Making the Midwest red again: a visual analytics approach to the United States presidential elections

Shawn Ban

December 22, 2017

1 Motivation, data and research questions

In the wake of the 2016 United States presidential elections, political commentary attributed the surprise victory of Republican candidate Donald Trump over Democratic candidate Hillary Clinton to strong support among rural voters with poor economic prospects. Much of the commentary [1] noted the high correlation between Trump's share of the vote and variables such as race, income, education and unemployment, but did not measure the effect of these variables. Popular commentary also tended to explain the effects of these variables on a national level, ignoring spatial variation in these effects [2].

In this project we aim to fill this knowledge gap by examining the relationship between Trump's share of the vote and several explanatory variables, and how this relationship varied at a county level in two states. For this analysis, we restrict our examination to two states, Michigan and Wisconsin, although the method could be extended to each of the 50 states. We believe these two states could prove particularly informative for a number of reasons:

- The margin of victory was small in both cases: 0.23% for Michigan and 0.77% for Wisconsin.
- Together they account for a sizeable 26 electoral college votes out of 538, or about 4.8% of the total.
- Both states typically voted Democrat in recent years: the Democrat candidate won the last six presidential elections in Michigan going back to 1992, and the last seven in Wisconsin going back to 1988.
- They are contiguous and represent two of the 12 states in the American Midwest as defined by the United States Census Bureau, so any findings could potentially be generalized to the rest of the region.

Data on the election results was gathered from the Secretary of State Offices of Michigan and Wisconsin, while demographic data and geographic boundaries are provided by the

Census Bureau. We choose eight explanatory variables to work with: median income, unemployment rate, veteran rate, percentage of male population, median age, percentage of white population, percentage of blue-collar workers, defined as workers employed in the construction, mining or transport sectors, and percentage of population with a bachelor's degree.

Some research questions we hope to answer are:

- How did Trump's share of the vote vary with the explanatory variables?
- How did this relationship vary over space - was it constant between the two states?
- Which counties in Michigan and Wisconsin are most like each other? What groupings can be produced of these counties?

2 Tasks and approach

As part of our exploratory data analysis, we first gather our explanatory variables and visualize them via histograms, performing transformations to approximate normality if necessary. We then explore how the dependent variable, Trump's percentage share of the vote, varied spatially in each state by plotting choropleths and cartograms.

Next, we investigate the relationship between the dependent variable and the explanatory variables by plotting a correlation heatmap and a series of scatterplots. The correlation heatmap also allows us to visualize how the explanatory variables vary with each other, and provide information on which variables may have collinearity.

As a further test for collinearity, we then perform principal component analysis on the explanatory variables, checking the eigenvalues of the correlation matrix. In particular, we compute the condition number, or the ratio of the largest to the smallest eigenvalue, to check that collinearity is within acceptable limits. At this stage we eliminate explanatory variables if necessary.

We then build a linear regression model with the remaining variables, proceeding in a step-wise manner by dropping one regressor at a time until all remaining regressors have a p-value below 0.05, indicating statistical significance. The residuals are then visualized via a choropleth, and if there seems to be spatial bias in the residuals, we implement a geographically-weighted regression model. The theory for the treatment of spatial nonstationarity was developed by Brunsdon et al. (1996) [3] and work on election data has been done previously by Beecham et al. (2017) [4]. For this analysis, we use the *GWModel* package in R developed by Gollini et al. (2015) [5].

Finally, we use the regressors from our models to implement K-means clustering and check if the counties can be divided into meaningful clusters. A choropleth of clusters is then built that serves as a map of characteristics of each state.

3 Analytical steps

We first plot histograms of the dependent and explanatory variables for Michigan (Figure 1) and Wisconsin (Figure 2). Due to skew, we perform log transforms on the population density variable.

