MWSUG 2017 - Paper DG03

Lag Models with Social Response Outcomes

David J Corliss, Peace-Work, Plymouth, MI

ABSTRACT

Lag models are a type of time series analysis where the current value of an outcome variable is modeled based, at least in part, on previous values of predictor variables. This creates new opportunities for the use of social media data, both as the result of previous events and as predictors of future outcomes. This paper demonstrates lag models with social media data to establish a connection between severe solar storms and subsequent hardware failures based on complaints recorded in Twitter. The methodology is then used to investigate the possibility of a statistical link between hate speech and subsequent acts of violence against persons targeted by the speech. These Data For Good analyses have been performed using SAS® University Edition, a free version of SAS available to students, professors and non-profit researchers.

INTRODUCTION

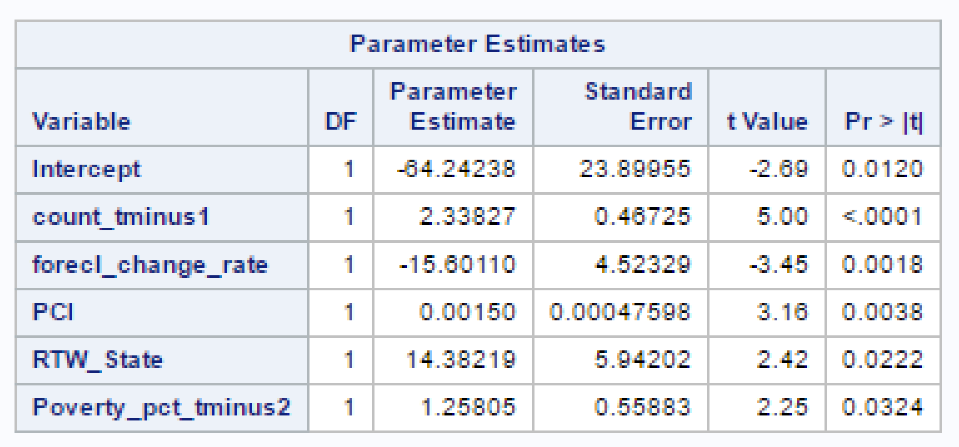
Leading variables are familiar to us from economics as factors whose changes are correlated with future events, whereas lagging indicators are correlated with previous activity. However, leading and lagging indicators appear in many areas of interest: changes in a person's diet, such as gaining or losing weight, correspond to risk group changes at a later date, or an increase in poverty over time corresponds to an increasing prison population later on. Lagging indicators may be used to infer the past, unobserved history of systems from evolving economic changes to medical histories to stellar explosions.

LAG MODELS

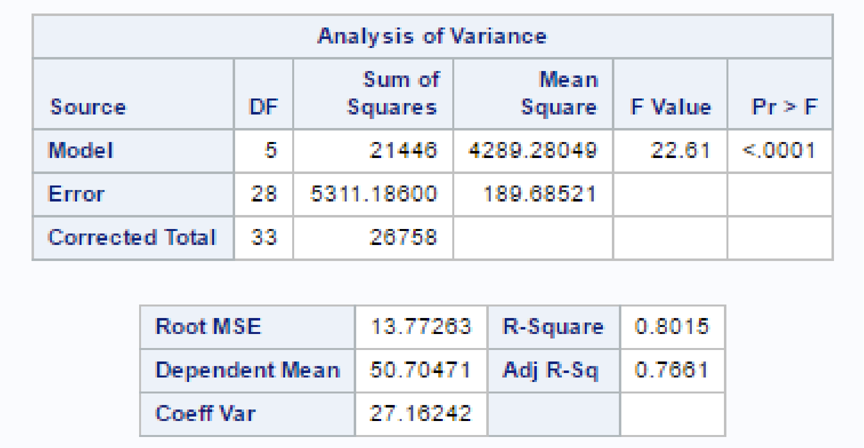
Lagged variables are a kind of repeated measure, when the same quantity is measured again and again at fixed intervals. A model is with lagged variables among the predictors is called a lag model. The regressors are not restricted only to lagged variables, only that one or more in included: non-temporally-displaced predictors can be included as well. It is important to use *evenly-spaced intervals* for the lagged variables. For example, there could be one variable for week in a series of – value of the variable during the current week, one week previous, two weeks previous, and so on - but not a combination of weeks and months. If data for a given time period is missing, it should be imputed if possible to produce a time series with values at every measured point in the series.

