Paper P0-9

Does the Percentage of College Students and Military Personnel Affect Political Contributions Per Zip Code? Visualization with PROC GMAP

Jamelle Simmons, Virginia Polytechnic Institute and State University

ABSTRACT

A geographical analysis was performed using SAS® v 9.3 to investigate the relationship between political contributions from a zip code and the percentage of college and military individuals within that zip code. For this analysis, data was merged from publicly available 2012 Federal Election Commission (FEC) contribution data and 2010 Census data. Additional variables were created using this combined data set. Other variables used in this model to help explain variability were the percentage of white, non-Hispanic or Latino individuals within a zip code, the percentage of contributions to the Democratic party from each zip code, and which way states swung in the 2008 elections. PROC GMAP was used to help visualize voting patterns nationally between the 2008 and 2012 elections as well as take an in-depth look at the state of North Carolina, as an example, to visualize the population percentages of college and military indicators and white, non-Hispanic or Latino population percentages per zip code

INTRODUCTION

Every election cycle, there is a buzz about which candidate will get the vote of students, minorities, military personnel, certain age groups as well as different gender. With each generation being affected by social issues differently, it makes it difficult to determine which way a group will vote during each election cycle. Issues such as ease of voting, the September 11th attacks, recessions, natural disasters, student loan debt, program funding and spending cuts, equality for all groups, and voter apathy all contribute to the views a person holds and ultimately can impact which party an individual will vote for [1-3]. As there are two areas that consistently are focused on each election cycle, military and college students, it is the goal of this paper to explore if these areas can help predict how much financial support will be given to a political party above and beyond knowing how the state voted in the previous cycle and the percentage of white, non-Hispanic or Latino groups within each state. In addition to building a model, data visualization techniques are used to help add validity to the model originally constructed which spans the nation as a whole and use the state of North Carolina as an example of how in-depth visualization can be.

METHODS

SOURCES OF DATA

Federal Election Commission (FEC) data was collected on the contributions to each candidate for the 2012 election cycle in addition to the zip code of each contributor. The 2010 Census data includes information on military housing quarters per zip code, college/university housing quarters per zip code, the total population per state, and the population of white, non-Hispanic or Latino persons per zip code. The University of Missouri has mapping codes (ZCTA) that are available for use which link the ZCTA5 codes given in the Census data zip codes for the entire US. Tiger/Line® shape files were downloaded for the state of North Carolina at the zip code level so that Census data could be merged one-to-one for each zip code. Additional sources of election data were retrieved from Politico and NBC websites where the 2008 election results and swing states were obtained from Politico and the 2012 election results obtained from NBC.

BUILDING THE FINAL DATA SETS

From the 2012 FEC data, only complete observations were kept where it was known which candidate a contributor gave to, the amount of the contribution, the zip code of the contributor, the occupation, and state of residence. Only the first 5 digits of the zip were read, which are later used to merge with the University of Missouri zip code file. With the 2010 Census data, only the SF1 summary file data was used with the associated GEOREF key. The file identification (FILEID), state abbreviation (STUSAB), summary level (SUMLEV), geographic component (GEOCOMP), character iteration (CHARITER), character iteration file sequence number (CIFSN), and the logical record number (LOGRECNO) are necessary merging keys that are needed to link observations between the GEOREF file and the SF1 files. Only files at the SUMLEV=860 are used as these are at the zip code level which and can be merged with the Missouri ZCTA mapping codes.

SF1 files data on, group quarters population by sex by age, for female and male College/University student housing quarters were obtained as well as the female and male military housing quarters for ages 18-64. As this paper is limited to the zip code level, it is not possible to link to individuals but rather total counts of individuals within quarters

per zip code for each group. Other variables collected at this level were the total population (not limited to 18-64 year range) per zip code as well as the total white, non-Hispanic or Latino (18-64 years of age). Gender was combined for college/university student housing as well as military quarters and converted into a percentage by dividing by the total population per zip code. The same was done for the white, non-Hispanic or Latino variable. Finally, all percentage values were merged with the FEC data set by zip code.

With this new data set (SWING_BRK3), the variables created were,

- Total contributions per political party per zip code
- Total contributions per state per party
- Percent contributions in a zip code to the Democratic party
- Percent contributions per party per state
- 2012 state swing (Democratic or Republican) based on which party received the most money per state
- 2012 actual swing (Democratic or Republican) based on the 2012 election results (from NBC)
- 2008 state swing (Solid Republican, Solid Democratic, or Swing State) based on 2008 polls (from Politico)
- Number of occurrences of college/university student housing quarters per state
- Number of occurrences of military quarters per state

