Exploratory Data Analysis (STAT - 670)
Mini Project - 2

April / 08 / 2022

Team Members:
Rajesh Sharma - rajeshar
Ashish Patidar -<username>

1. INTRODUCTION

1.1. Topic of Analysis

Many people studying politics in the United States believe that the country is becoming more ideologically polarized: either liberals are becoming more liberal, conservatives are becoming more conservative, or both. In either case, they believe that the middle of the political spectrum is being hollowed out.

1.2. Problem Statement

We need to analyze the political ideologies of the people studying politics in the United States who believe that the country is becoming more ideologically polarized.

1.3. Datasets description

This is a partially cleaned dataset[1]of roll call votes taken by the Senate and to know more about the dataset, tools are available at the github repository[2].

After downloading the data there is one subdirectory for each year between 1989 and 2014. Each folder contains two csvs: one giving each senator's vote on each of the bills for which there was a roll call vote in that year, and one describing the senators (name, id, state, party).

2. DEFINITIONS & PRELIMINARIES

2.1. PCA (Principal Component Analysis)

Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data. It can also be viewed as a variance maximization technique.

It tries to simplify the complexity of high dimensional data while keeping the same pattern. It is generally used to process data with high collinearity between columns. It can be used for dimensionality reduction, removing multicollinearity, and visualizing high dimensional data by reducing the dimension.

This technique does come with some drawbacks like low interpretability of principal components and there is always some loss of information while reducing data dimensionality.

2.2. MDS (Multi Dimensionality Scaling)

The purpose of multidimensional scaling (MDS) is to provide a visual representation of the pattern of similarities among a set of groups. Instead of measurements on variables as done in PCA we can try to make a map of the samples in a low-dimensional space (probably 2-dimensional space) so that the distances in that map match the distances between the samples as closely as possible.

The distances can be what was measured initially, or the distance could be constructed by the analyst from other variables that were measured directly.

Main drawback of the method is that a complex formula is required to convert raw data into a multidimensionality scale which can be a hectic and time consuming task.

2.3. Smoothing

It is a process to create an approximation function that attempts to capture important patterns in data while leaving out outliers/noise.

If we have a lot of data/noise, the smoother allows us to see what we can't in a scatterplot of the raw data. If we want to compare multiple sets of points, the smoother simplifies the description and allows us to make the comparison between the "main effects" in the data without our eye being distracted by the noise[3].

Smoothing might emphasize biasness by ignoring outliers/moise that may be meaningful.

2.4. Scree plot

A scree plot is a line plot of eigenvalues of principal components of a data in an multivariate statistics analysis. It is used to determine the number of principal components to keep in a principal component analysis to get the maximum variance.

Scree plots can be unreliable sometimes especially when they have multiple elbows and hence making it difficult to select the correct number of principal components to retain.

3. POLARIZATION INVESTIGATION

Polarization especially in politics is the extent to which opinions on an critical issue are opposed or can say it is a procedure of steps taken by whose result opposition increases over time. General discussions of polarization in the field of political science considers polarization in the context of political parties, senators or democratic systems of government.

Before doing any further analysis we did the following things in our code to enhance reusability of the code:

1. Read data of members and votes for each year using dataframe.
2. Create few functions to avoid repeated lines of code and enhance reusability of the code.
    a. recode_votes function : Function build to change votes ("yea","nay",etc.) to numerical format.
    b. mds_analysis function : Function coded to calculate MDS distance on first principle axis for each senator on basis of voting patterns
    c. mds_var : Function to calculate explained variance for two major parties for each respective year
    d. mds_mean : Function to generate average MDS distance for parties
3. Join members' data of every year into one vector to get the value of the senator who served for all(or maximum) years.

3.1. Polarization in two years (1990 and 2010)

Voting records of 1990 and 2010 are compared to see the changes in voting patterns for senators among different parties.

To compare these voting patterns we applied MDS on euclidean distances of senators calculated using dist() function of R.Scree plots of eigenvalues of both the years are generated to consider the number of components important for analysis.

