Article 1: Construct Validity and Correspondence of Google Trends

Kelsey Gonzalez

18 February, 2022

Contents

1 Abstract

Possible titles Digital Trace Data as Social Indicators: Indicators of Attention Rather Than Support

Keywords:

2 Intro

New big data sources have led to vast possibilities for social science research because they are bigger, cheaper, and already available (King 2011; Lazer et al. 2009; Salganik 2017). Before overenthusiastically embracing these sources into our workflows, social scientists must clearly establish parameters under which these data sources could be operationalized (Bail 2014; Lazer et al. 2014). As prior research outlined, the "quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data"

$$@lazerParableGoogleFlu2014, p.1203$$

. Building on this prior research outlining these various issues with big data (boyd and Crawford 2012; Lazer 2015), this paper tests the construct validity of Google Search Trends as an indicator of three different cases, namely cultural attitudes, disease prevalence and voting behavior. These three cases will be tested using cultural indicators from the NORC General Social Survey, United States county-level suicide rates from National Center for Health Statistics, US cases rates of Covid-19 from The New York Times, historical US Presidential Election results, and the American National Election Survey. Data will be analyzed against corresponding operationalized Google Trends longitudinal data using Pearson's r for pairwise correlations, testing for the strength of relationships between the Google Trend indicator and the respective comparison indicator. This paper will contribute to the creation of methodological norms and standards of how to use Google Trends as a big data source for societal research and serve as a critical inquiry into the adoption of big data without a critical eye for the ecological validity of the sources.

With the expansion of big data, some research has shown extremely innovative methods that lead to groundbreaking results that are shown to be reliable. As an example, Blumenstock, Cadamuro, and On (2015) use county-level cell-phone records to construct the distribution of poverty and wealth in Rwanda, a country where national surveys and censes are rare and costly. However, Blumenstock et al. (2015) go to great lengths to demonstrate that their operationalization of the cell phone data creates a reliable and valid construct; few social science papers utilizing big data dig into the construct validity of their metrics to this extent and even fewer publications focus on methodological guidelines of how to use sources of big data (Asseo et al. 2020; Stiles and Grogan-Myers 2018). However, research has shown the small adjustments to an algorithm or metric may void any research insight we are able to pull from such data (Lazer et al. 2014). Because of this, I propose a methodological validation study of the Google Search Trends source of big data to investigate how it is advisable to utilize this data in social scientific research.

I will use three categorizations of ways I propose Google Trends could be operationalized for social scientific usage. First, I'll test Google Trends as an operationalization of cultural attitudes with the General Social Survey. After Bail (2014)'s call for cultural sociologists to utilize the ever-expanding world of big data, Google trends as a data source began appearing in sociological and social science research. From research on mass shootings and firearms (Brownstein, Nahari, and Reis 2020; Semenza and Bernau 2020), protest and anti-Muslim sentiment (Bail, Merhout, and Ding 2018; Barrie 2020; Gross and Mann 2017), to analyzing country-level changes in social perception (Reyes, Majluf, and Ibáñez 2018), Google search trends are a new and innovative indicator of cultural interest. Extending into social networks and culture, Bail, Brown, and Wimmer (2019) even used Google trends to measure how culture spreads around the globe.

Google Search Trends have also been used continuously in estimations of disease prevalence and population health in journals like the Journal of Medical Internet Research. While much of this research has focused on the Covid-19 pandemic (Jimenez et al. 2020b, 2020a; Lim et al. 2020; Mavragani and Gkillas 2020; Nguyen et al. 2020; Todorova, Tsankova, and Ermenlieva 2021) , other research has investigated Google Trends as an indicator of wellbeing (Brodeur et al. 2021; Carpi et al. 2020; Du et al. 2020), suicidality (Burnett, Eapen, and Lin 2020), vaccination uptake (Dalum Hansen, Lioma, and Mølbak 2016), obesity (Sarigul and Rui 2014), and even insomnia (Zitting et al. 2020), to cover a few examples. For a partial review of other utilizations, see Nuti et al. (2014). According to Jaidka et al. (2021), the majority of studies profess a correlation of $> .70$, "demonstrating the vast potential of Google Search as a proxy for monitoring population health" (p. 3) based on assumptions that individuals search because of self-diagnosis and to identify possible courses of treatment (De Choudhury, Morris, and White 2014).

