A Hands-on Guide to Google Data

Seth Stephens-Davidowitz
Hal Varian
Google, Inc.

September 2014
Revised: March 7, 2015

Abstract

This document describes how to access and use Google data for social science research. This document was created using the literate programming system `knitr` so that all code in the document can be run as it stands.

Google provides three data sources that can be useful for social science: Google Trends, Google Correlate, and Google Consumer Surveys. Google Trends provides an index of search activity on specific terms and categories of terms across time and geography. Google Correlate finds queries that are correlated with other queries or with user-supplied data across time or US states. Google Consumer Surveys offers a simple, convenient way to conduct quick and inexpensive surveys of internet users.

1 Google Correlate

Economic data is often reported with a lag of months or quarters while Google query data is available in near real time. This means that queries that are contemporaneously correlated with an economic time series may be helpful for economic "nowcasting."

We illustrate here how Google Correlate can help build a model for housing activity. The first step is to download data for "New One Family Houses Sold" from FRED[1] We don't use data prior to January 2004 since that's when the Google series starts. Delete the column headers and extraneous material from the CSV file after downloading.

Now go to Google Correlate and click on "Enter your own data" followed by "Monthly Time Series." Select your CSV file, upload it, give the series a name, and click "Search correlations." You should something similar to Figure 1.

Note that the term most correlated with housing sales is [tahitian noni juice], which appears to be a spurious correlation. The next few terms are similarly spurious. However, after that, you get some terms that are definitely real-estate

[1] `http://research.stlouisfed.org/fred2/series/HSN1FNSA`.

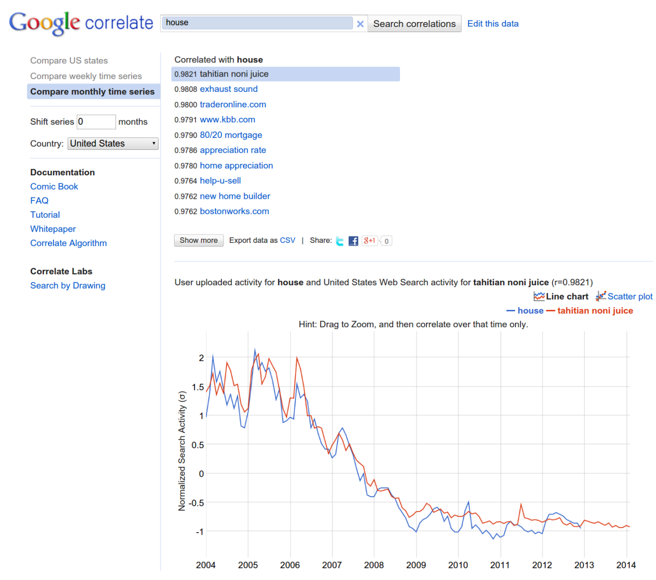


Figure 1: Screenshot from Google Correlate.

related. (Note that that the difference in the correlation coefficient for [tahitian noni juice] and [80/20 mortgage] is tiny.)

You can download the hundred most correlated terms by clicking on the "Export as CSV" link. The resulting CSV file contains the original series and one hundred correlates. Each series is standardized by subtracting off its mean and dividing by its standard deviation.

The question now is how to use these correlates to build a predictive model. One option is to simply use your judgment in choosing possible predictors. As indicated above, there will generally be spurious correlates in the data, so it makes sense to remove these prior to further analysis. The first, and most obvious, correlates to remove are queries that are unlikely to persist, such as [tahitian noni juice], since that query will likely not help for future nowcasting. For economic series, we generally remove non-economic queries from the CSV file. When we do that, we end up with about 70 potential predictors for the 105 monthly observations.

At this point, it makes sense to use a variable selection mechanism such as stepwise regression or LASSO. We will use a system developed by Steve Scott at Google called "Bayesian Structural Time Series," that allows you to model both the time series and regression components of the predictive model.[2]

[2]urlhttp://cran.r-project.org/web/packages/bsts/

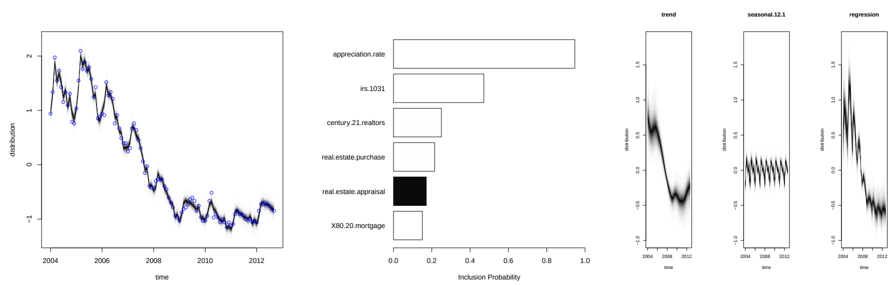


Figure 2: Output of BSTS. See text for explanation.

2 Bayesian structural time series

BSTS is an `R` library described in Scott and Varian [2012, 2014a]. Here we focus on how to use the system. The first step is to install the `R` package `bsts` and `BoomSpikeSlab` from CRAN. After that installation, you can just load the libraries as needed.

```
# read data from correlate and make it a zoo time series
dat <- read.csv("Data/econ-HSN1FNSA.csv")
y <- zoo(dat[,2],as.Date(dat[,1]))
# use correlates as possible predictors
x <- dat[,3:ncol(dat)]
# set a few parameters
numiter <- 4000
npred <- 5
# describe state space model consisting of
# trend and seasonal components
ss <- AddLocalLinearTrend(list(),y)
ss <- AddSeasonal(ss,y,nseasons=12)
# estimate the model
model <- bsts(y~.,state.specification=ss,data=x,
niter=numiter,expected.model.size=npred,ping=0,seed=123)
# Posterior distribution and actual outcome.
plot(model)
# Probability of inclusion of top predictors (p > .15)
plot(model,"coef",inc=.15)
# Contribution of trend, seasonal and regression components.
plot(model,"comp")
```

We now wait patiently while the model is estimated and then examine the results, shown in Figure 2. The first panel shows the fit, the second panel shows the most probable predictors, and third panel show the decomposition of the time series into three components: a trend, a seasonal component, and a

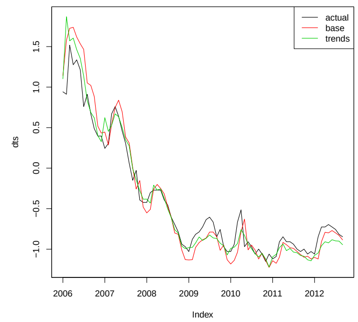


Figure 3: Out-of-sample forecasts

regression component. The last panel shows that the regression predictors are important.

