Spikes and Variance: Using Google Trends to Detect and Forecast Protests

Joan C. Timoneda, Purdue University
2230 Beering Hall, 100 N. University St.
West Lafayette, IN 47907
timoneda@purdue.edu

Erik Wibbels, Duke University
280 Gross Hall, 140 Science Drive
Durham, NC 27708

Abstract

Google search is ubiquitous, and Google Trends (GT) a potentially useful access point for big data on many topics the world over. We propose a new 'variance-in-time' method for forecasting events using Google Trends (GT). By collecting multiple and overlapping samples of GT data over time, our algorithm leverages variation both in the mean and the variance of a search term in order to accommodate some idiosyncracies in the GT platform. To elucidate our approach, we use it to forecast protests in the United States. We use data from the Crowd Counting Consortium between 2017 and 2019 to build a sample of true protest events as well as a synthetic control group where no protests occurred. The model's out-of-sample forecasts predict protests with higher accuracy than extant work using structural predictors, high frequency event data, or other sources of big data such as Twitter. Our results provide new insights into work specifically on political protests, while providing a general approach to GT that should be useful to researchers of many important, if rare, phenomenon.

1 Introduction

The big data revolution has brought the social sciences enormous opportunities. A growing body of work uses data from Twitter (Barberá, 2015; Barberá et al., 2015; Bahrami et al., 2018; Timoneda, 2018), large bodies of text (Aletras et al., 2016; Muthiah et al., 2015), and satellite imagery to study everything from terrorist funding to neighborhood demographics (Do et al., 2018; Min, 2015; Gebru et al., 2017). One rich source of big data that political scientists have under-appreciated, however, is Google Trends (GT). GT is a service provided by Google that analyzes the popularity of a search query over a given time period and geographical area. Since Google search is ubiquitous, capturing more than 90 percent of the search market in the vast majority of countries, GT is built on a huge amount of data that could help researchers. Indeed, it has proved useful in detecting upsurges in everything from retail sales to international travel (Choi and Varian, 2012). Despite its capacity to provide high-frequency data on everything from protests to political identities, GT has been deployed sparingly by political scientists. And while some have adopted techniques to generate topic-specific predictions using GT (Mavragani and Tsagarakis, 2016) and others have used creative search terms to analyze hard-to-survey populations (Chykina and Crabtree, 2018), their approaches are hard to generalize to other topics of interest.

In this article we set out to: first, provide clarity around generating, accessing and processing GT data for event detection; and second, show the utility of the approach by using GT to both detect protest events ex-post and predict events ex-ante. On point one, we emphasize the importance of understanding GT's indexing of searches for its application to any particular analytical question. Given GT's approach, any single search on the volume of a term such as 'protest' or 'police violence' can yield deeply misleading results–this is true whenever search terms are relatively rare. Thus, we emphasize the importance of variation *within* a given search query across time rather than comparisons to other search terms or to Google's overall volume. We propose a 'variance-in-time'

method that smooths GT scores over time and incorporates variance in interest into attempts to identify events. The intuition is that more consistent interest in an event may help us predict it as well or better than the mean GT score.

Second, we deploy our approach to detect and forecast protests, a rare but politically important event type that has solicited a substantial body of academic attention. We first illustrate our approach using the search term 'protest' during the lead up to the Baltimore protests of April 2015 surrounding the death of Freddie Gray at the hands of police. We show that the coefficient of variation in hourly GT searches for 'protest' in the Baltimore metro area over a period of two months does an excellent job of detecting and predicting the events of the 25th of April, the day of the riots. We then test the efficacy of our approach by relying on known protest events in the United States from the Crowd Counting Consortium (CCC) data.[1] We collect GT data for 130 known protest events and for a set of 133 synthetic controls when no protest occurred. We collect repeated samples of weekly data at the hourly level for a period of 3 month for each event in the dataset, yielding a total of 3.6 million data points. We then use machine learning to produce out-of-sample forecasts. Our approach is more accurate than standard approaches in the protest literature, including those that rely on other sources of big data, like Twitter.

In doing so, this article makes several contributions. First, it provides a better understanding of how GT works, the nature of the data it produces, as well as its strengths and limitations. We hope to make it easier for other researchers to deploy GT for their own purposes. Second, we contribute to efforts to detect and forecast bouts of protest, instability and crisis events (Bowlsby et al., 2019). We show that GT can usefully detect protests – even modest sized ones – at higher levels of temporal and spatial resolution than other, more traditional, data sources. Our 'variance-in-time' method is *hourly* and leverages regional and city-level data. The method smooths GT scores over

[1]See https://sites.google.com/view/crowdcountingconsortium/home?authuser=0 for the data.

time and incorporates *variance* in interest into the final forecast, increasing accuracy. Ultimately, these benefits are not simply empirical as such data forces researchers to think theoretically about causal processes at finer grain than common, quite structural analytical approaches. Third, and despite the limitations inherent in GT, we show that it has considerable potential for predicting rare events. Thus, GT can provide a means for evaluating out-of-sample model performance even as it offers promise as a tool for policymakers and researchers interested in promoting, preventing, or responding to events in the world.

2 Google Trends in Academic Research

GT has become popular in academic research in recent years. It was initially used in the late 2000s as a tool to analyze trends and monitor the evolution of diseases, viruses, or financial markets. Over time, the focus of research has shifted to forecasting (Jun et al., 2018). Pelat et al. (2009) track the evolution of the flu and diarrhea, while Carneiro and Mylonakis (2009) use GT to monitor disease outbreaks in real-time. Vosen and Schmidt (2011) build a new indicator of private consumption based on data from GT. Examples of forecasting include Preis et al. (2013), who try to find 'early warning signs' for stock market moves through changes in google search volumes, Teng et al. (2017) dynamic model to forecast Zika epidemics, and Zhang et al. (2018)'s forecast of seasonal infections.[2] Lazer et al. (2014), on the other hand, show the limitations of using GT to predict flu virus outbreaks.

In political science, we know of two noteworthy applications. First, Mavragani and Tsagarakis (2016) use GT to predict referenda outcomes in Europe from 2014-2017.[3] The authors calculate the share of 'yes' and 'no' searches (in each language) in the weeks leading up to the referendum and show that the percentages closely match the final outcome. The manner in which they transform GT

[2] Oddly, the term 'repression' is seasonal in the United States, with searches lowest in July and August every year.

[3] These were Scotland's Independence vote in 2014, the Greek anti-austerity referendum of 2015, the Brexit vote in 2016, Hungary's 2016 migrant quota referendum, Matteo Renzi's 2016 constitutional vote in Italy and the 2017 Turkish referendum.

data into percentages for or against a referendum question is not clear, and the approach is limited to dichotomous phenomena (such as yes/no votes). Nevertheless, the research shows the applicability of GT data to crucial democratic processes. Second, Chykina and Crabtree (2018) measure issue salience for populations that are very difficult to sample using traditional survey techniques. They show how GT data can identify concerns by unauthorized immigrants about deportation from the U.S. in the aftermath of proposed policy changes and Trump's election by using the search term "will i be deported". With this method, they are able to identify key anti-immigration events: the passage of Arizona's SB1070 law, Trump's election and inauguration, and the Muslim travel ban imposed by the new President in early 2017. The authors also exploit GT's geographic search tool to show that a majority of searches came from states with large illegal immigrant populations: California, Texas and New York.

This growing body of work has discussed many of the challenges inherent in GT's data generation process. First, Google search trends can be responsive to media trends rather than the underlying phenomenon of interest (Lazer et al., 2014). Second, GT results can be sensitive to apparently modest changes in search terms, and this is particularly stark when it comes to rare, but important, events like protests. Third, the underlying Google search and GT algorithms are under constant revision, and this can affect results. Despite these pitfalls, some extant work demonstrates the potential of Google searches to predict human behavior, such as Brigo et al. (2014)'s study showing that people search for epilepsy-related terms to aid initial self-diagnosis. Similarly, we suspect that people interested in protesting use Google's search engine to find information about where and when to do so.

