Identifying Winners & Losers of National Science Foundation Funding with Clustering

By Erik Lopez

Abstract

Trends in research funding by the Natural Science Foundation from 2010 to 2018 have affected distinct groups of congressional districts differently. By applying KMeans clustering onto grant funding data from Federal RePORTER we were able to identify two clusters of similarly affected congressional districts: districts improving in terms of funding, and districts receiving less funding. However, between these two clusters we did not find any significant differences in urbanization or density levels, nor did we find a significant difference between them in political leanings as measured by the voting outcome in the 2016 presidential election and the number of times each political party has held each congressional seat in the same time frame. These results would imply that there is no relation between the urbanization, density, or political lean and funding outcomes for the given time frame. Unfortunately, we find that these results may also been affected by the political nature of congressional districts, as they exhibit a huge variability in size and shape. Lastly, we observed that congressional districts with no clear funding trends partly obfuscated our clustering results as there is no clear way of grouping these districts. As an extension to our project we recommend rerunning the clustering analysis using a different unit of analysis with clustering algorithms better equipped at filtering noisy erratic data point.

Introduction

The federal government is a major player in funding scientific research throughout the United States. However, due to shifts in government budgets, the distribution and amount of that funding changes year to year and could very well be affecting its various recipients unevenly. Our main concern is that in the long term we risk permanently neglecting scientific research in regions of similar characteristics, creating a situation where innovation and research opportunities are considerably concentrated in more homogeneous areas of the country.

For this project we will try to identify the main federal funding trends at the congressional district level using the machine learning clustering algorithm KMeans. Once we have programmatically clustered similarly funded congressional districts, we will run some general descriptive analysis to determine if the districts between each cluster differ on any common characteristics like political lean, urbanization level, or population density. We will focus our efforts on research funding distributed by the National Science Foundation (NSF) from 2010 to 2018 to avoid noise from other government departments' own research funding changes and to limit the scope of our research.

Data and Methods

The most important dataset for our research comes from Federal RePORTER, which provides research funding award information such as the amount of the award, the year it was awarded, and the congressional district where a grant was delivered. From this platform we collected grant funding data from the year 2010 to 2018 and aggregated this data on the designated congressional district. This way each row in our cleaned funding dataset would correspond to a congressional district, and each column would correspond to the total funding that district received for a specific year. However, we could not match all funding records to a valid congressional district. As a result, out of the $35.39 billion of total NSF funding distributed from 2010 to 2018, we could not include into our analysis $2.02 billion, which represents 5.71% of the total funding.

For urbanization and density metrics we used CityLab's Congressional Density Index (CDI) dataset (Montgomery). This dataset contains each congressional districts' urban and density breakdowns, along with the presidential election 2016 voting results as a proportion. Additionally, to determine the political lean of each congressional district, we also programmatically scraped the Congress.gov website for the political party of each representative that has held that congressional seat from 2010 to 2018. We then counted how many Democrats or Republicans have held each district each year for the same timeframe. The proportion between Democrats and Republicans in each district will determine how each district leans in terms of political affiliation. Finally, to be able to visualize our clustering results onto a map of the country, we used Lewis et al. Congressional District Shapefiles as we found them to require no heavy preprocessing for our purposes.

Our first step in the analysis was to run Principal Component Analysis (PCA) on the normalized NSF congressional funding dataset. This process allows us to represent our 9-dimensional dataset (one dimension per year from 2010-2018) in two dimensions which will be helpful for visualizing our clustering results. As Figure 1 shows, the first two components of our PCA results can explain most of the variance in the funding data, with diminishing returns beyond two components. This means that PCA is a proper method to use to visualize our dataset. Figure 2 shows the resultant scatter plot from using the first two PCA components.