By default, the model computes the in-sample predictions. In order to evaluate the forecasting accuracy of the model, it may be helpful to examine out-of-sample prediction. This can be done with BSTS but it is time consuming, so we follow a hybrid strategy. We consider two models, a baseline autoregressive model with a one-month and twelve-month lag:

$$y_t = b_1 y_{t-1} + b_{12} y_{t-12} + e_t,$$

and the same model supplemented with some additional predictors from Google Correlate:

$$y_t = b_1 y_{t-1} + b_{12} y_{t-12} + a_t x_t + e_t.$$

We estimate each model through period t, forecast period $t+1$, and then compare the mean absolute percent error (MAPE).

```
# load package for out-of-sample-forecasts
source("oosf.R")
# choose top predictors
x1 <- zoo(x[,cbind("appreciation.rate","irs.1031","century.21.realtors",
               "real.estate.purchase")],as.Date(dat[,1]))
reg1 <- OutOfSampleForecast12(y,x1,k=24)
# mae.delta is the ratio of the trends MAE to the base MAE
MaeReport(reg1)
```

```
##   mae.base mae.trends  mae.delta
##  0.1451080  0.1115476  0.2312789
```

The three numbers reported are the mean absolute one-step ahead percentage prediction error (MAPE) using only the autoregressive model, the MAPE when we use the Google variables, and the ratio of the two. We see prediction error is substantially less when we use the Google predictors.

3 Cross section

We can also use Correlate to build models predicting cross-section data from US states. (Other countries are not yet available.)

3.1 House prices declines

To continue with the housing theme, let us examine cross-sectional house price declines. We downloaded the "eCoreLogic October 2013 Home Price Index Report" and converted the table "Single-Family Including Distressed" on page 7 to a CSV file showing house price declines by *state*. We uploaded it to Google Correlate and found the 100 queries that were most correlated with the price index.

```
dat <- read.csv("Data/correlate-housing_decline.csv")
d0 <- dat[,-1]
names(d0)[2:11]

##  [1] "short.sale.process"
##  [2] "short.sale"
##  [3] "underwater.mortgage"
##  [4] "seterus"
##  [5] "harp.3.0"
##  [6] "short.sale.package"
##  [7] "mortgage.forgiveness.debt.relief"
##  [8] "mortgage.forgiveness.debt.relief.act"
##  [9] "upside.down.mortgage"
## [10] "mortgage.short.sale"
```

Figure 3.1 illustrates the correlation between the price decline and the search `[short sale process]`.

If we take a linear combination of these queries (e.g., a regression) we can normally improved prediction performance. We use the `BoomSpikeSlab` package from CRAN to find good predictors.

```
library(BoomSpikeSlab)
reg0 <- lm.spike(housing.decline ~ .,niter=4000,data=d0,seed=123,ping=0)
plot(reg0,inc=.10)
```

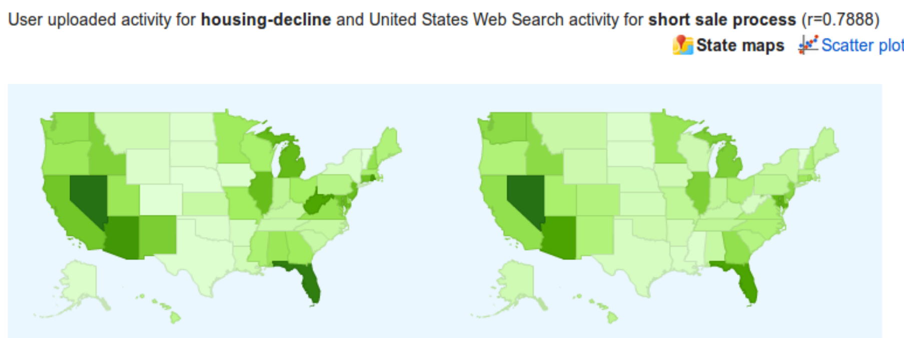
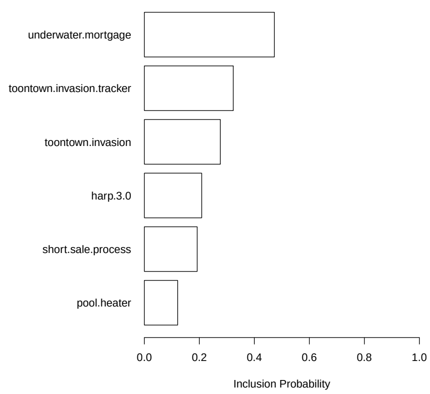


Figure 4: Price decline and `[short sale process]`.



The `[toontown]` queries appear to be spurious. To check this, we look at the geographic distribution of this query. Figure 5 shows a map from Google Trends showing the popularity of the `[toontown]` query in Fall 2013. Note how the popularity is concentrated in "sand states" which also had the largest real estate bubble.

Accordingly we remove the `toontown` queries and estimate again. We also get a spurious predictor in this case `club penguin membership` which we remove and estimate again. The final fit is shown in Figure 6.

```
d1 <- d0[,-grep("toontown",names(d0))]
d2 <- d1[,-grep("penguin",names(d1))]
reg2 <- lm.spike(housing.decline ~ .,niter=4000,data=d2,seed=123,ping=0)
plot(reg2,inc=.10)
```

Should we use `[solar pool heaters]` as a regressor? If the goal is to use this regression as an early warning signal for future housing starts, we might

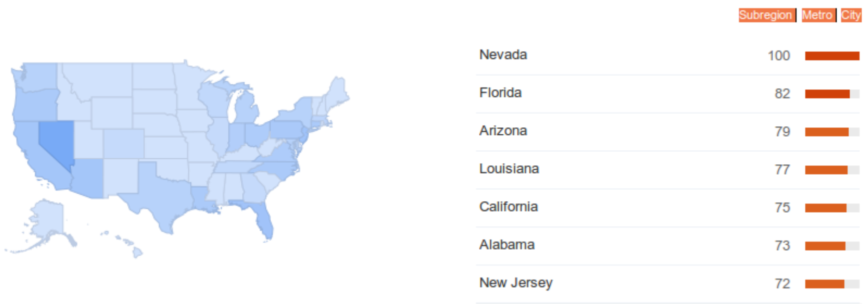


Figure 5: Searches on `toontown`

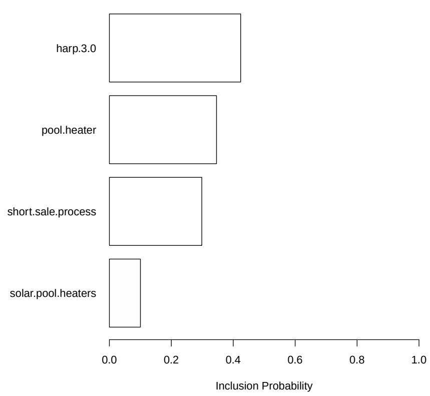


Figure 6: House price regression, final model.