3 How Google Trends Works

Any use of GT for research purposes requires careful consideration of how it is put together. It ostensibly tells us how popular a given term was on Google's main search engine for a specific time frame and location. GT allows users to download data at the country, state, regional, metro area and city levels. Metro area and city are only available in the US, while regional data are available around the world in a country's main first political subdivision. GT data is available from 2004 to the present. Users can specify the *time frame* (or window) for which they want to collect GT data –a few hours, a day, a week, a few months or years. *Time blocks*, or the frequency over which the data is analyzed across the time frame, can be hourly, daily, weekly, or monthly, depending on the length of the time frame.[4] For time frames of up to one week, GT reports hourly data; for time frames of 9 months or less, daily data; for time frames between 9 months and 5 years, weekly data; and for time frames greater than 5 years, monthly. Researchers can access GT data in multiple ways, including R, Python and Google's own Trends application programming interface (API).[5] Note that Google's search engine dominates in most markets worldwide, with market shares of 87.6% in the United States, 90% in Canada, and 97% in India, among others.[6] Data from some countries, most notably China, are not available due to internet restrictions.

To generate the GT data a user receives, Google first draws a random sample of all search traffic

[4] GT's API returns hourly data if the search is of 8 days or less, daily data if the search is for over 8 days and under a nine months, weekly data between nine months and five years, and monthly data for queries of over five years. Note that this applies to *non-realtime data* (historical archive), a random sample for any query requiring data between 2004 and 36 hours prior to the search. GT also provides *realtime data*, which is a random sample of searches from the last 7 days – time blocks are minutes for the most recent 4 hour period and hourly for the rest of the week. We do not use realtime data for this article.

[5] In R, the `gtrendsR` package provides daily, weekly and monthly data –hourly data are not available. The package returns different samples every 180 seconds. Python's `pytrends` library is similar in nature to `gtrendsR` with two important differences: (1) hourly data is available for any 7-day period in the historical archive. This is not against Google's terms of service. It is the company's policy to produce hourly data for any search shorter than 1 week in the historical archive. and (2) it maintains the same sample if accessed from the same IP address. The third way to access trends data is through the Google Trends API, which researchers can access for free for academic purposes. `gtrendsR` and `pytrends` are not officially sanctioned by Google but are commonly used in research. For this article, we have obtained official access to Google's API.

[6] Germany and Australia: 93%; France: 91%; Italy, Spain and Brazil: 95%; Korea: 85%; Japan: 70%. In terms of continents, Google's market share is 95% in Africa and Oceania and 92% in Europe, Asia and America.

on its service for the requested location and time period. Google's justifies producing results from samples instead of all search data on efficiency grounds.[7] This sampling process creates variability in the results from multiple, identical requests to Google's API. This variability is particularly severe for low-volume search terms, which may not register in some of the sample draws. Thus, the same GT search for the same place and hour can generate somewhat different results (Choi and Varian, 2012).

Second, from this sample, Google's algorithm calculates the 'share' of overall traffic that the term represents, but it does not make that data available.[8] Instead, it transforms these 'shares' of overall volume into a normalized index that ranges between 0 and 100.[9] A score of 100 indicates the time block when the search term was at its most popular during the time period under examination – that is, when the search term achieved its greatest share of overall Google search traffic (Choi and Varian, 2012). Thus, every search term will achieve a value of 100 at some point over the time frame of the GT search. GT then indexes the rest of the time blocks over the entire time period *in reference to this highest point.* If a time block has half of the highest hour's volume, it receives a score of 50; if another time block has 100 times lower volume or less, it receives a score of 0. All other values within a time frame are indexed in relation to the highest value of 100, and thus receive scores between 0 and 99, inclusive.[10] Thus, the GT indexed score for a search term reflects its popularity *in relation to itself* over the time frame in a given location.[11]

[7] Details on these samples are not published by Google. The company states on their website that "Google Trends data is an unbiased sample of Google search data. Only a percentage of searches are used to compile Trends data." See `https://support.google.com/trends/answer/4365533?hl=en`.

[8] The only exception to this is the Health API, which produces the actual share of Google's overall volume for a term during a period of time. However, this API is for health research only.

[9] Data are normalized by total searches within geographical areas. As Google explains in the support page linked above, "[e]ach data point is divided by the total searches of the geography and time range it represents to compare relative popularity. Otherwise, places with the most search volume would always be ranked highest."

[10] GT will multiply all other values by a given factor to scale them against the high point of 100. Say a given term was 5 percent of all Google traffic during a window (hour, day, etc) at its highest peak of popularity for a time frame. This 5 percent is indexed as 100. Another value, say 2.5, is then multiplied by a factor of 20 (100/5), and receives an indexed score of 50.

[11] To understand the term's relative size, we can compare it to another search term in the same query. When two terms are compared to each other, the time at which either was the most popular receives 100, and the rest are indexed in relation to this one single peak.

This indexing has three important implications for researchers. First, it makes GT a poor tool for assessing a search term's overall volume or popularity relative to other search terms. All search terms obtain a score of 100 at some point during a given search period even if the term represents a small share of overall search volume. Second, GT results can be very sensitive to the time frame over which the index is constructed. A typically low-volume search term, such as 'protest', will always generate a 100 score, even over time frames when protests are not salient. And those '100' scores are not easily distinguishable from '100' scores that the term 'protest' would achieve during the Arab Spring or the Freddie Gray riots in Baltimore. Thus, any single GT search can be deceptive with regards to a term's overall salience amongst Google searchers. Third and relatedly, while GT indexing does require that even low-volume searches achieve a 100-point score over a time frame, it does not necessarily assign the lowest volume time block for a search term a '0'. A search for popular terms such as 'Trump' or 'Gmail', for instance, will not yield a 0 GT score in any time block. Conversely, a search for a low-volume term will always yield a 100 score (assuming there is enough data to produce a trend). An important implication is that high volume searches will have less variance in GT scores than low volume searches on average, even when the underlying distribution of raw searches have equal variance. This feature means that GT is particularly well-suited for detecting rare events, such as protests, because exaggerated pre-event interest will be reflected in reduced variation in GT scores. We exploit this feature below.

Figure 1 provides an example to elucidate these points; it plots the GT score of the word 'protest' in the Baltimore metro area between April 20 and 27, 2015, when violent protests shook the city in the wake of Freddie Gray's death at the hands of the police. In this example, we have *hourly* time blocks over the time frame of April 21-27. Since the time blocks are hours, we have a total of 168 data points (24 hours for each of the 7 days) for the term 'protest' in Baltimore. The large spike in interest the night of April 25 roughly coincides with the most violent episodes of the week – the

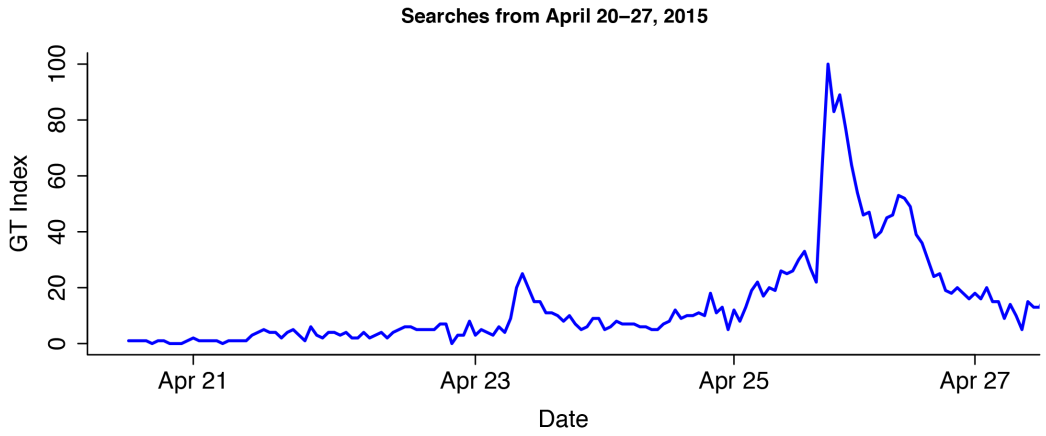

Figure 1: Hourly GT scores for 'protest' (April 20-27, 2015 – Baltimore Metro Area).

Baltimore 'uprising' or 'riots'. Interest was relatively low earlier in the week and picked up slightly on April 23rd, the day of the first large, but nonviolent, protest.

