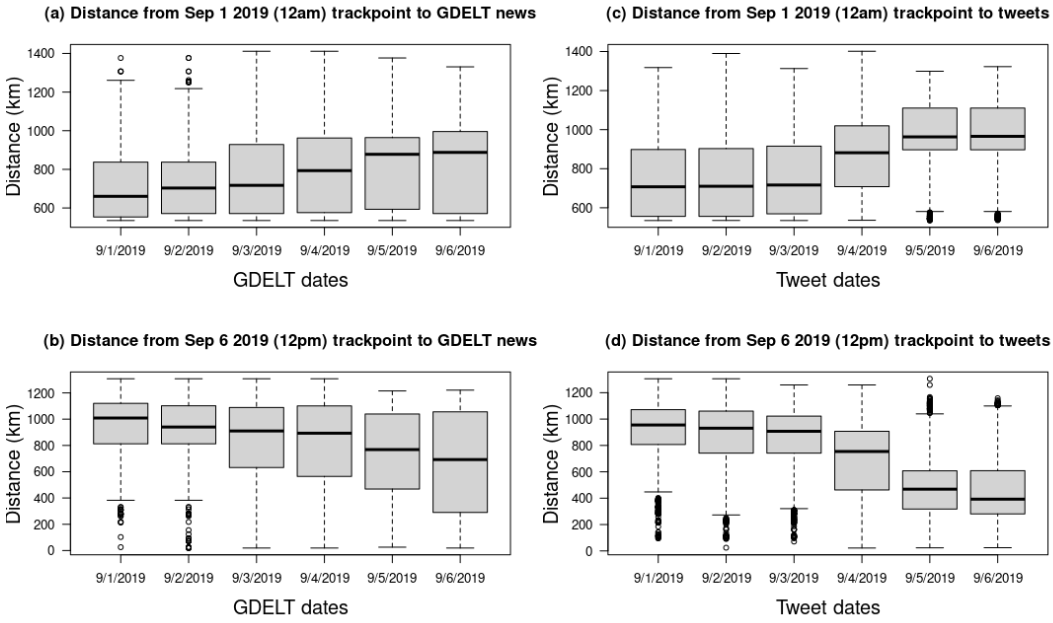
`Tweet_DorianMedian.r`

`Tweet_WilcoxMedianDifferences.r`

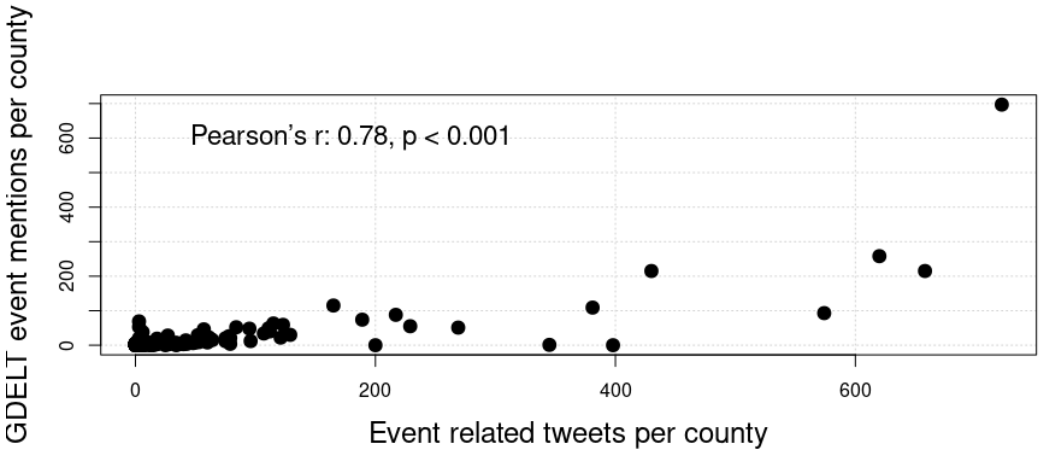
`Tweets_GdeltCountiesCorrelation.r` (only stepwise execution, skipping line causing an error)

Results

The maps seem impossible to recreate from data without clearer instructions as to how they were produced. A solution could be to provide the ArcGIS project files. The boxplots could be reproduced using the R script boxplots.r: It created a plot matching Figure 3 of the paper in labels, range and by visual inspection also matching data.



Tweets_GdeltCountiesCorrelation.r line 29 created a plot similar to Figure 6, while the remaining plots failed on my system.



It was not entirely clear at the time of the attempted reproduction what the supposed output of the other scripts is. The additional documentation provided in the meantime clarified this.

Some suggestions on improvements on the author's laudable efforts:

- Instead of using setwd() in a script, consider the {here} package for reading files from subdirectories

- Reusing the same variable name "tbv" all throughout the scripts is confusing, instead a clear link between code and paper (e.g. figures, tables) would clarify a lot.