STATISTICAL MODEL

Numeric variables were converted to percentages to produce a similar scale between zip codes, due to the variation in total population size. With some observations producing values of zero after conversion to percentages, the value 0.001 was added to each observation so that PROC TRANSREG could run properly. PROC UNIVARIATE was used to help find the best transformation of each variable prior to inclusion into the PROC TRANSREG procedure (figure 1, left). The final independent variables included in the model were, WN_PER2= Percent white, non-Hispanic or Latino per zip, ICOL= Inverse college/university student housing percentage per zip, and IMIL= Inverse military quarters percentage per zip. The Box-Cox output (figure 1, right) produced from the TRANSREG procedure shows that with the current independent variables, the dependent variable DEM_PCT2 (Percent contributions in a zip code to the Democratic party) should not be transformed with a lambda value equal to 1. With ODS Graphics on, graphical output can be specified which can be viewed in SAS HTML output, but will also be written to the current ODS destination.

```
LIBNAME READTO 'C:\USERS\LENOVO\
DESKTOP\SAS PROC GMAP\PERMANENT
FILES';

ODS GRAPHICS ON;

PROC UNIVARIATE
DATA=READTO.SWING_BRK3 NOPRINT;
HISTOGRAM DEM_PCT2 WN_PER2 ICOL IMIL;
RUN;

PROC TRANSREG DATA=READTO.SWING_BRK3
TEST;
MODEL BOXCOX(DEM_PCT2 / LAMBDA=-2 -1
-0.5 TO 0.5 BY 0.05 1 2 CONVENIENT
PARAMETER=2 ALPHA=0.00001) =
IDENTITY(WN_PER2 ICOL IMIL);
RUN;
```

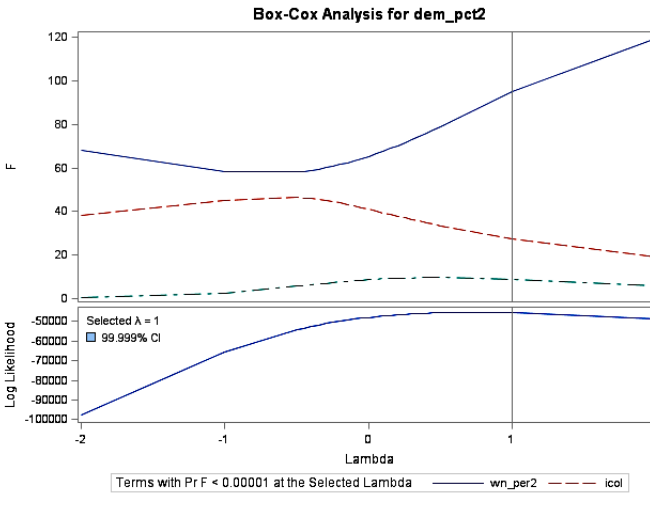

Figure 1. PROC TRANSREG code for the Box-Cox procedure to choose an appropriate lambda for the specified model (left). Box-Cox result for transformation of dependent variable DEM_PCT2 (right).

