Monthly Report on "An Extensible Model for Deduplication of the GDELT Events Database"

Andee Kaplan (andee.kaplan@colostate.edu; 832-526-7947)

November 2019

Progress and Plans

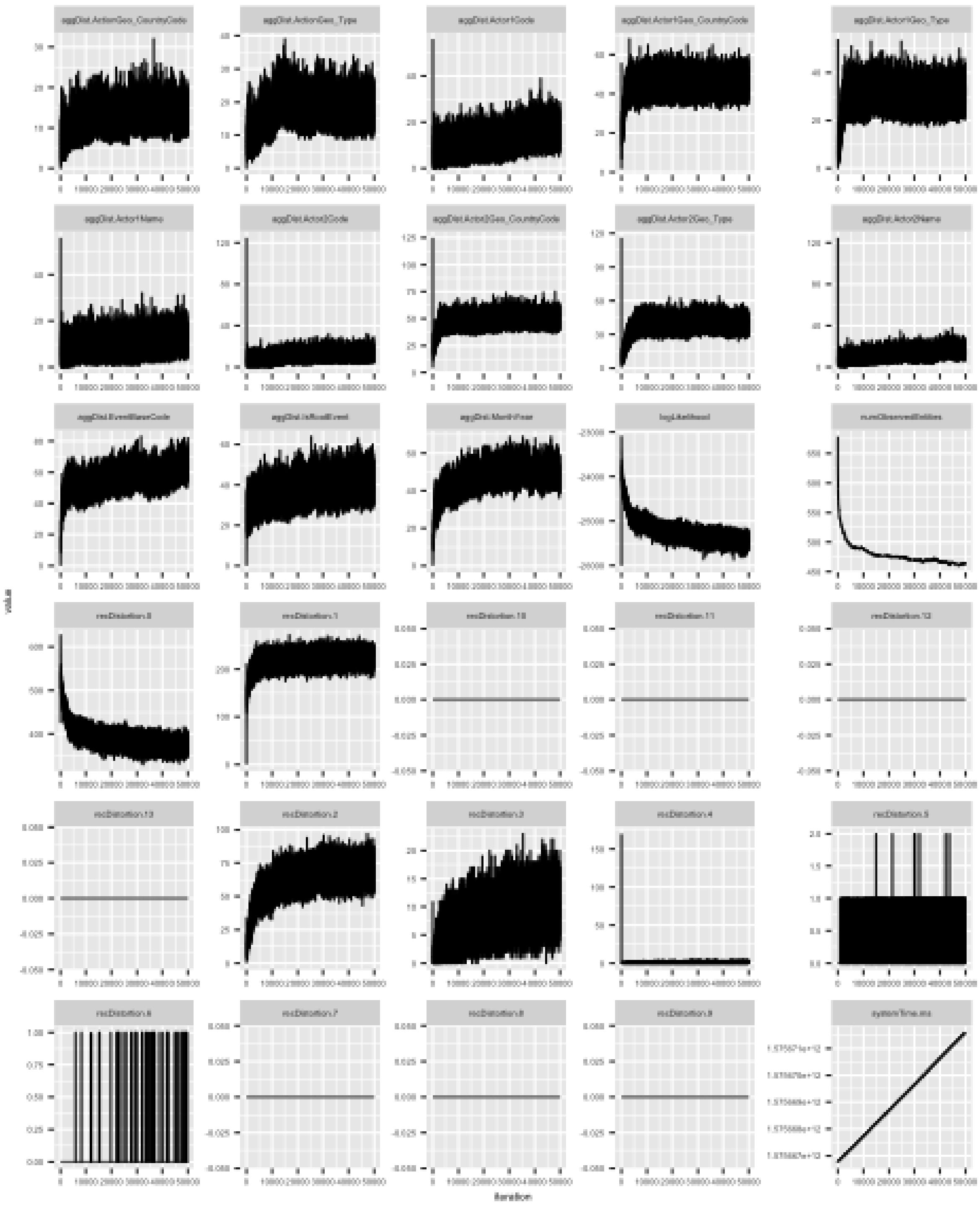
1. Description of progress made against each deliverable during the reporting period

   We have run a record linkage model (`dblink`) on a small portion of the GDELT data (West Africa events that occured in 1979) after blocking by year. The variables we have used for the linkage are: month of occurence, actors names and codes, event code, and geographic codes. We are currently inspecting the results before running the same model on the remaining years.

2. Brief description of significant results

   The figure below shows trace plots of 50,000 draws for different functions of the draws:

   - aggregate distance for each value,
   - log-likelihood function,
   - number of observed entities,
   - number of records with 1, 2, 3, 4 distortions, and
   - system time for running the record linkage model.



From the diagnostics, it looks like the model should be run longer to sample from the target and modifications to the variables should be investigated.

3. Planned activities for the following reporting period

   We will look at a string-concatenated virsion of the two actors, instead of actor 1 and actor 2 variables separated. This is because it is reasonable to assume these actors could become switched in order as the data is scraped from the news articles. Additionally, we will hand inspect a subset of the linkages to ensure reasonable results before running `dblink` on the entire West Africa GDELT dataset (blocked by year).

4. List of any LAS-funded trips during reporting period, with description of work presented

   None.

5. Description of any significant meetings/events held (e.g., a focused discovery activity, LAS Symposium) or conducted related to this task

   Casey and I have continued to meet weekly.

Issues

1. Progress on/results obtained related to previously identified problem areas

   Casey's surgery was a success and he was able to resume work quite quickly with no impact to the schedule. We were also able to navigate the use of CSU's cluster and can run `dblink` (a spark scala package) on the cluster.

2. Proposed significant changes to your methodology, goals, milestones, or deliverables

   We are moving forward with using the `dblink` model on the West Africa GDELT data, blocked by year.

3. New challenges affecting technical performance or schedule, with background

   None.

4. Has anything happened to impact your anticipated schedule?

   No.