Article 1: Construct Validity and Correspondence of Google Trends

Kelsey Gonzalez

23 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 1 outlines which data sources are used for this project and which trends are matched to each source.

Table 1: New York Air Quality Measurements

Validated.Data.Source	Type	Dates	Google.Trends.Used
Behaviors and Attitudes			
General Social Survey	Cross-Sectional	2010 - 2020	
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	
Presidential Election Results	Cross Sectional	2016 & 2020	'Hilary Clinton', 'Donald Trump', 'Joe Biden'

3.1 Measures

3.1.1 Google Trends

This paper focuses on a validation of Google search trends (Google 2022). Google trends portray the search frequency for specific search terms across designated media markets areas (DMAs), a nonoverlapping aggregation of U.S. counties to 210 media markets based on similar population clusters (Sood 2016). Raw data is on a scale from 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 as time-series (a single geo-location across time). To remedy this and build a longitudinal dataset of each search topic for the longitudinal datasets, 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). Missing datapoints in longitudinal datasets were filled in with interpolated values using `zoo::na.approx()` (Zeileis and Grothendieck 2005).

3.1.2 attitudinal

One measure of attitudinal indicators is the Vaccine Hesitancy for COVID-19 (Center for Disease Control and Prevention 2021). The CDC uses the U.S. Census Bureau’s Household Pulse Survey (HPS) and the 2019 American Community Survey (ACS) 1-year Public Use Microdata Sample (PUMS) to measure U.S. residents’ intentions to receive the COVID-19 vaccine if available during May 26, 2021 – June 7, 2021. This dataset consists of 3148 observations, one for each U.S. County. The variable measures the precent of adults in the county who describe themselves as “unsure”, “probably not”, or “definitely not” going to get a COVID-19 vaccine once one is available to them. The variable ranges from 4.99% to 32.33%.

Another attitudinal indicator I use is the Mask-Wearing Survey Data conducted by Dynata for the New York Times from July 2 through 14, 2020 (The New York Times 2020). 250,000 survey respondents were asked, “How often do you wear a mask in public when you expect to be within six feet of another person?”. The NYT weighted each response to create a county level measure of what percent of the county never, rarely, sometimes, frequently, and always wore a mask when in public. This dataset consists of 3148 observations, one for each U.S. County. This measure represents the percent of adults in the county who never or rarely wear a mask (range = 0.10% to 55.80%)

3.1.3 health

I also use U.S. Covid-19 rates to validate health and disease related topics. I retrieve U.S. county-level Covid-19 rates from by The New York Times (2022), who compile this data based on reports from state and local health agencies. It is widely acknowledged that there are biases in this data due to inconsistencies and availability in testing as well as different community propensity to test (CDC 2020; Gu n.d.) However, it is the best measure we have of actual case rates. Case Rate is measured as number of cases per 100,000 population. Observations vary from a minimum of 0 to a maximum of 1460.46 for each Monday from January 27, 2020 through December 27, 2021. There are 285986 cases across 3136 counties and 101 dates. Missing data were interpolated using `zoo::na.approx()` (Zeileis and Grothendieck 2005).

I also use county-level suicide rates from the US Centers for Disease Control and Prevention (2022). Data is grouped by year from 2010-2020. Raw death rates are scaled by population size for each year and can be interpreted as the death rate by suicide for every 1000 people. There are 34683 total cases, resulting from 34617 observations of `r nrow(count(suicide, fips))` counties. Missing data were interpolated using `zoo::na.approx()` (Zeileis and Grothendieck 2005). Measures range from 0.034 to 53.254.

3.1.4 political

Finally, I test Google Trends as an operationalization of political attitudes by first looking at actual voting outcomes in historical US Presidential Election results. This data comes from McGovern et al. (2020), who scraped the results from Townhall.com, Fox News, Politico, and the New York Times. Data on presidential outcomes were avaiable for 3147 counties in 2016 and 3118 counties in 2020. Each variable measures the percent of votes for the candidate, with the lowest percent at 3.09% and the highest at 92.15%.

In addition, I use data on political opinions from the the American National Election Survey 2020 Time-Series Study (The University of Michigan 2020). This data is paired with restricted data provided by ICPSR, the Inter-university Consortium for Political and Social Research, that contains the geo-ids of the 5,441 survey respondents. The study interviewed respondents in a pre-election survey that was conducted between August 18, 2020 and the day of the US Presidential election day, November 3, 2020.