With the ODS GRAPHICS still on, the PROC GLM procedure (figure 2, left) is run with the inclusion of the 2008 state swing variable (identified in both the MODEL and CLASS statements). Plot diagnostics and output options are specified which determine what appears in the output (figure 2, right, residual diagnostics not shown).

```
PROC GLM
DATA=READTO.SWING_BRK3
PLOT=DIAGNOSTICS RESIDUALS
PLOTS(MAXPOINTS=20000);
CLASS SW_2008;
MODEL DEM_PCT2=SW_2008
WN_PER2 ICOL IMIL/
INTERCEPT SOLUTION
TOLERANCE;
RUN;
QUIT;
ODS GRAPHICS OFF;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	36280498.56	6046749.76	11198.0	<.0001
Error	14030	7575969.28	539.98		
Uncorrected Total	14036	43856467.84			

R-Square	Coeff Var	Root MSE	dem_pct2 Mean
0.165968	46.68651	23.23755	49.77358

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Intercept	1	34772916.27	34772916.27	64396.3	<.0001
sw_2008	2	1401010.57	700505.28	1297.27	<.0001
wn_per2	1	72895.51	72895.51	135.00	<.0001
icol	1	29203.31	29203.31	54.08	<.0001
imil	1	4472.92	4472.92	8.28	0.0040

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Intercept	1	787185.977	787185.977	1457.80	<.0001
sw_2008	2	1426316.263	713158.132	1320.70	<.0001
wn_per2	1	78853.837	78853.837	146.03	<.0001
icol	1	19922.789	19922.789	36.90	<.0001
imil	1	4472.916	4472.916	8.28	0.0040

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	59.24503275	B	1.59127544	37.23	<.0001
sw_2008 Solid Democratic	13.35256827	B	0.45723703	29.20	<.0001
sw_2008 Solid Republican	-11.93339906	B	0.52057002	-22.92	<.0001
sw_2008 Swing State	0.00000000	B	.	.	.
wn_per2	-0.15957889		0.01320548	-12.08	<.0001
icol	-0.00964077		0.00158718	-6.07	<.0001
imil	-0.00417392		0.00145024	-2.88	0.0040

Figure 2. PROC GLM model code to predict percent democrat contributions per zip code (left). Model output statistics (right).

The ANOVA output shows statistical significance of the overall model with an F-value of 11,198 and a p-value <0.0001 with all parameter estimate p-values below 0.05. The model used a total of 14,036 observations and was run with a default alpha value of 0.05. It should be noted that only 16.60% of the total variability in the model can be explained by this model so there are other factors not included in the model that could help explain the remaining variance. The distribution of the residuals (output not shown) were normally distributed and the studentized residuals vs. predicted values displayed a random pattern (output not shown). Overall, after adjusting for which way states swung in the 2008 elections and the percentage of white, non-Hispanic or Latino, the percentage of college/university and military housing quarters do affect the percentage of political contributions per zip code. It should be noted that in order to interpret the parameter estimates of this model transformations should be used.

VISUALIZATION OF MODEL COMPONENTS

While statistical models are informative, it is also beneficial to get an idea of how the data looks. Visualizing data may also help reveal patterns that are not easily detectable when working with numbers alone. One way of visualizing data is with PROC GMAP which is useful if you have a way to link your data to mapping coordinates. All components of the GLM model can be linked on the state level, and can also be linked at the zip code level with the exception of the 2008 voting patterns. One of the first model components to view is the 2008 voting patterns. In the SWING_BRK3 data set, there are thousands of observations with many per state at the zip code level and all carry the 2008 outcomes for each state each zip code resides in, but we only need one observation per state. First, all graphics options are reset using GOPTIONS RESET=ALL, the ODS graphics are turned on and colors are specified for each possible state outcome (figure 3, left). Pattern colors can be specified by name (patterns 1 and 2), RBG value (pattern 3), or by HLS value. The option V=MS produces a solid color pattern for each state. To use PROC GMAP, a response data set is needed as well as a mapping data set, and a key that is common between the two so that the files can be linked to produce mapped data. For figure 3, the key is STATECODE which are state abbreviations such as AL, TX, SC, CA, etc. that are found in MAPS.US, but had to be created in the SWING_BRK3 data set. It is important that both variables are spelled the same in both data sets and are of the same length.

```
GOPTIONS RESET=ALL;
ODS GRAPHICS ON;
PATTERN1 V=MS C=BLUE;  PATTERN2 V=MS C=RED;
PATTERN3 V=MS C=CXD3D3D3;

TITLE '2008 STATE SWING';
FOOTNOTE1 'OUTCOMES FROM POLITICO WEBSITE ';

PROC GMAP DATA=READTO.SWING_BRK3
MAP=MAPS.US;
ID STATECODE;
CHORO SW_2008/STATISTIC=FIRST LEVELS=3;
LABEL SW_2008='STATE SWING';
RUN;
QUIT;
ODS GRAPHICS OFF;
TITLE;
FOOTNOTE;
```

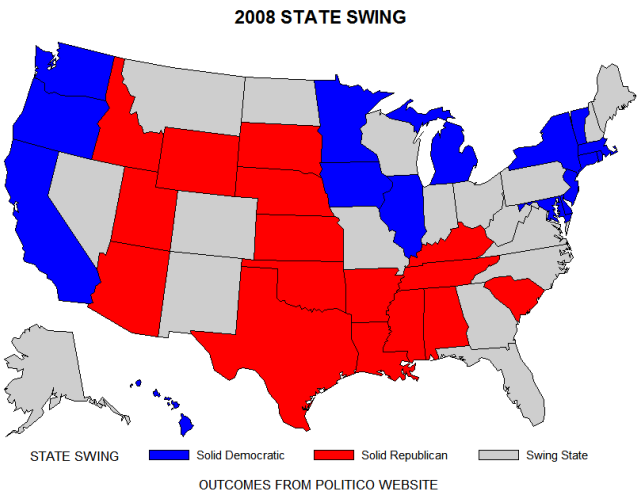

Figure 3. PROC GMAP code to produce 2008 state leanings (left). PROC GMAP output (right)

Above, the MAPS.US STATECODE variable is of length $2 so the STATECODE variable created in the response set must be of the same length and name or warnings and errors will be produced in the SAS log. When using the PROC GMAP statement, the key found in both data sets must be identified with the ID statement. The CHORO statement is used to color the map in solid colors based on the pattern statements and the option STATISTIC=FIRST only uses the first observation of each state. The LEVELS= option specifies the color levels that are used. Above, three pattern statements are used and three levels were specified. The resulting PROC GMAP output can be seen in figure 3.

From the FEC data that was collected, it was decided to map out how a state may swing in the 2012 election cycle based on how much was contributed to each party for each state. For example, the state of North Carolina had more money contributed to the Democratic Party, based on FEC data with complete observations, so this state was declared Democratic. As figure 4 is mapped at the state level, but contributions were at the zip code level, SAS was used to sum all contributions per party across each state. The coding is similar to that seen in figure 3, but only two levels are specified in figure 4.