Lagged variables can be valuable contributors to many type of models – wherever a measurement taken at a previous point in time is predictive of the outcome of interest. In this example, a simple regression model predicts the frequency of an economic event at the present moment – a process sometimes called "now-casting". The regressors include the value of the same measure from the immediately previous time period and also the reported percent of persons living below the poverty line measured during two time intervals previous to the modeled outcome, along with other factors that do not change over the time period of interest:

```
proc reg data=ht_data;
   model count = count_tminus1  forecl_change_rate
                 PCI  rtw_state  Poverty_pct_tminus2 ;
run;
```



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-64.24238	23.89955	-2.69	0.0120
count_tminus1	1	2.33827	0.46725	5.00	<.0001
forecl_change_rate	1	-15.60110	4.52329	-3.45	0.0018
PCI	1	0.00150	0.00047598	3.16	0.0038
RTW_State	1	14.38219	5.94202	2.42	0.0222
Poverty_pct_tminus2	1	1.25805	0.55883	2.25	0.0324



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	21446	4289.28049	22.61	<.0001
Error	28	5311.18600	189.68521		
Corrected Total	33	26758			

Root MSE	13.77263	R-Square	0.8015
Dependent Mean	50.70471	Adj R-Sq	0.7661
Coeff Var	27.16242		

Figure 1. Lag Model of Reported Human Trafficking Victims by State

In this model, we find contributions from both fixed and lagged variables, with different optimal amounts of time lag for different variables.

When continuous time series data is supplied, a variable can be converted into a lagged series of variables using the LAG function. This macro converts two named variables into series of lagged variables, and can be expanded to accommodate as many variables as desired.

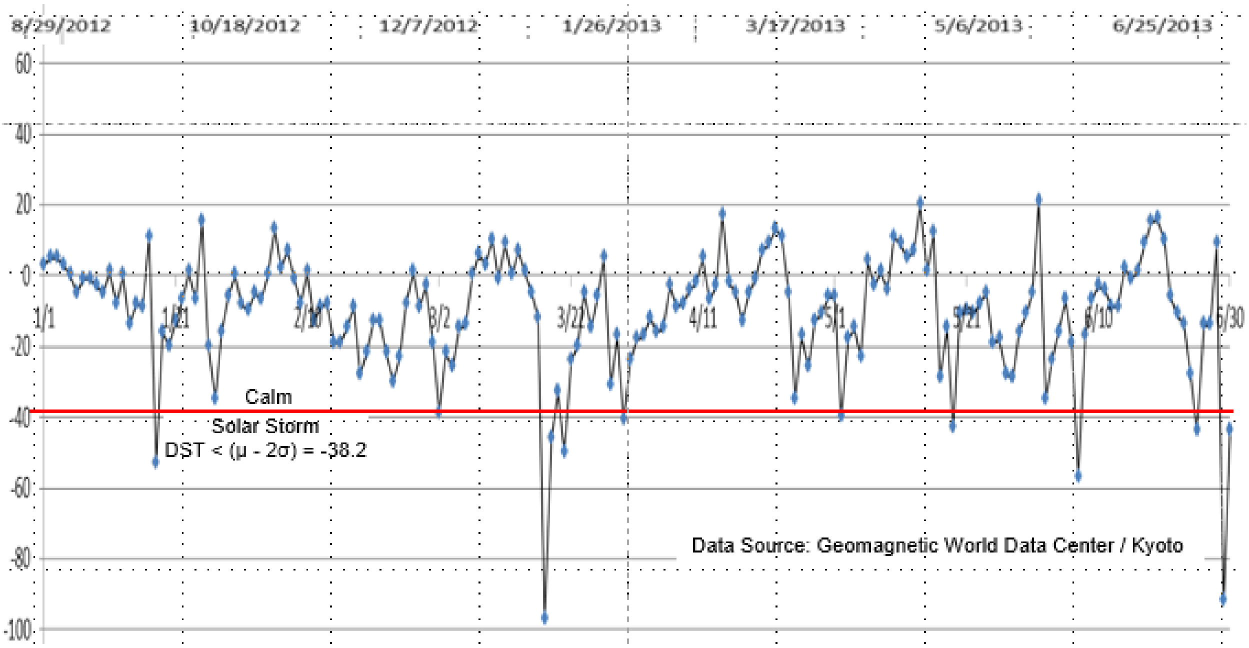
```
%macro lag(var1,var2,lag_n);
data lag_data;
   set source_data;
   x1_lag0 = &var1;
   x2_lag0 = &var2;

   %do i = 1 %to &lag_n;
      %let j = %eval(&i - 1);
      x1_lag&i = lag(x1_lag&j);
      x2_lag&i = lag(x2_lag&j);
   %end;
run;
%mend;
%lag(gdp,unemployment,3);
                 PCI  rtw_state  Poverty_pct_tminus2 ;
run;
```

