Data Mining Report

by Fabian Walocha

17 March 2019

Topics of discussion

This analysis was performed and documented as a project in the *Data Mining* as part of the *Machine Learning and Data Mining* curriculum at *Université Jean Monnet*.

The main question that is supposed to be answered in this work is whether or not the political opinion in Germany experienced a tonal shift after the German reunification in the year 1990 in the short and in the long run. Moreover, this analysis choses to focus on West German federal states. While the causal effect and repercussions in Eastern German states are contentious and well researched, the same cannot be said about issues regarding only the Western German part of the country. This question is mainly to be answered in a data-driven fashion, i.e. by analysing and transforming at raw data and drawing conclusions from this analysis, rather than second hand accounts and literature. Since the question posed is multifacetted and can be answered in various equally acceptable ways, the following analysis focuses on the following indicators which highlight things about the political climate in Germany.

1. The immediate reaction after the fall of the Berlin Wall in 1989 and the reunification in October of 1990.
2. The general trend of left-wing vs center vs right-wing association across time.
3. Regional differences of political tendencies.
4. Predictiveness of voting behavior based on other factors.

It is believed that by merging information indicating political trends and culture in Germany using a wide variety of methods, the analysis is able to find synthesis on the topic.

About the dataset

The dataset used in this study was created and provided by the research group *Wahlen* (Forschungsgruppe Wahlen (2019)), on behalf of *ZDF*, a publicly funded German broadcasting station. The dataset consists of surveys containing questions about social facts, electoral intent and political affiliation of German citizens. The survey was conducted via the phone and spans across a timeframe of 41 years between 1977 and 2017. The questionaire used was kept relatively the same across the years, every time a question was changed, discarded or adjusted, the change was noted. The dataset consists of 82 variables, among others the age and livelihood of the person, their opinion about the current state of politics and economy in Germany, and their last and prospected electoral vote for federal elections.

Out of these 82 variables, 24 variables were hand-selected for this analysis based on their assumed relevancy for the questions at hand. The full list of variables and a short description for each of them can be found in Appendix A. Furthermore a full description of all variables used in the dataset can be found on the website of the GESIS foundation (Forschungsgruppe Wahlen (2019)).

Data cleaning, preprocessing

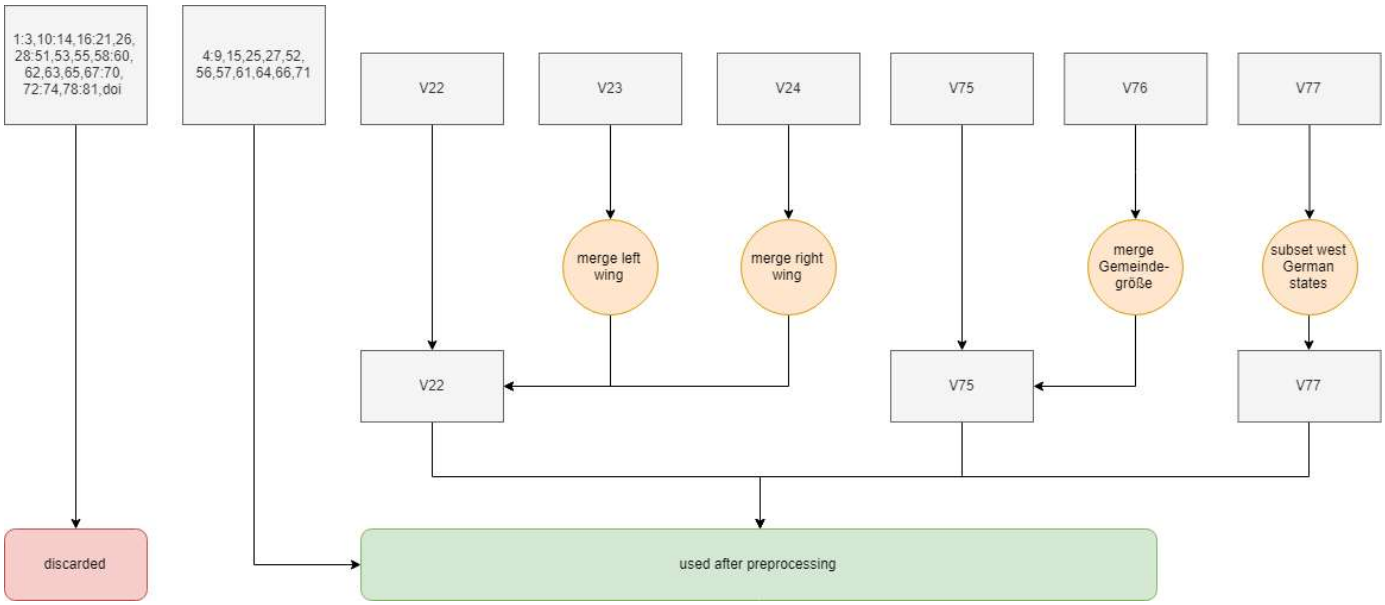
The data was prepared uniformly for all further processing steps in the following way. First, the variables *v22*, *v23*, *v24* containing information about the subjects stance in a left-right-paradigm, their left-wing tendencies and their right-wing tendencies respectively, were merged into one variable. The variables *v23* and *v24* were used from 1989 to 1996 instead of *v22* to ask the subject for their political tendencies. In these years, the