```
GOPTIONS RESET=ALL;
ODS GRAPHICS ON;
PATTERN1 V=MS C=BLUE;  PATTERN2 V=MS C=RED;

TITLE 'STATE SWING DETERMINED BY 2012 FEC
CONTRIBUTIONS';

PROC GMAP DATA=READTO.SWING_BRK3
MAP=MAPS.US;
ID STATECODE;
CHORO ST_SWING/ STATISTIC=FIRST LEVELS=2;
LABEL ST_SWING='STATE SWING';
RUN;
QUIT;
ODS GRAPHICS OFF;
TITLE;
FOOTNOTE;
```

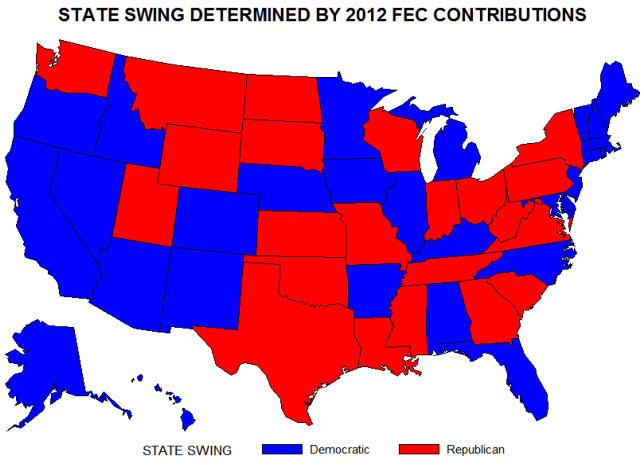

Figure 4. PROC GMAP code to produce 2012 state leanings based on FEC data (left). PROC GMAP output (right).

VISUALIZATION WITH FORMATS

Another way to obtain more control over the mapping output is to apply formats. To gain a more national view of the number of college/university student housing quarters, SAS code was written to count each unique occurrence of student housing quarters per state. The format COUNT (figure 5) was created to separate the totals into 6 possible groups and applied in the PROC GMAP code to the variable C_COUNT. Titles, footnotes, labels, and patterns were also added, but not shown below.

```
PROC FORMAT;
VALUE COUNT
      LOW-19.999='<20 QUARTERS'
      20-39.999='20 - <40 QUARTERS'
      40-59.999='40 - <60 QUARTERS'
      60-79.999='60 - <80 QUARTERS'
      80-99.999='80 - <100 QUARTERS'
            100-HIGH='100+ QUARTERS'
            ;
PROC GMAP DATA=COUNTING MAP=MAPS.US;
ID STATECODE;
CHORO C_COUNT/STATISTIC=FIRST LEVELS=6;
FORMAT C_COUNT COUNT.;
RUN;
QUIT;
```

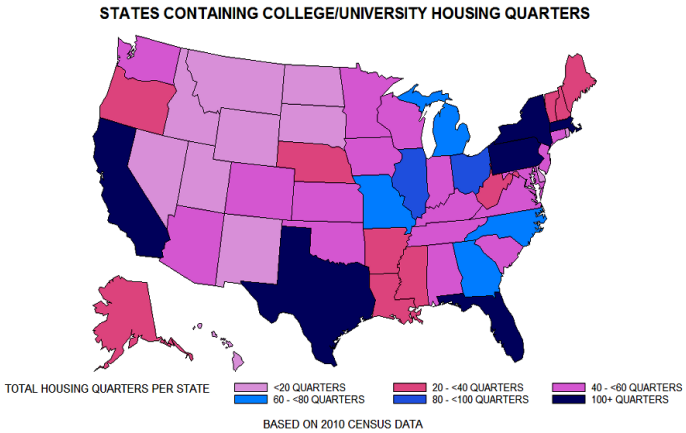

Figure 5. PROC GMAP code to produce state totals of college/university student housing quarters (left). PROC GMAP output (right).

The same approach was used for totaling up the number of occurrences of military quarters per state (figure 6) and the format COUNTS is applied. Titles, footnotes, labels, and patterns were also added, but not shown below.