SOCIAL MEDIA DATA AND SOLAR STORM ACTIVITY

Social media data can be a valuable resource for investigating a wide variety of subjects. While direct inferences from individual posts are common – sentiment analysis, location, purchase activity, and so forth – time series analysis of trends in social media activity can reveal underlying factors affecting populations as a whole. In instances where a latent root cause drive social media activity, even the people posting may be unaware of the event that may be discerned from their collected posts.

An example of this is seen in longitudinal analysis of the frequency of specific terms. As one example, the hashtag #techfail has become the neo-expletive of choice for many people to complain about sudden unexpected failures in devices and systems, random accidents, and mysterious slowdowns and failures with no apparent cause.

Radiation from significant solar storms impacting the Earth is a known cause of device and system failure. Time series analysis of the daily number of #techfail tweets indicates a connection to solar storm activity, which can be described with a lag model.



Data Source: Geomagnetic World Data Center / Kyoto

Figure 2. Time Series of Solar Activity From 8/29/2012-7/6/2013

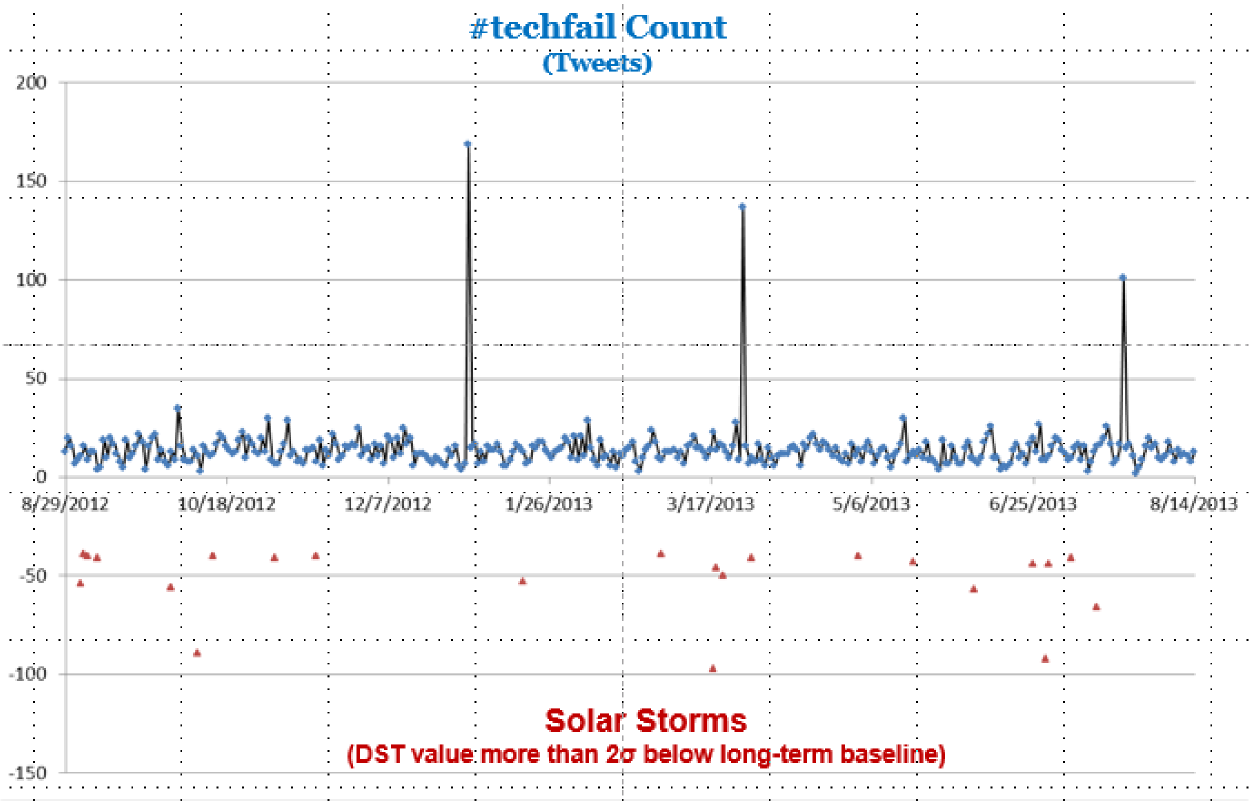


Figure 3. Time Series of Solar Activity and spikes in #techfail counts

The World Data Center for Geomagnetism collects and archives data on their Disturbance Storm Time index or DST, measuring the intensity of the effects solar storms have on earth. Analyzing the data using a lag model, the strongest correlation between storm events and #techfail spikes occurs with peaks in the Twitter data following storms by lag of 2 to 4 days.

Solar storms are not constant: only 16.0% of all days follow a > 2 σ solar storm by 2 to 4 days. Looking at just the top days for #techfail tweets, 57.9% follow a solar storm. Bursts of #techfail tweets can happen for lots of different reasons – the number 1 day in the data by far was New Year's Day 2013. Also, not every strong solar storm shoots up the rate of #techfail tweets – only some of them do. Still, with the solar storms we can detect now, the chance of a burst of #techfail tweets increases by a factor of 3.6 times during the period of 2-4 days following a storm.