3.1.5 Other

In addition to specific outcomes of interest, I also gathered various county controls to investigate possible variable confounding. The first Seven of these variables come from the 2010-2019 5-year American Community Surveys (US Census Bureau 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019). These data include total population, population density, unemployment rate for those 16 and over, median county income, average commute time, percent of households living under the poverty line, and percent above 65 years old. In addition, some models employ the U.S. Current Population Survey & American Community Survey Geographic Estimates of Internet Use, 1997-2018 (Tolbert and Mossberger 2020) to estimate households with broadband internet subscriptions. This final variable is an attempt to capture the latent propensity to use the internet for information search.

3.2 Analysis

Google Search Trends data and additional demographic data are merged with each individual indicator based on County FIPS Codes and date. After creating these different datasets, I use the Pearson correlation formula (formula (1)) for cross-sectional numeric data to calculate the strength of the relationship between each Google Trend and the respective data source.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (1)$$

To address longitudinal correlations, I employ Repeated Measures Correlation using the `rmcorr` package in R (Bakdash and Marusich 2017; Bland and Altman 1995). Repeated Measures Correlation is useful for determining the within-county association for paired measures across time and counties. Results of these correlations can be seen in table 2.

As an additional test of the relationships, I employ multiple linear regression (for cross-sectional datasets) and random intercept hierarchical linear models (Pinheiro et al. 2021) (for longitudinal datasets) to identify the strength of relationships across locales: I include city-data like county population size, broadband rates, and median income to attempt to disentangle possible confounders of the relationship between Google search Trends and verified indicators. Identifying these possible confounded relationships will help to explain why some articles find relationships between the trends and outcomes while others did not. For linear regression models, I normalize independent variables, i.e.variables have been centered and scaled to have a mean of 0 and standard deviation 1.

4 Results

4.1 Cultural

Table 2: Correlation Results

measure	variable	trend1	trend2	trend3	trend4
Vaccine Hesitancy					
		covid_conspiracy	covid_19_vaccine	coronavirus	covid_19
Pearson's R Correlation	Vaccine Hesitancy	0.1084	-0.2016	-0.4021	-0.3154

measure	variable	trend1	trend2	trend3	trend4
Mask Attitudes					
		face_mask	mask	cloth_face_mask	
Pearson's R Correlation	Mask Rare	0.0809	0.0877	0.296	
Covid Rates					
		covid_19	smell_loss	taste_loss	
Pearson's R Correlation	Covid Rate	-0.1012	0.2835	0.2851	
rmcorr	Covid Rate	-0.099***	0.322***	0.313***	
Suicide Rates					
		suicide	depression	suicide_hotline	
Pearson's R Correlation	Suicide Rate	0.0534	0.0219	0.0719	
rmcorr	Suicide Rate	0.055***	0.142***	0.131***	
2016 Presidential Votes					
		Hilary Clinton	Donald Trump		
Pearson's R Correlation	2016 Votes for Clinton	0.1919	0.2282		
Pearson's R Correlation	2016 Votes for Trump	-0.2254	-0.2154		
		Joe Biden	Donald Trump		
Pearson's R Correlation	2020 Votes for Biden	0.2473	0.2748		
Pearson's R Correlation	2020 Votes for Trump	-0.2481	-0.2744		

The first attitudinal indicators is Vaccine Hesitancy for COVID-19 vaccines (Center for Disease Control and Prevention 2021). When comparing the measure of vaccine hesitancy to four different Google Search

Trends, the correlation does not exceed -0.4021 (‘Coronavirus’ and vaccine hesitancy). Correlation’s under |0.40| are considered to be weak according to the common rules of thumb. When running these correlations in multiple linear regression (see the results in table @ref(tab:vacc_hes_analysis)), I see an r^2 of 0.221 (Model 1) and 0.346 (Model 2), indicating that the Google Trends are able to explain about 22% of the variation in Vaccine Hesitancy alone. Demographic characteristics like the percentage of households with broadband internet and the population density are able to explain about 35% of the variation (model 3), outperforming the first two models. The Trend coefficients themselves, however, are significant and remain significant when controlling for demographics. This reinforces the finding from the Pearson Correlation that there is a significant but weak relationship between the Google Trends and Vaccine Hesitancy.

