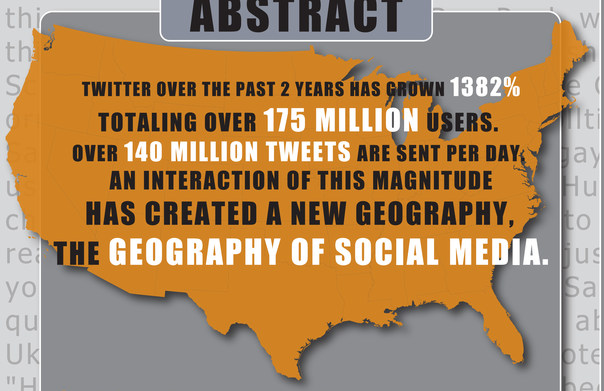
The Geography of Super Tuesday In Ohio through Twitter

Meghan Kelly & Jon Bowen with mentors Christina Hupy and Ryan Weichelt

ABSTRACT



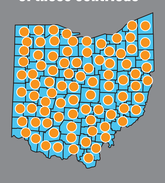
TWITTER OVER THE PAST 2 YEARS HAS GROWN 1382%
TOTALING OVER 175 MILLION USERS.
OVER 140 MILLION TWEETS ARE SENT PER DAY.
AN INTERACTION OF THIS MAGNITUDE
HAS CREATED A NEW GEOGRAPHY,
THE GEOGRAPHY OF SOCIAL MEDIA.

Social media has changed the way we communicate by fostering interactive dialog whether it it is on personal, buisness, political, news or pop culture topics. Many analysts are eager to use social media to gain a better understanding of various markets, for example, Facebooks usage of targeted advertisements. Another such market is voters. This research explored the relationship between social media data and election outcomes. More specifically, this research examined the statistical relationship between Twitter posts and the Super Tuesday results for the 2012 GOP primary in Ohio. Twitter users are offered a "GeoTag" setting that allows individual Tweets to be search based on location. The GOP primary results and the corresponding count of "GeoTagged" Twitter posts from March 4th-6th for each candidate were collected for 89 counties in Ohio. Regression Analysis explored the relationship between the independant variables, Twitter post counts, various demographic variables from the US Census Bureau (population, income, race, and education), and the dependant variable, primary results by candidate. This project explores the importance of Twitter data as a tool for politicians and political geographers alike, as a new way to track the popularity of candidates.

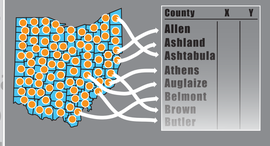
DATA RETRIEVAL

1. Obtain Ohio counties from US Census Bureau

2. Determine the centroid of each county in order to create an 8.5 mile buffer of those centriods



3. Generate a list of X,Y coordinates from centroid buffers for geo-locational Twitter search in"R"



County	X	Y
Allen		
Ashland		
Ashtabula		
Athens		
Auglaize		
Belmont		
Brown		
Butler		

4. Write a script that R can use to scrap the Twitter-base for dates March 4th-6th based on the keywords Santorum, Gingrich, Romney, Ron Paul.

```
searchTwitter("gingrich", n=1000000,
geocode='41.677,-83.649,8.5mi',since=
'2012-03-04',until='2012-03-06')
```

5. Download and normalize 2010 US Census demographic data from the publically available Fact-Finder. These variables included;

Percent High School
Percent Bachlor Degree
Percent non-White
Percent Female
Median Age
Median Income

6. Perform regression analysis using the statistical program SPSS and geostatistical program GEODA to determine linear and spatial relationships between percent votes and percent Tweets per candidate.

7. After determining linear and spatial relationships existed, multiple regression was used to identify the demographic characteristics of Ohio voters by candidate.

REGRESSION RESULTS

Using Pearson's Correlation and scatterplot analysis, a positive correlation was found between the percent of votes for Romney and the percent of Tweets for Romney. This relationship is demonstrated with the positive slope of the trend line (Figure 2). Although the R value is not considered high at 0.420, the 0.000 significance value indicates there is a linear relationship between the number of tweets for Romney and the percent of primary votes cast for him across Ohio counties.

A negative correlation was found between the percent of votes for Santorum and the percent of Tweets for Santorum. This relationship is demonstrated with a negative slope of the trend line (Figure 3). While the R value of -0.224 was weaker than Romney's correlation, there is sitll a significant linear relationship with a 0.036 probablity value.

