**NWS Project Notes**

## 2/8/2022

ACTIONS:

* Used R, the Google Trends API and Gtrends package in R to pull data on weather related search terms along with placebos.
* Built a coding structure that allowed within and across state comparison for entire US from 2005-2017, excluding Puerto Rico and the US Virgin Islands.
* BlairRAWork->NWS Forecast Dropbox contains 2 files:
  + Trends\_WithinState.csv
  + Trends\_AcrossState.csv

CODEBOOK:

|  |  |
| --- | --- |
| **Variables in Trends\_WithinState.csv and Trends\_AcrossState.csv** | **Description** |
| Geo/ State/ Abrev/ FIPS | Various state identifiers |
| Date/year/month | Search month and year identifiers. 2/1/2005 indicates to the entire of month of February 2005 |
| hits\_Weather | Google trends index number (0-100) for the search term “weather” |
| hits\_Forecast | Google trends index number (0-100) for the search term “forecast” |
| hits\_Temperature | Google trends index number (0-100) for the search term “temperature” |
| hits\_NOAA | Google trends index number (0-100) for the search term “noaa” |
| hits\_Dogs | Google trends index number (0-100) for the search term “dogs”. Intended as a uncorrelated placebo term. |
| hits\_Headache | Google trends index number (0-100) for the search term “headache”. Intended as a uncorrelated placebo term. |

**Trends\_WithinState.csv:** contains the trends, or “hits”, for each search term’s (see codebook) popularity as a fraction of total searches in each state, with the largest normalized to = 100. Here, every state in the data will have one month per search term where (search)\_hits=100, and the remainder of the monthly within-state, series describes variation relative to 100. For example, the figures below show Arizona search history for “forecast” in red and Alaska history for the same search term “forecast” in blue. Note that the index =100 occurs in early 2010 for Arizona and early 2017 for Alaska. This provides a good deal of within-state variation but tells us little about across state differences.

Graphical user interface, chart, line chart

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

**Trends\_AcrossState.csv:** contains the trends, or “hits”, for each search term (see codebook) over time as a fraction of that term’s highest state-months compared to other US state-months . This means that only one state-month in the US per search term will have an index=100, and the remainder of the state-month series describes variation relative to 100. For example, the figure below shows Arizona search history for “forecast” in red and Alaska history for the same search term “forecast” in blue as before, but here we see that compared to Alaska’s search history for “forecast”, Arizona’s search history has been scaled downward in relative terms (Alaska 9/2016 = 100 for this search term).

Graphical user interface, chart, line chart

Description automatically generated

The figure above is smoothed, but we can zoom in to see that Arizona’s index in Trends\_AcrossState.csv is the same shape as Trends\_WithinState.csv before, but looking at the y axis, has been scaled down.

Graphical user interface, chart, application, line chart

Description automatically generated

**Summary and Technical notes:** Working with and understanding google trends can be challenging because their scaling, term and volume restrictions.

One issue is Google’s [trends website](https://trends.google.com/trends/?geo=US) does not provide bulk downloads of single state timeseries required to produce a country wide data set. If your interested in 5 different search terms over 50 states over time it would require 250 individual downloads. Instead to generate Trends\_WithinState.csv I’ve created a single function in R, based on Philippe Massicotte R package “[gtrendsR](https://github.com/PMassicotte/gtrendsR)”, that produces state-by-state data by running 51 individual queries at a time per search and can be found in our Dropbox or [Github](https://github.com/blairlo/GoogleTrendsWeatherExtract).

Another big hurdle, [documented by others](https://towardsdatascience.com/using-google-trends-at-scale-1c8b902b6bfa#:~:text=Currently%2C%20the%20public%2Dfacing%20Google,of%20all%20the%20major%20candidates.), is that regardless of using Google’s website or an API based package in R or python, Google limits the number of time series you can request at once to 5. This means 5 search terms, 5 geographies, or any combination to equal 5. While we might first think to download whatever we need in batches, we must remember that each search is scaled to the highest volume series in the query. Each 5 would be comparable within query, but not across query in any consistent way.

One way around this to produce the Trends\_AcrossState.csv is ensure that each query of 5 (in our case 5 states), has the highest search volume state in it (so batched of 4 states, indexed to the baseline state). To find the appropriate baseline, I request each search term (weather, forecast etc.) in a US wide query for 2005-2017. While this won’t give us state time series, it tells us the state with the highest aggregate search volume over that period—not nessicary the highest state-month, but a good first guess. I’ve created a across-state function in the same R markdown, which accepts this baseline state as an argument passed to 14 separate state queries. I check each series and if multiple state-months =100 I select the next state as a baseline, rerun, recheck and so on until only one state-month=100.