Hacking Google Trends to Predict Voter Turnout

Aaron DeVera, Joe O’Brien, & Dr. Robert Hume

Introduction:

Americans’ access to the internet and their use of search engines like Google has grown tremendously over the past decade. As the common facets of everyday life are performed more frequently on the internet, producers and marketing research firms have found search trend data, like the data provided by Google Trends, to be helpful in forecasting sales and consumer preferences (Boone, Ganesham, & Hicks, 2015). Even outside of market research, Google Trends are indicative of public interest in a wide variety of topics. They have a demonstrated statistical relationship with some social and economic trends (Choi & Varian, 2012) and may even help predict certain behavior (Silver, 2015). Since Google search trends provide insight into levels of public interest and are statistically related to certain behaviors they might also provide a window into public interest in voting in a given election as well as the likelihood of individuals in a certain area to do so. Therefore, we hypothesize that the amount of Google searches for the location of polling places is positively correlated to voter turnout, given the parameters of time period and geolocation..

Methodology:

•Collect total turnout figures for each November general election in all 50 states and Washington, D.C. from 2004 to 2014 from The United States Election Project , check it against official state figures from each Department of State or Board of Elections website

•Since there is no public API (Application Programming Interface) for Google Trends, we used aggressive automated collection methods to obtain Google Trends data on every state between the years 2004 to 2014 for people entering the query “where to vote”.

- We assumed the act of Google searching “where to vote” is a popular method used by voters to discover the location of their nearest precinct or polling place

•Compare relative behavior of Google's search trends against the relative behavior of the total turnout figures, first using R Studio and then using Python to generate our automated tests and compilation.

- We normalized data from both datasets by measuring the data points’ standard deviations from the mean to accelerate comparable analysis in a variety of tests.
- Strong correlations=R-Values between 1.0 & 0.5, moderate correlations= between 0.3 & 0.5, weak correlations= R-Values between 0.3 & 0.1, no correlation = R-Values between 0 & 0.1
- Statistical significance = P-Value $\leq$ .05

Results:

•National Aggregate:

- R-Val = 0.84
- P-Val=0.036
- 31 states showed strong correlations
- 18 were statistically significant while 16 were not

•5 states showed moderate & statistically insignificant correlations

•6 states showed weak & statistically insignificant correlations

Results (Continued):

- 2 states showed no correlation
- 2 states lacked search trend data & two more displayed errors in correlation and P-Values