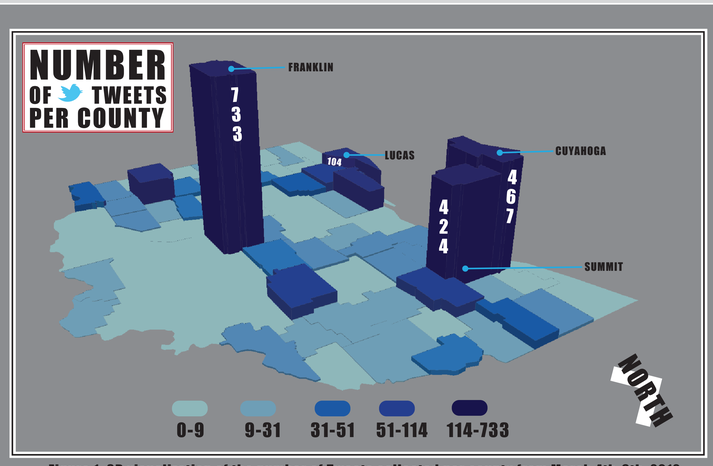


Figure 1: 3D visualization of the number of Tweets collected per county from March 4th-6th, 2012

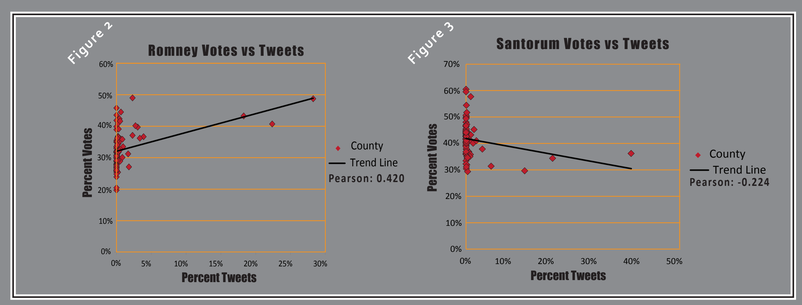


Table I

Ordinary Least Squares								
Variable	Coefficient	Std. Error	T-Value	Probability	R^2	Log Likelihood	Akaike	Schwarz
Constant	0.3204	0.0061	52.5115	0.000	0.176543	130.911	-257.822	-252.868
Tweet_Romney	0.5964	0.1389	4.2939	0.000				
Constant	0.4183	0.0068	61.1523	0.000	0.050225	120.19	-236.38	-231.425
Tweet_Santorum	-0.2951	0.1384	-2.1325	0.0358				

Table II

Spatial Lag								
Variable	Coefficient	Std. Error	Z-Value	Probability	R^2	Log Likelihood	Akaike	Schwarz
W_Romney	0.7479	0.0748	9.99	0.000	0.608	156.714	-307.429	-299.997
Constant	0.0777	0.0246	3.16	0.0015				
Tweet_Romney	0.3642	0.0952	3082	0.0013				
W_Santorum	0.7713	0.0728	10.5839	0.000	0.571156	147.604	-289.207	-281.775
Constant	0.0976	0.0307	3.1741	0.0015				
Tweet_Santorum	-0.1947	0.0919	-2.1168	0.0342				

Table III

Stepwise Regression Analysis								
Dependent	Independent	Model Summary	R^2	SEE	B	BETA	t	Significance
Percent Vote Romney	Percent High School	Percent Bachelors	0.585	0.108	0.597	0.489	5.531	0.000
	Percent Bachelors	Percent Not White		0.079	0.366	0.404	4.646	0.000
	Percent Female	Median Age		0.150	0.485	0.237	3.233	0.002
	Percent Not White							
	Median Age							
	Median Household Income							

Table IV

Stepwise Regression Analysis								
Dependent	Independent	Model Summary	R^2	SEE	B	BETA	t	Significance
Percent Vote Santorum	Percent High School	Percent Bachelors	0.358	0.141	-0.482	-0.375	-3.413	0.001
	Percent Bachelors	Percent Not White		0.103	-0.307	-0.323	-2.982	0.004
	Percent Female	Median Age		0.196	-0.449	-0.209	-2.292	0.024
	Percent Not White							
	Median Age							
	Median Household Income							

Table I illustrates the Ordinary Least Squares Regression analysis for the number of Tweets per candidate and their percentage of the vote cast per county. In both instances, weak R-squared values were influenced by the large number of counties with no Tweets. To account for this limitation, Spatial Regression or Spatial Lag was used; results shown in Table II. The Spatial Regressions resulted in dramatically improved model performance. This improvement is indicated by the large increases in the R-squared (coupled with increases in the log likelihood and decreases in the Akaike and Schwartz variables) for each model. Romney outperformed Santorum in counties with Tweets mentioning Romney more often; while, Santorum performed worse in counties that mentioned Santorum more often. This discrepancy may be accounted for by either the context of the Tweets, i.e. negative Tweets, and or by the demographics of Santorum's constituents and possible lack of Twitter activity.