MININIG TWITTER DATA WITH SAS

Twitter allows independent developers to create their own applications with search capability using an API. available at no cost. Twitter encourages these new applications by making the API available at no cost to registered developers, providing a web page to register for an API, and instructions analyzed implementation. SAS Support has a page on importing tweets that can be found here: http://blogs.sas.com/content/sascom/2013/12/12/how-to-import-twitter-tweets-in-sas-data-step-using-oauth-2-authentication-style/. An excellent paper on setting up a Twitter API in a SAS program has been written by Isabel Litton and Rebecca Ottesen of California Polytechnic State University. The tweets analyzed in the following section were pulled using a modification of the Litton-Ottesen macro %GrabTweet.

In the mining and analysis of Twitter feeds, a five-step process is recommended:

1. Register with Twitter as a developer to get access
2. Create an API to deploy search terms and receive tweets
3. Create a program to deploy the API, obtain tweets, and parse the Twitter stream into a SAS dataset

4. Perform exploratory data analysis on a few tweets to find the best search terms for the investigation
5. Extract and analyze a large number of tweets pulled using the preferred search terms

The process begins with registering with Twitter as an app developer at https://dev.twitter.com/oauth/overview. Fill out the on-line form to register. Once this is complete, Twitter will send three important articles by email:

- consumerKey – this serves as a Twitter User ID
- consumerSecret – this is the login password to the Twitter system
- Bearer Token: This combines the roles of consumerKey and consumerSecret to provide a single text string to be deployed in SAS code to access tweets.

It is strongly advised to keep all three highly secure. When deploying this information in SAS code, a FILENAME statement should be used to read this information from a secure location. These should not be hard-coded into SAS programs!