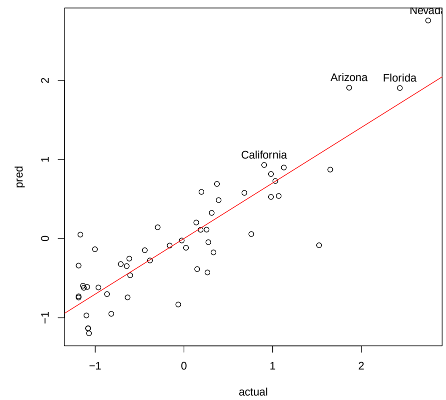


Figure 7: Acutal versus fitted housing data.

drop the **`[solar pool heater]`** predictor as it is unlikely that the next housing crisis would start in the "sand states." On the other hand, if this query showed up as a predictor early in the crisis, it may have helped attention to focus more attention on those geographies where **`[solar pool heater]`** was common.

Finally, Figure 7 plots actual versus predicted, to give some idea of the fit.

```
temp <- predict(reg2,newdata=d2)
pred <- rowMeans(temp)
actual <- d2$housing.decline
plot(pred~actual)
reg3 <- lm(pred ~ actual)
abline(reg3,col=2)
states <- dat[,1]
z <- states[c(3,5,10,29)]
text(y=pred[z],x=actual[z],labels=states[z],pos=3)
```

3.2 Life expectancy

Suppose we want to look at life expectancy by state.[3] In this case, it turns out that it is more interesting to find queries associated with abnormally *short* lifespans, so we put a minus sign in front the entries in the CSV file. (We will refer to the negative of lifespan as "morbidity.")

We upload the file to Google Correlate, now using the "US States" option; this gives us a heat map showing the queries correlated with short lives. Note

[3] urlkff.org/other/state-indicator/life-expectancy/

that short life expectancy and the queries associated with short life expectancy are concentrated in the Deep South and Appalachia.

We download the series of correlates as before and then build a predictive model. Since this is cross sectional data, we use the package BoomSpikeSlab.

```
# library(BoomSpikeSlab)
dat <- read.csv("Data/correlate-negative_life_expectancy.csv")
d <- dat[,-1]
reg <- lm.spike(negative.life.expectancy ~ .,niter=4000,data=d)
plot(reg,inc=.10)
```

The predictors are interesting. The “Obama says” predictor seemed strange so we tried it in Google Suggest. On April 17, 2014, the suggested completions of “Obama says ... ” were 1) he is god, 2) there are 57 states, 4) constitution is dead, 4) serve satan, 5) uh too much. Most of these searches seem to express negative sentiment Obama.

Finally Figure 9 shows the actual morbidity compared to fitted. The big negative outlier is the District of Columbia. In fact, we find that District of Columbia is often an outlier. This could be because many searches likely come from commuters.

```
temp <- predict(reg,newdata=d)
neg.life <- rowMeans(temp)
plot(neg.life~d$negative.life.expectancy)
reg1 <- lm(neg.life~d$negative.life.expectancy)
abline(reg1,col=2)
```

4 Google Trends

We turn now to Google Trends. This tools used the same data used in Correlate and provides an index of search activity by query or query category. Suppose you are interested in the search activity on the Los Angeles Lakers. You can go to Google Trends and enter the term `Lakers`. You get a chart showing the time series, the geographic distribution of searches on that term, related searches, and so on.

Using the navigation bar at the top of the page, you can restrict the index to particular geographies (countries, states, metro areas), particular time periods, and particular categories. If you choose a time period that is 3 months or shorter you get daily data, otherwise you get weekly data. If the time period is 3 years or longer, the monthly data is plotted, otherwise it is weekly data.

Categories are helpful when there is ambiguity in the search term. For example, enter the term `apple` and restrict the geography to the United States. Now select the category `Computer & Electronics`. Compare this to the pattern to that when you use the category `Food & Drink`. Quite a difference!

3/17/2014 negative life expectancy - Google Correlate

hal@google.com | Manage my Correlate data | Sign out

negative life expectancy | Search correlations | Edit this data

Compare US states
Compare weekly time series
Compare monthly time series

Documentation
Comic Book
FAQ
Tutorial
Whitepaper
Correlate Algorithm

Correlate Labs
Search by Drawing

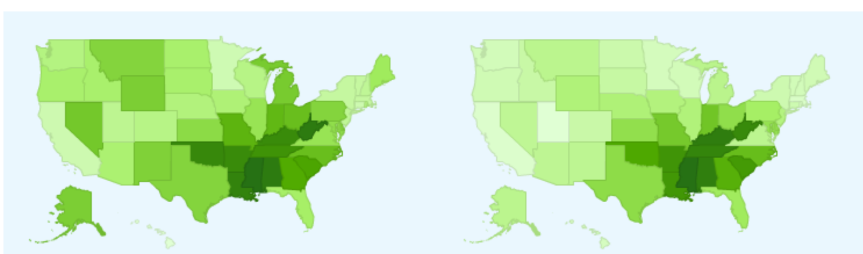
Correlated with **negative life expectancy**

0.9092 blood pressure medicine
0.8985 obama a
0.8978 major payne
0.8975 against obama
0.8936 king james bible online
0.8935 about obama
0.8928 prescription medicine
0.8920 40 caliber
0.8919 .38 revolver
0.8916 reprobate
0.8911 performance track
0.8910 lost books of the bible
0.8905 glock 40 cal
0.8898 lost books
0.8896 the mark of the beast
0.8892 obama says
0.8891 obama said
0.8882 sodom and
0.8882 the antichrist
0.8865 globe life
0.8858 the judge
0.8834 hair pics
0.8833 medicine side effects
0.8829 momma
0.8828 james david
0.8823 flexeril

https://www.google.com/trends/correlate/search?e=id:vs1wQ1crZMx&t=all#default,90 1/4

User uploaded activity for **negative life expectancy** and United States Web Search activity for **blood pressure medicine** (r=0.9092)