Notice that the y-axis is bound between 0 and 100. As described above, this is a function of how Google builds the GT index. First, GT assigns a value of 100 to the hour with the highest level of search volume for 'protest' within the overall time frame. In this case, 'protest' was most googled in Baltimore between 6 and 7pm Eastern on April 25. We cannot know 'protest's share in overall Google search volume in Baltimore during that hour, but we know it was at its highest share of all searches that week. Hence, GT assigns that one hour a score of 100. GT indexes the rest of the hourly scores for the week *in reference to this highest point.* For instance, on April 23rd between 2 and 3pm – the first small peak of the week – searches for 'protest' represented one third of the search volume of the week's most popular hour (again, April 25 6-7pm).

This example also helps clarify why GT data is poor for assessing a search term's overall popularity relative to all search terms. For rare search terms, an index of 100 is achieved even with low absolute search volume. Imagine, for instance, that searches for 'protest' were 23 percent of all search traffic on Google in Baltimore between 6 and 7pm on April 25, 2015. 23 percent is reflective of an objectively popular search term across all Google searches, but that is not what GT reports.

Because 23 percent is also the peak popularity of 'protest' searches over the April 21-27 period, GT produces a score of 100 for April 25 between 6 and 7pm and all other time blocks are indexed against it. However, some other objectively unpopular term – say 'David Foster Wallace' – could also receive a score of 100 for April 25 between 6 and 7pm, even if it peaked at only 2 percent of total search, as long as that hour was when 'David Foster Wallace' was most searched over the time period.[12]

Figure 2, on the other hand, shows how GT indexing makes any single GT search very sensitive to the time frame of analysis, and thus, potentially deceptive. The figure shows hourly output on GT for the weeks of April 13 to 21 (a), April 17 to 24 (b), April 20 to 27 (c) and all of 2015 in the Baltimore metro area for the search term 'protest'. Plot (a), like the others, has one peak of 100 on April 17, but the variation is high across the period. Two other peaks are close to the 100 mark on early April 13 and late April 20. Notice that the peak from (a) disappears in (b), as April 17 search volume now pales in comparison to the peak of April 24. Variation is much lower, indicating that the surge of April 24 could potentially constitute an event. However, as shown in (c), that peak itself is much smaller than the one on April 25 at night (7pm EST), the day of the riots. Thus, one snapshot of GT data will always produce a high point of 100, but whether that peak is constitutive of an important event is difficult to tell from any single capture. This indexing adds a layer of difficulty in utilizing GT data. Our approach, described below, exploits variance in a search term's own trend to circumvent the limitations inherent in one-off GT results.[13]

[12] The DFW 100-point GT peak is actually on April 24 for that week. The GT 'compare' option, which allows one to compare two search terms, on the other hand, shows that the overall volume of searches for 'protest' was much higher than 'David Foster Wallace' across the whole week.

[13] A way to conceive of this is as a 'fixed-effect', where we analyze within-term variation rather than contrast differences across terms or solely in relation to overall search volume.

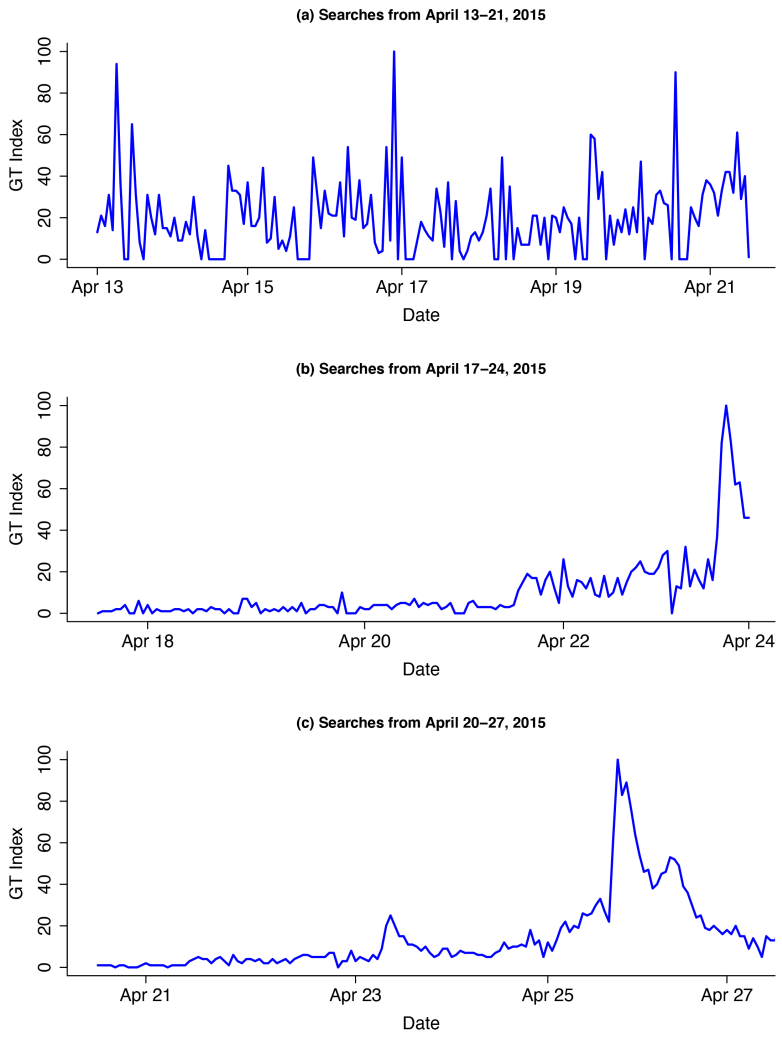

Figure 2: Hourly GT scores for 1protest' over three different search periods (April 2015, Baltimore Metro).

4 Event Detection and Forecasting

Potentially one of the most promising applications of GT for the purposes of political scientists involve 'event' detection and prediction. An event is an instance of a given phenomenon that takes place in a specific location for a finite period of time. In the event detection and prediction literature, researchers typically code events as involving an initiator (which could be anything from

a person to the police to a country), a target (ditto), a magnitude and a location (Schrodt and Gerner, 1994; Gerner et al., 1994; King and Lowe, 2003). There are large machine learning projects, including GDELT (Leetaru and Schrodt, 2013) and ICEWS (Boschee et al., 2015), that aim to turn the world's news into event data on everything from international conflict to mentions of specific world leaders. Other approaches rely on human coding of news and qualitative research to detect everything from protests (Fisher et al., 2019) to insurgent attacks (Raleigh et al., 2010) to international crises (Brecher et al., 2016). Event data has been used to model an enormous range of outcomes in the social sciences. Given Google search's ubiquity around the world, we believe it offers some important advantages for event detection and prediction.

Most importantly, GT can overcome some important challenges and limitations in current sources for event data. First, GT can improve on the frequency of most event data. While some such data is reported at the annual level (Sarkees and Wayman, 2010), others are more fine-grained, but rarely get better than daily (Leetaru and Schrodt, 2013). Properly tuned GT can report data at the *hourly* level. While this might not be useful for all social scientific enterprises, there are some (such as protests or financial markets) where hours can be crucially important; in some areas of research, the availability of higher frequency data might unveil analytical puzzles heretofore unexplored. Second, much event data in political science is reported at the national level, and thus lacks geographic nuance. In many cases, GT allows researchers to access at least the first subnational divisions within countries. Third and perhaps most importantly, GT overcomes some of the challenges inherent in more curated big data approaches that rely on the news. The challenges are myriad, including bias in underlying news sources, choosing which sources to rely on, developing and updating actor dictionaries, and de-duplicating reporting of events across multiple sources (Grimmer and Stewart, 2013; Leetaru and Schrodt, 2013; Lucas et al., 2015; Schrodt, 2012). These challenges have resulted in a number of disagreements in empirical work on everything from the extent of reporting bias

(Weidmann and Ward, 2010; Hollenbach and Pierskalla, 2017; Weidmann, 2016) to the utility of rules-based approaches over more advanced natural language processing techniques (Schrodt et al., 2014; Schrodt, 2012). By relying on billions of Google searches and a clear understanding of how GT indexes, GT side-steps many of these problems and offers the possibility to detect all manner of events that individuals around the world care about, including general elections, protests, police crackdowns, important legal changes, etc.