Tweets captured using a Twitter API are transmitted as a continuous text stream and must be parsed. The comma-delimited fields appear in the following order:

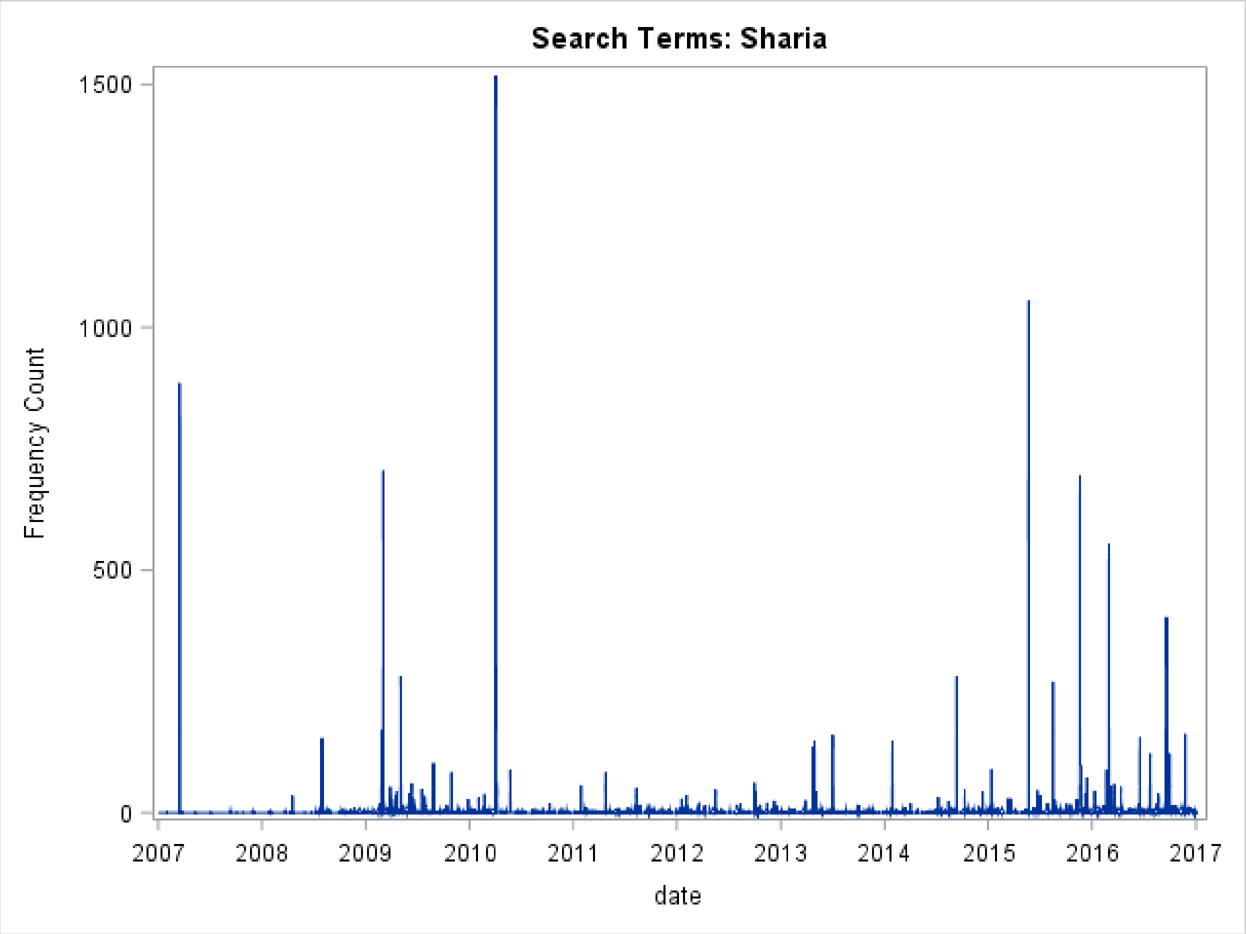
- Creation timestamp of original tweet
- Tweet ID (unique to the tweet)
- Tweet text
- Twiitter coded User ID
- User actual name
- User screen name (Twitter “handle”)
- Location of user when Tweet was sent
- Tweet creation timestamp
- Language in which the tweet is written
- Retweet flag

While the Twitter API is limited to 100 tweets, much large numbers may be obtained by writing a macro to loop through repeated iterations of Twitter extraction. To perform analysis, it is recommended first to perform exploratory data analysis on a small number of tweets, examining individual tweets, to identify search terms yielding tweets best suited to the desired analysis. For example, in extracting hate tweets, it is found the search terms Sharia, America, and Hate produce a very pure stream of anti-Muslim hate speech.

