Defining Opinion Space:

A K-means Cluster Analysis of Pew's Political Typology Survey

Jonathan Campbell

Introduction

Discussion of the polarization of American political opinion has been the hot topic of political science in recent years. Data has suggested that Americans are more and more inclined to identify with either one of the main political parties, leading to a sense of a widening philosophical gap in the center of public opinion as moderates and independents appear to disappear. Beyond that, the focal points of our political parties have come to embody viewpoints that lie either far right or far left on the Conservative-Liberal scale. Pundits and commentators take these diverging figureheads and use them as models to infer the distribution of American public opinion.

The issue with this trend is that it tends to discount moderate viewpoints as insignificant due to the center's apparent decline in size and importance. Despite this sense, the spectrum of political ideology is much more diverse than political discourse suggests. Social and political interaction has been shown to play an important role in the formation of a person's ideology as does news and media. These interactions can account for the growing number of individuals associating their values with a political party but not necessarily for any substantive change in ideology. Therefore, in order to truly value whether and how polarization is affecting American society, it is necessary to map ideological groupings that can more accurately portray the spectrum of public opinion.

This research looks at the space of political ideology analyzing the data used in Pew Research Bureau's Political Typology Survey (2011) and running K-means cluster analyses. My research goal would be to look at the political ideology of Americans in order to classify and identify areas (clusters, silo's etc) of public opinion in the United States. The identification of clusters has many practical implications. First and foremost, changes in the underlying distribution greatly affects the assumptions of political organizers and politicians concerning their views of the electorate. Analyses such as these give insight to the mentalities of citizens and voters' issue attitudes go together on an ideological level. Additionally, it would provide insight into the existence of a polarization trend in American politics.

Design and Methodology

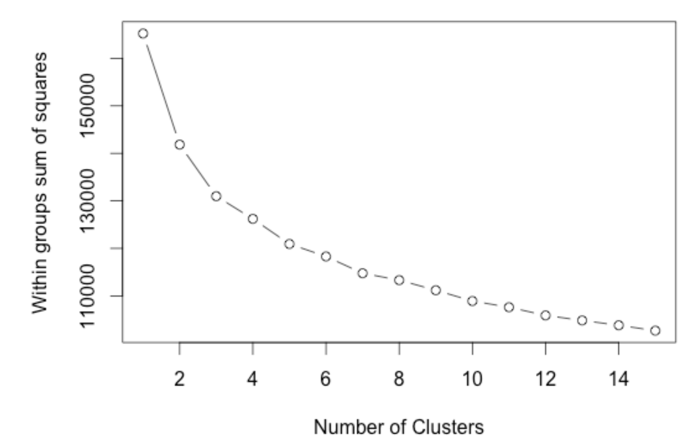
In the typology survey, respondents were asked to associate with one of two opposed statements, and then asked to value that association as either "Strongly" or "Not Strongly." They were given the options to respond neither, both equally, or with nonresponse. In order to define the clusters on only politically salient individuals, the study also used measures of political activism based on voter registration, voter frequency, and whether they follow government and public affairs to define a separate cluster of bystanders.