Visually on Figure 2 we can infer that there seems to exist three possible clusters, but the naked eye is not a great tool for finding the best number of clusters. Additionally, since the clustering algorithm KMeans needs the user to pre-define the goal number of clusters, we need to run our clustering process with the number that best separates the given data. To find this optimal number of clusters we ran KMeans using numbers between 2 to 10 and scored each clustering result with the Silhouette Scoring algorithm. This scoring algorithm measures how good our clusters separate the data points from each other, where a score of 1 represents a perfect separation of data points. After scoring all nine results, we kept the cluster number that yielded the best score. In our case, we found the highest score of 0.27 when we performed KMeans clustering with 2 clusters. Figure 3 visualizes the clustering results from using only 2 clusters onto the PCA datapoints.

Finally, once we found the cluster assignments for all congressional districts, we started our descriptive analysis. To do so, we joined both the Congress.gov and the Congress Density Index dataset with the proper cluster label according to its congressional district. By adding the cluster label to both datasets, we can filter by the cluster label to calculate the mean and error of each relevant category. The mean and error values found in this process will be compared against each other to find if any significant differences between both clusters.

Results

Our results found that, as shown in Figure 4, the two clusters exhibit two different trends: cluster #0 showing a decreasing funding trend, while cluster #1 showing a gradual increase in funding. However, we do see that the error bars do overlap in several years, making our findings less certain. We hypothesize this could be due to the inclusion of congressional districts with inconsistent funding trends that obscure our results when included in either cluster.

Next, after finding that the two clusters showed different trend lines in funding, we moved to check if these two clusters differed in other aspects. Using the Congressional Density Index dataset, we summarized each cluster in terms of density and urbanization levels. Figure 5 shows the average urban levels between the two clusters; however, the error bars make it clear that none of the differences between the clusters are significant. It is a similar story for Figure 6 when we look for differences in density between both clusters. Additionally, we also examined the political differences between the two clusters. Figure 7 shows how heavily Democratic or Republican each cluster tends to be. Figure 8 shows the results on how each congressional district voted on the 2016 presidential election. In both cases we also do not see any significant differences between both clusters.

Finally, we also plotted the results of our cluster assignments onto a map of the continental United States, Figure 9. Although it seems that some regions in the US seem to be disproportionately a part of cluster #0 (the cluster with declining NSF funding) we need to remember that Congressional Districts are much bigger in rural areas. These differences in sizes and shapes (due to gerrymandering) makes it hard to objectively determine if there are any regional discrepancies in funding without being biased.

These results would point that the changes in funding are not related or correlated to political differences, at least when aggregated at the congressional district level. If this is the case, we then need to do more studies about how districts have been impacted by the reduction in grant awards and find the underlying reasons as to why we’re seeing less money go to these districts. Once after we find these reasons and the impact of lower funding, we can suggest proper policy solutions to address this lack of funding.

Limitations & Biases

When we ran our descriptive analysis of both clusters for density and urbanization levels we were limited by the accuracy of our dataset. Since the Congressional Density Index dataset is just a snapshot of all congressional districts, we assumed that districts did not change in demographics too steeply to make this dataset obsolete. To make our analysis more accurate we would have to find a way to track urban and density changes on all districts for the same time period of our analysis.

Additionally, since some grants did not have a proper congressional district and thus could not be matched, this funding was not included in our analysis. The omission of around 5.71% of the total funding awarded by the NSF from 2010 to 2018 potentially skewed the accuracy of our clustering as some districts could have been underrepresented in terms of funding. Additional background information about how Federal RePORTER tags a grant to a congressional district would be needed to determine if it is even possible to incorporate these unmatched records.

Finally, the accuracy of our results was limited to the shortcomings related to KMeans clustering. The KMeans clustering algorithm has a bias for circularly shaped clusters, and as we saw on Figure 2 our data is elongated along the y-axis. To get around this issue we could potentially try other clustering algorithms, like Gaussian Mixture, to handle irregularly shaped data much better. Furthermore, given that some congressional districts had an erratic funding pattern, we also could benefit from rerunning this analysis using clustering algorithms that can “filter out” noisy datapoints like the DBSCAN clustering algorithm.