```
PROC FORMAT;
VALUE COUNTS
      LOW-9.999='<10 QUARTERS'
      10-19.999='10 - <20 QUARTERS'
      20-29.999='20 - <30 QUARTERS'
      30-39.999='30 - <40 QUARTERS'
      40-49.999='40 - <50 QUARTERS'
      50-HIGH='50+ QUARTERS'
            ;
PROC GMAP DATA=COUNTING MAP=MAPS.US;
ID STATECODE;
CHORO M_COUNT/ STATISTIC=FIRST LEVELS=6;
FORMAT M_COUNT COUNTS.;
RUN;
QUIT;
```

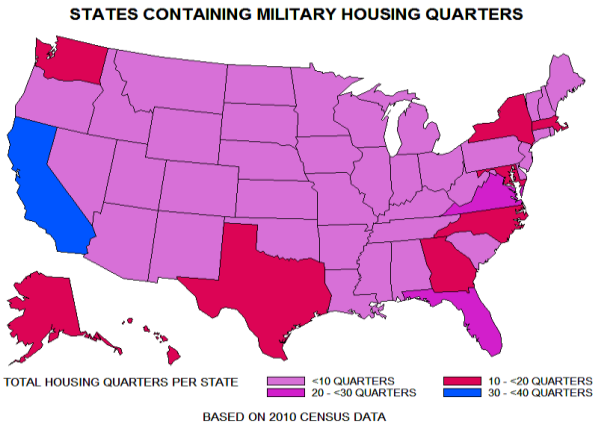

Figure 6. PROC GMAP code to produce state totals of military quarters (left). PROC GMAP output (right).

VISUALIZATION PER ZIP CODE (NORTH CAROLINA)

It is also possible to move from the national level to the state level, and even further, with PROC GMAP. From figure 3, the state of North Carolina was a swing state in the 2008 elections. Figure 4 shows North Carolina as Democratic based on 2012 FEC data. Figures 5 and 6 show there are enough student and military quarters that it may be useful to take a closer look at this state. A Tiger/Line® shape file for the 2010 Census was downloaded to merge with the SWING_BRK3 data set (where statecode='NC'). PROC MAPIMPORT was used to point to the downloaded shapefile using the DATAFILE= statement and the OUT= statement was used to name the output file. The OUT file and the SWING_BRK3 file were first sorted by ZCTA5CE10 (zip code) and then merged by that key. It is important to note that the lengths of these variables must be of the same length or a warning message will appear in the SAS log. With the new data set, conditional statements were used to assign values of zero to observations that did not have a one-to-one merge. Omission of this step will result in PROC GMAP only mapping zip codes with data, resulting in an incomplete map of North Carolina.

The PROC GMAP steps are the same as that seen in the above figure, with the exception of the map used. Instead of using MAPS.US (a predefined map included with SAS software), the OUT file from the PROC IMPORT step NC_ZSHP is used. The ID has also changed from the state level to the zip code level. The code in figure 7 (left) can be used to produce both graphics in figure 7 (right) just by changing the title, footnote, label, and variables used.

```
GOPTIONS RESET=ALL;
ODS GRAPHICS ON;
TITLE1 "NORTH CAROLINA COLLEGE/UNIVERSITY
HOUSING QUARTERS BY ZIP";
FOOTNOTE1 "COMBINED MALE AND FEMALE
PERCENTAGES";

PATTERN1 V=S C=CXEEE8AA; PATTERN2 V=S
C=CXD9892B; PATTERN3 V=S C=CXDA70D6;
PATTERN4 V=S C=CX2E8B57; PATTERN5 V=S
C=CXB22222; PATTERN6 V=S C=CX45173E;