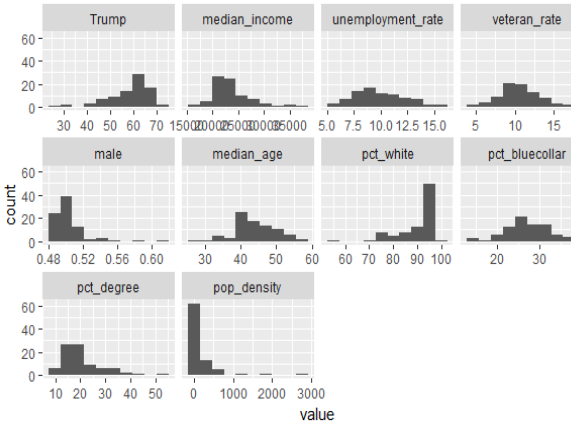

Figure 1: Histograms of variables, Michigan

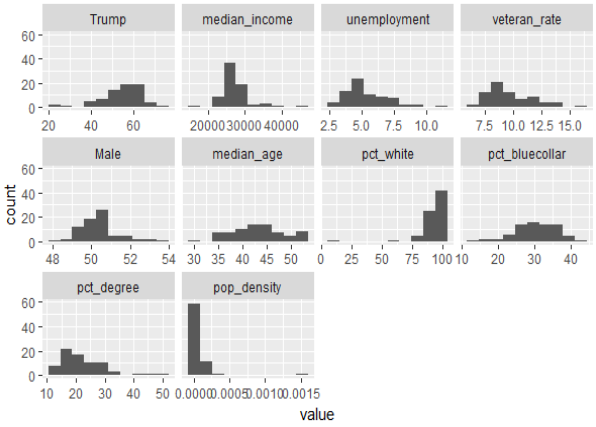

Figure 2: Histograms of variables, Wisconsin

We then plot choropleths of Trump's percentage share of the vote for Michigan (Figure 3) and Wisconsin (Figure 4). The areas with low support for Trump are represented by lighter colours, and we note that these areas are mostly urban centres, with one exception in Wisconsin - Menominee County, in the northeast of the state, is a protected Native American reservation and very sparsely populated. We believe this county could prove to be an outlier when building our linear model. As urban areas typically have higher minority proportion and population density, and more service sector jobs that require a degree, the choropleth analysis supports a number of the explanatory variables.

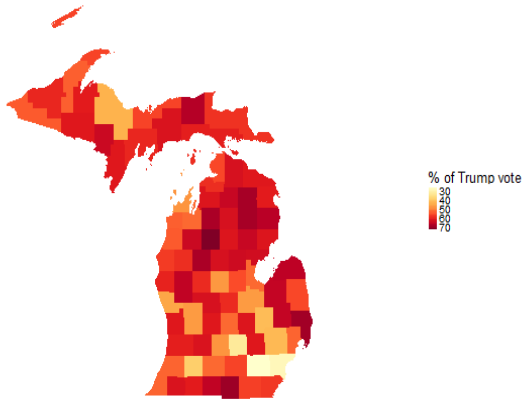

Figure 3: Trump's support by county, Michigan

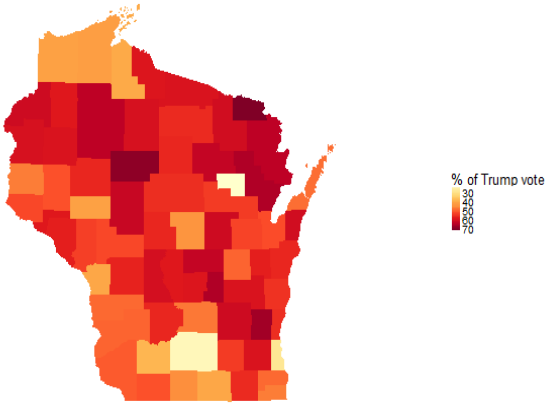

Figure 4: Trump's support by county, Wisconsin

We next plot a correlation heatmap of all the variables for Michigan (Figure 5) and Wisconsin (Figure 6). There seems to be a high correlation between the dependent and independent variables, as shown by the bottom row, which is promising. We note also the high correlation between several of the independent variables, such as income and degree percentage, and veteran rate and median age, which serves as a warning about potential multicollinearity. Finally, the correlation between Trump's support and the independent variables mostly have the same polarity in Michigan and Wisconsin, with the exception of income and unemployment.