Conclusion & Possible Extensions

To summarize, our analysis found that from the years 2010 to 2018, congressional districts seem to be split into two groups: districts that saw an increase in NSF funding awards, and districts that saw a decline. However, we did not find any significant differences between the two groups in terms of density, urbanization, and political lean, signifying that the trends in funding are not related to population demographics nor politics. Before making policy suggestions, we do suggest rerunning the project using a different unit of analysis that is more consistent than a congressional district in terms of size and borders shape. We also suggest as a future expansion to try other clustering algorithms that can better fit to the shape of our data, like Gaussian Mix clustering, or an algorithm that can filter out outliers, like DBSCAN clustering. Finally, we suggest adding geospatial analysis to determine if there is a relation between funding trends and the map location of the grant award.

REFERENCES

Federal Reporter. (2020). Retrieved March 15, 2020, from https://federalreporter.nih.gov/

Lewis, J. B., DeVine, B., & Pritcher, L. (2018). United States Congressional District Shapefiles. Retrieved April 1, 2020, from http://cdmaps.polisci.ucla.edu/

Lopez, Erik (2020, April 15). Big Data for Public Policy's Project Repository, found in https://github.com/lobodemonte/big-data-for-public-policy

Montgomery, D. (2018, November 28). CityLab's Congressional Density Index. Retrieved April 1, 2020, from https://www.citylab.com/equity/2018/11/citylab-congressional-density-index/575749/

United States Congress. (2020). Congress Members. Retrieved March 15, 2020, from https://www.congress.gov/