PROC GMAP DATA=NC_BRK MAP=NC_ZSHP;
ID ZCTA5CE10;
CHORO COL_PERC/STATISTIC=FIRST LEVELS=6;
LABEL COL_PERC='BASED ON 2010 CENSUS DATA
(VALUES IN %)';
RUN;
QUIT;
ODS GRAPHICS OFF;
TITLE;
FOOTNOTE;
```

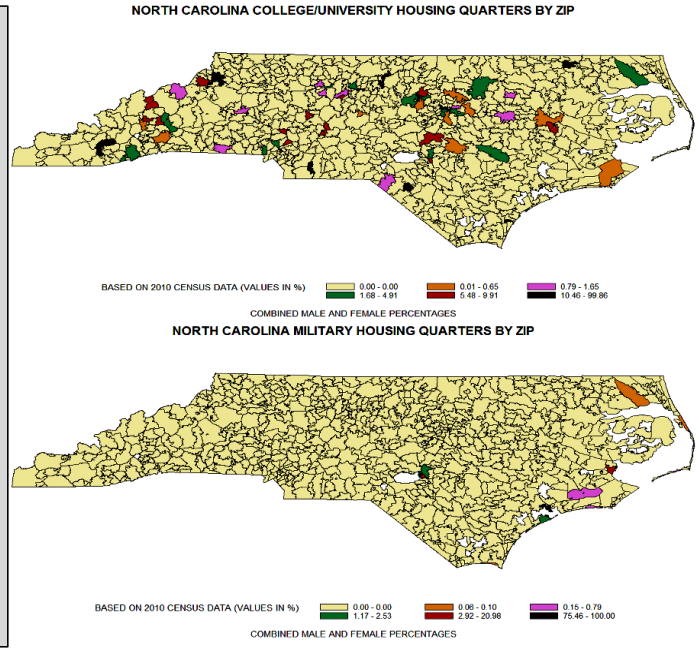

Figure 7. PROC GMAP code to produce state zip code identification of student and military quarters (left). PROC GMAP output (right).

By producing the output in the same color scheme, it is possible to stack output side by side or top to bottom to see if high areas of student housing quarters also have high counts of military quarters. It is worth noting that areas of white shading are bodies of water in the state of North Carolina. Similar coding in figure 7 was used to produce the output in figure 8 for the white, non-Hispanic or Latino percentage as well as the percentage of Democratic contributions per zip code using the same coloring scheme for both maps.

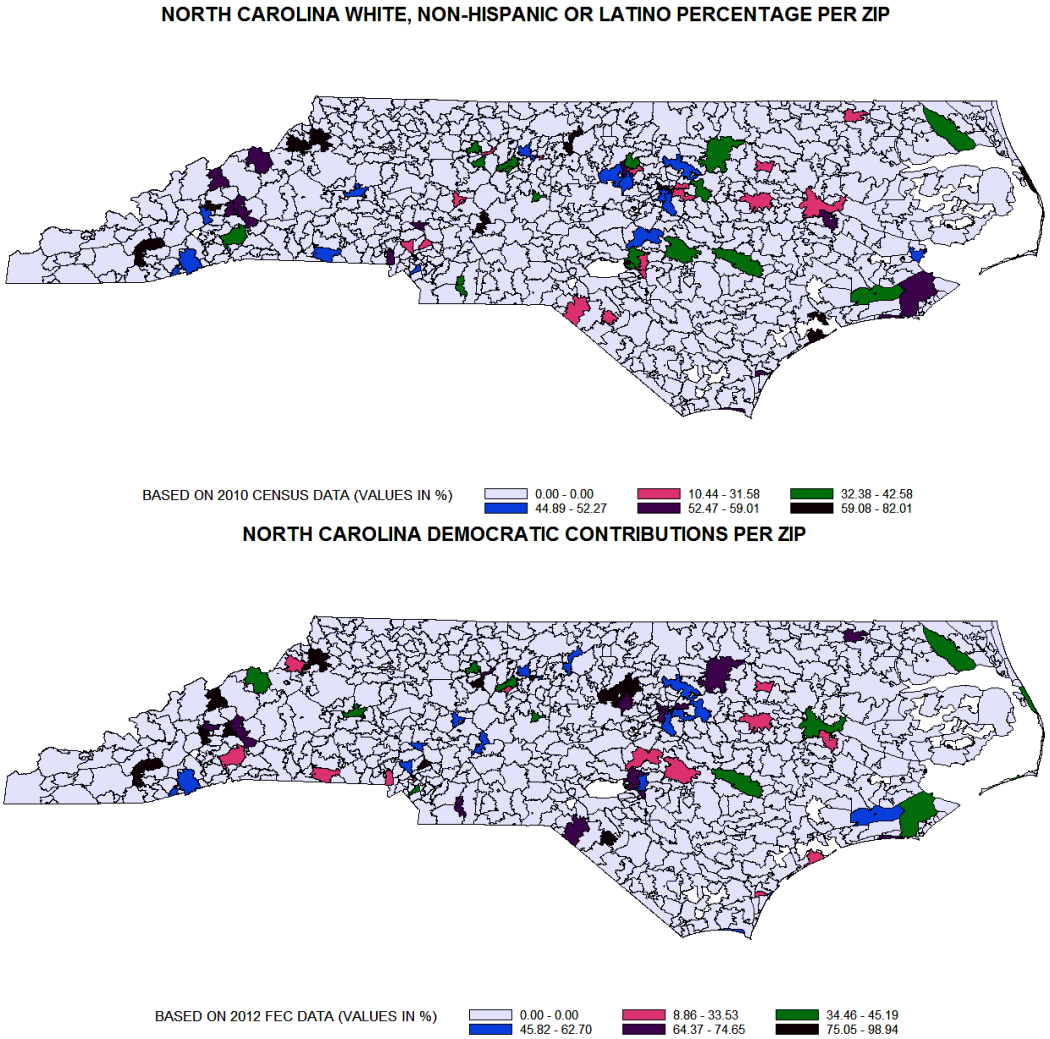

Figure 8. PROC GMAP output of white non-Hispanic or Latino percentage per zip code in North Carolina (top). PROC GMAP output of percentage of Democratic contributions per zip code (bottom).

COMPARING GLM MODEL TO 2012 FEC OUTCOMES (NORTH CAROLINA)

Using the GLM model that was built in previous sections, the following output statement is added,