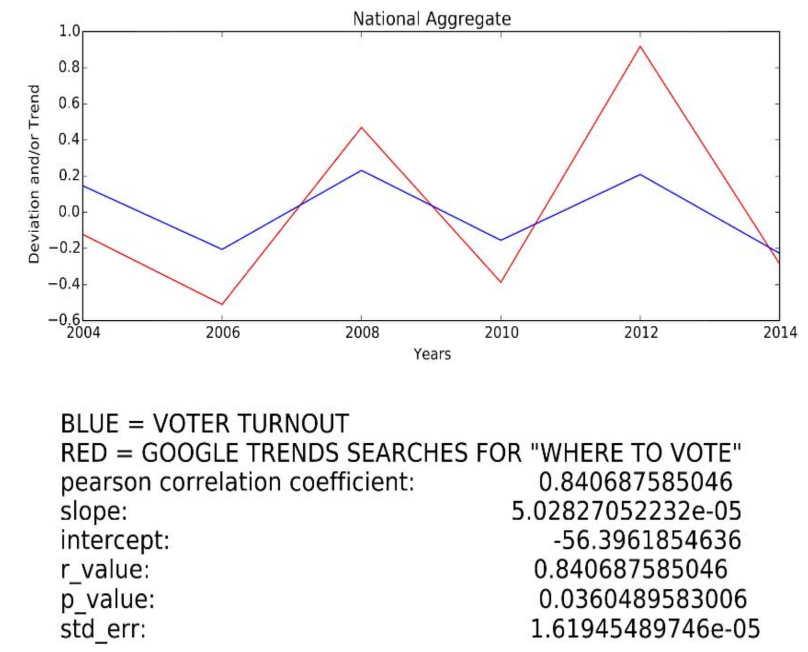


Figure 1: National Aggregate

1	State	Correlation (R-Value)	Measure of significance (P-Value)
2	Alabama	0.812	0.050
3	Arizona	0.929	0.007
4	California	0.851	0.031
5	District of Columnbia	0.908	0.012
6	Florida	0.977	0.001
7	Georgia	0.816	0.048
8	Illinois	0.951	0.004
9	Kentucky	0.975	0.001
10	Maryland	0.908	0.012
11	Massachusetts	0.942	0.005
12	Michigan	0.918	0.010
13	Minnesota	0.977	0.000
14	Missouri	0.889	0.018
15	New York	0.959	0.002
16	North Carolina	0.915	0.011
17	Ohio	0.932	0.006
18	Pennsylvania	0.935	0.006
19	Washington	0.905	0.010

Figure 2: Statistically Significant Strong Correlations

1	State	Correlation (R-Value)	Measure of significance (P-Value)
2	Alaska	0.665	0.150
3	Hawaii	0.659	0.155
4	Idaho	0.627	0.183
5	Indiana	0.802	0.055
6	Kansas	0.548	0.260
7	Louisiana	0.551	0.257
8	Nebraska	0.595	0.212
9	Nevada	0.708	0.115
10	New Hampshire	0.653	0.160
11	New Jersey	0.627	0.183
12	North Dakota	0.581	0.226
13	South Carolina	0.759	0.080
14	Tennessee	0.709	0.110
15	Texas	0.724	0.100
16	Virginia	0.638	0.170
17	Wisconsin	0.662	0.150

Figure 3: Statistically Insignificant Strong Correlations

1	State	Correlation (R-Value)	Measure of significance (P-Value)
2	Arkasas	0.358	0.486
3	Iowa	0.433	0.391
4	Maine	0.468	0.349
5	Mississippi	0.372	0.468
6	Rhode Island	0.338	0.510

Figure 4:Statistically Insignificant Moderate Correlations

1	State	Correlation (R-Value)	Measure of significance (P-Value)
2	Montana	0.206	0.695
3	New Mexico	0.155	0.770
4	Oklahoma	0.297	0.567
5	South Dakota	0.189	0.710
6	Utah	0.167	0.750
7	West Virginia	0.231	0.660