The results of Stepwise Regression characterize the demographics of "Romney and Santorum Voters" and are shown in Figures 2 and 3, as well as Tables I-IV. Romney outperformed Santorum in counties with increasing proportions of persons with higher education, minority populations, and older populations. The strongest variable defining Romney constituents was percent bachelor's degree as indicated by the larger Beta Value. In contrast, Santorum performed better in counties with smaller proportions of bachelor's degrees, minority populations, and younger populations. As with Romney, the proportion of bachelor's degree has the strongest Beta Value. All in all, Romney's greatest support was in urban areas, while Santorum had stronger support in rural areas of Ohio. This pattern is corroborated by the clustering evident in the Lisa Maps (Figures 4 and 5).

LISA MAPS illustrate the spatial clustering of candidate constituents in Ohio counties. Voter patterns in Ohio demonstrate Tobler's Law, which states that nearer places are more similar than places that are further away. Romney constituents are concentrated in the Northeast and Southwest corners of the state, both largely urban counties (Figure 4). Santorum's constituents are also clustered, but in opposite regions as Romney's (Figure 5). These maps highlight the spatial polarity between candidate constituents. Moran's I, a statistic which measures the degree of spatial autocorrelation, reflects this clustering.

Figure 4

Romney
Morans I: 0.5578

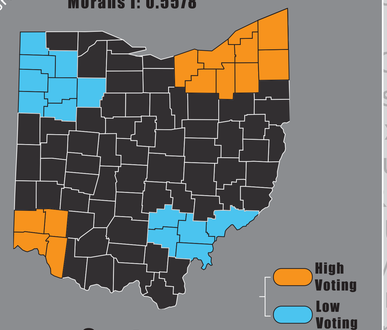
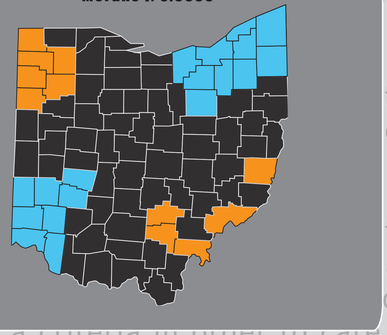


Figure 5

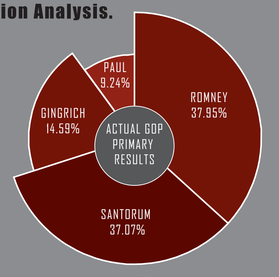
Santorum
Morans I: 0.5639

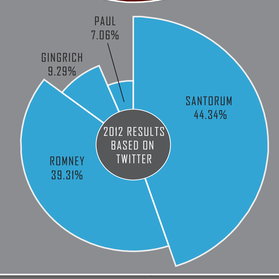


SUPER TUESDAY

Super Tuesday was held March 6, 2012 as a presidential nominating process for the GOP. Typically a swing state, Ohio uses the process of primary voting to nominate a candidate for the 2012 Presidential Election.

The pie charts below show the actual GOP Primary results in red and the Primary results based on Tweets per candidate in blue. Both depict a contentious battle between Romney and Santorum to be further looked at with Regression Analysis.





TWITTER APPLICATIONS

The relationships found in this research between Primary voting, Twitter, and the demographics of voters, highlight the important role of communication through social media. This analysis opens a new market that is highly important in today's society. A 140 character Tweet has proven to be a valuable expression of voter preference which can be related to election outcomes. This research demonstrates a new Geography of elections through Twitter and social media. This study is important to political Geographers and politicians alike, opening a new door into the minds of voters. Because of the clear prevalence of "GeoTagged"social media the formalized definition of place can be re-evaluated and re-defined to incorporate cyberspace into the normal idea of place.

The methods provided in this research are powerful tools for future elections studies as social media continues to spread at rapid speeds. With slight modifications, this methodology could predict future voter markets for candidates and election outcomes. Spatial analysis through Twitter could further define voting markets and the primary characteristics of constituents for particular candidates.

Acknowledgements:

Office of Research and Sponsored Programs (ORSP)
Professor Christina Hupy
Professor Ryan Weichelt
Scott Drzyzga
US Census Bureau: FactFinder

Department of Geography and Anthropology