Below we focus specifically on identifying and forecasting protests using GT. There are two reasons for this choice. First, there is a lot of ground-truthed data on protest events, which aids with the development of models for event detection and forecasting (Fisher et al., 2019). Second, a large body of work has explored the predictors of protest, instability and civil violence using in-sample and out-of-sample forecasting (Fearon and Laitin, 2003; Gurr and Lichbach, 1986; Bowlsby et al., 2019; Goldstone et al., 2010; Chenoweth and Ulfelder, 2017; Hegre et al., 2013; Cederman and Weidmann, 2017; Ward et al., 2010). These works use different structural and micro-level predictors to issue forecasts, but the predictive power of these variables inevitably varies over time. A strong predictor in 1980 may not be a strong predictor today (Bowlsby et al., 2019). One advantage of using data from GT is parsimony. Accurately predicting protest outbreaks using intensity and variation in search interest over time for only one keyword simplifies the model substantially and, most importantly, should be consistent across time (Bowlsby et al., 2019). It also circumvents challenges inherent in text-based approaches using other sources of big data, such as Twitter, to predict protest outbreaks (Korolov et al., 2016; Bahrami et al., 2018). Despite recent progress, tweets remain especially difficult for machines to filter and classify (Zhou et al., 2015). Given its ubiquity and often superior geolocalization,[14] Google search data might also offer more advance notice of impending events than Twitter, which has proven fairly accurate at forecasting a day or

[14]Since most Twitter users do not share their exact coordinates, researchers are forced to use self-reported location, which users often fail to update.

two ahead of protests. All told, our focus on temporal variation in a single search term on GT approach is more parsimonious than current models.

Lastly, while our focus is on event detection and forecasting, there are other creative applications of GT in political science worth mentioning. One is the ability to survey populations using search strings that only specific groups of people are likely to use as in Chykina and Crabtree (2018). Another application involves creating proxies for common, important, but infrequently and/or misreported variables such as unemployment. Gülenay-Chadwick and Şengül (2013), for instance, show that GT is successful at nowcasting (i.e. predicting present values) of the non-agricultural unemployment rate in Turkey, which is only officially reported at monthly frequency. Finally, GT can reveal public attitudes without the expense of large sample surveys. Mavragani and Tsagarakis (2016), for instance, show how search volume on Google can reveal people's preferences on political referenda and predict their result. These are a few examples of many potential applications of GT in political science, and we encourage further research along these lines. Whatever the application, it is important to approach GT with a clear understanding of how it works, since naive applications can produce misleading results.

5 The Variance-in-Time Approach to GT data

We propose a 'variance-in-time' approach to identifying and forecasting events using GT. The intuition behind the approach is that an analysis of high frequency, repeated samples of search terms can both successfully detect politically relevant events, even rare ones, and provide an important tool for forecasting them. Taking repeated samples of GT data over time allows us to incorporate variance into the analysis. Given the 'black box' nature of GT indexing, focusing on *consistency* of interest rather than changes in the GT index itself can better reveal latent interest in an event (particularly a rare event) before it happens.

The 'variance-in-time' approach consists of downloading GT data for all possible weekly combinations over a given time period. For our purposes, we collect data in 3-month periods, which have 82 full weeks, on average, with overlapping days.[15] In the case of Freddie Gray, the riots took place on April 25, 2015 between approximately 4pm and 10pm. We collect data between March 1 and May 31. The first week in the sample is March 1 to March 7, the second is March 2 to March 8, and so on until the last week of May 24 to May 31. As described in a previous section, since data is collected in weekly windows, GT returns hourly data, so every week contains 168 data points after taking the mean for each hour.[16] After seven days, each hour in the sample has been observed a total of seven times.[17] GT scores can change slightly, substantially or stay the same for each data point every week, depending on whether the 100-point peak changes or stays the same. Each 3-month collection produces 13,776 data points.[18]

After collecting the raw weekly data, we first take the mean μ GT index score for each hour in the sample. Calculating the mean serves to minimize the signal provided by any given '100' score that might result from the combination of low search volume and GT's requirement that one hour in a week be indexed to 100.[19] We obtain a mean hourly GT index score for a total of 2,160 observations and calculate the standard deviation σ in GT scores in 12 hour intervals. We use 12 hour intervals to maximize granularity while also allowing us to calculate meaningful standard deviations.[20] In the context of GT, 12-hour periods are often long enough to detect meaningful change –many events see large increases and decreases in interest in windows of 12 hours or less.[21]

[15] For instance, July 1-7, 2-8, 3-9, etc. In our application in the next section we collect data for 2 months before protests and 1 month after them. The time window should be sufficient to establish a baseline for each protest before interest picks up.

[16] There are 168 observations per week (24*7) after taking the mean for each hour. Each full week contains 1176 total data points in the raw data (168*7).

[17] The first and last week have fewer than seven observations because they fall out of the collection sooner. March 1, for instance, is only collected once, in the week between March 1 and 7. Thus, March 7 is the first day for which we have seven observations for each hour. Collection must start somewhere, and this will always happen at the beginning and end of the full collection interval.

[18] 82 weeks * 168 hours/week.

[19] If a '100' is a result of noise in any given sample, it is likely to revert to a much smaller number in other samples.

[20] These intervals go from midnight to noon and noon to midnight every day.

[21] For instance, once an event is triggered, it only takes a few hours for search interest to skyrocket. Variation in

Using the GT means per hour and the standard deviation, we then compute the *coefficient of the variation* for each hour, i.e. σ/μ.[22]

The coefficient of the variation helps us incorporate the variance in GT scores into our calculations and go beyond mean scores, which are inherently noisy when trying to detect rare events (like protests) in contexts of low search volume. The intuition here is that *consistency* of interest, and not just the level of interest, can both indicate as well as forecast an event. Given the nature of GT indexing, a topic may become more popular without its mean increasing very much, because another –and unrelated– peak of '100' may be drowning out the increase in the mean for the new event. Lower variance, on the other hand, will always reveal more consistent interest in a search topic, which in turn suggests that the topic is becoming more popular even if the mean GT score does not reflect that. Incorporating the variance helps us paint a fuller picture of search interest. Definitionally, holding the mean constant, the coefficient of the variation will be smaller when variation in search interest is low (smaller σ). When variance is high, the coefficient will be larger. Therefore, we expect a lower coefficient of variation to be associated with a higher probability of observing a given event –in our case, a protest.

Part of the reason for this expectation lies with the behavior of the mean μ. With a single sample of GT data, a high GT score can indicates broad interest in a topic, but it can also result from a small increase in interest in a context of low search interest. Recall that GT *always* produces a score of 100 for every collection, as it assigns 100 to the time block that had the highest share of searches among overall traffic on Google. If one collects data for a week where interest was quite low throughout, scores are likely to be abnormally *high*, (say, around 50), because many hours will

interest may have been informative prior to the trigger (our focus in this article) but major changes occur within just a few hours, often much fewer than 12 hours. This means that 12-hour periods, while relatively short, are meaningful in the context of GT and allow us to calculate the standard deviation in search interest confidently. More importantly, they provide granular data that allow us to detect meaningful changes in interest more precisely. We also run this calculation for 24-hour periods.

[22] We do not multiply by 100.

achieve '100' scores even without serious search interest. The hour with the most searches will receive a score of 100, but with rare events search volume in other hour blocks will be similar. This means that many hour blocks will receive relatively high GT scores, as they are indexed against the peak hour. Some will go to zero because no search volume occurred. Thus weeks with little interest have high GT means, rather than low ones, but most of this variation is a function of many different low search blocks actually achieving a '100' score. Conversely, when interest becomes more consistent, the mean will initially *decrease*. Consistent interest reduces the number of 100-point peaks driven by low-volume searches and lowers the values of other data points around a true spike. Hence, the mean will decrease when a search topic begins to garner interest. It will then increase again when interest is consistent and high, even though there will be fewer peaks of 100.

This fluctuation in the mean creates interesting dynamics in the coefficient of variation. A high mean will make the coefficient smaller. We expect higher means both in periods with low interest (noise) and when a rare event is likely to take place (i.e. true high interest), and lower means when the topic begins to attract attention on GT. Therefore, the coefficient should be largest when the issue receives very little attention on GT, become smaller as the issue becomes more popular, and be at its lowest values when attention is at its peak. In other words, the numerator (σ) decreases as interest increases, and the mean evolves in a u-shape pattern: it is higher with low interest, lower with increasing interest, and high again with true high-volume interest.