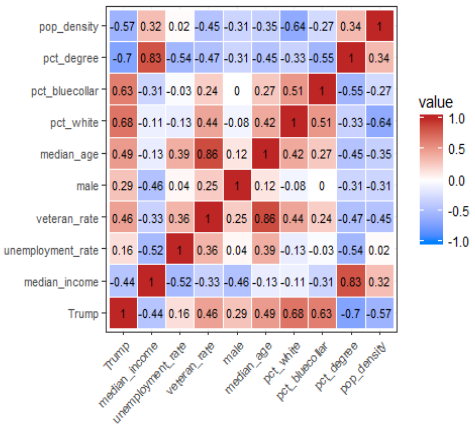

Figure 5: Correlation heatmap, Michigan

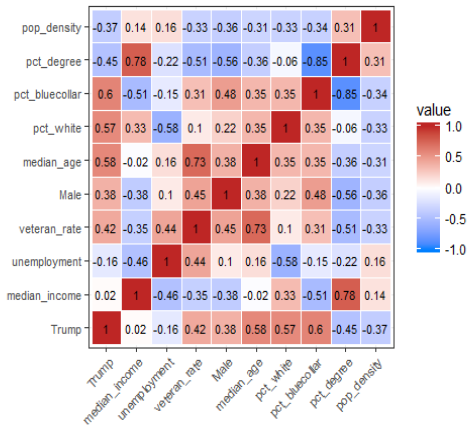

Figure 6: Correlation heatmap, Wisconsin

A scatterplot analysis for the states (Figures 7, 8) confirms these relationships. We note that income is negatively correlated with Trump's support in Michigan but has low correlation in Wisconsin. The scatterplots also provides visual evidence for the outlier of Menominee County, represented by the small circle with a low support for Trump.