person was first asked whether or not they identify themselves left-wing, right-wing or center, if they answered either left- or right-wing, the surveyor then asked for their belief intensity from a scale of 1 to 5. In comparison, in all other years, the subject was asked how they would place themselves in a left-right spectrum on a scale from 1 (meaning very left-leaning) over 6 (meaning neither left- nor right-leaning) to 11 (meaning very right-leaning). The structure and scaling of the responses thus lend themselves to be merged for further analysis. It is assumed that the rephrasing of the question does not introduce a significant bias into the results.

Secondly, the variables *v76* and *v77* were also merged. The question posed in *v76* was phrased to get the number of people living in the subjects *locality* (Ortsgröße) which the question posed in *v77* which was used from 2011 to 2017 asked about the size of the *municipality* (Gemeindegröße). In coloquial German, both terms are interpreted very similarly. Because of this, it was decided to merge these two variables. In this process, the difference in ranges was taken into account and changed accordingly (e.g. the range 5-10k inhabitans was merged with the range 5-20k inhabitants in *v77*).

Lastly, since this analysis only concernes itself with Western German opinions, we discard all subjects which say to reside in eastern German federal states. Furthermore, the data collected from West-Berlin was ommited for this analysis.

The remaining 21 variables were then used for all further analysis. The following diagram visualizes the process.



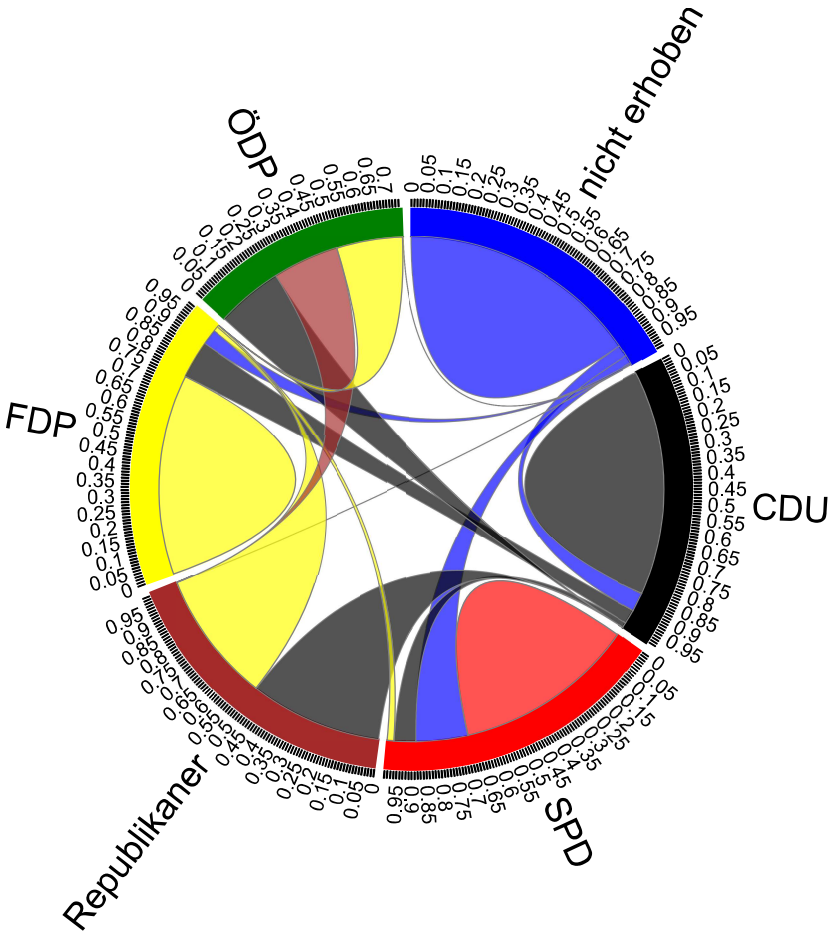
Modeling and Evaluation

The modeling was done in multiple parts to deal with all the above mentioned points respectively.