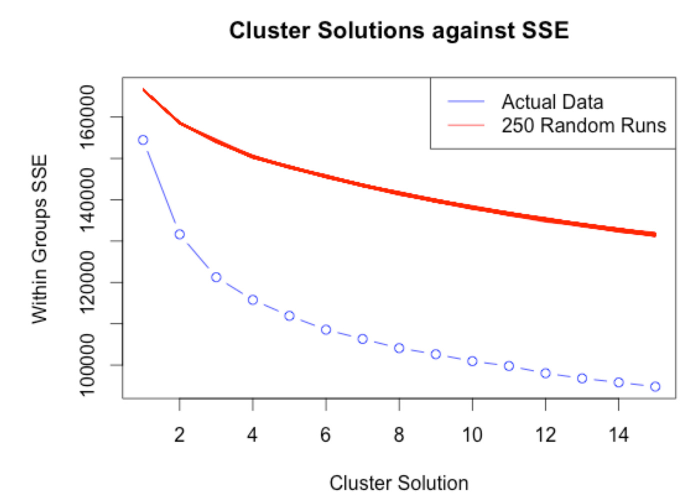
Nineteen survey responses were recoded on an ordinal scale from -2 to 2, and each used as a level in a k-means cluster analysis. "K-means analysis is a divisive, non-hierarchical method of defining clusters. This is an iterative process, which means that at each step the membership of each individual in a cluster is reevaluated based on the current centers of each existing cluster." (Peebles, 2011) This process can be repeated until a desired number of clusters is reached, or a level of clusters can be chosen such that the addition of an additional cluster does not significantly contribute to a decrease in error. This process is non-hierarchical due to the ability of each individual point to be reassigned to a different cluster at each stage in the analysis. "Clusters are defined based on Euclidean distances so as to reduce the variability of individuals within a cluster, while maximizing the variability between clusters. (Peebles, 2011) This method of analysis will provide cluster assignments to each data in addition to defining average values that can be used to define each cluster's ideological markings

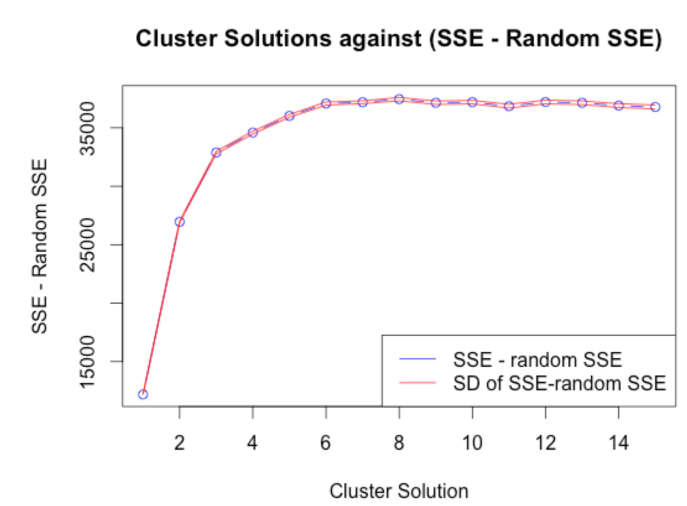
In order to find the best method for the cluster analysis, I used four data transformations. The first data transformation used all nineteen survey response questions as dimensions for the clusters and coded nonresponse as NA, causing for some data loss. The second transformation, averaged the survey responses in the same category and used the 9 categories as the dimensions of the cluster analysis, decreasing the amount of data loss. The third transformation used the nineteen survey responses, but coded the nonresponse as 0. The fourth and final transformation again averaged the survey responses again, but while using the 0-coded nonresponse.

Choosing an Appropriate Cluster Number

One method for identifying an appropriate number of clusters is to compare the sum of squared error (SSE) for a number of different cluster solutions. An increase in the number of clusters corresponds with decreasess in the SSE. Therefore, without making assumptions about the data, we can attempt to see the effect of additional cluster solutions, and choose an appropriate number of clusters by looking for a point at which the decrease in SSE slows dramatically. Unfortunately, from this first plot, it is not fully clear at which particular point we achieve an optimal solution.