The second attitudinal measure I test is how Google Trends relates to rare mask usage. As with vaccine hesitancy, the Pearson correlations are weak, if not negligible. I introduce these trends in multiple linear regression in table @ref(tab:mask_analysis). Model 1 demonstrates that these three Google Search Trends can explain about 9% of the variance in mask usage across U.S. counties, reinforcing the conclusion that the relationship is quite weak. The coefficients themselves are somewhat significant in Model 1. However, after including the demographic varaibles in Models 2 and 3, we see that the relationship between the Google Search Trends and mask usage is strengthed in magnitude and in significance, likely indicating a repression effect due to background relationships between the trends and demograhic variables. While the trends are significant and match the magnitude of the demographic variables, the low r^2 for the trends still provides evidence that Google Search Trends data cannot replace survey analysis when trying to measure rare mask usage.

4.2 Medical

```
## Random effect variances not available. Returned R2 does not account for random effects.
## Random effect variances not available. Returned R2 does not account for random effects.
```

pearson correlations rmcorr rsquared coefficients conclusion

pearson correlations rmcorr rsquared coefficients conclusion

4.3 Political

pearson correlations rsquared coefficients conclusion

pearson correlations rsquared coefficients conclusion

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.189***	0.179***	0.191***	0.189***
	(0.001)	(0.002)	(0.001)	(0.001)
covid_19_vaccine	−0.003***	−0.019***		−0.003**
	(0.001)	(0.003)		(0.001)
coronavirus	−0.022***	−0.015***		−0.012***
	(0.001)	(0.002)		(0.001)
covid_19	−0.004***	−0.010***		−0.003***
	(0.001)	(0.003)		(0.001)
covid_conspiracy		−0.004*		
		(0.002)		
total_pop			−0.005***	−0.005***
			(0.001)	(0.001)
pop_density			−0.004***	−0.003***
			(0.001)	(0.001)
unemployment_rate			−0.002*	0.000
			(0.001)	(0.001)
over_65			−0.013***	−0.010***
			(0.001)	(0.001)
poverty_rate			0.005**	0.005**
			(0.002)	(0.002)
median_income			−0.021***	−0.017***
			(0.001)	(0.001)
broadband			−0.008***	−0.005***
			(0.001)	(0.001)
Num.Obs.	2720	913	3133	2711
R2	0.221	0.346	0.350	0.432
R2 Adj.	0.220	0.344	0.349	0.430
Log.Lik.	4510.726	1625.586	5405.127	4926.819
F	256.921	120.339	240.876	205.656

* p < .05. ** p < .01. *** p < .001 (two-tailed test).

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

	Model 1	Model 2	Model 3
(Intercept)	0.159***	0.163***	0.160***
	(0.002)	(0.002)	(0.002)
face_mask	−0.009*		−0.022***
	(0.004)		(0.004)
mask	0.009*		0.017***
	(0.004)		(0.004)
cloth_face_mask	0.029***		0.023***
	(0.002)		(0.002)
total_pop		−0.012***	−0.010***
		(0.002)	(0.002)
pop_density		−0.003+	−0.003*
		(0.002)	(0.002)
unemployment_rate		−0.028***	−0.026***
		(0.002)	(0.002)
over_65		−0.006**	−0.005*
		(0.002)	(0.002)
poverty_rate		−0.010**	−0.008**
		(0.003)	(0.003)
median_income		−0.036***	−0.032***
		(0.003)	(0.003)
broadband		−0.006*	−0.008**
		(0.003)	(0.003)
Num.Obs.	2827	3133	2819
R2	0.089	0.176	0.236
R2 Adj.	0.088	0.174	0.233
Log.Lik.	2755.688	3184.261	2994.427
F	92.248	95.547	86.633

* p < .05. ** p < .01. *** p < .001 (two-tailed test).