State maps Scatter plot



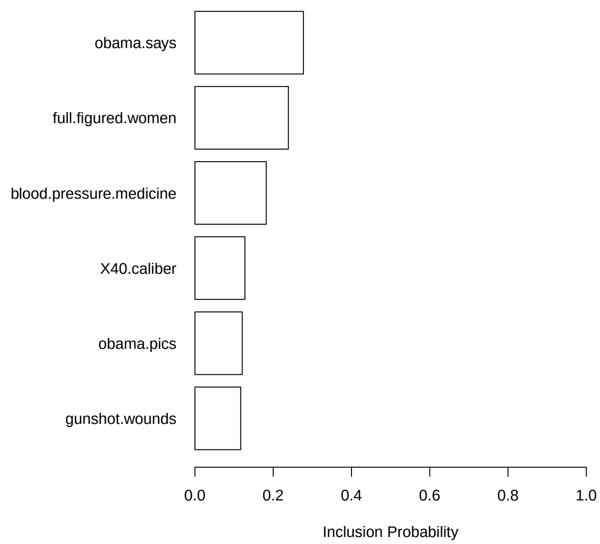


Figure 8: Predictors of short life expectancy

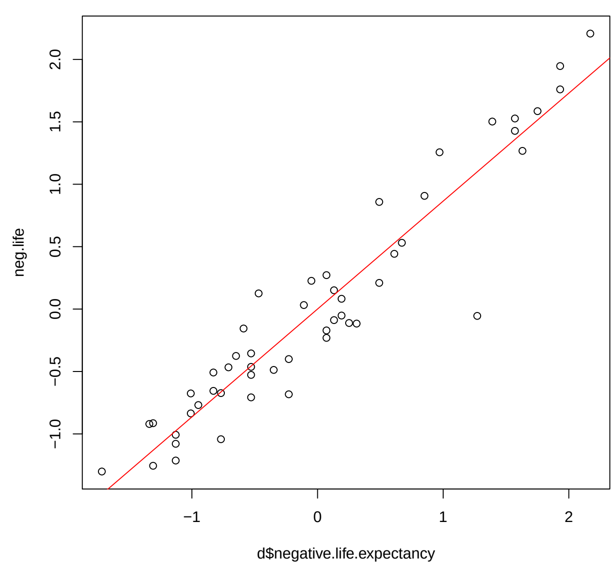


Figure 9: Actual vs. fitted morbidity

You can also look at an index of all searches in a category. For example, choose the category `Sports` and the geography `Worldwide` and leave the search term blank. This shows us an index for sports-related queries. The four-year cycle of the Olympics is apparent.

Another way to disambiguate queries is to use the *entities* selection. Google attempts to identify entities by looking at searches surrounding the search in question. For example, if someone searches `apple` in conjunction with `[turkey], [sweet potato], [apple]` they are probably looking for search results referring to the fruit. Entities are useful in that they bind together different ways to describe something—abbreviations, spelling, synonyms and so on.

4.1 Match types

Trends uses the following conventions to refine searches.

- + means "or." If you type `Lakers+Celtics`, the results will be searches that include either the word `Lakers` or the word `Celtics`.
- - means to exclude a word. If you type `jobs - steve`, results will be searches that include `jobs` but do not include `steve`
- A space means "and." If you type `Lakers Celtics`, the results will be searches that include both the word `Lakers` and the word `Celtics`. The order does not matter.
- Quotes force a phrase match. If you type `''Lakers Celtics''`, results will be searches that include the exact phrase `Lakers Celtics`.

4.2 What does Google Trends measure?

Recall that Google Trends reports an *index* of search activity. The index measures the fraction of queries that include the term in question in the chosen geography at a particular time relative the total number of queries at that time. The maximum value of the index is set to be 100. For example, if one data point is 50 and another data point is 100, this means that the number of searches satisfying the condition was half as large for the first data point as for the second data point. The scaling is done separately for each request, but you can compare up to 5 items per request.

If Google Trends shows that a search term has decreased through time, this does not necessarily mean that there are fewer searches now than there were previously. It means that there are fewer searches, as a percent of all searches, than there were previously. In absolute terms, searches on virtually every topic has increased over time.

Similarly, if Rhode Island scores higher than California for a term this does not generally mean that Rhode Island makes more total searches for the term than California. It means that as a percent of of total searches, there are relatively more searches in Rhode Island than California on that term. This is the more meaningful metric for social science, since otherwise bigger places with more searches would always score higher.

Here are four more important points. First, Google Trends has an unreported privacy threshold. If total searches are below that threshold, a 0 will be reported. This means that not enough were made to advance past the threshold. The privacy threshold is based on absolute numbers. Thus, smaller places will more frequently show zeros, as will earlier time periods. If you run into zeros, it may be helpful to use a coarser time period or geography.

Second, Google Trends data comes from a sample of the total Google search corpus. This means samples might differ slightly if you get a different sample. If very precise data is necessary, a researcher can average different samples. That said, the data is large enough that each sample should give similar results. In cases where there appear to be outliers, researchers can just issue their query again on another day.

Third, Google Trends data is averaged to the nearest integer. If this is a concern, a researcher can pull multiple samples and average them to get a more precise estimate. If you compare two queries, one of which is very popular and the other much less so, the normalization can push the unpopular query to zero. The way to deal with this is to run a separate request for each query. The normalized magnitude of the queries will no longer be comparable, but the growth rate comparison will still be meaningful.

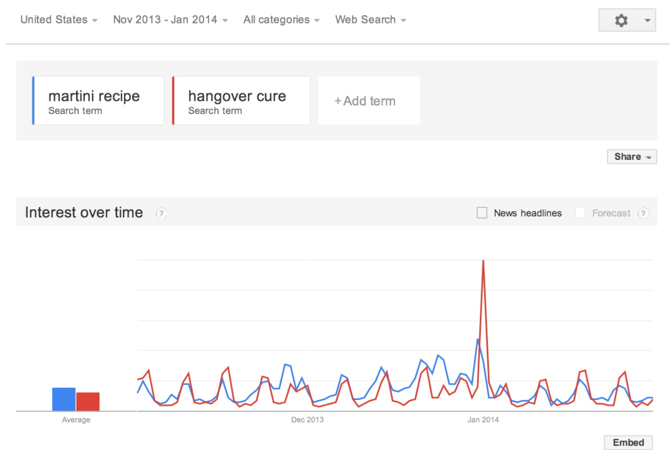
Fourth, and related to the previous two points, data is cached each day. Even though it comes from a sample, the same request made on the same day will report data from the same sample. A researcher who wants to average multiple samples must wait a day to get a new sample.

It is worth emphasizing that the sampling generally gives reasonably precise estimates. Generally we do not expect that expect that researchers will need more than a single sample.

4.3 Time series

Suppose a researcher wants to see how the popularity of a search term has changed through time in a particular geo. For example, a researcher may be curious on what days people are most likely to search for `[martini recipe]` between November 2013 and January 2014 in the United States. The researcher types in `martini recipe`, chooses the United States, and chooses the relevant time period. The researcher will find that a higher proportion of searches include `[martini recipe]` on Saturdays than any other day. In addition, the searches on this topic spike on December 31, New Year's Eve.