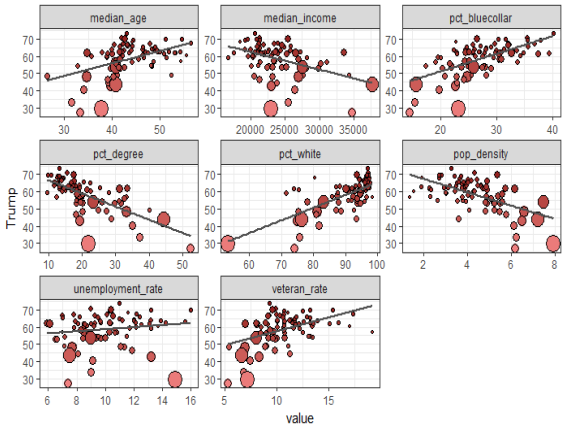

Figure 7: Trump's support vs. explanatory variables, Michigan

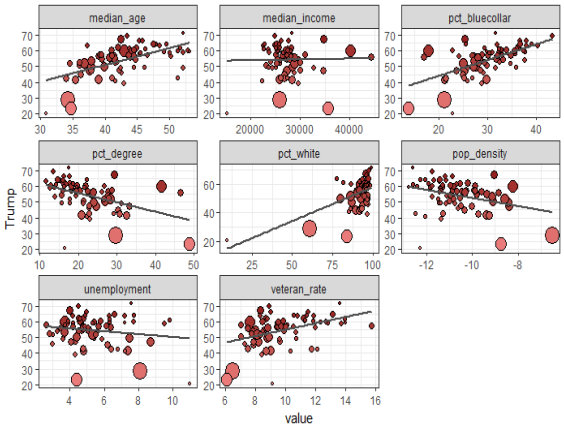

Figure 8: Trump's support vs. explanatory variables, Wisconsin

The analysis so far suggests that we are likely to use different regressors in building linear models for the two states. For Michigan, we choose five explanatory variables: income, age, percentage of white population, percentage of degree education and population density. We first run a principal component analysis by checking the eigenvalues of the correlation matrix and note the condition number, or the ratio of the largest to the smallest eigenvalue, is 57.4. As a rule of thumb, we consider 100 as a threshhold for the condition number, and conclude that multicollinearity is within acceptable limits. The linear regression yields the following output:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.089e+00  9.026e-01  -4.530 2.12e-05 ***
median_income -1.728e-04  3.808e-05  -4.538 2.06e-05 ***
median_age     3.682e-01  1.049e-01   3.510 0.000753 ***
pct_white      3.538e-01  6.487e-02   5.453 5.77e-07 ***
pct_degree    -1.062e+00  1.407e-01  -7.550 7.46e-11 ***
pop_density   -4.643e-01  1.046e-01  -4.441 2.96e-05 ***
---
Residual standard error: 0.4734 on 77 degrees of freedom
Multiple R-squared:  0.7895,Adjusted R-squared:  0.7759
F-statistic: 57.77 on 5 and 77 DF,  p-value: < 2.2e-16
```

The model has statistical significance and explains 77.6% of the variation in Trump's support between counties, according to the adjusted R-squared. A quantile-quantile plot of the residuals suggests they are mostly normally distributed, with some bias in the tails (Figure 9). We also examine a choropleth to detect any spatial bias (Figure 10) in the residuals, and note the cluster of blue in the middle of the state and the Upper Peninsula, as well as the cluster of red in the south, around Ann Arbor and Detroit.

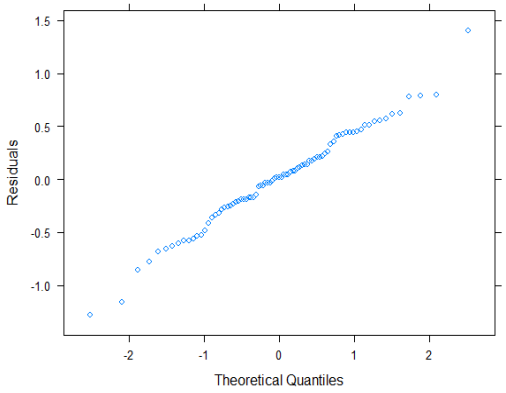

Figure 9: Q-Q plot of model residuals, Michigan

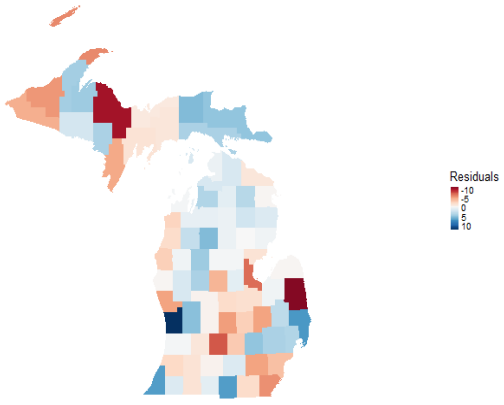

Figure 10: Choropleth of residuals, Michigan

For Wisconsin, we perform a similar analysis, but drop income as a regressor due to low correlation. We initially run the regression with four explanatory variables: age, percentage of white population, percentage of degree education and population density. We note the condition number is lower at 7.5, indicating lower multicollinearity. The initial run of the regression shows that population density has too high of a p-value for Wisconsin. Dropping this leaves a linear model with three explanatory variables:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.19638    8.85876   0.474 0.637232
median_age   0.58157    0.16728   3.477 0.000889 ***
pct_white    0.37222    0.07197   5.172 2.22e-06 ***
pct_degree  -0.41118    0.11165  -3.683 0.000459 ***
---
Residual standard error: 6.46 on 68 degrees of freedom
Multiple R-squared:  0.5745,Adjusted R-squared:  0.5558
F-statistic: 30.61 on 3 and 68 DF,  p-value: 1.226e-12
```

The model is statistically significant, but its explanatory power is just 55.6%. We note also the high p-value of the intercept. We similar examine the residuals for bias via a q-q plot (Figure 11) and a choropleth (Figure 12). There is strong evidence of spatial bias in the Wisconsin model, with clusters of red in the northwest and southwest of the state.

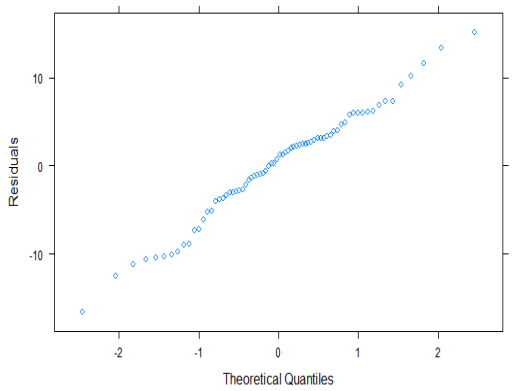

Figure 11: Q-Q plot of model residuals, Wisconsin

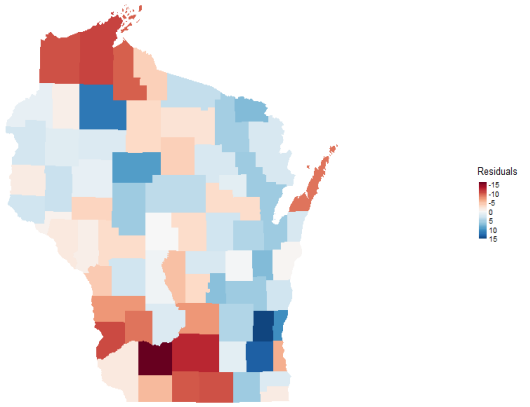

Figure 12: Choropleth of residuals, Wisconsin

We then implement geographically-weighted regression models, or local models, for both states, first solving for the optimal bandwidth that minimizes the small-sample size corrected Akaike information criterion (AICc). The solver produces an optimal bandwidth of 71 for Michigan and 26 for Wisconsin. Both models benefit from the geographical weighting, reporting a higher adjusted R-squared of 80.7% for Michigan and an impressive 75.8% for Wisconsin. Figures 13 and 14 show the correlation coefficients of Trump's support against the explanatory variables and demonstrate the smoothing effect of considering local models.