Below are the Scree plots for year 1990 and 2010:

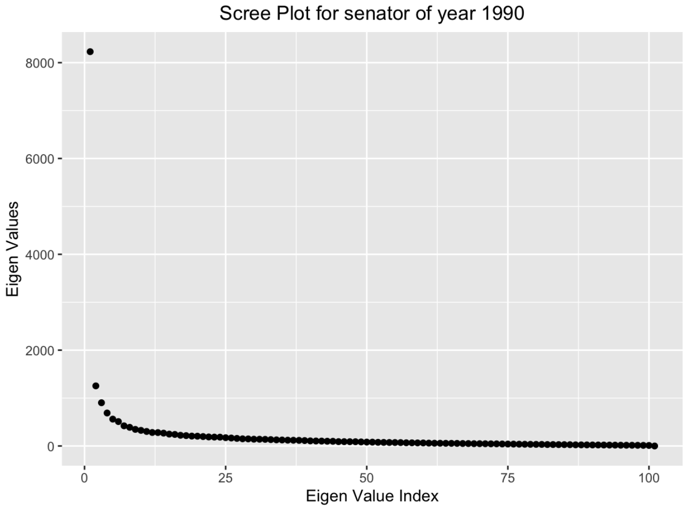


Figure: 1

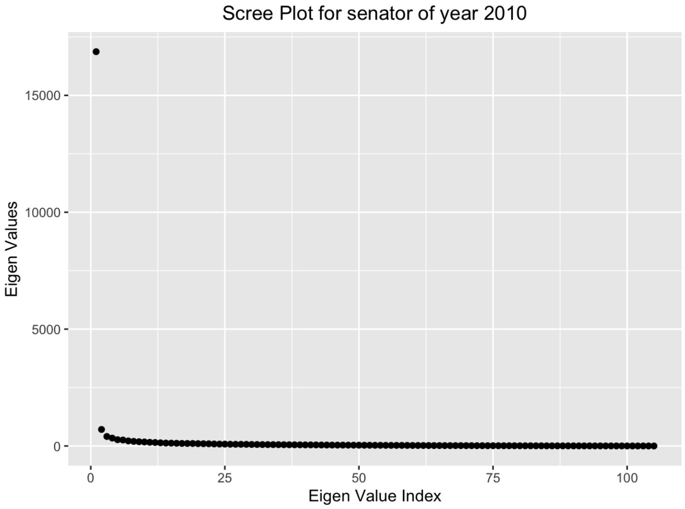


Figure: 2

From both the Scree plots it can be said that the first principal axis is most important and we can clearly see it as an outlier in both the plots. Second principal axis could be considered as important but not as important as the first one.

After considering insights from the Scree Plot for both the years, the distance of senators are plotted with respect to the first and second principal axis. Senators are color separated based on parties they belong to.

Below are the point plots for distances of senators for year 1990 and 2010:

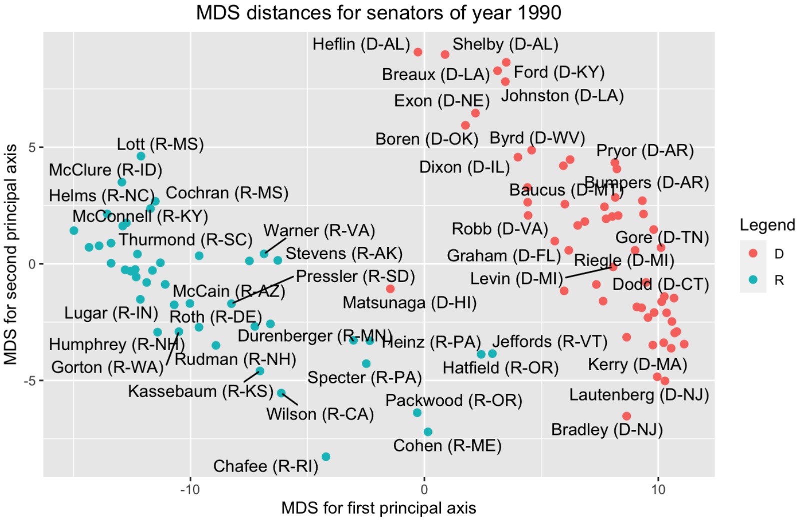


Figure: 3

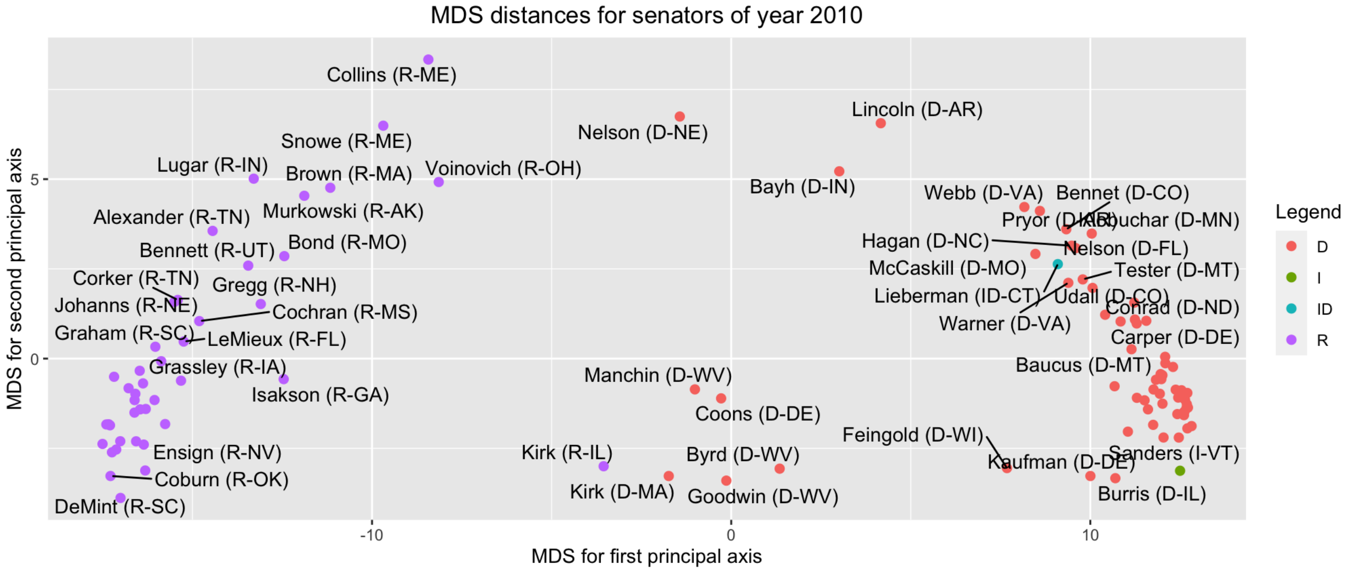


Figure: 4

Note- as parties I and ID are so less in number that it is very hard to conclude anything from their visualization and also they came into existence in later years. So are ignoring them for our analysis.

Major questions answered

Q. Does it look like they fall on a one-dimensional liberal/conservative axis?

Ans. For both the years we can say that all the senators fall on liberal/conservative axis but there are some outliers which are mentioned in second question.

Q. Are there outliers?

Ans. For the year 1990, there is only one outlier Matsunga (D-HI), Jeffords(R-VT) and Hatfield(R-OR) cannot be exactly considered as outliers because they are on the tail of the distribution of senators of the republican party.

For year 2010, there were 6 outliers out of which 5 were from democratic party whose names are as follows:

- Manchin(D-WV)
- Coons(D-DE)
- Byrd(D-WV)
- Goodwin(D-WV)
- Kirk(D-MA)

Above list gives us an interesting insight about senators of West Virginia being outliers and having different politics than the party.

Kirk(R-IL) from republican is the only outlier from the party in 2010.

Q. Are the patterns similar in the two years that you chose?

Ans. Patterns are completely different for the selected years; for earlier years(1990) there were fewer outliers but for later years(2010) the number of outliers had increased. There were two new parties in later years whose politics aligned with democratic party. Even the distance value of senators has increased on both sides in later years, hinting to increase polarization with time**.**

Q. What are the differences in the PCA/MDS plots in the two years that you chose?
Ans. There is a difference between the lines of fit for major parties(democrat and republican) in both the years. In 1990 line of fit for both parties had almost the same slope, but in 2010 they almost had the opposite slope. If comparing parties only, there is a difference between the slope for the republic party and the democratic trend is almost similar in both the years.

3.2. Polarization over time

For better analysis of polarization, explained variance of both parties is calculated on first principal axis using the eigen values generated from cmdscale() method of R.
The below formula is used to calculate explained variance:

Explained Variance on first principal axis = first eigenvalue/sum of all eigenvalues

We have plotted the variance for both parties from 1989-2014 to answer major questions like: **‘** Do you think that polarization has increased over time? ‘, **‘** Is something more complicated Happening? ‘

Below are the plots for variance of both the parties & of difference between variances of both parties over the years:

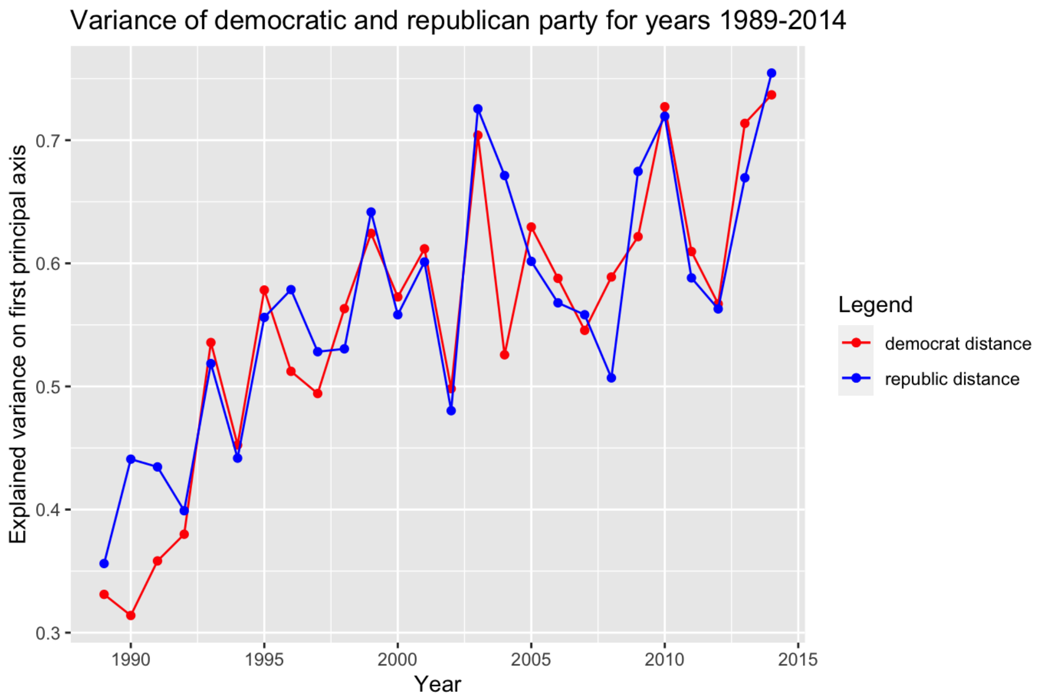


Figure: 5

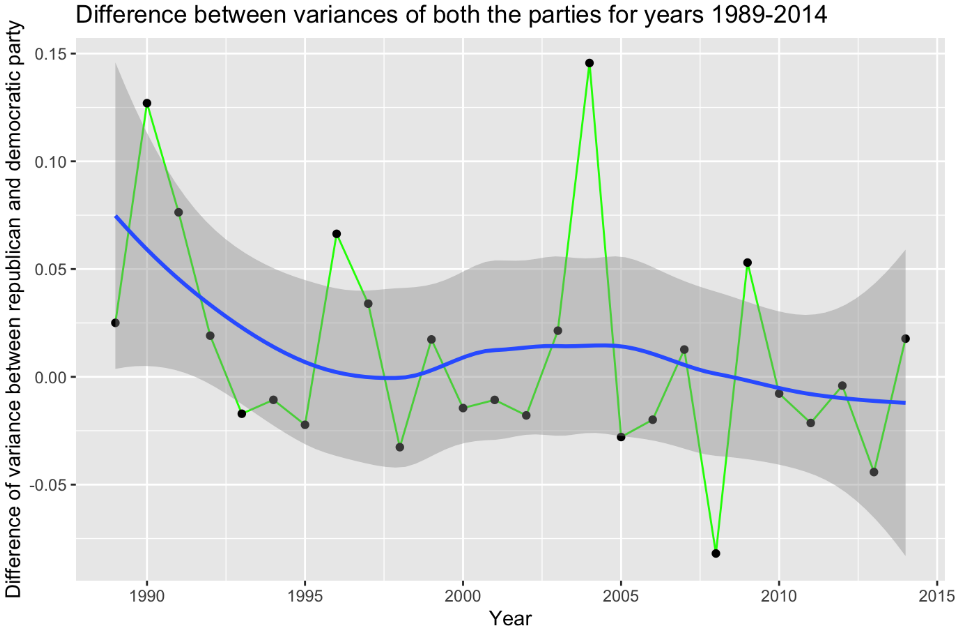


Figure: 6

By looking at the graph(Figure: 6), we can say that variance for both parties has increased over time. It is going farther from its origin, which means that both parties' polarization has increased over time. One more interesting insight from the variance graph is that increase in polarization is following the same almost similar pattern for both parties except for some years.