Various sources have also used Google Trends as a way to forecast political elections and political attitudes (Wolf 2018). For instance, Swearingen and Ripberger (2014) investigate how U.S. Senate Elections relate to attention measured by search traffic. Prado-Román, Gómez-Martínez, and Orden-Cruz (2020) compare how Google Search trends are able to predict presidential election results in both the United States and Canada. Finally, the OECD Development Centre is investigating how Google data can help elucidate governments' approval in Latin America (Montoya et al. 2020).

Research Question - How can we operationalize Google Search Trends as a valid indicator for uses in social science research?

https://journals.sagepub.com/doi/10.1177/0894439316631043 § What constructs might google trends capture and not capture well? Capture attention but not attitudes

Also Asseo - "we assumed that media coverage may potentially decouple the search popularity from the number of cases, since searches would result not only from self-symptoms, but also from interest elicited by media coverage."

3 Research Methodology

To investigate the construct and criterion validity of the use of Google Trends in these three areas, I gathered geo-located social science data across multiple sources to address the three areas of inquiry for this paper. Table **??** outlines which data sources are used for this project and which trends are matched to each source.

Behaviors and Attitudes			
General Social Survey	Cross-Sectional	2010 - 2020	NA
Vaccine Hesitancy for COVID-19	Cross-Sectional	March 3 – 15, 2021	covid conspiracy', 'COVID-19 vaccine', 'Coronavirus', 'Covid-19'
Mask-Wearing Survey Data	Cross-Sectional	July 2 - 14, 2020	'Face Mask', 'Mask', 'Cloth Face Mask'
Health			
Covid Rates	Longitudinal	Every Monday, 2020 - 2021	'Covid-19', 'Coronavirus', 'Taste Loss', 'Smell Loss'
County Suicide Rates	Longitudinal	Yearly 2010-2020	'Suicide', 'Depression',' Suicide Hotline'
Political			
American National Election Survey	Cross-Sectional	2020	NA
Presidential Election Results	Cross Sectional	2016 & 2020	'Hilary Clinton', 'Donald Trump', 'Joe Biden'

3.1 Measures

3.1.1 Google Trends

I use Google search trends (Google 2022), over the study period across individual designated media markets areas (DMAs), a nonoverlapping aggregation of U.S. counties to 210 media markets based on similar population clusters (Sood 2016).

To investigate the rate of *searching* for norms online in both cases, I use the following search topics: 'Social Distancing' (2020, stay-at-home case only), 'Covid-19 Vaccine' (2021, vaccination uptake case only)

and Covid Conspiracy (2020-2021, both cases). Search topics are a more robust measurement than a single search term: topics are aggregations of the rates of multiple, highly correlated search terms together into a cohesive topic. For example, while 'Beyoncé,' 'Beyonce' and 'beyonce knowles' are all separate search terms, 'Beyoncé Knowles' encompasses all of these into a single search topic. While the data is originally on a scale of 0 to 100, with 100 being the maximum search popularity out of all DMAs, Google Search Trend are now only available cross-sectionally (a single time period across a geography) or time-series (a single geo-location across time). To remedy this and build a longitudinal dataset of each search topic, I follow the method proposed in Park, Kwak, and An (n.d.) (p. 5). This method involves building a dataset of unscaled cross-sectional values, selecting a DMA to use to establish the rescaling ratio (I use 'Los Angeles CA'), and then finding the time-series values for the one DMA. To find the rescaling ratio for each week in the time-series, you divide the time-series value for each week by the cross-sectional value for each week, resulting in a rescaling vector to be used for all weeks in the dataset across geographies. To rescale each longitudinal value, multiply the respective week's rescaling ratio by the cross-sectional value. Rescaled longitudinal data was compared against time-series data for multiple test counties and was equivalent. For a more in-depth explanation of this procedure, see Park et al. (n.d.) (p. 5).

3.1.2 attitudinal

(NORC at the University of Chicago 2020)"

- Vaccine Hesitancy (Center for Disease Control and Prevention 2021)

(The New York Times 2020)

- Mask Usage New York Times Survey

3.1.3 health

For disease prevalence, I will use United States county-level suicide rates from National Center for Health Statistics, US cases rates of Covid-19 from The New York Times, and obesity rates to compare search trends to observable data. • Covid Rates (The New York Times 2022)

- Suicide https://wonder.cdc.gov/mcd.html Multiple Cause of Death Data 2019 (see notes for which variables count) (Center for Disease Control and Prevention 2022)

3.1.4 political

Finally, I'll test Google Trends as an operationalization of political attitudes by looking at actual voting outcomes in historical US Presidential Election results and the American National Election Survey and Google Trends about candidates.