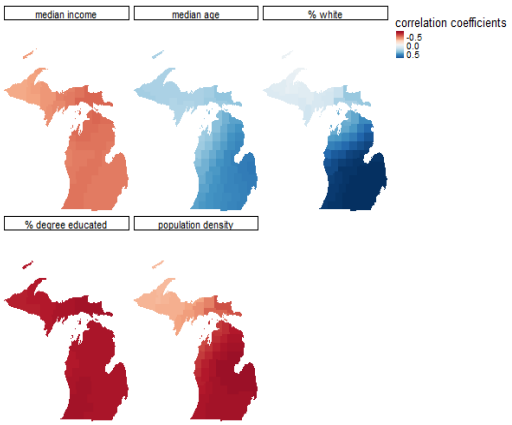

Figure 13: Geographically weighted correlation coefficients, Michigan

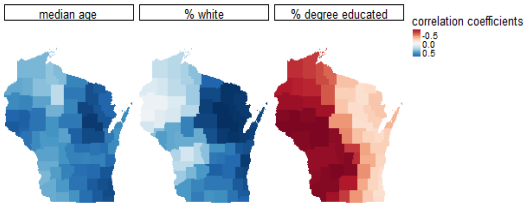

Figure 14: Geographically weighted correlation coefficients, Wisconsin

Finally, we cluster the counties based on the geographically-weighted correlation coefficients via K-means clustering. The elbow method suggests an optimal number of 3 clusters for Michigan (Figure 15) and 4 clusters for Wisconsin (Figure 16).

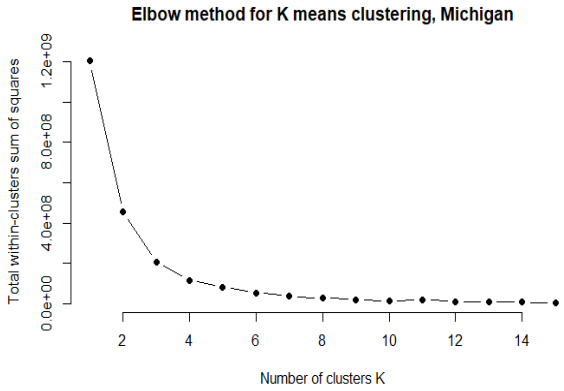

Figure 15: Elbow method for clustering, Michigan

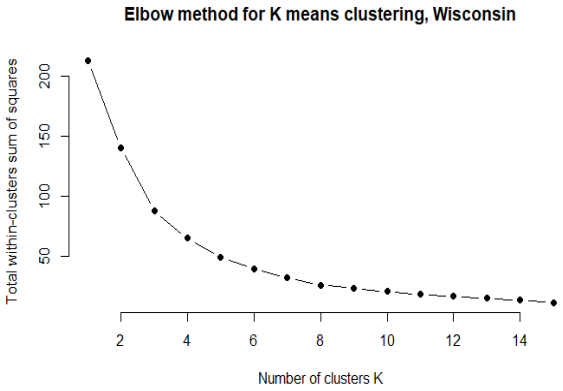

Figure 16: Elbow method for clustering, Wisconsin

Figures 17 and 18 show the output from the K-means algorithm. It is difficult to immediately characterize the clusters, but the dark blue clusters in Michigan indicate the urban centres of Lansing, Ann Arbor and Detroit. It is surprising several counties in the northwest share some characteristics with these cities. For Wisconsin, the urban centres are located in the light blue clusters, and we note the uniqueness of Menominee County. We also note that the dark blue clusters are mostly located in the north of the state and correspond with heavily forested areas.

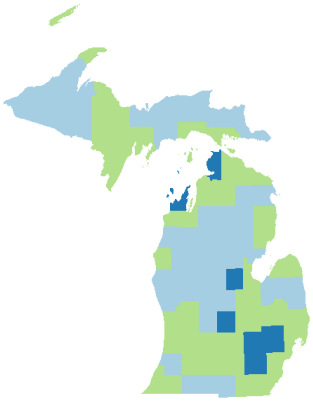

Figure 17: K-means clustering, Michigan

Figure 18: K-means clustering, Wisconsin

4 Findings

The evaluation criteria of the linear models is summarized in the following table:

	Michigan models		Wisconsin models	
Criteria	Global	Local	Global	Local
Adjusted R^2	77.6%	80.7%	55.6%	75.8%
AIC	482.4	92.9	478.9	87.7

The higher adjusted R^2 and lower AIC demonstrates the improvement of local over global models. In answering our initial research questions, we note that Trump's support varied significantly with income, age, percentage of white population, percentage of population with a degree, and population density in Michigan. In Wisconsin, however, it varied significantly with just age, percentage of white population, and percentage of population with a degree. The relationship was not static, and displayed some spatial bias, more so in Wisconsin than Michigan. By considering geographically-weighted models, we were able to control for some of the spatial bias and improve model performance by a noteworthy amount, particularly in Wisconsin.

