Hacking Google Trends to Predict Voter Turnout

Aaron Stephen DeVera

adevera@fordham.edu

@aaronsdevera

Fordham College Rose Hill

Economics Major, Computer Science Minor

Joseph Michael O'Brien

jobrien81@fordham.edu

@_JoeyO

Fordham College Rose Hill

Political Science & History Major

Acknowledgements: We would like to thank Benjamin Shull and Jason Almeida for their material support.

Hacking Google Trends to Predict Voter Turnout

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. We next discuss the complexity of predicting voter turnout in political science literature, our methodology, and finally test our hypothesis within the state of New Jersey. After testing this hypothesis, we conclude with a discussion of project implications.

Literature Review

Voter turnout is one of the most complex and important areas of study in political science. Factors like the weather and the perceived competitiveness of an election as well as closelyrelated social variables ranging from, but by no means limited to, one's race to whether or not they own a home may influence an individual's decision to go to the polls on Election Day (McGrath, 2012). This wide range of factors makes it extremely difficult to predict or manipulate levels of voter turnout. This poses a problem for the Democratic and Republican parties which have poured millions of dollars into constructing proprietary turnout forecasts to help them better deploy their limited resources (Aceves, 2014). These proprietary models are the most common predictive models and remain closely-guarded secrets by their respective owners, which has limited the ability of academics to forecast turnout. Aceves' 2014 attempt was a notable exception. It retroactively predicted county-level turnout for the 2008 and 2012 U.S. presidential elections with relative accuracy from a wide variety of climate, socio-economic, and demographic data. Interestingly, the model also incorporated Google Trends data, which seems to have increased its accuracy considerably (Aceves, 2014). Inspired by his attempt and Nate Silver's (2015) correlation tests that demonstrated a link between Google Trends and hockey ticket sales, the goal of this experiment is to measure whether people using Google search to find their nearest polling location had any influence on voter turnout.

Methodology

We selected a single state to serve as a test subject in order to appreciate the possibility of a pattern before we conduct data mining and analysis operations on all of our datasets for all fifty states. We tested our hypothesis in New Jersey because we wanted a contained environment with diverse demographics and widespread internet access. The first step was to collect the total turnout figures from each November general election in the state from 2004 to 2014. The November general elections provided a wide data set that corresponded with the increased use of google over the past decade and, unlike non-general elections, had positions that were decided on by a statewide electorate. This made their total turnout figures more representative of the entire state's motivation to vote, the dependent variable we want to observer These total turnout figures for the general election were derived from the official results compiled by New Jersey's Department of State available on its website. The total turnout numbers were then manually organized into a spreadsheet and checked for accuracy by comparing them against the sums of all 21 counties' total votes in each election; we encountered no problems with the reported state totals. With this phase complete, we were ready to acquire the Google Trends data and run our statistical tests.

Google Trends shows popular searches that people are asking Google web search service. It filters on the parameters of geolocation and time, which allowed us to access an approximation of what people were searching in a given area at a given time. While the Google Trends site is public, the software and code that runs the site is not. Unlike many other Google services, there is no public API (Application Programming Interface) that allows developers to build apps using the Trends data. Using aggressive automated collection methods, we collected Google Trends data on every state between the years 2004 to 2014 for people entering the query "where to vote". We assume 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.

Results and Discussion

After we acquired the Google trends data for each state from 2004 to 2014 of people searching "where to vote" we compared the relative behavior of people's Google's searches against the relative behavior of voter turnout reflected in the total turnout figures we collected. This was first done in R Studio, but our automated tests and compilation for publication was generated in Python. We normalized the data from both datasets by measuring the data points' standard deviations from the mean to accelerate comparable analysis in a variety of tests.

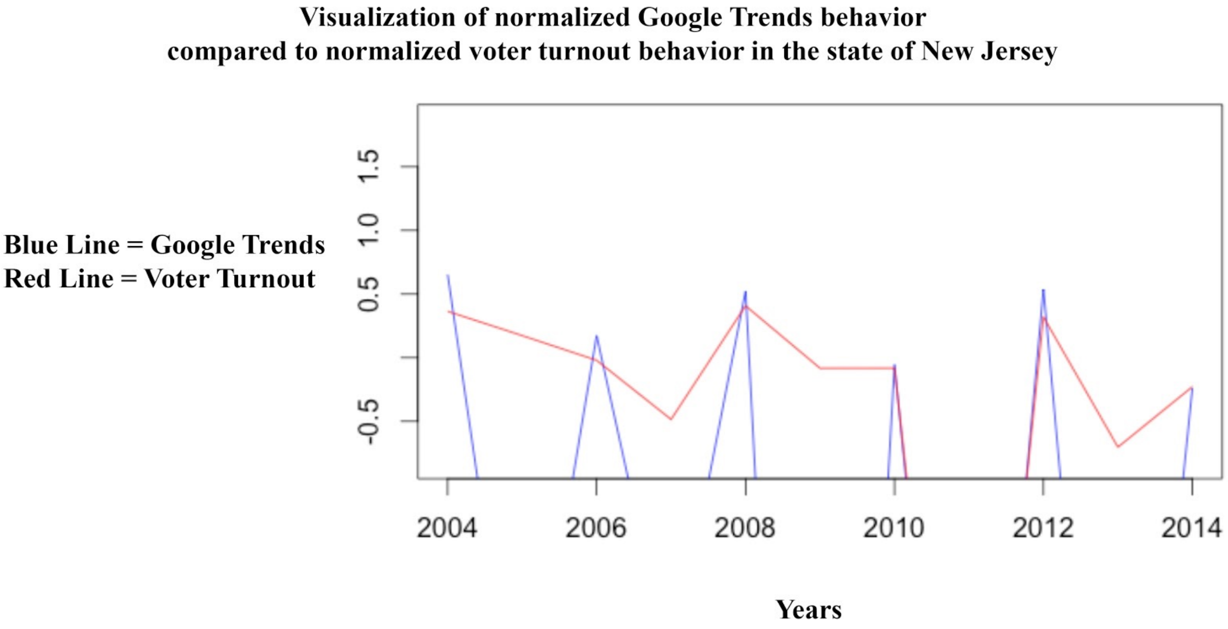

Figure 1. A visualization of voter turnout and Google search behavior in New Jersey from 2004 to 2014.

Our results from the New Jersey data showed preliminary signs of correlation. By the Pearson correlation test, the data is correlated to 80.97%. By testing our linear model, we saw that the R-squared, a separate indicator used for correlation and the model's goodness of fit is 0.657. We deem this test significant due to our low standard error, which is 0.2522%. A graph of the Google Trends search data compared to the voter turnout data in the state of New Jersey demonstrates a possible dependency (see Figure 1).

These results show promise for similar correlations in other state data. If other states show a similar pattern to New Jersey's, there could be a national correlational trend between voting-related Google search queries and voter participation in elections. Currently with only the New Jersey data, there is a possibility of that the Google searches might be influenced by the increased adoption of Google searching over time. While the behavior we have observed doesn't show signs of this influencing our result, additional data from more states will ensure the strength of the test and the accuracy of our model. Finally, our results from New Jersey alone can be developed into a model for predicting voter turnout in future elections. Testing and analysis on additional states' data will improve our ability to predict voter turnout in such a model. The New Jersey test subject data has strongly suggested the existence of a pattern and compels us to continue data mining, parsing, and analysis operations on all states. This is an ongoing project. and we hope to continue to communicate with the political science community.

References

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/.