ANES (The University of Michigan 2020)

- Presidential results (McGovern et al. 2020)

3.1.5 Independent Variables

3.2 Analysis

For cross-sectional numeric data, I used X to calculate

testing for the strength of relationship between the Google Trend indicator and the respective comparison indicator. To address longitudinal correlations, I employed

I test both the raw trends, as well a combination of trends created through factor analysis as a method of dimensionality reduction.

As an additional test of correlation, I employ multiple linear regression to identify the strength of relationships across locales: I include city-data like county population size, internet usage rates, and Z to attempt to disentangle why some cases may be better predicted by Google Trends while others may not.

4 Results

4.1 Cultural

4.2 Medical

4.3 Political

5 Discussion

While I expect these tests to show high correlation between observed indicators and Google search trends, there will be three important questions that surface. First, just because something is correlated, does that mean it can replace the collection of other types of data? Second, how correlated does a trend need to be for social scientists to justifiably rely on it to indicate some outcome? And finally, how can we construct analyses like this to be robust to changes in the terms used across time and location?

The purpose of this paper is more methodological than theoretical, and I see this paper having an impact on the social sciences and computational social science as researchers pursue more projects using this source of big data. Google Trends are relatively underutilized in the field compared to in the health sciences and business. Once I assess how this data can be used, I would like to be able to join Bail (2014) in encouraging

social scientists to pursue more research with big data while taking into account the potential pitfalls with any source of big data (McFarland and McFarland 2015).

6 Conclusion

References

Asseo, Kim, Fabrizio Fierro, Yuli Slavutsky, Johannes Frasnelli, and Masha Y. Niv. 2020. "Tracking COVID-19 Using Taste and Smell Loss Google Searches Is Not a Reliable Strategy." *Scientific Reports* 10(1, 1):20527. doi: 10.1038/s41598-020-77316-3.

Bail, Christopher A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43(3):465–82. doi: 10.1007/s11186-014-9216-5.

Bail, Christopher A., Taylor W. Brown, and Andreas Wimmer. 2019. "Prestige, Proximity, and Prejudice: How Google Search Terms Diffuse Across the World." *American Journal of Sociology* 124(5):1496–1548. doi: 10.1086/702007.

Bail, Christopher A., Friedolin Merhout, and Peng Ding. 2018. "Using Internet Search Data to Examine the Relationship Between Anti-Muslim and Pro-ISIS Sentiment in U.S. Counties." *Science Advances* 4(6):eaao5948. doi: 10.1126/sciadv.aao5948.

Barrie, Christopher. 2020. "Searching Racism After George Floyd." *Socius* 6:2378023120971507. doi: 10.1177/2378023120971507.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350(6264):1073–76. doi: 10.1126/science.aac4420.

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15(5):662–79. doi: 10.1080/1369118X.2012.678878.

Brodeur, Abel, Andrew E. Clark, Sarah Fleche, and Nattavudh Powdthavee. 2021. "COVID-19, Lockdowns and Well-Being: Evidence from Google Trends." *Journal of Public Economics* 193:104346. doi: 10.1016/j.jpubeco.2020.104346.

Brownstein, John S., Adam D. Nahari, and Ben Y. Reis. 2020. "Internet Search Patterns Reveal Firearm Sales, Policies, and Deaths." *Npj Digital Medicine* 3(1, 1):1–9. doi: 10.1038/s41746-020-00356-6.

Burnett, Dayle, Valsamma Eapen, and Ping-I. Lin. 2020. "Time Trends of the Public's Attention Toward Suicide During the COVID-19 Pandemic: Retrospective, Longitudinal Time-Series Study." *JMIR Public Health and Surveillance* 6(4):e24694. doi: 10.2196/24694.

Carpi, Tiziana, Airo Hino, Stefano Maria Iacus, and Giuseppe Porro. 2020. "Twitter Subjective Well-Being Indicator During COVID-19 Pandemic: A Cross-Country Comparative Study." *Preprint* 31.