```
OUTPUT    OUT=COUNT_RESID    P=DEM_HAT    R=DEM_RESID    STUDENT=DEM_STUDENT;
```

where an output data set OUT_RESID is created which holds the predicted values for the democratic contribution percentage per zip code (DEM_HAT), the residuals (DEM_RESID) and the studentized residuals (DEM_STUDENT). This data set is restricted to the state of North Carolina where two maps are produced (figure 9). The first map (top) uses the 2012 FEC data and the contributions per zip code to determine its political leaning by which party received the most contributions. The second map (bottom) is determined by the predicted Democratic contributions per zip code where if the percentage is 51% or higher, the area is determined to be Democratic, otherwise it is determined to be Republican. In both maps, areas without color represent zip codes that did not have reported FEC data. To help visualize how accurate the GLM predictions of Democratic percentages are to the actual Democratic percentages, determined with FEC data, both maps are positioned side by side.

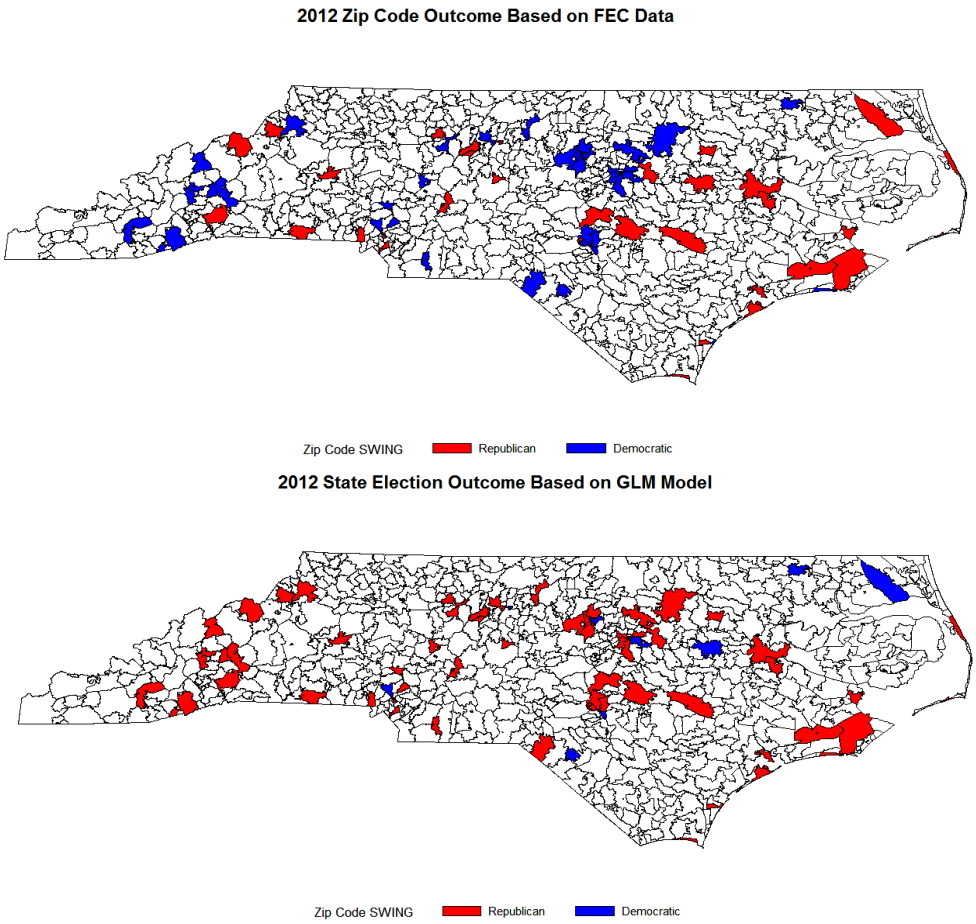

Figure 9. PROC GMAP output of political leanings per zip code based on 2012 FEC data (top). PROC GMAP output of political leanings per zip code based on PROC GLM model predictions (bottom).

With a quick glance, it is easily noticed that the FEC contributions data has a split between Republican and Democratic for the different zip codes. By using the PROC GLM model, there are more zip codes that are expected to be Republican than Democratic in North Carolina. If we assume that this pattern continues throughout the state, ultimately the state will be Republican dominated when it comes to contributions in the 2012 Election cycle. Now in 2013, it is possible to compare which way the state of North Carolina is predicted to have leaned in the 2012 elections (figure 9, bottom) to the actual outcome (figure 10). In both figures the results are the same, North Carolina has leaned Republican.

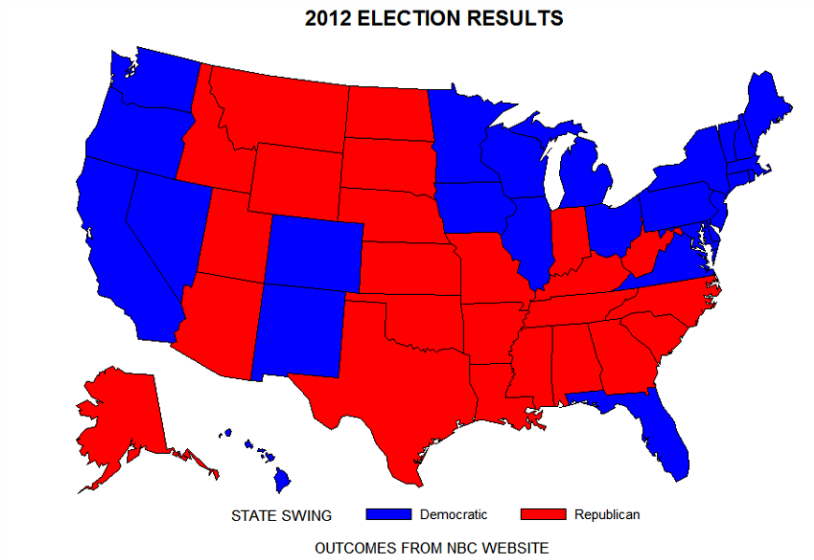

Figure 10. 2012 election cycle results. Data retrieved from NBC website.

CONCLUSION

PROC GMAP is capable of producing graphics which can help visualize data as a whole as well as help identify any patterns that may be present which may not be noticeable from the numeric data alone. Above it has been shown that 2012 election data could be viewed at the national and state levels to get a feel for how each variable looks, observable patterns, and comparing actual outcomes to model predictions. Two areas that could be explored in future papers would be adding more model predictors and additional graphics options and coding techniques to help present visual information in a more compact, interactive, and easy to digest format. The first area would improve model performance and also improve the accuracy of the visual information presented. The second area would allow for quicker detection of patterns and reduce pages of output individuals would have to search through.