Figure 3 shows this graphically for the case of Freddie Gray. The orange line plots the mean GT scores for all hours between March 1 and May 31, 2015, as well as the mean and the variance for each day in the sample (red and black lines).[23] The variance starts decreasing substantially around April 12, the day of Freddie Gray's arrest. The hourly means follow a similar trajectory, before they increase again as the protests take place. Note the many cases of 100 point scores, even early

[23]The daily mean is intended to guide the reader visually. The computation of the coefficient of the variation is done using the hourly mean (orange line).

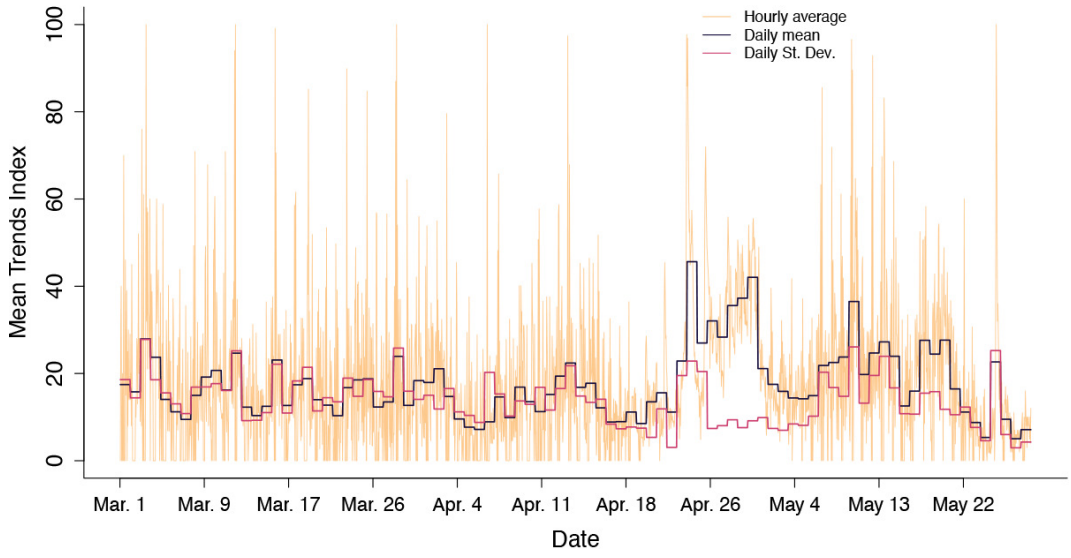

Figure 3: Means and Variance over time in Baltimore, March 1 - May 31, 2015.

in the month before any of the incidents bearing on the Gray protests took place; many of these high scores reflect GT indexing on a low volume search. There is only one mean peak greater than 70 between April 12 and May 6 (the peak of April 25), but at least six in the month prior to April 12.

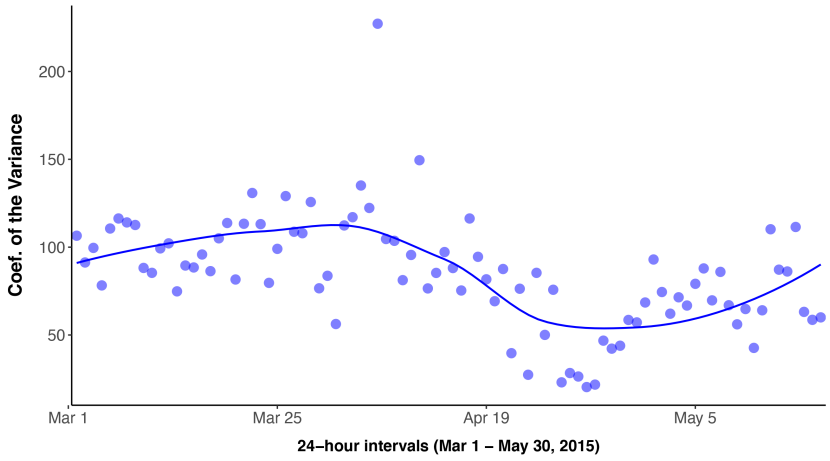

Figure 4: Evolution of the coefficient of the variation in Baltimore, March 1 - May 31, 2015.

Figure 4 plots the resulting coefficient of the variation – the standard deviation over the mean. Points are much farther apart on the left side of the plot before April 12 (day 43). After that,

points are first closer to one another (lower variance) and then decrease sharply before and during the protest. Figures 3 and 4 show the potential of the coefficient of variation to *detect* and *forecast* a protest event days before the protest actually takes place. We can detect an event by observing multiple points in a sequence in which the coefficient of the variation is much lower than the expected trend (see Figure 4). In terms of forecasting, there are two aspects of Figure 4 that we can exploit: variance overall decreases after an initial positive shock to variance around April 4. We cannot issue a reliable forecast based on one event and build a generalizable model from it, but we think there is sufficient evidence in the Freddy Gray event to suggest that our approach to GT data may yield accurate forecasts. We develop a generalizable forecasting strategy in the next section.

6 Forecasting Protests in the U.S.

6.1 Data

To systematically test the forecasting capacity of our variance-in-time approach, we apply it to an expanded sample of 130 protest events and 133 non-events between January 2017 and May 2019 in metro areas in the United States.[24] We identify known protest events using data from the Crowd Counting Consortium (CCC), which tracks all protests in the United States since January 2017.[25] To build our sample of known protests, we select all protests in CCC sample larger than 2,000 participants (39 events), take a random sample of 42 events of between 1,000 and 2,000 participants, and another random sample of 48 small protests between 1 and 100 protesters. The 130 protest sample is thus well-balanced in terms of protest size.[26] The largest protest event has

[24] There are 130 unique metro areas in the sample.

[25] https://sites.google.com/view/crowdcountingconsortium/home.

[26] We selected protests of varying sizes to ensure that there was variety in the sample regarding the potential level of attention that a protest received. The numbers are uneven across sample groups because: 1. there were only 39 protest events with more than 2,000 participants; and 2. some metro areas in the CCC data are not reported in GT. For instance, the CCC identified a protest in Salem, Oregon, but GT does not report data for the city of Salem and its metro area. We have not been able to include these protests in the sample.

an estimated count of 60,000 participants and the smallest has 1.[27]

We construct a synthetic control group by first taking a random sample of metro areas and dates from the CCC data for which we know a protest did not take place for a period of 3 months. We select 130 periods randomly and, within the 3 months for each period, we create a false protest date at day 60, i.e. two months in. Since no protest occurred on day 60 or within the third month, we expect the coefficient of the variation to fail to predict these protests.

We then apply our variance-in-time approach to each of the 263 protest and control 'events' in the sample. We collect data from GT for a period of 3 months –2 months prior to the event and 1 month after it. For example, for a protest that takes place in 'Austin, TX' on '9/1/2018', we request weekly data for the Austin metro area (code US-TX-635) from July 1 until September 30, 2018.[28] The search term used in all collections is 'protest'. The raw data consists of 3.623 million observations.[29] After calculating the mean for each hour, the sample contains 568,080 total observations. The unit of observation is the city-hour. There are a total of 280,800 observations for protest events and 287,280 for non-events. Lastly, note that protests are a rare event, as they account for only 1.11 percent of the data.

We apply (1) a statistical model to perform in-sample prediction with our full dataset and (2) a forecasting model using logistic regression and K-fold cross-validation for out-of-sample prediction on aggregated data.[30]

6.2 In-Sample Results

First, we fit a generalized linear model using the full sample of 568,080 observations with a set of lagged values of the coefficient of variation as the main predictors. We also add metro-area fixed

[27] Estimates by the CCC. Only 2 protests have a size of 1 and that's the result of random sampling of protests under 100 participants. A list of all protests included in the sample, with participant estimates and their claims, is included in the Appendix.

[28] Again, protests do not overlap within periods. For the forecasts we use data only from the first two months.

[29] 168 hours/week * 82 unique weeks * 263 events = 3.623M.

[30] The results from the models in table format are reported in the Appendix.

effects. The coefficient of variation is lagged up to five hours before the protest event and then daily until 7 days before the protest date. Protest is dichotomous: days with protests are set to 1, and days without protests to 0.[31] The results are identical using a linear probability model – the sample size is so large that logistic regression coefficients remain largely unbiased even in the presence of a large number of parameters (Beck, 2020).