APPENDIX

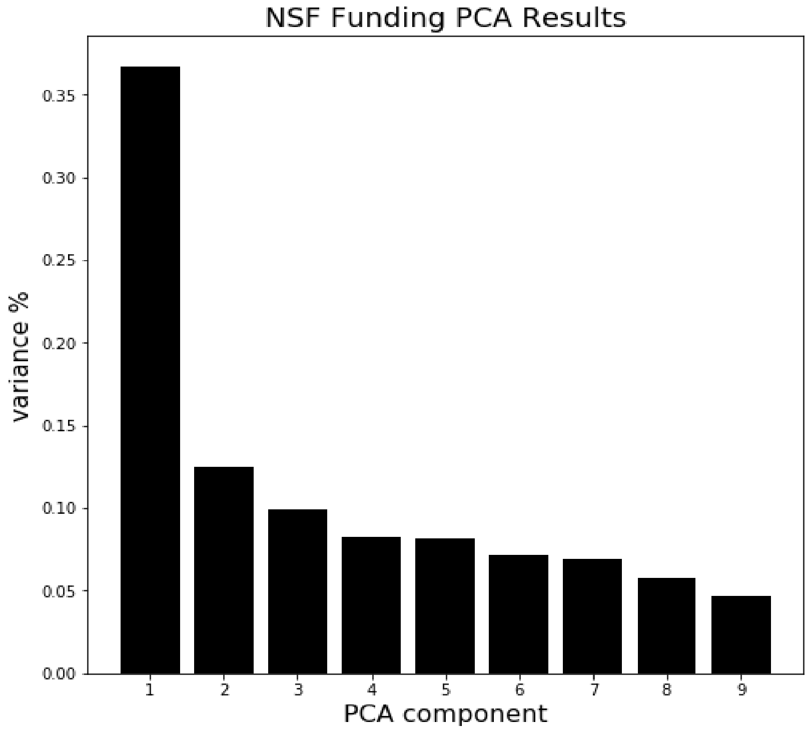

Figure 1

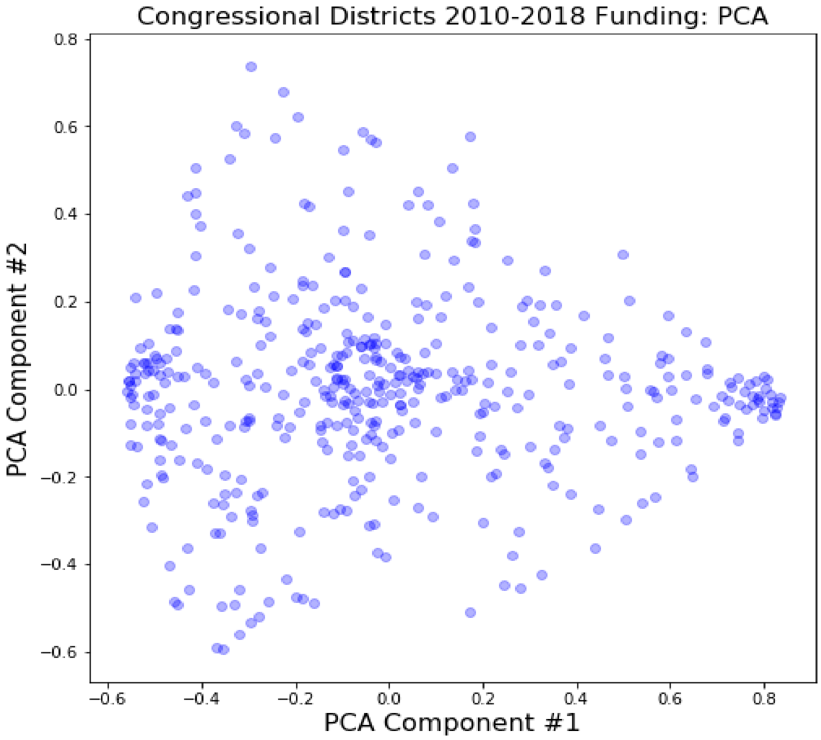

Figure 2

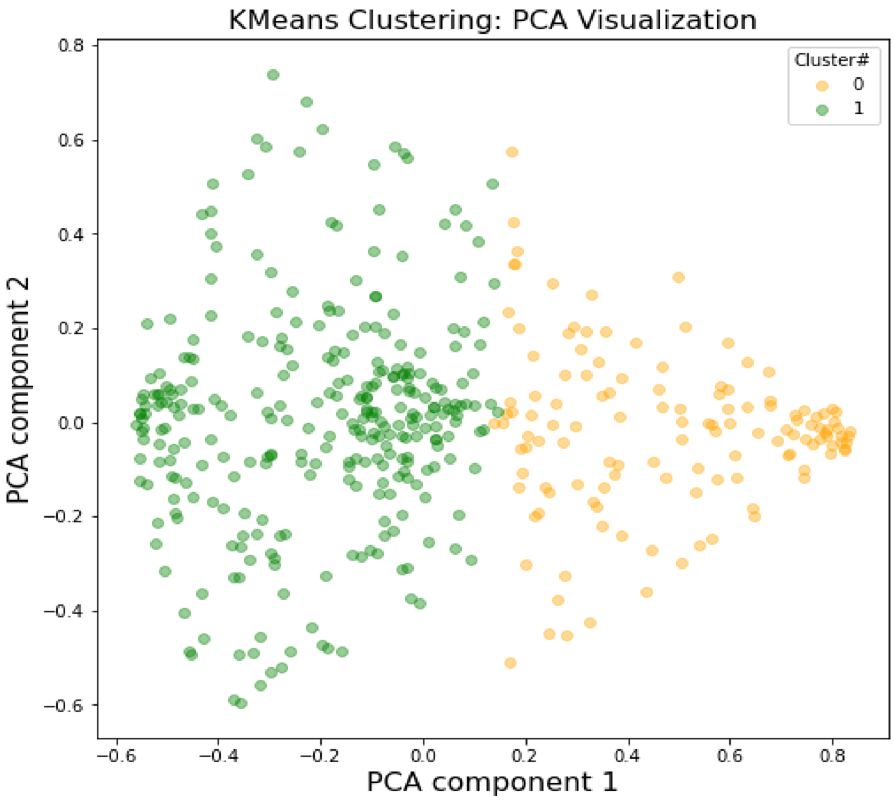