REFERENCES

[1.] Reinhard, Beth; Baron, Kevin. “A Military Vote That Doesn’t Really Exist.” National Journal. May 29, 2013. Available at http://www.nationaljournal.com/2012-election/analysis/a-military-vote-that-doesn-t-really-exist-20120528

[2.] “The 2012 Youth Vote.” The Center for Information and Research on Civic Learning and Engagement. July 17, 2013. Available at http://www.civicyouth.org/quick-facts/youth-voting/

[3.] Nelson, Libby A. “Will College Students Show Up.” Inside Higher Ed. September 6, 2012. Available at http://www.insidehighered.com/news/2012/09/06/democratic-national-convention-key-question-will-college-students-vote

[4.] “Politico’s 2008 Swing State Map.” Politico. 2009. Available at http://www.politico.com/convention/swingstate.html

[5.] Politico. 2009. Available at http://www.politico.com/electionmap2008/

[6.] “Presidential Election Results”. NBC News. August 19, 2013. Available at http://elections.nbcnews.com/ns/politics/2012/all/president/#.UhGcrZLCaSp

[7.] Federal Election Commission. Available at http://www.fec.gov/portal/download.shtml

[8.] The University of Missouri Mapping Codes. Available at http://mcdc.missouri.edu/pub/data/corrlst/Zip_to_ZCTA_crosswalk_2010_JSI.csv

[9.] “2010 Census Data”. United States Census. Available at http://www.census.gov/2010census/data/

[10.] “2010 Tiger/Line® Shapefiles”. U.S. Department of Commerce. Available at http://www.census.gov/cgi-bin/geo/shapefiles2010/main

[11.] Okerson, Barbara B. 2013. “Creating Zip-Code Level Maps with SAS®.” *Proceedings of the SAS Global Forum 2013 Conference*. San Francisco, California. Rick Mitchell. Available at http://support.sas.com/resources/papers/proceedings13/214-2013.pdf

[12.] Zdeb, Mike. 2004. “Creating Maps with SAS/GRAPH® - Drill Downs, Pop-Ups, and Animation.” *Proceedings of the SAS Users Group International 2004 Conference.* Montreal, Canada. Duke Owen. Available at http://www2.sas.com/proceedings/sugi29/120-29.pdf

[13.] “The TRANSREG Procedure.” SAS. 2013. Available at http://support.sas.com/documentation

[14.] “The GLM Procedure.” SAS. 2013. Available at http://support.sas.com/documentation

[15.] “GMAP Procedure.” SAS. 2013. Available at http://support.sas.com/documentation

ACKNOWLEDGMENTS

Thanks to Rebecca Ottesen for help with resources and feedback.

CONTACT INFORMATION

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

Name:	Jamelle Simmons
Address:	Virginia Polytechnic Institute and State University
City, State ZIP:	Blacksburg, VA 24060
E-mail:	jmsimm13@vt.edu