A researcher can also compare two search terms over the same time period, in the same place. The researcher can type in `[hangover cure]` to compare it to `[martini recipe]`. See Figure 4.3 for the results. The similarity of the blue and red lines will show that these searches are made, on average, a similar amount. However, the time patterns are different. `[Hangover cures]` is more



popular on Sundays and is an order of magnitude more common than [martini recipe] on January 1.

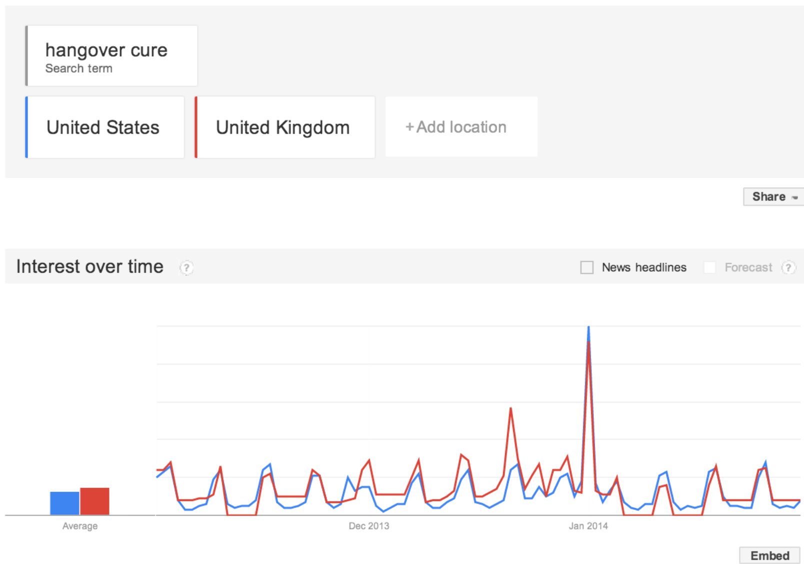
You can also compare multiple geos over the same time period. Figure 10 shows search volume for [hangover cure] during the same time period in the United States. But it also adds another country, the United Kingdom. On average, the United Kingdom searches for [hangover cure] more frequently during this time period. But apparently the United States has bigger New Years parties, as Americans top the British for [hangover cure] searches on January 1.

4.4 Geography

Google Trends also shows the geography of search volumes. As with the time series, the geographic data are normalized. Each number is divided by the total number of searches in an area and normalized so that the highest-scoring state has 100. If state A scores 100 and state B scores 50 in the same request, this means that the percentages of searches that included the search term was twice as high in state A as in state B. For a given plot, the darker the state in the output heat map, the higher the proportion of searches that include that term. It is not meaningful to compare states across requests, since the normalization is done separately for each request.

Figure 11 shows the results for typing in each of `Jewish` and `Mormon`. Panel (a) shows that search volume for the word `Jewish` differs in different parts of the country. It is highest in New York, the state with the highest Jewish population. In fact, this map correlates very highly ($R^2 = 0.88$) with the proportion of a state's population that is Jewish. Panel (b) shows that the map of `Mormon` search rate is very different. It is highest in Utah, the state with the highest Mormon population, and second highest in Idaho, which has the second-highest Mormon population.

Figure 10: Hangovers, United States versus United Kingdom



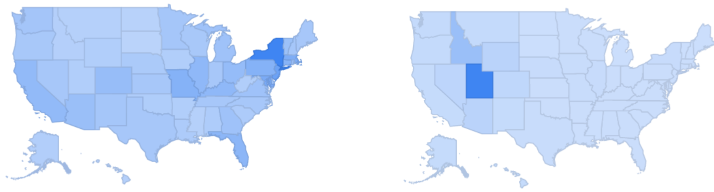
4.5 Query selection

We believe that Google searches may be indicative of particular attitudes or behaviors that would otherwise not be easy to measure. The difficulty is that there are literally trillions of possible searches. Which searches should you choose? A major concern with Google Trends data is cherry-picking: the researcher might consciously or subconsciously choose the search term that gives a desired result.

If there is clearly a single salient word this danger is mitigated. In Stephens-Davidowitz [2012], the author uses the unambiguously most salient word related to racial animus against African-Americans. Stephens-Davidowitz [2013] uses just the words `[vote]` and `[voting]` to measure intention to vote prior to an election. Swearingen and Ripberger [2014] use a Senate candidate's name to see if Google searches can proxy for interest in an election.

Be careful about ambiguity. If there are multiple meanings associated with a word, you can use a minus sign to take out one or two words that are not related to the variable of interest. Baker and Fradkin [2013] uses searches for `jobs` to measure job search. But they take out searches that also include the word "Steve." Madestam et al. [2013] use searches for `Tea Party` to measure interest in the political party but take out searches that also include the word `Boston`.

Figure 11: Search for "Jewish" versus "Mormon"



(a) Jewish (b) Mormon

4.6 Applications

Google Trends has been used in a number of academic papers. We highlight a few such examples here.

Stephens-Davidowitz [2012] measures racism in different parts of the United States based on search volume for a salient racist word. It turns out that the number of racially charged searches is a robust predictor of Barack Obama's underperformance in certain regions, indicating that Obama did worse than previous Democratic candidates in areas with higher racism. This finding is robust to controls for demographics and other Google search terms. The measured size of the vote loss due to racism are 1.5 to 3 times larger using Google searches than survey-based estimates.

Baker and Fradkin [2013] uses Google searches to measure intensity of job search in different parts of Texas. They compare this measure to unemployment insurance records. They find that job search intensity is significantly lower when more people have many weeks of eligibility for unemployment insurance remaining.

Mathews and Tucker [2014] examine how the composition of Google searches changed in response to revelations from Edward Snowden. They show that surveillance revelations had a chilling effect on searches: people were less likely to make searches that could be of interest to government investigators.

There are patterns to many of the early papers using Google searches. First, they often focus on areas related to social desirability bias—that is, the tendency to mislead about sensitive issues in surveys. People may want to hide their racism or exaggerate their job search intensity when unemployed. There is strong evidence that Google searches suffer significantly less from social desirability bias than other data sources (Stephens-Davidowitz [2012]).

Second, these studies utilize the geographic coverage of Google searches. Even a large survey may yield small samples in small geographic areas. In contrast, Google searches often have large samples even in small geographic areas. This allows for measures of job search intensity and racism by media market.

