From: Andreas Beger
To: V-Dem Institute
Date: 15 April 2021
Re: DemSpaces and PART 2021 forecast updates

Both PART and DemSpaces have updated forecasts covering 2021–2022, and these have been integrated into the respective dashboard apps. This note summarizes in more detail the work that went into the updates and some changes that I made during the course of producing the updated data and forecasts.

In general, I tried to streamline code, data, and the overall updating process so that the work flow is easier to follow and more transparent, and so that future updates will require less effort.

Common changes to both projects include:

- All the code and data to produce the forecasts and dashboards are now on V-Dem GitHub repositories, at https://github.com/vdeminstitute/part, https://github.com/vdeminstitute/demspaces, and https://github.com/vdeminstitute/demspacesR (support R package for DemSpaces).
- Both repos/projects have `2021-update` sub-folders with material relevant to this data update.
- I added data cleaning scripts for 6 external data sources (ACD, GDP, population, infant mortality, Powell & Thyne coups, and state age/time since independence). I had previously just copied over updated data I had from other projects; now the code needed to update these external sources is present as well.
- I added CHANGELOD and UPDATING notes in both projects to track changes and give an overview of the updating process, respectively. I also tried to mark all places that require manual changes, e.g. changing references to specific years in the dashboard map legends, with the phrase "UPDATING:" so that they can be more easily found.
- Improving the dashboards: I went over both dashboards and generally tried to reduce dependence on big packages (e.g. tidyverse), pre-calculate a couple more tables that previously were calculated each time someone tried to use a dashboard, and remove redundant columns in data tables to reduce their size. I did profiling for both and couldn't find any other obvious bottlenecks that are not inherently related to Shiny. The size of the dashboard apps are smaller now (e.g. for DemSpaces it's half of the previous size now), and anecdotally the dashboards start up quicker; I'm not sure whether the calculations when you click various things are faster or not.

Below are more project-specific changes.

DemSpaces

- One big substantive change is that I reduced the number of external data sources and features dramatically, cutting the number of colums in the final data in half. To do this, I calculated variable importance scores from the random forest forecasting models to assess which sets of variables contribute to the models' predictive accuracy. The impetus was to reduce the number of columns in the data and see whether some external data sets could be eliminated, in order to speed up the time required to run the forecasting models and also make the updating process quicker. Based on the results I dropped 3 of 7 external data sources as well as some V-Dem variable transformations, leading to final merged data that went from more than 450 columns to ~230. This cut the time needed to run the forecast models by more than half, and it seems to have resulted actually in a slight increase in test forecast accuracy. More details on this are at https://github.com/vdeminstitute/demspaces/blob/main/2021-update/variable-importance.md.
- In addition to V-Dem, I updated 4 external data sources (Powell & Thyne coups, state age from the Gleditsch & War state list, GDP and infant mortality from WDI, and UN population data).
- Regarding the dashboard changes I discussed above, the DemSpaces dashboard tarball for example has gone down from 1.3Mb to 0.7MB as a result of streamlining the data that it uses.
- Fixed two small bugs in the DemSpaces dashbard (GitHub issues #2 and #5).
- Accuracy report for IRI. This is in the `2021-update` subfolder in the project repo on GitHub.

PART

- In addition to V-Dem and the 4 external data sources I updated for DemSpaces, I also added updated ACD conflict data. EPR also goes into the merge data but has not been updated, so I re-used the existing data with an additional year lag.
- The PART forecasts are based on an ensemble of 3 other models, all of which were self-tuning hyperparameters using cross-validation. As a result running the forecasting models used to take something on the order of 36 hours. I did fairly extensive tuning experiments for those models instead, in order to idenfity "good enough" fixed hyparameters and eliminate all the cross-validation. The models now run in under 1 hour, and test forecast accuracy is similar (actually slightly higher).
- The PART dashboard tarball went from 1.5MB to 0.7MB in size as a result of the streamlining I mentioned above. One important change that should speed it up a lot is that I simplified the map data it uses, similar to what Laura M. did for the DemSpaces dashboard when we developed it.