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

	Model 1	Model 2	Model 3
(Intercept)	−532.306***	−712.023***	−534.608***
	(8.271)	(6.009)	(8.268)
covid_19	1.031***		1.024***
	(0.082)		(0.082)
smell_loss	7.373***		7.374***
	(0.082)		(0.082)
taste_loss	6.344***		6.347***
	(0.078)		(0.078)
covid_rate_fips_mean	0.813***	0.982***	0.879***
	(0.019)	(0.010)	(0.020)
date	0.029***	0.038***	0.029***
	(0.000)	(0.000)	(0.000)
SD (Intercept)	5.196	0.002	4.984
SD (Observations)	29.478	31.736	29.441
total_pop		0.064	0.409**
		(0.066)	(0.140)
pop_density		0.010	0.463***
		(0.067)	(0.123)
unemployment_rate		0.049	0.725***
		(0.080)	(0.155)
over_65		−0.116	−0.055
		(0.074)	(0.139)
poverty_rate		0.054	0.351
		(0.118)	(0.221)
median_income		0.031	0.907***
		(0.114)	(0.214)
broadband		0.073	0.437*
		(0.097)	(0.182)
Num.Obs.	248 346	284 666	247 655
R2 Marg.	0.201	0.081	0.200
R2 Cond.		0.081	

Random intercept per county

* p < .05. ** p < .01. *** p < .001 (two-tailed test).

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*

	Model 1	Model 2	Model 3
(Intercept)	−9.798***	−10.447***	−9.052***
	(0.306)	(1.037)	(1.182)
suicide	−0.001		0.000
	(0.000)		(0.000)
depression	−0.002**		0.001
	(0.001)		(0.001)
suicide_hotline	0.000		0.001
	(0.000)		(0.001)
death_rate_fips_mean	1.000***	1.008***	1.008***
	(0.000)	(0.001)	(0.001)
year	0.005***	0.005***	0.004***
	(0.000)	(0.001)	(0.001)
SD (Intercept)	0.000	0.000	0.000
SD (Observations)	0.056	0.056	0.054
total_pop		−0.011*	−0.010*
		(0.005)	(0.005)
pop_density		0.004	0.004
		(0.005)	(0.005)
unemployment_rate		−0.002**	−0.003**
		(0.001)	(0.001)
over_65		−0.001	−0.002
		(0.002)	(0.002)
poverty_rate		−0.001	−0.001
		(0.001)	(0.001)
median_income		−0.002*	−0.002*
		(0.001)	(0.001)
broadband		0.002**	0.003***
		(0.001)	(0.001)
Num.Obs.	28 287	13 018	11 854
R2 Marg.	0.996	0.995	0.994
R2 Cond.	0.996	0.995	0.994

Random intercept per county

* p < .05. ** p < .01. *** p < .001 (two-tailed test).

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

	Model 1	Model 2	Model 3
(Intercept)	0.319***	0.318***	0.318***
	(0.003)	(0.002)	(0.002)
‘Donald Trump‘	0.027***		0.009***
	(0.003)		(0.003)
‘Hilary Clinton‘	0.016***		0.017***
	(0.003)		(0.003)
total_pop		0.028***	0.025***
		(0.002)	(0.002)
pop_density		0.022***	0.019***
		(0.002)	(0.002)
unemployment_rate		0.047***	0.043***
		(0.003)	(0.003)
over_65		−0.012***	−0.015***
		(0.003)	(0.003)
poverty_rate		0.034***	0.037***
		(0.004)	(0.004)
median_income		0.054***	0.047***
		(0.004)	(0.004)
Num.Obs.	3131	3145	3129
R2	0.061	0.314	0.330
R2 Adj.	0.061	0.313	0.329
Log.Lik.	1533.665	2037.065	2066.480
F	101.965	239.652	192.316

Results predicting Donald J. Trump percentage largely equivalent and available upon request.
* p < .05. ** p < .01. *** p < .001 (two-tailed test).

the Relationship Between Anti-Muslim and Pro-ISIS Sentiment in U.S. Counties.” *Science Advances* 4(6):eaao5948. doi: 10.1126/sciadv.aao5948.

Bakdash, Jonathan Z., and Laura R. Marusich. 2017. “Repeated Measures Correlation.” *Frontiers in Psychology* 8:456. doi: 10.3389/fpsyg.2017.00456.

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

Bland, J. Martin, and Douglas G. Altman. 1995. “Statistics Notes: Calculating Correlation Coefficients with Repeated Observations: Part 1—Correlation Within Subjects.” *Bmj* 310(6977):446. doi: 10.1136/bmj.310.6977.446.