Third, researchers often use Google measures that correlate with existing measures. Stephens-Davidowitz [2012] shows that the Google measure of racism correlates with General Social Survey measures, such as opposition to interracial marriage. Baker and Fradkin [2013] shows that Google job search measures correlate with time-use survey measures. While existing measures have weaknesses motivating the use of Google Trends, zero or negative correlation between Google searches and these measures may make us question the validity of the Google measures.

There are many papers that use Google Trends for "nowcasting" economic variables. Choi and Varian [2009] look at a number of examples, including automobile sales, initial claims for unemployment benefits, destination planning, and consumer confidence. Scott and Varian [2012, 2014b] describe the Bayesian Structure Time Series approach to variable selection mentioned earlier and present models for initial claims, monthly retail sales, consumer sentiment, and gun sales.

Researchers at several central banks have built interesting models using Trends data as leading indicators. Noteworthy examples include Arola and Galan [2012], McLaren and Shanbhoge [2011], Hellerstein and Middeldorp [2012], Suhoy [2009], Carrière-Swallow and Labbé [2011], Cesare et al. [2014], and Meja et al. [2013].

4.7 Google Trends: potential pitfalls

Of course, there are some potential pitfalls to using Google data. We highlight two here.

First, caution should be used in interpreting long-term trends in search behavior. For example, U.S. searches that include the word `[science]` appear to decline since 2004. Some have interpreted that this is due to decreased interest in science through time. However the composition of Google *searchers* has changed through time. In 2004 the internet was heavily used in colleges and universities where searches on science and scientific concepts were common. By 2014, the internet had a much broader population of users.

In our experience, abrupt changes, patterns by date, or relative changes in different areas over time are far more likely to be meaningful than a long-term trend. It might be, for example, that the decline in searches for `science` is very different in different parts of the United States. This sort relative difference is generally more meaningful than a long-term trend.

Second, caution should be used in making statements based on the relative value of two searches at the national level. For example, in the United States, the word `Jewish` is included in 3.2 times more searches than `Mormon`. This does not mean that the Jewish population is 3.2 times larger than the Mormon population. There are many other explanations, such as Jewish people using the internet in higher proportions or having more questions that require using the word `Jewish`. In general, Google data is more useful for relative comparisons.

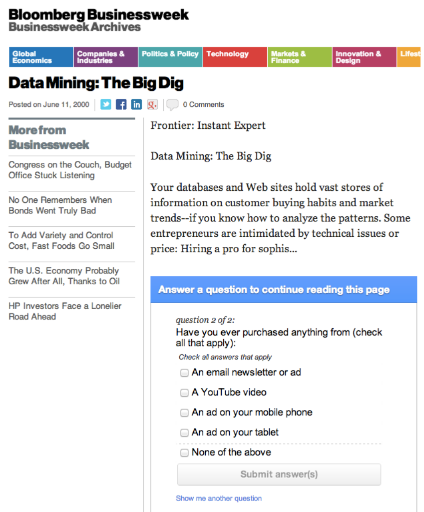


Figure 12: Example of survey shown to user.

5 Google Consumer Surveys

This product allows researchers to conduct simple one-question surveys such as "Do you support Obama in the coming election?" There are four relevant parties. A *researcher* creates the question, a *publisher* puts the survey question on its site as a gateway to premium content, and *user* answers the question in order to get access to the premium content. *Google* provides the service of putting the survey on the publishers' site and collecting responses.

The survey writer pays a small fee (currently ten cents) for each answer, which is divided between the publisher and Google. Essentially, the user is "paying" for access to the premium content by answering the survey, and the publisher receives that payment in exchange for granting access. Figure5 shows how a survey looks to a reader.

The GCS product was originally developed for marketing surveys, but we have found it is useful for policy surveys as well. Generally you can get a thousand responses in a day or two. Even if you intend to create a more elaborate survey eventually, GCS gives you a quick way to get feedback about what responses might look like.

The responses are associated with city, inferred age, gender, income and a few other demographic characteristics. City is based on IP address, age and gender are inferred based on web site visits and income is inferred from location and Census data.

Here are some example surveys we have run.

- Do you approve or disapprove of how President Obama is handling health care?

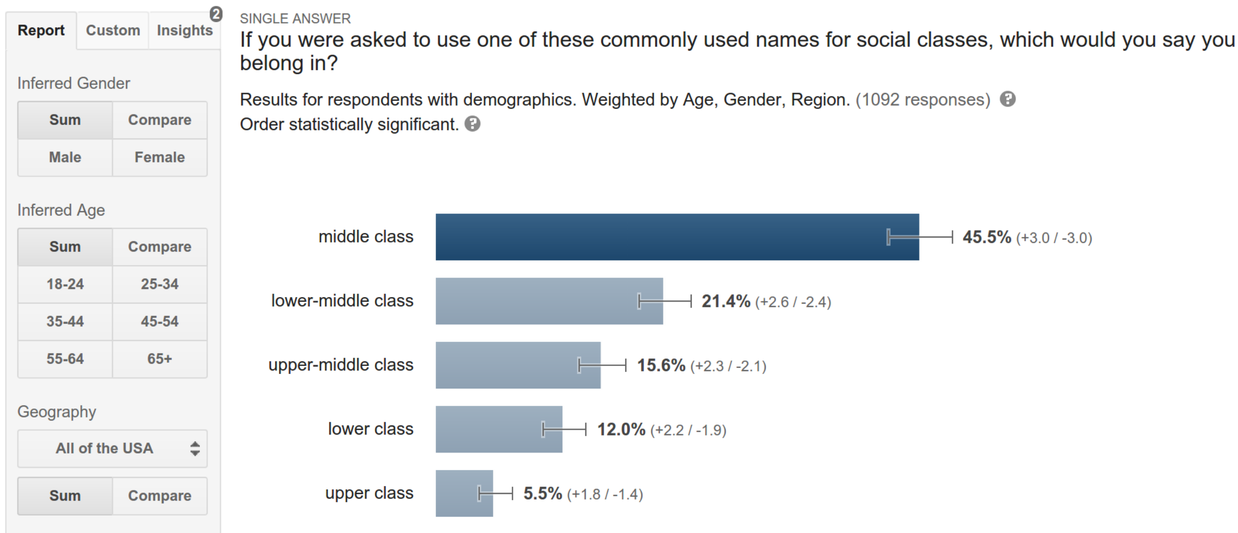


Figure 13: Output screen of Google Consumer Surveys

- Is international trade good or bad for the US economy?
- I prefer to buy products that are assembled in America. [Agree or disagree
- If you were asked to use one of these commonly used names for social classes, which would you say you belong in?

Some of these cases were an attempt to replicate other published surveys. For example, the last question about social class, was in a survey conducted by Morin and Motel [2012]. Figure 5 shows a screenshot of GCS output for this question.