Center for Disease Control and Prevention. 2021. *CDC Vaccine Hesitancy for COVID-19*. Retrieved from: https://data.cdc.gov/Vaccinations/Vaccine-Hesitancy-for-COVID-19-County-and-local-es/q9mh-h2tw.

Center for Disease Control and Prevention. 2022. *County Suicide Rates*. Retrieved from: https://wonder.cdc.gov/mcd.html.

Dalum Hansen, Niels, Christina Lioma, and Kåre Mølbak. 2016. "Ensemble Learned Vaccination Uptake Prediction Using Web Search Queries." Pp. 1953–56 in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16*. New York, NY, USA: Association for Computing Machinery.

De Choudhury, Munmun, Meredith Ringel Morris, and Ryen W. White. 2014. "Seeking and Sharing Health Information Online: Comparing Search Engines and Social Media." Pp. 1365–76 in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. Toronto, Ontario, Canada: ACM Press.

Du, Hongfei, Jing Yang, Ronnel B. King, Lei Yang, and Peilian Chi. 2020. "COVID-19 Increases Online Searches for Emotional and Health-Related Terms." *Applied Psychology: Health and Well-Being* 12(4). doi: 10.1111/aphw.12237.

Google. 2022. *Google Trends*. Retrieved from: https://www.google.com/trends.

Gross, Neil, and Marcus Mann. 2017. "Is There a 'Ferguson Effect?' Google Searches, Concern about Police Violence, and Crime in U.S. Cities, 2014–2016." *Socius* 3:2378023117703122. doi: 10.1177/2378023117703122.

Jaidka, Kokil, Johannes Eichstaedt, Salvatore Giorgi, H. Andrew Schwartz, and Lyle H. Ungar. 2021. "Information-Seeking Vs. Sharing: Which Explains Regional Health? An Analysis of Google Search and Twitter Trends." *Telematics and Informatics* 59:101540. doi: 10.1016/j.tele.2020.101540.

Jimenez, Alberto, Rosa M. Estevez-Reboredo, Miguel A. Santed, and Victoria Ramos. 2020a. "COVID-19 Symptom Google Search Surges, Precede Local Incidence Surges: Evidence from Spain." *JMIR Preprints* 22(12). doi: 10.2196/23518.

Jimenez, Alberto, Rosa M. Estevez-Reboredo, Miguel A. Santed, and Victoria Ramos. 2020b. "Individuals' Concerns, Predict the Spread of the Coronavirus (COVID 19): The Case of Spain (Preprint)." doi: 10.2196/preprints.23518.

King, Gary. 2011. "Ensuring the Data-Rich Future of the Social Sciences." *Science* 331(6018):719–21. doi: 10.1126/science.1197872.

Lazer, David. 2015. "Issues of Construct Validity and Reliability in Massive, Passive Data Collections." *Cities Papers* 4–7.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343(6176):1203–5. doi: 10.1126/science.1248506.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. "Computational Social Science." *Science* 323(5915):721–23.

doi: 10.1126/science.1167742.

Lim, Jin Lee, Chong Yau Ong, Beiqi Xie, and Lian Leng Low. 2020. "Estimating Information Seeking-Behaviour of Public in Malaysia During COVID-19 by Using Google Trends." *Malaysian Journal of Medical Sciences* 27(5):202–4. doi: 10.21315/mjms2020.27.5.16.

Mavragani, Amaryllis, and Konstantinos Gkillas. 2020. "COVID-19 Predictability in the United States Using Google Trends Time Series." *Scientific Reports* 10(1, 1):20693. doi: 10.1038/s41598-020-77275-9.

McFarland, Daniel A., and H. Richard McFarland. 2015. "Big Data and the Danger of Being Precisely Inaccurate." *Big Data & Society* 2(2):2053951715602495. doi: 10.1177/2053951715602495.

McGovern, Tony, Stephen Larson, Bill Morris, and Matt Hodges. 2020. *US County-Level Presidential Election Results*. Retrieved from: https://github.com/tonmcg/US_County_Level_Election_Results_08-20. doi: 10.5281/zenodo.3975765.

Montoya, Nathalia, Sebastián Nieto-Parra, René Orozco, and Juan Vázquez Zamora. 2020. "Using Google Data to Understand Governments' Approval in Latin America." *OECD Development Centre Working Papers*. doi: 10.1787/89ed5e8f-en.