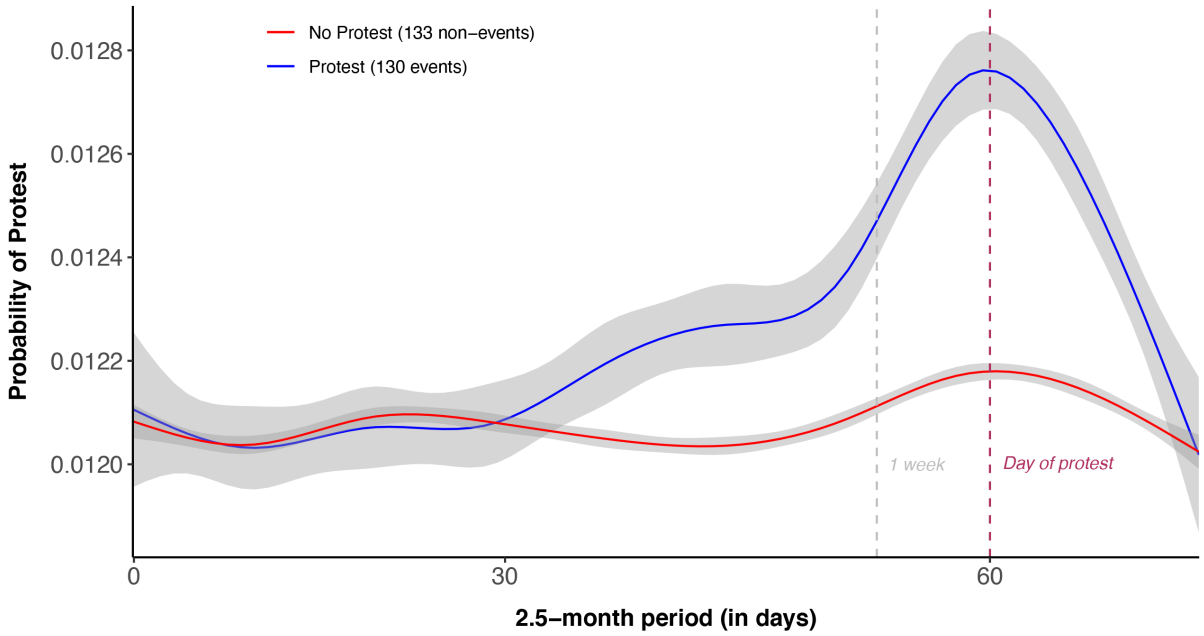

Figure 5: Effect of the coefficient of the variation in GT scores on protest.

Figure 5 shows the predicted probabilities for protest and no protest groups based on the logistic model.[32] We refer to the protest subset as 'positive' and the synthetic control as 'negative'. The x-axis are days and the predicted probabilities of protest are on the y-axis. All protests in the sample are set to take place on day 60. The results are computed using the full model, not a split sample.[33]

The contrast between the predictions for our true protest sample and the synthetic control is

[31] This applies to all 24 hours, since we cannot know exactly at what time the protest started.
[32] As for statistical significance, lags of 5 and 7 days are especially significant along with more immediate lags of up to 24 hours.
[33] They are the averages of the predictions for the two groups from the complete model.

striking. The model places the highest probability of protest in the positive group at the day of the protest. The probability of a protest is 7 percent higher in cities where a protest occurred in the data than in our control group, a statistically significant difference. Moreover, as Figure 5 shows, the probability curves for these two groups begin to diverge much earlier. Significant differences can be observed one and two weeks before the day of the protest. This creates an opportunity for predicting protests further ahead than current models (Bahrami et al., 2018). On the other hand, the probability of protest for the control group is mostly flat for the entire period. It is also indistinguishable from the treatment group's probability of protest between one and two months prior to the protest date.

6.3 Out-of-Sample Results

We apply a logistic regression classifier and use K-fold cross-validation to produce out-of-sample forecasts. We follow the literature in the selection of the model (Korolov et al., 2016; Bahrami et al., 2018).[34] The output of the logistic regression is the probability that a protest takes place in a metro area in a given time period. The classifier transforms the probability for each protest event into a binary value. We allow the algorithm to select a cutoff point to perform the conversion. The model assigns a 1 if the predicted probability is above the cutoff point and a 0 otherwise. Thus the model's output is a dichotomous variable that indicates whether a protest is predicted or not.

We use K-fold cross-validation to perform out-of sample forecasts setting the number of folds at 10.[35] K-fold cross-validation divides the sample into ten random splits, keeping nine as training data and one as test data. For each unique group, we fit a model on the training data and evaluate it on the test set. We obtain a set of ten different results, which we use to summarize the skill of the

[34] Results obtain using other classifiers, such as linear regression and random forest.

[35] We select k=10, a common choice in applied machine learning, as it is has been "shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance" (James et al., 2013). The model shows similar results using a wider range of k-values.

model. This procedure is particularly useful to test the skill of our model considering the relatively small sample size of protests. Recall that our final dataset for the forecasting model consists of 263 observations –130 protest events and 133 null events. A simple train/test split, in which we split the sample into training and test sets only once, generally yields more biased and optimistic estimates of model fit, a problem that is exacerbated in smaller samples (see Brownlee, 2018).

Our models test the predictive capacity of the coefficient of the variation on protest. The first model uses the maximum and minimum values of the coefficient of the variation between 1 and 3 weeks prior to the protest. These variables are aggregated at the protest level. For instance, if a protest took place on June 1, we would calculate the maximum and minimum values of the coefficient of the variation between May 7 and May 24. This effectively lags our prediction one full week (our previous in-sample model indicates this is highly predictive of a forthcoming protest). This lag shows that our model can forecast protests further in advance than current work, which usually issue three-, two- and one-day forecasts (Korolov et al., 2016; Bahrami et al., 2018). In the model, we also include the maximum and minimum values between 3 and 5 weeks and 5 and 7 weeks prior to the protest.

Figure 6 shows the area under the Receiver Operating Characteristic (ROC) curve for the first model. The ROC curve illustrates the predictive capacity of a model, showing the trade-off between true positive and false positive rates as we vary the decision threshold. The area under the ROC tells us accurately a model classifies true positives. The closer it is to 1, the better the model performance. Here, Figure 6 provides strong evidence that the coefficient of variation can be highly predictive of protest events. The area under the ROC score for the main model is 0.92, while out-of-sample accuracy stands at 0.85, both high scores.[36] Using the coefficient of variation, we are correctly predicting 85 percent of protests on unseen data. This translates into 110 protests

[36]In-sample accuracy is 0.87 and the AUC is 0.96.

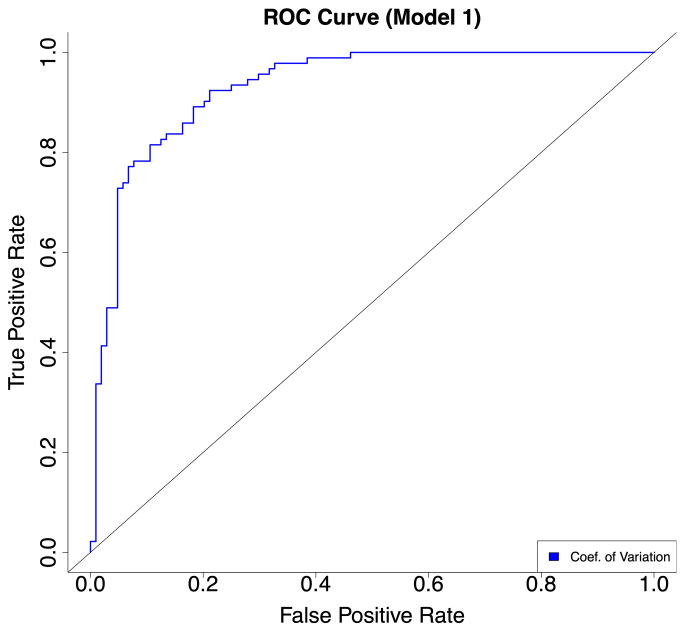

Figure 6: Effect of the coefficient of the variation in GT scores on protest.

correctly identified by the model for a recall of 0.88. We fail to predict 13 protests that did take place. Conversely, we're identifying 114 true negative events and incorrectly identifying 26 protests that never happened (false positives).