Figure 5: Statistically Insignificant Weak Correlations

1	State	Correlation (R-Value)	Measure of significance (P-Value)
2	Delaware	0.063	0.905
3	Oregon	0.092	0.862

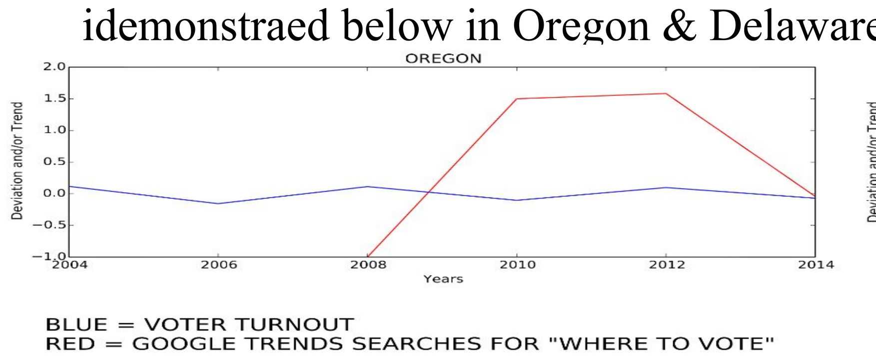
Figure 6: No Correlation

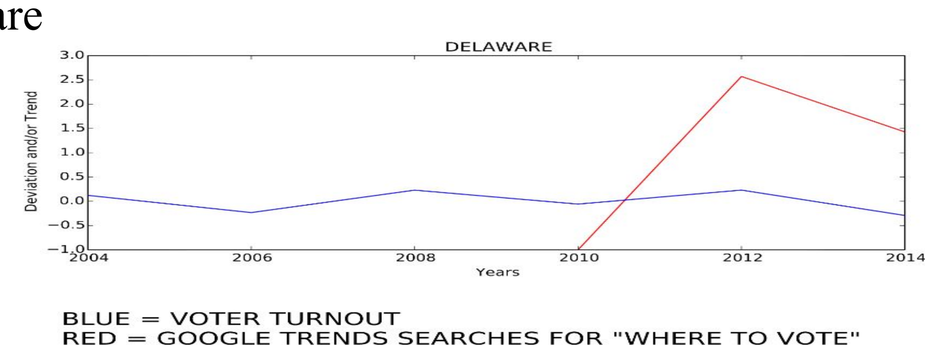
1	State	Correlation (R-Value)	Measure of significance (P-Value)
2	Colorado	5.168	0.999
3	Connecticut	0.992	9.923
4	Vermont	N/A	N/A
5	Wyoming	N/A	N/A

Figure 7: Problem States

Issues:

- 2 problem states (Vermont & Wyoming) lacked any search trend data while another two experienced an error during correlation calculations (Connecticut with P-Value & Colorado, with R-Value)
- Incomplete data sets for search terms obstructed our ability to test the correlation as States that had this issue demonstrated lower correlations and higher P-values as idemonstraed below in Oregon & Delaware





Conclusions:

- Google Trends have a strong statistically significant correlation to total voter turnout nationally
- The correlation varies due to population size and access to internet at a given time

Conclusions (Continued):

•Predictive value of these correlations will be stronger in areas with a highly concentrated population and easy access to the internet

•As access to the internet increases over time, states that showed weaker and insignificant correlations should show stronger and significant ones

•Google Trend and total turnout data should similarly increase in its predictive value over time in areas with less access to internet, showing promise for the future

Suggestions for Future Research:

•Integrate census data, county-level search data and turnout figures into a predictive model

•Investigate Google protocol for tracking search terms

•Trace history of internet connectivity across states, regions, and country

Bibliography:

Aceves, R. (2014). *Predicting Change in County-Level Presidential Election Voter Turnout using Data Mining Methods* (Masters Dissertation). Retrieved from Central Connecticut State Thesis, Dissertation, and Special Projects Collection. http://content.library.ccsu.edu/cdm/ref/collection/ccsutheses/id/2006.

Boone, T., Ganesham, R., & Hicks, R.L. (2015). Incorporating Google Trends Data Into Sales Forecasting. *Foresight: The International Journal of Applied Forecasting*, *38*, 9-14.

Choi, H. & Varian, H. (2012). Predicting the Present with Google Trends. *The Economic Record, 88* (Special Issue), 2-9.

Google Trends - Web Search interest - Worldwide, 2004 - present. (n.d.).From http://www.google.com/trends/explore.

McDonald, M.P. (2016, 20 January.) 1980-2014 November General Election State Turnout Rates. *United States Elections Project*. Retrieved From https://docs.google.com/spreadsheets/d/1or-N33CpOZYQ1UfZo0h8yGPSyz0Db-xjmZOXg3VJi-Q/edit#gid=1670431880.

McGrath, M. (2012). Election Reform and Voter Turnout: A Review of the History. *National Civic Review*, *101* (3), 38-43.

Silver, N. (2015, 21 Apr.) How Google Searches Can Predict Hockey Ticket Sales. *FiveThirtyEight*. Retrieved From http://fivethirtyeight.com/

datalab/how-google-searches-can-predict-hockey-ticket-sales/.

Acknowledgements:

We thank our faculty mentor, Dr. Robert Hume, for his input and support.