Figure 14 shows the distribution of responses for the Pew survey and the Google survey for this question. As can be seen the results are quite close.

We have found that the GCS surveys are generally similar to surveys published by reputable organizations. Keeter and Christian [2012] is a report that critically examines GCS results and is overall positive. Of course, the GCS surveys have limitations: they have to be very short, you can only ask one question, the sample of users is not necessarily representative, and so on. Nevertheless, they can be quite useful for getting rapid results.

Recently has released a mobile phone survey tool called the *Google Opinions Rewards* that targets mobile phone users who opt in to the program and allows for a more flexible survey design.

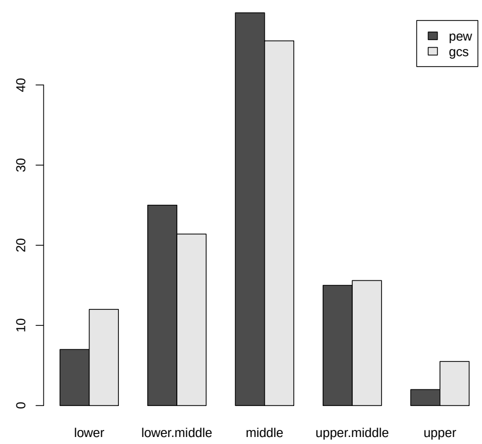


Figure 14: Comparing Pew and GCS answers to social class question.

5.1 Survey amplification

It is possible to combine the Google Trends data described in the previous section with the GCS data described in this section, a procedure we call *survey amplification*.

It is common for survey scientists to run a regression of geographically aggregated survey responses against geographically aggregated demographic data, such as that provided by the Bureau of the Census. This regression allows us to see how Obama support varies across geos with respect to age, gender, income, etc. Additionally, we can use this regression to predict responses in a given area once we know the demographics associated with that area.

Unfortunately, we typically have only a small number of such regressors. In addition to using these traditional regressors we propose using Google Trends searches on various query categories as regressors. Consider, for example, Figure 5.1 which shows search intensity for `[chevrolet]` and `[toyota]` across states. We see similar variation if we look at DMA, county, or city data.

In order to carry out the survey amplification, we choose about 200 query categories from Google Trends that we believe will be relevant to roughly 10,000 cities in the US. We view the vector of query categories associated with each city as a "description" of the population of that city. This is analogous to the common procedure of associated a list of demographic variables with each city. But rather than having a list of a dozen or so demographic variables we have the (normalized) volumes of 200 query categories. We can also supplement this data with the inferred demographics of the respondent that are provided as part of the GCS output.

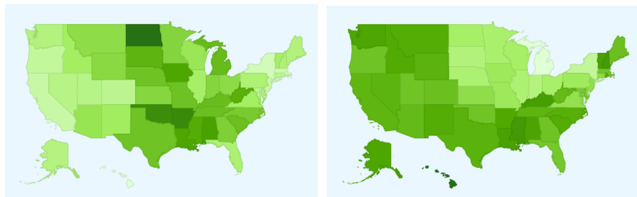


Figure 15: Panel (a) shows searches for `chevrolet`, while Panel (b) shows searches for `toyota`
.

5.2 Political support

To make this more concrete, consider the following steps.

1. Run a GCS asking "Do you support Obama in the upcoming election?"
2. Associate each (yes,no) response in the survey data to the city associated with the respondent.
3. Build a predictive model for the responses using the Trends category data described above.
4. The resulting regression can be used to extrapolate survey responses to any other geographic region using the Google Trends categories associated with that city.

The predictive model we used was a logistic spike-slab regression, but other models such as LASSO or random forest could also be used.[4] The variables that were the "best" predictors of Obama support are shown in Figure 5.2.

Using these predictors, we can estimate Obama's support for any state, DMA, or city. We compare our predictions to actual vote total, as shown in Figure 5.2.

5.3 Assembled in America

Consider the question "I prefer to buy products that are assembled in America." Just as above we can build a model that predicts positive responses to this question. The "best" predictive variables are shown in Figure 5.3.

The cities that were predicted to be the most responsive to this message are Kernshaw, SC; Summersville, WV; Grundy, VA; Chesnee, SC . . . The cities that were predicted to be the least responsive to this message are Calipatria, CA; Fremont, CA; Mountain View, CA; San Jose, CA, ...

[4]See Varian [2014] for a description of these techniques.

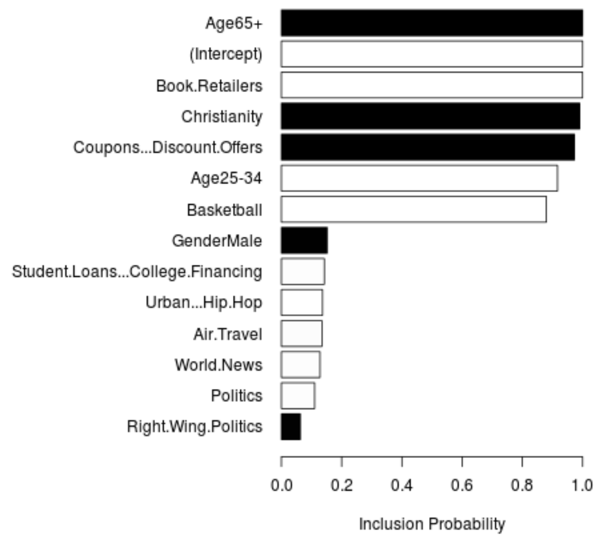
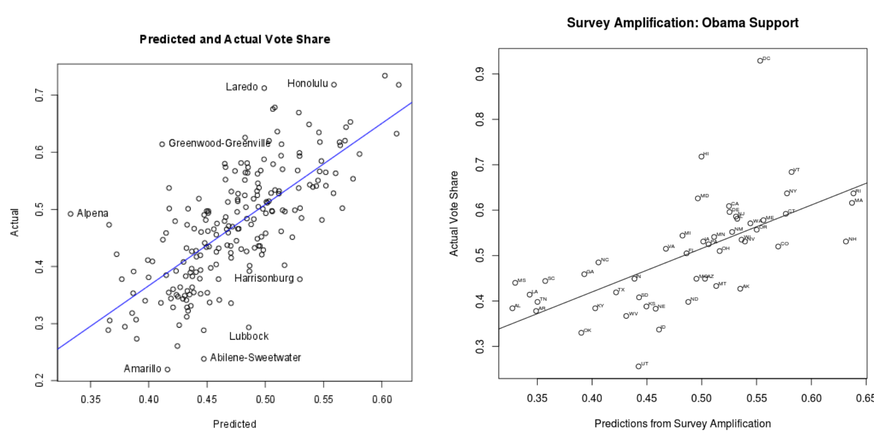