Since the increase in variance for parties is not linear, it could be said that there is something much more complicated because between year 2002 & 2003, variance increased a lot for both the parties but for 2008 & 2009, the increase in variance for republican much more than compared to increase in the variance of democratic. There are many small details regarding the increase in polarization for each year between 1989-and 2014, making it more complicated.

4. Absolute changes in Ideological position over time

The analyses in the previous parts told us about the relative ideological positions and the amount of separation between the parties, but not about absolute changes from year to year. Now we want to know about absolute differences in the ideological positions of the parties so that we can assign blame to one or both parties for polarization, but it’s a tricky examination to get at because both the senators voting on the bills and the bills being voted on change from year to year.

As we know from the previous part that polarization between both the parties had increased with time, we will now compare the MDS distance on the first principal axis of the longest-serving senator between the years 1989-2014 with the average MDS distance of both the parties and see if the politics of senator stayed same over the years.

Mr. Wyden of Democratic party from Orlando had served for 18 years from 1996-2014 and is the longest serving senator during the given period.

Below(Figure:7) is the graph of MDS distances for senator Wyden (D-OR) and both parties:

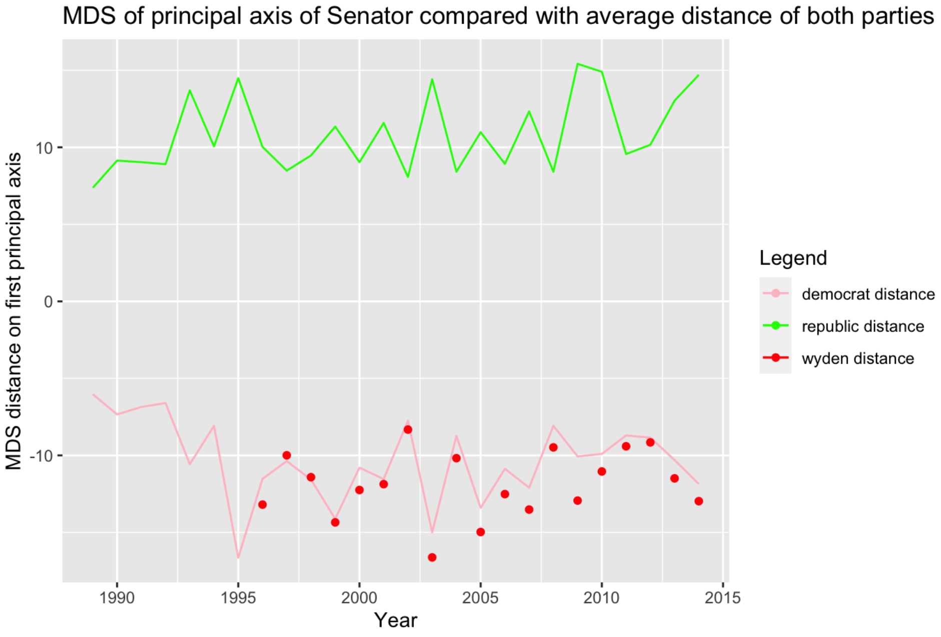


Figure: 7

The below two plots are the difference between the distances of the senator with his party and the other major party. These plots are constructed so that we can see the relative change in the politics of senators of both parties.