Nguyen, Hoang Long, Zhenhe Pan, Hashim Abu-gellban, Fang Jin, and Yuanlin Zhang. 2020. "Google Trends Analysis of COVID-19." Retrieved November 13, 2020 (http://arxiv.org/abs/2011.03847).

NORC at the University of Chicago. 2020. *The General Social Survey, 2020 Cross-Sectional*. Retrieved from: https://gssdataexplorer.norc.org/.

Nuti, Sudhakar V., Brian Wayda, Isuru Ranasinghe, Sisi Wang, Rachel P. Dreyer, Serene I. Chen, and Karthik Murugiah. 2014. "The Use of Google Trends in Health Care Research: A Systematic Review." *PLOS ONE* 9(10):e109583. doi: 10.1371/journal.pone.0109583.

Park, Taeyong, Haewoon Kwak, and Jisun An. n.d. "Building Longitudinal Google Trends to Measure Dynamic Local-level Issue Attention." *Preprint* 18.

Prado-Román, Camilo, Raúl Gómez-Martínez, and Carmen Orden-Cruz. 2020. "Google Trends as a Predictor of Presidential Elections: The United States Versus Canada." *American Behavioral Scientist* 65(4):666–80. doi: 10.1177/0002764220975067.

Reyes, Tomas, Nicolás Majluf, and Ricardo Ibáñez. 2018. "Using Internet Search Data to Measure Changes in Social Perceptions: A Methodology and an Application." *Social Science Quarterly* 99(2):829–45. doi: 10.1111/ssqu.12449.

Salganik, Matthew J. 2017. *Bit By Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Sarigul, Sercan, and Huaxia Rui. 2014. "Nowcasting Obesity in the U.S. Using Google Search Volume Data." in *Agricultural and Applied Economics Association & Canadian Agricultural Economics Society*

& European Association of Agricultural Economists. Montreal, Canada.

Semenza, Daniel Charles, and John A. Bernau. 2020. "Information-Seeking in the Wake of Tragedy: An Examination of Public Response to Mass Shootings Using Google Search Data." *Sociological Perspectives* 1–18. doi: 10.1177/0731121420964785.

Sood, Gaurav. 2016. *Geographic information on designated media markets* [computer program]. Version 9.1. https://doi.org/10.7910/DVN/IVXEH.

Stiles, Elizabeth A., and Patrick E. Grogan-Myers. 2018. "Assessing Criterion Validity of Using Internet Searches as a Measure of Public Attention." *American Review of Politics* 36(2, 2):1–18. doi: 10.15763/issn.2374-779X.2018.36.2.1-18.

Swearingen, C. Douglas, and Joseph T. Ripberger. 2014. "Google Insights and U.S. Senate Elections: Does Search Traffic Provide a Valid Measure of Public Attention to Political Candidates?" *Social Science Quarterly* 95(3):882–93. doi: 10.1111/ssqu.12075.

The New York Times. 2020. *Mask-Wearing Survey Data*. Retrieved from: https://github.com/nytimes/covid-19-data/tree/master/mask-use.

The New York Times. 2022. *County Covid Rates*. Retrieved from: https://github.com/nytimes/covid-19-data/tree/master/rolling-averages.

The University of Michigan. 2020. *American National Election Survey 2020 Time Series Study*. Retrieved from: https://electionstudies.org/data-center/.

Todorova, Tatina T., Gabriela S. Tsankova, and Neli M. Ermenlieva. 2021. "Internet Based Data As A Powerful Tool For Tracking Infectious Disease Dynamics." *Journal of IMAB - Annual Proceeding (Scientific Papers)* 27(1):3493–96. doi: 10.5272/jimab.2021271.3493.

Wolf, Jordan Taylor. 2018. "Trending in the Right Direction: Using Google Trends Data as a Measure of Public Opinion During a Presidential Election." Thesis, Virginia Tech.

Zitting, Kirsi-Marja, Heidi M. Lammers-van der Holst, Robin K. Yuan, Wei Wang, Stuart F. Quan, and Jeanne F. Duffy. 2020. "Google Trends Reveal Increases in Internet Searches for Insomnia During the COVID-19 Global Pandemic." *Journal of Clinical Sleep Medicine*. doi: 10.5664/jcsm.8810.

7 Appendix