Figure 3

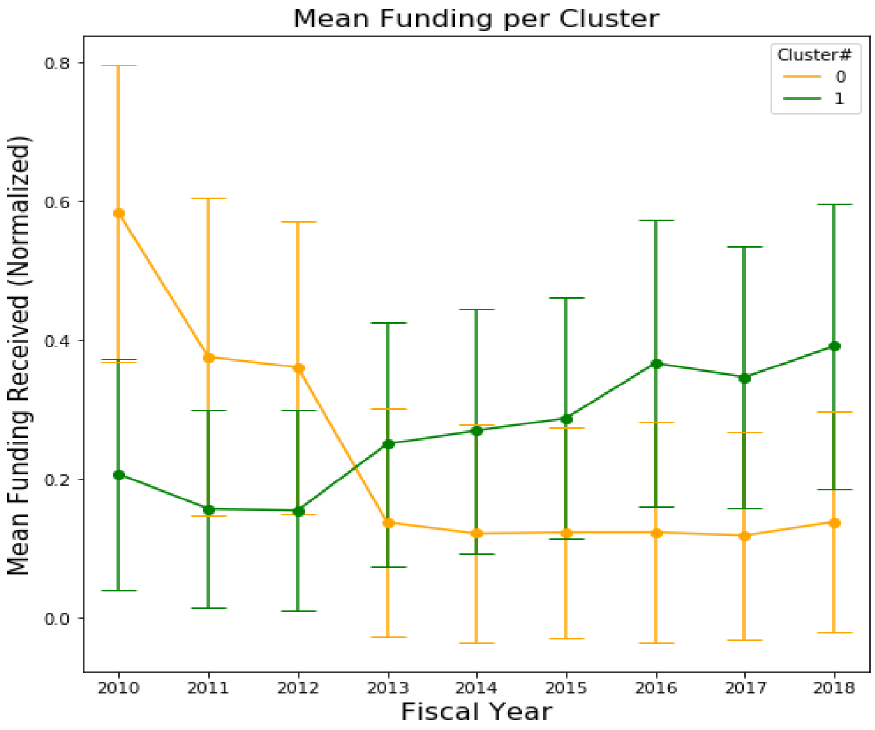

Figure 4

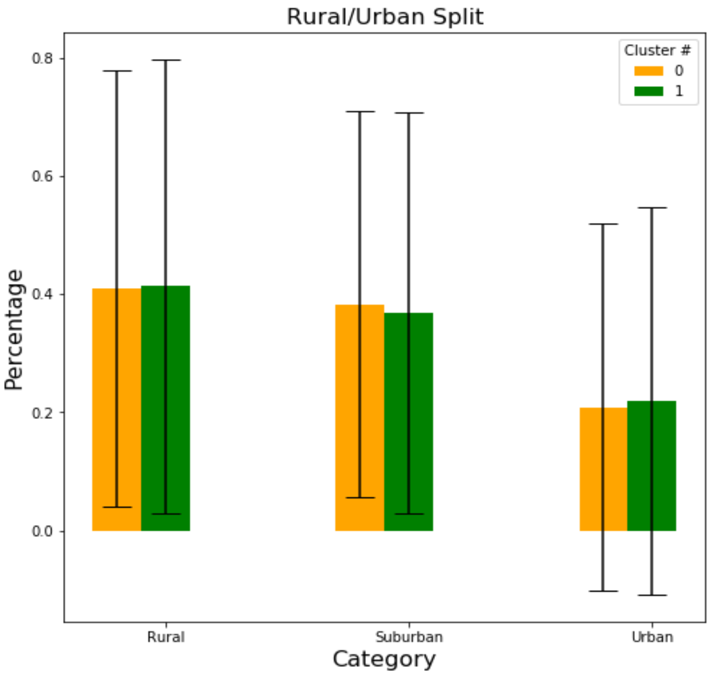

Figure 5

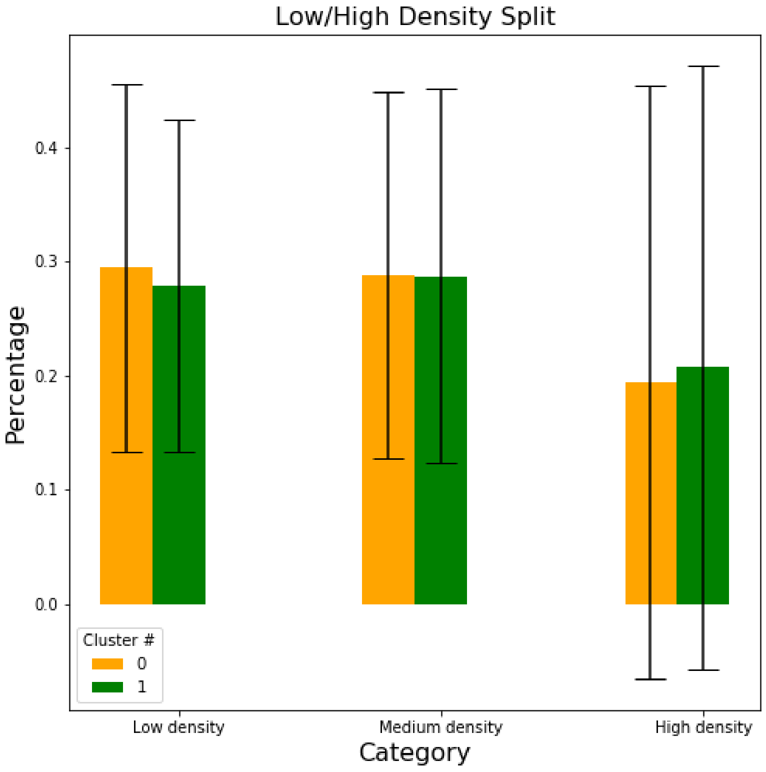