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.

	Model 1	Model 2	Model 3
(Intercept)	0.333***	0.333***	0.333***
	(0.003)	(0.002)	(0.002)
‘Donald Trump‘	0.036***		0.009*
	(0.005)		(0.004)
‘Joe Biden‘	0.009+		0.021***
	(0.005)		(0.004)
total_pop		0.030***	0.027***
		(0.003)	(0.003)
pop_density		0.021***	0.019***
		(0.003)	(0.003)
unemployment_rate		0.044***	0.041***
		(0.003)	(0.003)
over_65		−0.007*	−0.010***
		(0.003)	(0.003)
poverty_rate		0.032***	0.036***
		(0.004)	(0.004)
median_income		0.069***	0.058***
		(0.004)	(0.004)
Num.Obs.	3118	3118	3118
R2	0.077	0.295	0.322
R2 Adj.	0.076	0.294	0.320
Log.Lik.	1414.094	1835.433	1895.677
F	129.169	217.239	184.551

Results predicting Donald J. Trump percentage largely equivalent and available upon request.

* p < .05. ** p < .01. *** p < .001 (two-tailed test).

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.

CDC. 2020. "COVID-19 Pandemic Planning Scenarios." Retrieved February 3, 2022 (https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html).

Center for Disease Control and Prevention. 2021. *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.

Gu, Youyang. n.d. "Estimating True Infections - Revisited." Retrieved February 3, 2022 (https://covid19-projections.com/estimating-true-infections-revisited/).

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).

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.

Pinheiro, José, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, Siem Heisterkamp, Bert Van Willigen, and Johannes Ranke. 2021. *Nlme: Linear and Nonlinear Mixed Effects Models*. Retrieved from: https://cran.r-project.org/web/packages/nlme/.

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.* Harvard Dataverse. Retrieved from: https://doi.org/10.7910/DVN/IVXEH. doi: 10.7910/DVN/IVXEHT.

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.

Tolbert, Caroline, and Karen Mossberger. 2020. *U.s. Current Population Survey & American Community Survey Geographic Estimates of Internet Use, 1997-2018.* Retrieved from: https://techdatasociety.asu.

edu/broadband-data-portal/dataaccess/countydata.

US Census Bureau. 2010. *American Community Survey: 5-Year Estimates*. Retrieved from: https://www.census.gov/programs-surveys/acs/news/data-releases/2010/release.html.

US Census Bureau. 2011. *American Community Survey: 5-Year Estimates*. Retrieved from: https://www.census.gov/programs-surveys/acs/news/data-releases/2011/release.html.

US Census Bureau. 2012. *American Community Survey: 5-Year Estimates*. Retrieved from: https://www.census.gov/programs-surveys/acs/news/data-releases/2012/release.html.

US Census Bureau. 2013. *American Community Survey: 5-Year Estimates*. Retrieved from: https://www.census.gov/programs-surveys/acs/news/data-releases/2013/release.html.

US Census Bureau. 2014. *American Community Survey: 5-Year Estimates*. Retrieved from: https://www.census.gov/programs-surveys/acs/news/data-releases/2014/release.html.

US Census Bureau. 2015. *American Community Survey: 5-Year Estimates*. Retrieved from: https://www.census.gov/programs-surveys/acs/news/data-releases/2015/release.html.

US Census Bureau. 2016. *American Community Survey: 5-Year Estimates*. Retrieved from: https://www.census.gov/programs-surveys/acs/news/data-releases/2016/release.html.

US Census Bureau. 2017. *American Community Survey: 5-Year Estimates*. Retrieved from: https://www.census.gov/programs-surveys/acs/news/data-releases/2017/release.html.

US Census Bureau. 2018. *American Community Survey: 5-Year Estimates*. Retrieved from: https://www.census.gov/programs-surveys/acs/news/data-releases/2018/release.html.

US Census Bureau. 2019. *American Community Survey: 5-Year Estimates*. Retrieved from: https://www.census.gov/programs-surveys/acs/news/data-releases/2019/release.html.

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.

Zeileis, Achim, and Gabor Grothendieck. 2005. "Zoo: S3 Infrastructure for Regular and Irregular Time Series." *Journal of Statistical Software* 14(6):1–27. doi: 10.18637/jss.v014.i06.

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