We build a second model that calculates daily protest probabilities per urban area and predicts the day when a protest is likely to take place. Our predictors are the coefficient of variation (logged) lagged from 1 up to 14 days before the protest. Since protest events are rare in the training data (0.8%), we use gradient boosting (gbm) to circumvent known pitfalls of logistic regression with highly imbalanced outcome variables (King and Zeng, 2001; Beck, 2020).[37] Given rare events, accuracy scores are not a great tool to evaluate the model as high train and test accuracies are solely driven by correctly predicted zeroes.[38] We thus turn again to area under the ROC scores and report the confusion matrix, both of which are more instructive of model performance. Model 2 has a mean AUC score of 0.90, which again provides strong evidence of the power of the coefficient of

[37] We use the `gbm` package in `R` and specify an optimal shrinkage term (learning rate) of 0.2, an interaction depth of 12, and 10,000 decision trees.

[38] Model 1 aggregates the data at the level of the urban area, leaving only 263 observations. Model 2 uses daily scores of the CV for each urban area, leaving 15,780 observations (60 days * 263 urban areas).

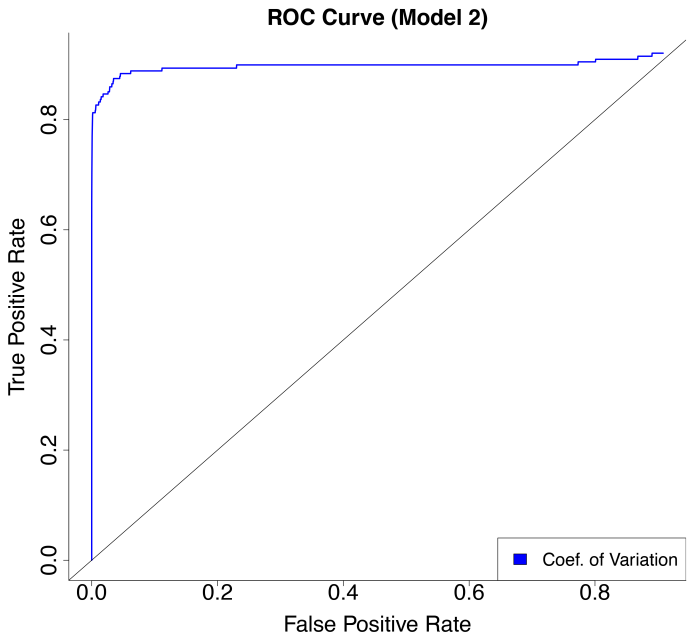

Figure 7: Effect of the coefficient of the variation in GT scores on protest.

the variation to predict protest events.

The confusion matrix confirms the strength of the model. A confusion matrix compares the predicted values of an outcome from a model versus the actual values in the data. We look for (1) a high number of matches between predicted true positives and actual true positives, and (2) a high number of matches between predicted true negatives and actual true negatives in the data. Over 10 folds,[39], which means that we need to sample on urban area and we cannot use independent observations for our folds. the model correctly predicts an average of 23.6 protest days in the test data and only fails to predict 6.4 known protest events (false negatives). 78.4 percent of protests are thus correctly identified, as are all non-protest observations. Moreover, the model produces no false positives on average, leaving an F_1 score of 0.88. These results using gbm are especially strong considering how rare protest events are in the second dataset (130 out of 15,780 observations). The gbm model greatly improves on the performance of the logistic regression classifier, which does no better than random chance.

[39]Folds are custom-made to sample by *urban area* rather than by observation. We are predicting protest in a given day and urban area using lagged independent variables

These findings show that the coefficient of variation can predict protest from GT data earlier than current models. For reference, we use innovative work by Bahrami et al. (2018) that leverages Twitter data to forecast protest events. The authors collected tweets between November 9 and 16, 2016 with hashtags related to protests against Donald Trump's election and used machine learning to forecast when a protest was likely to take place. Their model predicts protests perfectly for November 15 and 16, dates in which protest events were common throughout the United States. Three days prior to these major events, however, their model issues protest probabilities around coin-toss levels. Our model manages to issue much more accurate forecasts up to a week in advance, leveraging small but detectable changes in Google search activity.

7 Conclusion

In this paper we have (1) explained how GT data is generated and why its characteristics make it a difficult tool to use; (2) showed how variation in GT results for search terms (and especially low-volume search terms) in repeated samples can be used to detect events like protests; and (3) tested the efficacy of our approach to forecast protests in metro areas in the United States. We also provide practical advice on the appropriate time frame for searches and what to look for in the coefficient of variation. Ultimately, any use of GT for research purposes will require some tailoring for the specific research question at hand, but we believe GT offers enormous potential for further work in the social sciences. It is an easily available tool that aggregates information about the everyday curiosities of many, many people across the world that is encoded in billions of their daily searches. To the extent those searches bear on difficult-to-study and rare, but pressing, events, we encourage others to explore GT as a source for truly big, promising data.

While our specific application is to "protest" searches in the U.S., it is worth considering the broader set of events that GT might prove useful for researching. First, by exploiting GT's dis-

tinctive indexing, our approach is only likely to be useful for events that have relatively low search volume in normal times. Thus, searches for 'deportation' and 'MS 13' are promising because they might provide insight into locally salient, but infrequent, events like ICE raids or a gang offensive. On the other hand, high-volume searches have fewer spikes and few '0's to exploit. We suspect one could develop a GT-based approach to detecting events even for high-volume search terms, but our approach would not help. Second, our efforts suggest that as one increases the number of words in a search and/or introduce different search terms for the same concept, results can vary a lot. Thus, GT is its most powerful when users share a common, simple search language.

Finally, our work with GT points to two important paths for future research. First, the research community has devoted enormous resources to the coding of event data. In some cases (as with ACLED), that involves a large number of human coders reading and classifying an enormous amount of news, policy reports, etc. In other cases, as with ACLED and ICEWS, machines code digital media according to rule-based approaches to natural language processing. Both approaches have generated enormous bodies of research and important advances in knowledge. GT relies on a different source of data–user searches–but could provide insight into at least some those same events. An explicit comparison of these different approaches would help clarify the strength and limitations of each. Second, the promises of GT extend beyond event data. As discussed earlier in the paper, there are important applications that involve creating proxies for important, but infrequently reported, misreported, or expensive-to-gather measures. Examples span the gamut from nowcasting intermittently reported unemployment to learning the incidence of political opinions in a population without the cost of running sample surveys. The wide range of potential applications underscores how useful GT can be when researchers approach it with a detailed understanding of how it works.

Acknowledgements: The authors would like to thank Serkant Adiguzel, Pablo Beramendi, Spencer

Dorsey, David Dow, Ana María Montoya, Soomin Oh, and Jeremy Springman for comments on previous versions.

Data availability statement: The replication materials for this paper can be found at Timoneda, J. C. and E. Wibbels (2020). Replication Data for: "Spikes and Variance: Using Google Trends to Detect and Forecast Protests". Harvard Dataverse, https://doi.org/10.7910/DVN/WXZH8C.

Conflicts of Interest: There is no conflict of interest to disclose.

References

Aletras, N., D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos (2016). Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science 2*, e93.

Arora, V. S., M. M. McKee, and D. Stuckler (2019). Google trends: Opportunities and limitations in health and health policy research. *Health Policy*.

Bahrami, M., Y. Findik, B. Bozkaya, and S. Balcisoy (2018). Twitter reveals: Using twitter analytics to predict public protests. *arXiv preprint arXiv:1805.00358*.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis 23*(1), 76–91.

Barberá, P., J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science 26*(10), 1531–1542.

Beck, N. (2020). Estimating grouped data models with a binary-dependent variable and fixed effects via a logit versus a linear probability model: The impact of dropped units. *Political Analysis 28*(1), 139–145.

Boschee, E., J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, and M. Ward (2015). Icews coded event data. *Harvard Dataverse 12*.

Bowlsby, D., E. Chenoweth, C. Hendrix, and J. D. Moyer (2019). The future is a moving target: Predicting political instability. *British Journal of Political Science*, 1–13.

Brecher, M., J. Wilkenfeld, K. Beardsley, P. James, and D. Quinn (2016). International crisis behavior data codebook, version 11. *URL: http://sites. duke. edu/icbdata/data-collections*.