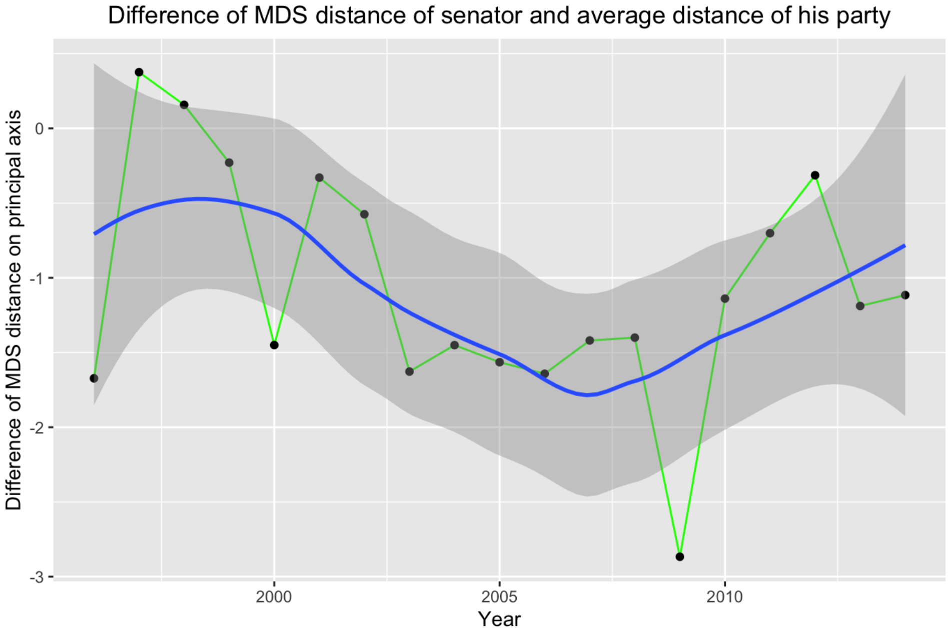
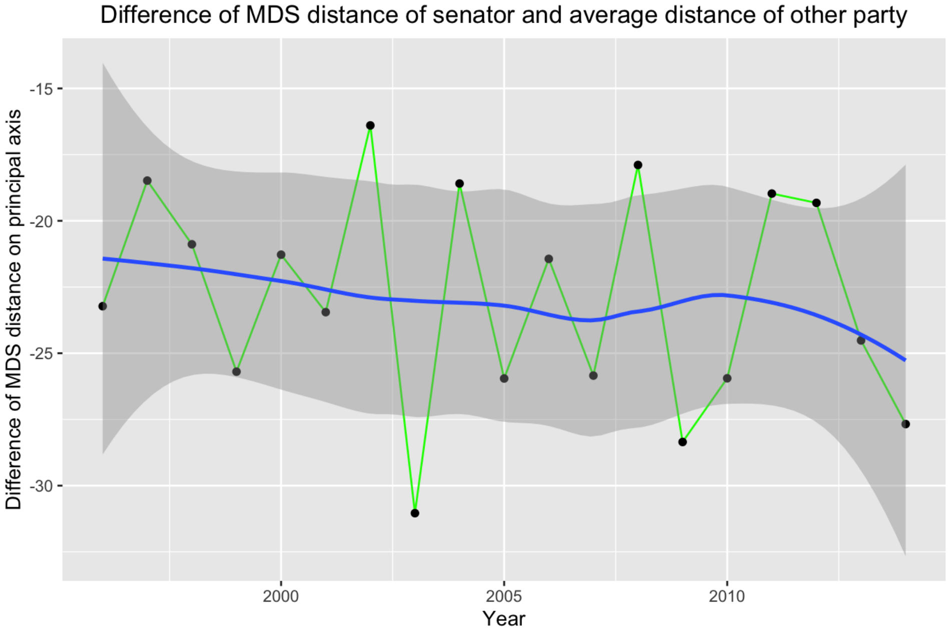


Figure: 8



By looking at the above graphs(FIgure: 7-9) following questions can be answered:
Ans. The senator's position with respect to both the parties was almost the same, there is no linear relation, but we have used loess smoother to see an approximate relation with regard to both the parties.

Q. Describe the implications of your plot either in terms of which party is more to blame for polarization or relating to the question of whether that senator's politics has changed over time?
Ans. The politics of the senator we have selected has stayed the same for the years, and his politics were almost similar to that of his party, but it can be seen that Mr. Wyden's distance is more than the average distance of the democratic party for the most of his tenure. Therefore, we can partly hold him responsible for polarization. Out of the two parties, it is very difficult to blame one of them because the above graphs are not conclusive enough to put the blame on any particular party.

Conclusion

Although we tried our best to create a perfect analysis report with the techniques we know, there can be some limitations to our analysis because we can see that there is much noise in MDS plots and a generic trend cannot be explained. Therefore, we are taking an approximation using loess smoother. In the first section of the report, we talked about outliers for the years 1990 and 2010, but we do not have any explanation for those senators being outliers. In the second section, where we calculated variance, we do not see a linear or exponential growth in the variance of any of the parties patterns. There is much more distortion, so we do not have an explanation for the pattern being so zigzag. Based on these observations, it is pretty clear that there can be various other explanations for the phenomena we have suggested.

References

1. https://jfukuyama.github.io/teaching/stat670/assignments/congress.zip
2. https://github.com/unitedstates/congress
3. https://en.wikipedia.org/wiki/Political_polarization
4. https://en.wikipedia.org/wiki/Scree_plot