Meanwhile, a cluster analysis revealed stark differences among counties in each state, particularly between urban and rural areas, and between the upper and lower halves of both states. Some rural counties in upper Michigan seemed to demonstrate the characteristics of urban areas, while Menominee County in Wisconsin was a unique outlier.

5 Critical reflection

Choropleths are a tried and tested method of visualizing election and census data. However, we are unable to measure statistical significance merely through visual inspection,

particularly when multiple variables are involved, and where multicollinearity exists. Linear models are useful for explaining variation, but may display spatial bias. In such an event, geographical weighting can be useful for improving model performance. We believe this method can be generalized to most types of census data where spatial heterogeneity exists, but it should not be seen as a panacea for all linear regression models.

One potential limitation of this study is that it assumes that the electorate in each county is representative of that county's demographics. The demographic data is from the 5-year estimates of the American Community Survey conducted by the United States Census Bureau, covering a period of 2012-2016, and migrations in that time period are not captured in this analysis. Urban areas in particular may have a large number of temporary residents, such as migrant factory workers or university students. Additionally, nationwide voter turnout was just 55.5% in 2016, and some residents may be ineligible to vote due to non-citizen status or prior felony convictions.

A further caveat is that linear models typically have strong explanatory power but poor predictive accuracy. Variables that had a significant effect in the 2016 election may have little predictive power for future elections.

Nevertheless, the findings in this study at least demonstrate the differences between two states and between counties within each state. Variables that have an explanatory effect in one area may not have a significant effect in another. The improved performance of the local models over the global models reveals the special importance of geography in election data. We should therefore be skeptical of blanket claims that any one variable or collection of variables, such as income and unemployment, explained Trump's surprise victory, especially nationwide, particularly as we cannot even make that claim between two states.

The cluster analysis could also be useful for political strategists planning for subsequent elections. Being able to aggregate large number of counties into a small number of clusters could be valuable in efficient allocation of resources, such as advertising campaigns and voter targetting efforts. Campaigns could potentially tailor messages to different areas, and run a number of parallel advertising strategies within each state. Cluster analysis could be readily extended to other domains, such as businesses deciding where to locate retail outlets, or how to tailor their advertising messages. We caution, however, that multivariable clustering can be very susceptible to feature selection. A different set of explanatory variables could have produced a vastly different set of clusters for each state.

For future work, there remains the 20-25% variation that is unexplained by the local models. One option might be to gather more explanatory variables to improve the explanatory power of the models. Furthermore, the analysis in this study could readily be extended to other states and to the entire country as a whole. Other methods of clustering could also be chosen, such as self-organizing maps, and the results compared to the K-means clusters produced here.

References

[1] Silver, N. (2017) *Education, Not Income, Predicted Who Would Vote For Trump.* [online] FiveThirtyEight. Available at: http://fivethirtyeight.com/features/education-not-income-predicted-who-would-vote-for-trump/ [Accessed 22 Dec. 2017].

[2] Balz, D. (2017) *Rural America lifted Trump to the presidency. Support is strong, but not monolithic.* [online] Washington Post. Available at: https://www.washingtonpost.com/politics/rural-america-lifted-trump-to-the-presidency-support-is-strong-but-not-monolithic [Accessed 22 Dec. 2017].

[3] Brunsdon, C., Fotheringham, A., & Charlton, M. (1996) Geographically Weighted regression: A Method for Exploring Spatial Nonstationarity. *Geograhpical Analysis*, 28 (4), 281-298.

[4] Beecham, R. , Slingsby, A. & Brunsdon, C. (2017) Locally-varying explanations behind the United Kingdom's vote to leave the European Union. *Journal of Spatial Information Science*, ISSN 1948-660X (In Press).

[5] Gollini, I., Lu, B., Charlton, M., Brunsdon, C. & Harris, P. (2015) GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. *Journal of Statistical Software*, 63 (17), 1-50.