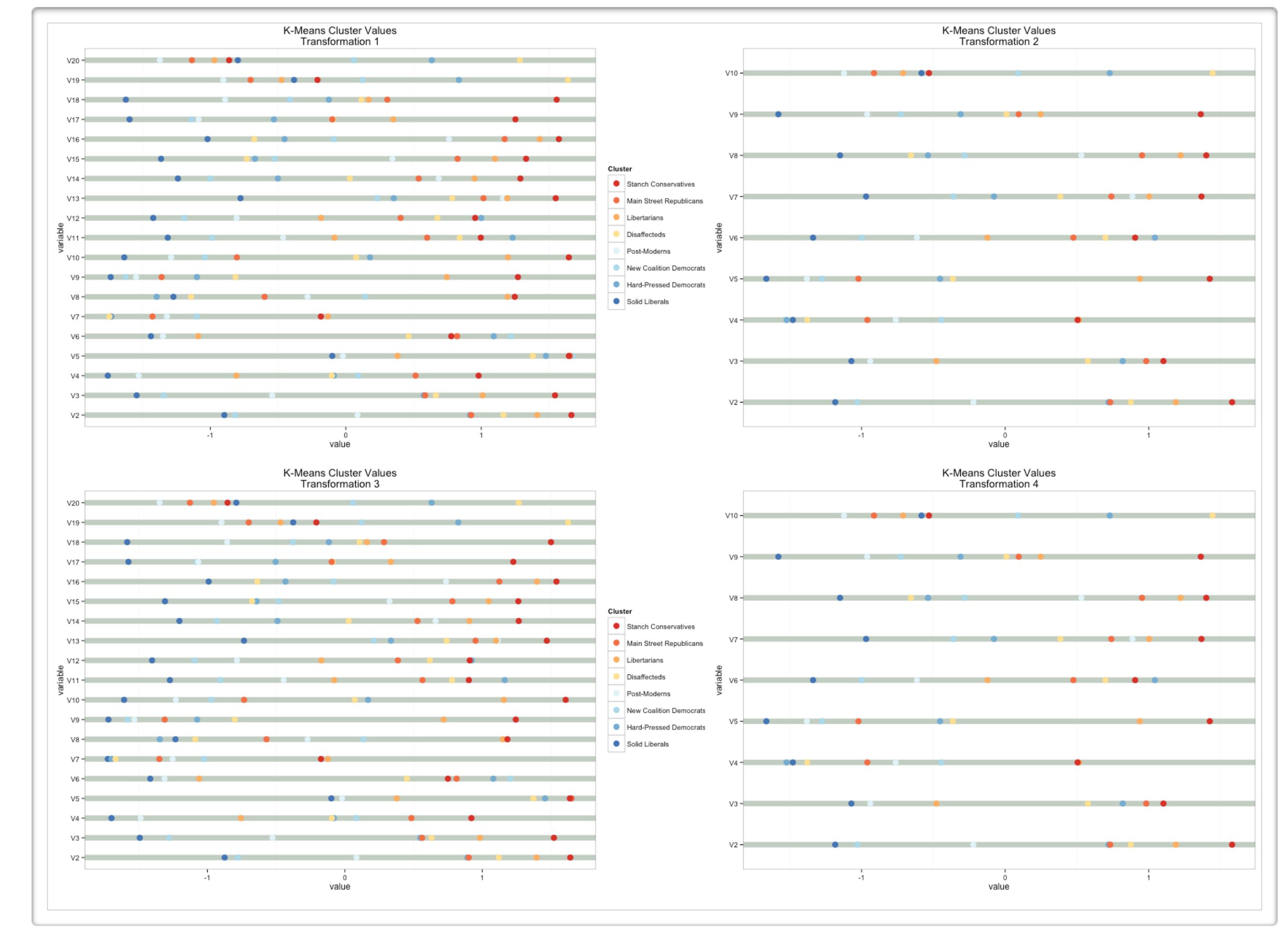




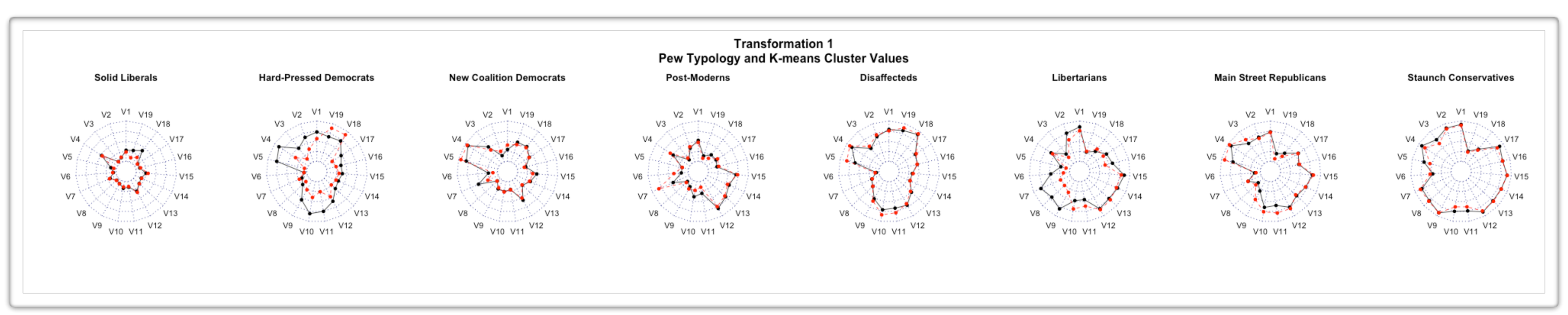
The second plot visualizes the original plot along with 250 versions of the original data input, randomized by column. The intuition of this plot being that if there are strong clusters present, the SSE of the actual data should decrease more quickly than the random data as the cluster levels go up, which is the case.

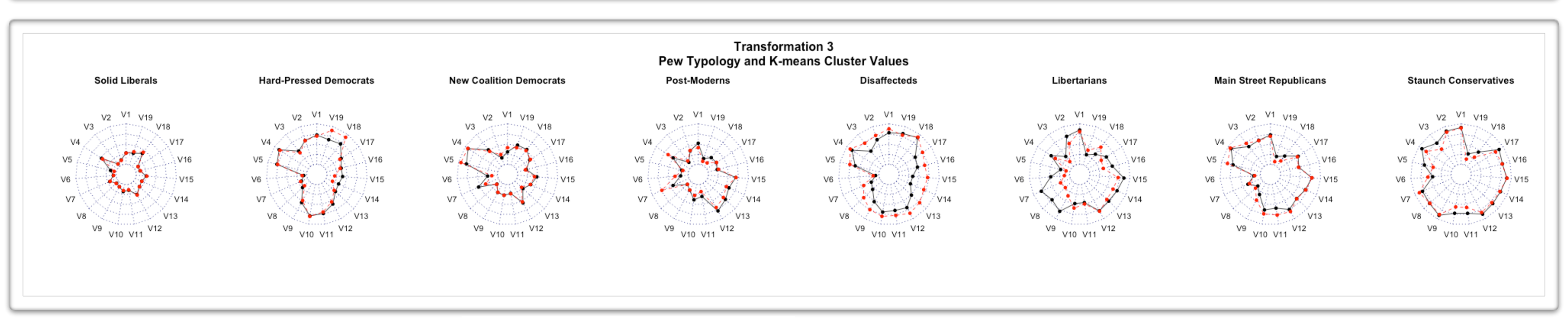
Finally, an appropriate cluster solution could also be defined as one where the actual SSE differs the most from the mean of the random SSE. The final plot shows that the difference level stabilizes around around a maximum value at the 6 cluster solution.

Discussion / Conclusions



With the completion of the cluster analysis, I reoriented the the sign of the cluster means such that liberal means lie solely on the negative end of the scale. With a standard visualization (as above), the results themselves do not look so surprising. The values across the clustering parameters seem to slowly transition from left to right, with values of contiguous groups bleeding into one another's space. However, the conventional left-right scale is somewhat deceiving. Because the variables are all on the same standard scale, radar plots can visualize each of the groups as a whole. My first observation is while looking at the radar plots we can see that as you move from left to right, the mean responses of groups do not gradually swell outward as suggested by the conventional wisdom, but instead push out and in depending on the salience of particular issues to each group. Additionally, when comparing the pew values to those of the analysis, we see that the groups in the middle are much less well defined. While allowing for some variation in the implementation of the clustering algorithm, we see consistent levels of disagreement between the methods.