Figure 6

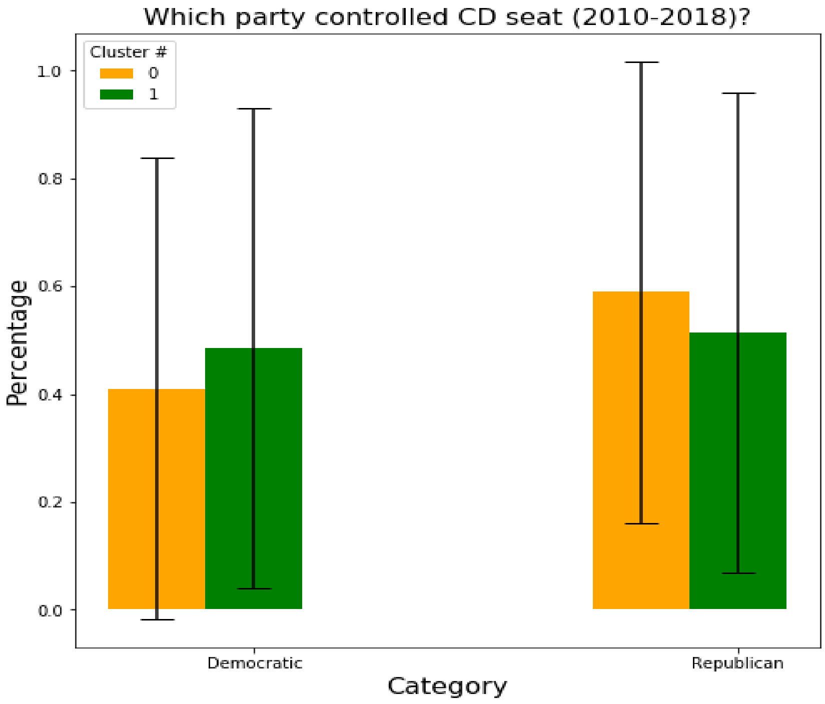

Figure 7

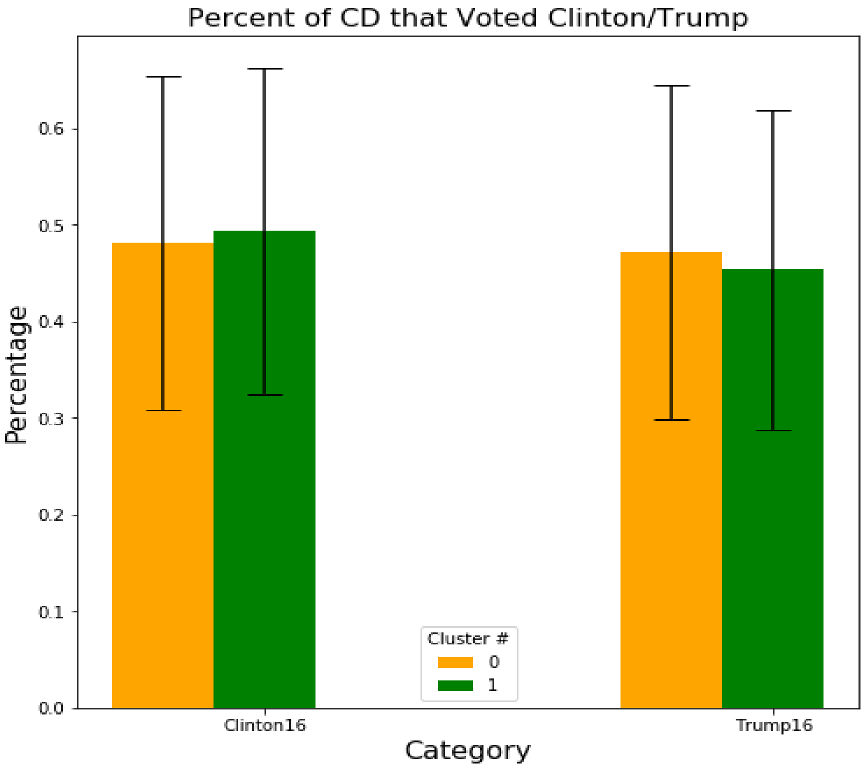

Figure 8

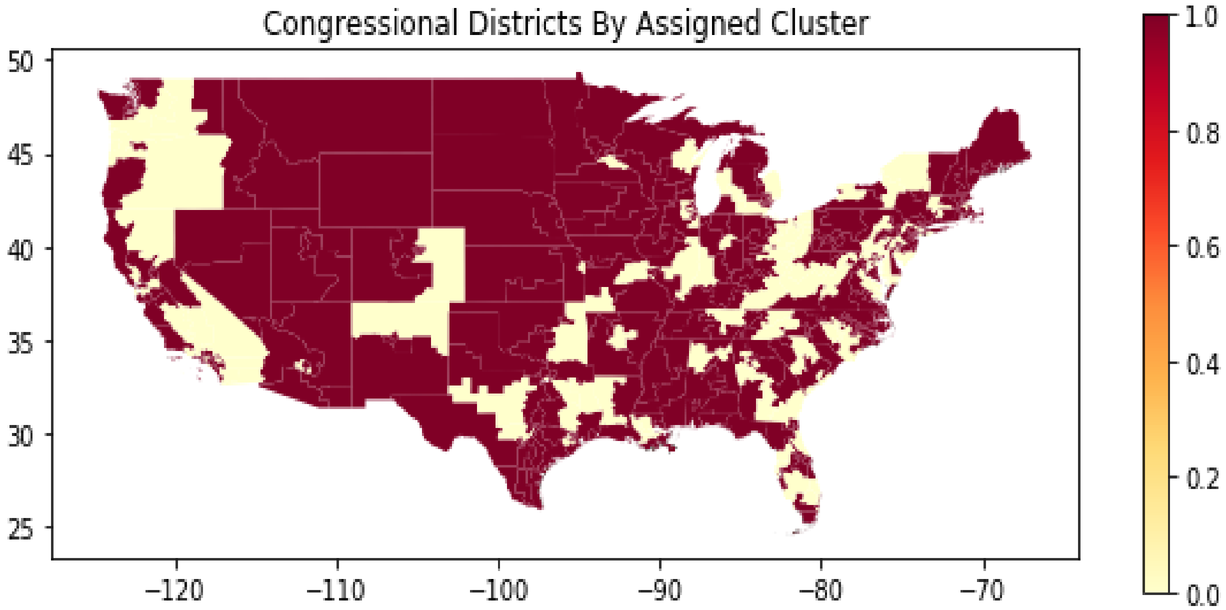
Congressional Districts By Assigned Cluster
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70
1.0
0.8
0.6
0.4
0.2
0.0

Figure 9