Figure 16: Predictors of Obama supporters



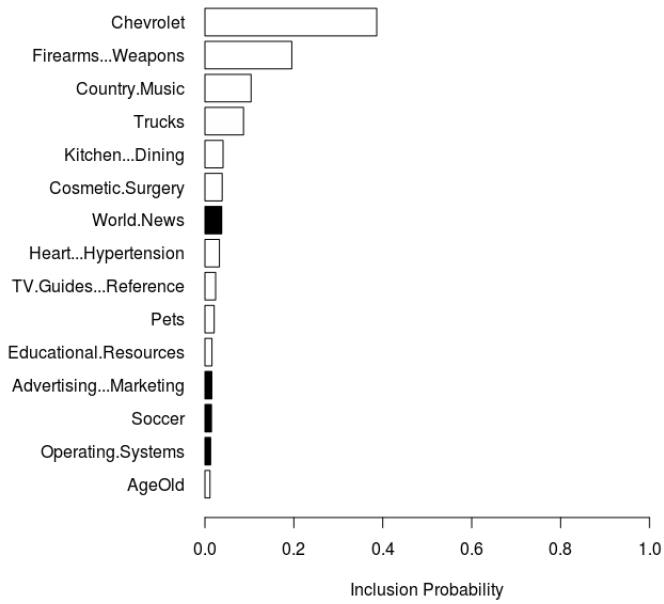


Figure 17: Predictors for “assembled in America” question

6 Summary

We have described a few of the applications of Google Correlate, Google Trends, and Google Consumer Surveys. In our view, these tools for data can be used to generate several insights for social science and there a many other examples waiting to be discovered.

References

Concha Arola and Enrique Galan. Tracking the future on the web: Construction of leading indicators using internet searches. Technical report, Bank of Spain, 2012. URL `http://www.bde.es/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/12/Fich/do1203e.pdf`.

Scott R. Baker and Andrey Fradkin. The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data. *SSRN Electronic Journal*, 2013.

Yan Carrière-Swallow and Felipe Labbé. Nowcasting with Google Trends in an emerging market. *Journal of Forecasting*, 2011. doi: 10.1002/for.1252. URL `http://ideas.repec.org/p/chb/bcchwp/588.html`. Working Papers Central Bank of Chile 588.

Antonio Di Cesare, Giuseppe Grande, Michele Manna, and Marco Taboga. Recent estimates of sovereign risk premia for euro-area countries. Technical report, Banca d'Italia, 2014. URL http://www.bancaditalia.it/pubblicazioni/econo/quest_ecofin_2/qef128/QEF_128.pdf.

Hyunyoung Choi and Hal Varian. Predicting the present with Google Trends. Technical report, Google, 2009. URL http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.

Rebecca Hellerstein and Menno Middeldorp. Forecasting with internet search data. *Liberty Street Economics Blog of the Federal Reserve Bank of New York*, Jan 4 2012. URL http://libertystreeteconomics.newyorkfed.org/2012/01/forecasting-with-internet-search-data.html.

Scott Keeter and Leah Christian. A comparison of results from surveys by the pew research center and google consumer surveys. Technical report, Pew Research Center for People and the Press, 2012. URL http://www.people-press.org/files/legacy-pdf/11-7-12%20Google%20Methodology%20paper.pdf.

Andreas Madestam, Daniel Shoag, Stan Veuger, and David Yanagizawa-Drott. Do Political Protests Matter? Evidence from the Tea Party Movement. *The Quarterly Journal of Economics*, 128(4):1633–1685, August 2013. ISSN 0033-5533. doi: 10.1093/qje/qjt021. URL http://qje.oxfordjournals.org/content/128/4/1633.full.

Alex Mathews and Catherine Tucker. Government surveillance and internet search behavior. Technical report, MIT, 2014. URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2412564.

Nick McLaren and Rachana Shanbhoge. Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, June 2011. URL http://www.bankofengland.co.uk/publications/quarterlybulletin/qb110206.pdf.

Luis Fernando Meja, Daniel Monsalve, Santiago Pulido Yesid Parra, and ngela Mara Reyes. Indicadores ISAAC: Siguiendo la actividad sectorial a partir de google trends. Technical report, Ministerio de Hacienda y Credito Pulico Colombia, 2013. URL http://www.minhacienda.gov.co/portal/page/portal/HomeMinhacienda/politicafiscal/reportesmacroeconomicos/NotasFiscales/22%20Siguiendo%20la%20actividad%20sectorial%20a%20partir%20de%20Google%20Trends.pdf.

Rich Morin and Seth Motel. A third of americans now say they are in the lower classes. Technical report, Pew Research Social & Demographic Trends, 2012. URL http://www.pewsocialtrends.org/2012/09/10/a-third-of-americans-now-say-they-are-in-the-lower-classes/.

Steve Scott and Hal Varian. Bayesian variable selection for nowcasting economic time series. Technical report, Google, 2012. URL `http://www.ischool.berkeley.edu/~hal/Papers/2012/fat.pdf`. Presented at Joint Statistical Meetings, San Diego.

Steve Scott and Hal Varian. Predicting the present with Bayesian structural time series. *Int. J. Mathematical Modeling and Numerical Optimisation*, 5 (1), 2014a. URL `http://www.ischool.berkeley.edu/~hal/Papers/2013/pred-present-with-bsts.pdf`. NBER Working Paper 19567.

Steven L. Scott and Hal R. Varian. Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modeling and Numerical Optimization*, 5(1/2):4–23, 2014b. URL `http://www.sims.berkeley.edu/~hal/Papers/2013/pred-present-with-bsts.pdf`.

Seth Stephens-Davidowitz. The Cost of Racial Animus on a Black Presidential Candidate: Evidence Using Google Search Data. *Working Paper*, 2012.

Seth Stephens-Davidowitz. Who Will Vote? Ask Google. Technical report, Google, 2013.

Tanya Suhoy. Query indices and a 2008 downturn: Israeli data. Technical report, Bank of Israel, 2009. URL `http://www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.pdf`.

C. Douglas Swearingen and Joseph T. Ripberger. Google Insights and U.S. Senate Elections: Does Search Traffic Provide a Valid Measure of Public Attention to Political Candidates? *Social Science Quarterly*, pages 0–0, January 2014. ISSN 00384941. doi: 10.1111/ssqu.12075. URL `http://doi.wiley.com/10.1111/ssqu.12075`.

Hal R. Varian. Big data: New tricks for ecoometrics. *Journal of Economic Perspectives*, 28(2):3–28, 2014.