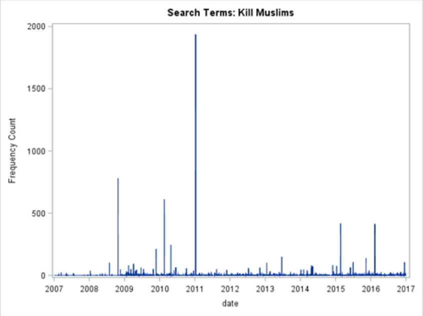
ANALYSIS OF HATE SPEECH IN TWITTER DATA

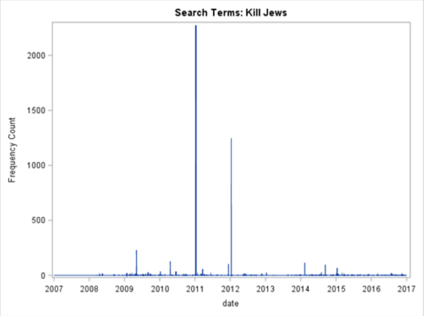
Once a substantial body of tweets has been extracted, the time series of the number of tweets per day can be compared to either a list of discrete events or event counts. As in the case of solar storm events and daily counts #techfail tweets were found t be related, lag models can be used investigate the hypothesized relationship between hate speech and subsequent violence against persons targeted by the speech. (Note: readers inexperienced analysis of hate speech are strongly cautioned regarding the use of language in this analysis.)

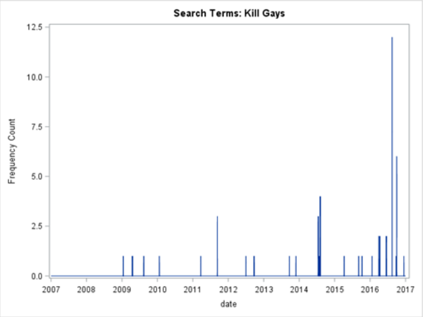
Tweets are characterized by sharp, short-term spikes in the time series of search term counts. For example:

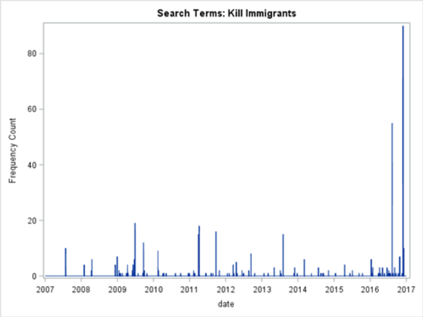


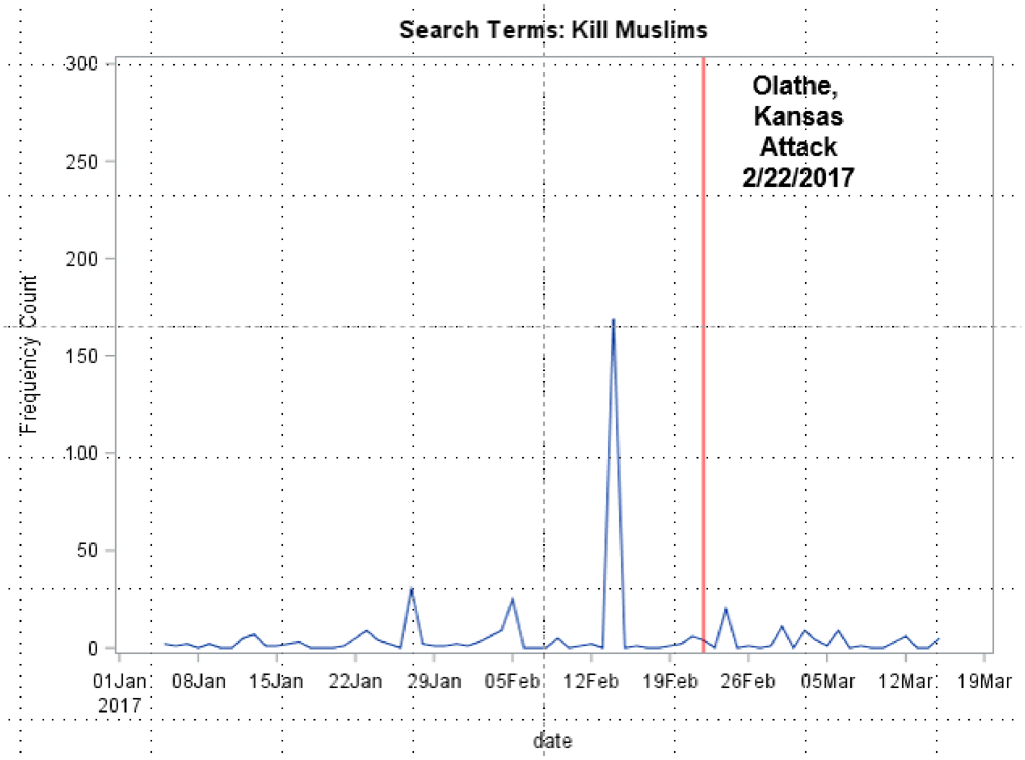
While the search term “Sharia” is strongly associated with hate speech, other terms can be searched as well, with similar results:



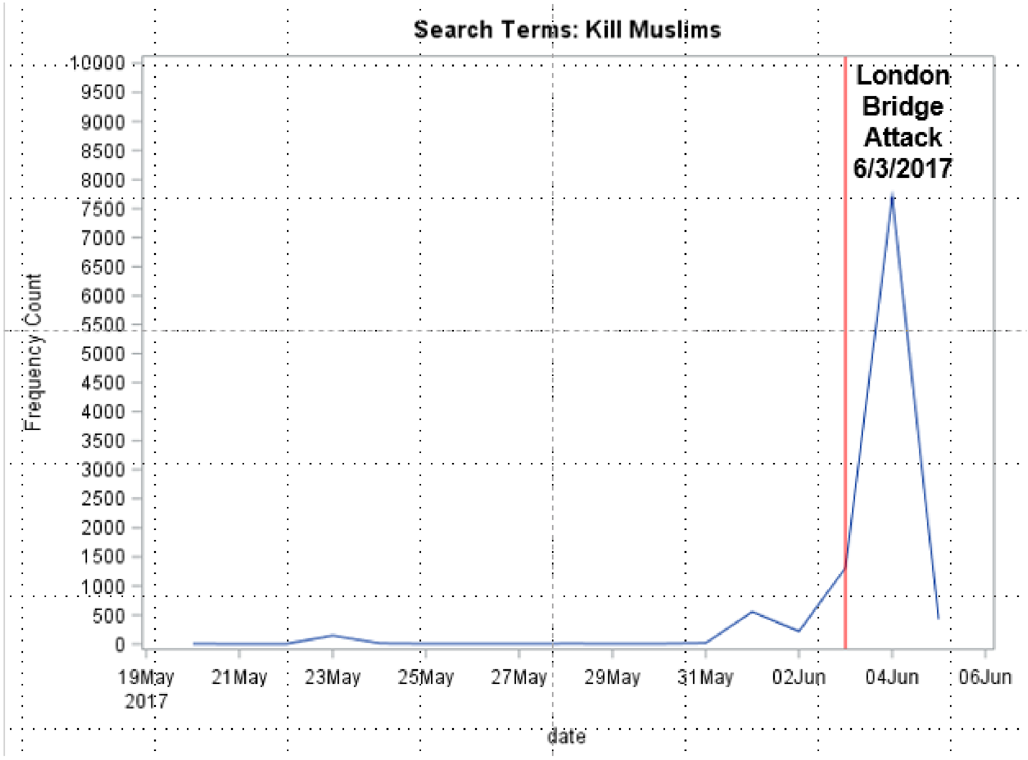


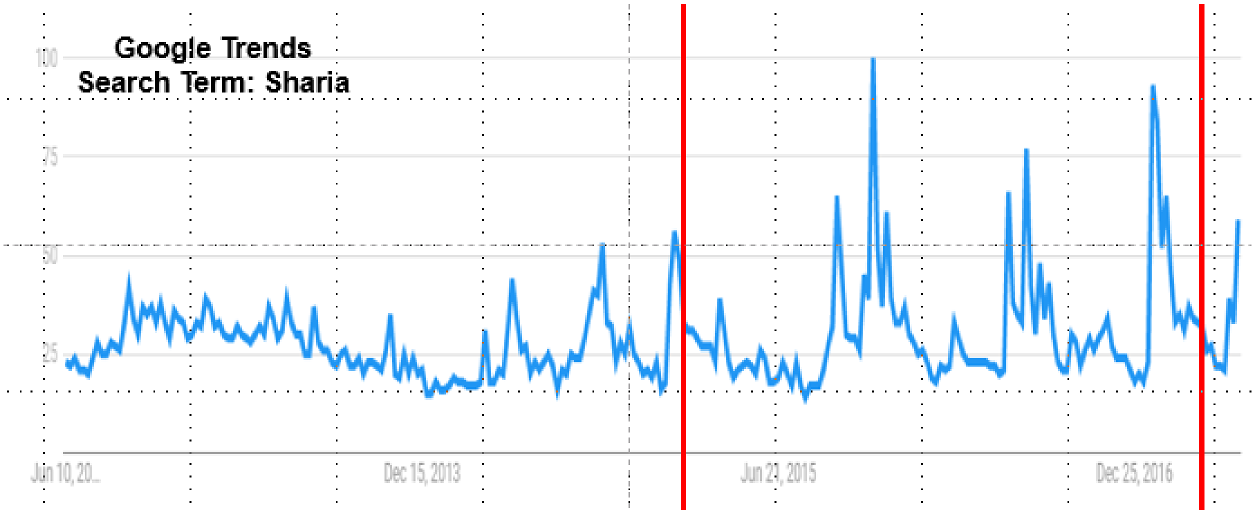






In this analysis, spikes in Twitter data are not seen to closely precede attacks inn minorities and other groups targeted by violence. A highly reactive pattern is seen instead, with sharp spikes in hate speech following terrorist attacks. The size of these spikes may be predictive of the degree to which reprisal attacks are seen.





Contrasting with the time series pattern of hate speech in Twitter, preliminary evidence may indicate peaks hate speech search terms in Google Trends can precede attacks on targeted groups. More data and analysis are need to investigate this.

CONCLUSION

Social media data can be matched to event data for Time Series Analysis, including lead / lag models. Twitter data can be mined and analyzed at high volume. As a result, analytic methods optimized for big data methods may be necessary.

Different social media channels have particular characteristics which should be remembered when selecting sources for analysis:

- Twitter: a short-duration lagging medium, highly reactive to new events. In the case of hate speech, it may be predictive of reprisal attacks
- Facebook is a medium-term channel with extended conversation threads. A general lack of access to data doe to Facebook privacy terms limits its use in time series analysis o in-house threads.
- Google Trends: a leading Indicator, much like a Social Media equivalent of Durable Goods. Google Trends data may be predictive of subsequent events.
- YouTube – the #2 Social Media channel - and #3 Instagram: ofrer no effective means to mine images at this time

Lag model analysis of Twitter data Twitter data indicates a link between solar storm activity and spikes in tweets using the term #techfail. Initial analysis of Hate Crimes with social media data shows promise but better crime data is needed.

REFERENCES

State-level counts of reported human trafficking victims collected and published by the National Human Trafficking Hotline – www.polarisproject.org

Solar activity data collected and published by the Geomagnetic Data Service of the Geomagnetic World Data Center / Kyoto - kugi.kyoto-u.ac.jp

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

David J Corliss, PhD
Peace-Work
734-837-9323
davidjcorliss@peace-work.org
peace-work.org