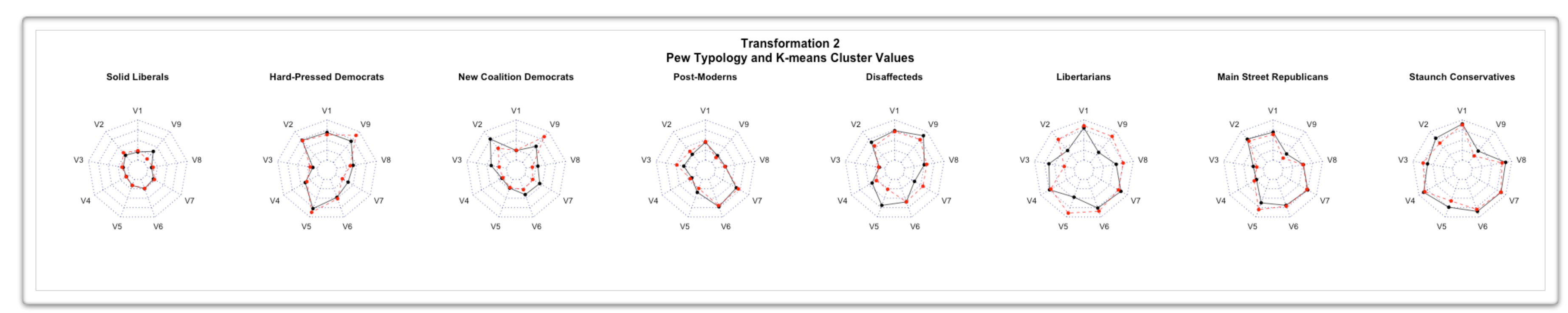


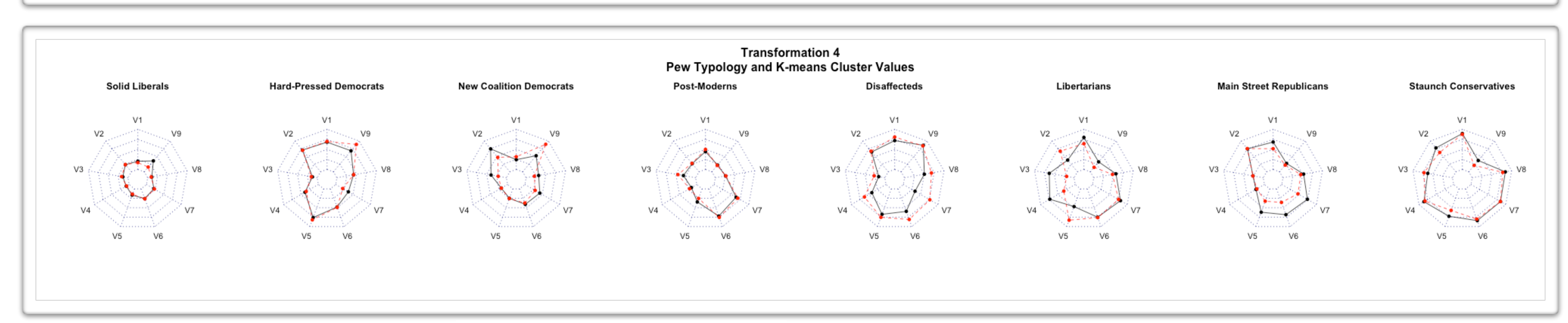


K-means Analysis Pew Typology

Summary

The cluster analysis does produce some independent subgroups, but -- surprisingly -- they are (a) not quite as distinct (or coherent) as analysts may have previously suggested and (b) the transition from more liberal to more conservative groupings is not quite as fluid as the customary scale would imply. While the individuals without highly polarized opinions are more sensitive to identify, their groupings in both methods show their beliefs diverging from the standard left-right heuristic. Radar plots with standardized values make these results apparent. Overall, this research provides evidence as to the limitations of the liberal-conservative scale as a baseline for evaluating opinion space and the benefits of identifying silos of opinion through ideological realities.





K-means Analysis Pew Typology