Brigo, F., S. C. Igwe, H. Ausserer, R. Nardone, F. Tezzon, L. G. Bongiovanni, and E. Trinka (2014). Why do people google epilepsy?: An infodemiological study of online behavior for epilepsy-related search terms. *Epilepsy & behavior 31*, 67–70.

Brownlee, J. (2018). *Statistical Methods for Machine Learning: Discover how to Transform Data into Knowledge with Python*. Machine Learning Mastery.

Carneiro, H. A. and E. Mylonakis (2009). Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases 49*(10), 1557–1564.

Carrière-Swallow, Y. and F. Labbé (2013). Nowcasting with google trends in an emerging market. *Journal of Forecasting 32*(4), 289–298.

Cederman, L.-E. and N. B. Weidmann (2017). Predicting armed conflict: Time to adjust our expectations? *Science 355*(6324), 474–476.

Chadwick, M. G. and G. Sengül (2015). Nowcasting the unemployment rate in turkey: Let's ask google. *Central Bank Review 15*(3), 15.

Chenoweth, E. and J. Ulfelder (2017). Can structural conditions explain the onset of nonviolent uprisings? *Journal of Conflict Resolution 61*(2), 298–324.

Choi, H. and H. Varian (2012). Predicting the present with google trends. *Economic Record 88*, 2–9.

Chykina, V. and C. Crabtree (2018). Using google trends to measure issue salience for hard-to-survey populations. *Socius 4*, 2378023118760414.

Do, Q.-T., J. N. Shapiro, C. D. Elvidge, M. Abdel-Jelil, D. P. Ahn, K. Baugh, J. Hansen-Lewis, M. Zhizhin, and M. D. Bazilian (2018). Terrorism, geopolitics, and oil security: Using remote sensing to estimate oil production of the islamic state. *Energy research & social science 44*, 411–418.

Fearon, J. D. and D. D. Laitin (2003). Ethnicity, insurgency, and civil war. *American political science review 97*(1), 75–90.

Fisher, D. R., K. T. Andrews, N. Caren, E. Chenoweth, M. T. Heaney, T. Leung, L. N. Perkins, and J. Pressman (2019). The science of contemporary street protest: New efforts in the united states. *Science Advances 5*(10), eaaw5461.

Gebru, T., J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences 114*(50), 13108–13113.

Gerner, D. J., P. A. Schrodt, R. A. Francisco, and J. L. Weddle (1994). Machine coding of event data using regional and international sources. *International Studies Quarterly 38*(1), 91–119.

Goldstone, J. A., R. H. Bates, D. L. Epstein, T. R. Gurr, M. B. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward (2010). A global model for forecasting political instability. *American Journal of Political Science 54*(1), 190–208.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis 21*(3), 267–297.

Gurr, T. R. and M. I. Lichbach (1986). Forecasting internal conflict: A competitive evaluation of empirical theories. *Comparative Political Studies 19*(1), 3–38.

Hegre, H., J. Karlsen, H. M. Nygård, H. Strand, and H. Urdal (2013). Predicting armed conflict, 2010–2050. *International Studies Quarterly 57*(2), 250–270.

Hollenbach, F. M. and J. H. Pierskalla (2017). A re-assessment of reporting bias in event-based violence data with respect to cell phone coverage. *Research & Politics 4*(3), 2053168017730687.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*, Volume 112. Springer.

Jun, S.-P., H. S. Yoo, and S. Choi (2018). Ten years of research change using google trends: From the perspective of big data utilizations and applications. *Technological Forecasting and Social Change 130*, 69–87.

King, G. and W. Lowe (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization 57*(3), 617–642.

King, G. and L. Zeng (2001). Logistic regression in rare events data. *Political analysis 9*(2), 137–163.

Korolov, R., D. Lu, J. Wang, G. Zhou, C. Bonial, C. Voss, L. Kaplan, W. Wallace, J. Han, and H. Ji (2016). On predicting social unrest using social media. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pp. 89–95. IEEE.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). The parable of google flu: traps in big data analysis. *Science 343*(6176), 1203–1205.

Leetaru, K. and P. A. Schrodt (2013). Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, Volume 2, pp. 1–49. Citeseer.

Lucas, C., R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley (2015). Computer-assisted text analysis for comparative politics. *Political Analysis 23*(2), 254–277.

Mavragani, A. and K. P. Tsagarakis (2016). Yes or no: Predicting the 2015 greferendum results using google trends. *Technological Forecasting and Social Change 109*, 1–5.

Min, B. (2015). *Power and the vote: Elections and electricity in the developing world*. Cambridge University Press.

Muthiah, S., B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz, and N. Ramakrishnan (2015). Planned protest modeling in news and social media. In *Twenty-Seventh IAAI Conference*.

Nuti, S. V., B. Wayda, I. Ranasinghe, S. Wang, R. P. Dreyer, S. I. Chen, and K. Murugiah (2014). The use of google trends in health care research: a systematic review. *PloS one 9*(10), e109583.

Pelat, C., C. Turbelin, A. Bar-Hen, A. Flahault, and A.-J. Valleron (2009). More diseases tracked by using google trends. *Emerging infectious diseases 15*(8), 1327.

Preis, T., H. S. Moat, and H. E. Stanley (2013). Quantifying trading behavior in financial markets using google trends. *Scientific reports 3*, 1684.

Raleigh, C., A. Linke, H. Hegre, and J. Karlsen (2010). Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research 47*(5), 651–660.

Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in puerto rico using google trends data. *Tourism Management 57*, 12–20.

Sarkees, M. R. and F. W. Wayman (2010). *Resort to war: a data guide to inter-state, extra-state, intra-state, and non-state wars, 1816-2007*. Cq Pr.

Schrodt, P. A. (2012). Precedents, progress, and prospects in political event data. *International Interactions 38*(4), 546–569.

Schrodt, P. A., J. Beieler, and M. Idris (2014). Three'sa charm?: open event data coding with el: Diablo, petrarch, and the open event data alliance. In *ISA Annual Convention*.

Schrodt, P. A. and D. J. Gerner (1994). Validity assessment of a machine-coded event data set for the middle east, 1982-1992. *American Journal of Political Science 38*(3), 825–854.

Shen, J. K., N. A. Seebacher, and S. D. Morrison (2019). Global interest in gender affirmation surgery: A google trends analysis. *Plastic and reconstructive surgery 143*(1), 254e–256e.

Teng, Y., D. Bi, G. Xie, Y. Jin, Y. Huang, B. Lin, X. An, D. Feng, and Y. Tong (2017). Dynamic forecasting of zika epidemics using google trends. *PloS one 12*(1), e0165085.

Timoneda, J. C. (2018). Where in the world is my tweet: Detecting irregular removal patterns on twitter. *PloS one 13*(9), e0203104.

Timoneda, J. C. and E. Wibbels (2020). Replication Data for: "Spikes and Variance: Using Google Trends to Detect and Forecast Protests". *Harvard Dataverse DRAFT VERSION*, https://doi.org/10.7910/DVN/WXZH8C.

Tkachenko, N., S. Chotvijit, N. Gupta, E. Bradley, C. Gilks, W. Guo, H. Crosby, E. Shore, M. Thiarai, R. Procter, et al. (2017). Google trends can improve surveillance of type 2 diabetes. *Scientific reports 7*(1), 4993.

Vosen, S. and T. Schmidt (2011). Forecasting private consumption: survey-based indicators vs. google trends. *Journal of Forecasting 30*(6), 565–578.

Ward, M. D., B. D. Greenhill, and K. M. Bakke (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of peace research 47*(4), 363–375.

Weidmann, N. B. (2016). A closer look at reporting bias in conflict event data. *American Journal of Political Science 60*(1), 206–218.

Weidmann, N. B. and M. D. Ward (2010). Predicting conflict in space and time. *Journal of Conflict Resolution 54*(6), 883–901.

Yu, L., Y. Zhao, L. Tang, and Z. Yang (2019). Online big data-driven oil consumption forecasting with google trends. *International Journal of Forecasting 35*(1), 213–223.

Zhang, X., S. Dang, F. Ji, J. Shi, Y. Li, M. Li, X. Jia, Y. Wan, X. Bao, and W. Wang (2018). Seasonality of cellulitis: evidence from google trends. *Infection and drug resistance 11*, 689.

Zhou, D., L. Chen, and Y. He (2015). An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.