1. The immediate reaction after the fall of the Berlin Wall in 1989 and the reunification in October of 1990.

In order to find out what the general populace thought about the way that the reunification was dealt with by the government and their overall approval, the elections of 1990 and the preceding federal election in West-Germany of 1987 were looked at. More specifically, the gross number an indicator of how much opinion changed is the gross number of people who decided to switch parties during this election, especially away from the ruling party of the time (a coalition of the christian-conservative party CDU and the liberal party FDP).

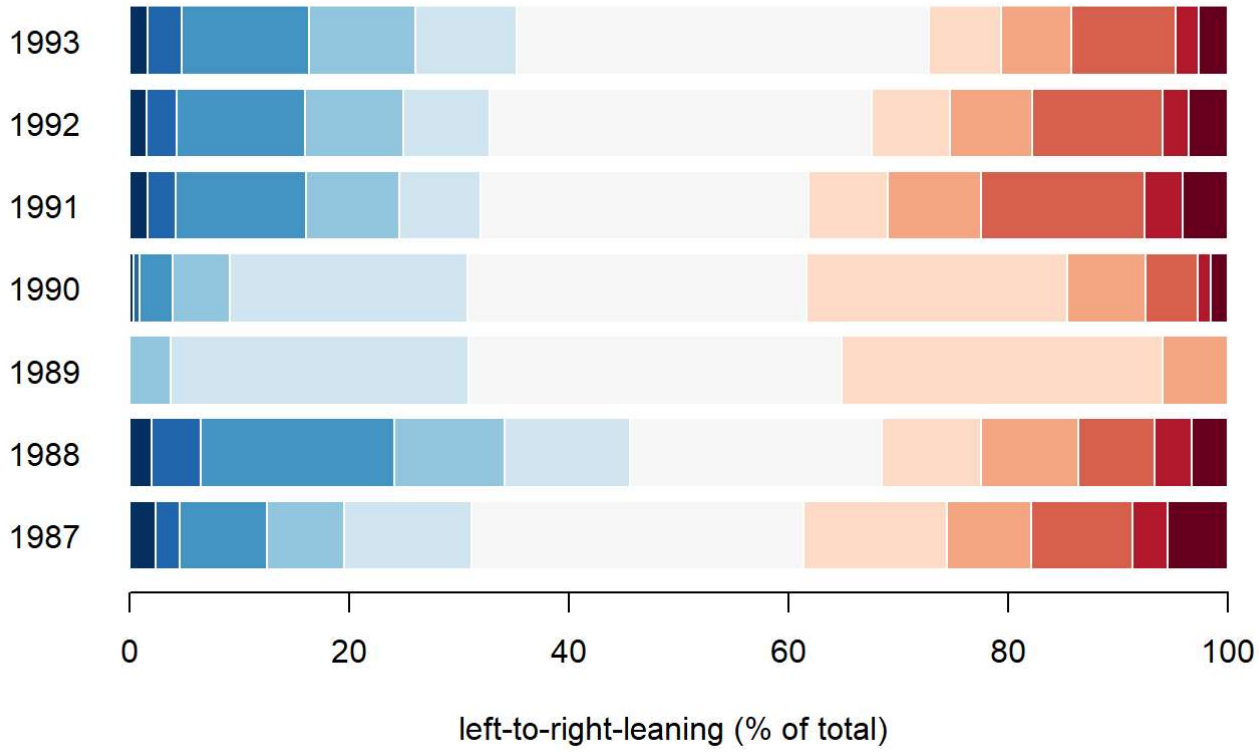
The results of this are shown in the ribbon plot below.



The ribbon plot was created using a custom plot ([Chord and Zoonekynd (2019)), it shows the following features. The colors at the rim of the circle represent the color associated with the outside mentioned party, the inner ribbon show the change in voting behavior as percentage of all voters in the receiving party for the election in 1990. The ribbons are bidirectional but the color refers to the bigger directional influence. For example, a big yellow ribbon coming from *FDP* to *Republicaner* means that about 50% of the voters in the election came from people who voted *FDP* in the last election. As can be seen, a lot of changes between parties happened. The newly found hard-right party *Republikaner* gained their votes mainly from the ruling parties of the time *CDU* and *FDP*. On the other hand, a not-insignificant part of undecided voters (named as *nicht erhoben*) voted for the main social democrat oposition party SPD. These are already indicators that immediate reaction after the vote were not uniformly positive.

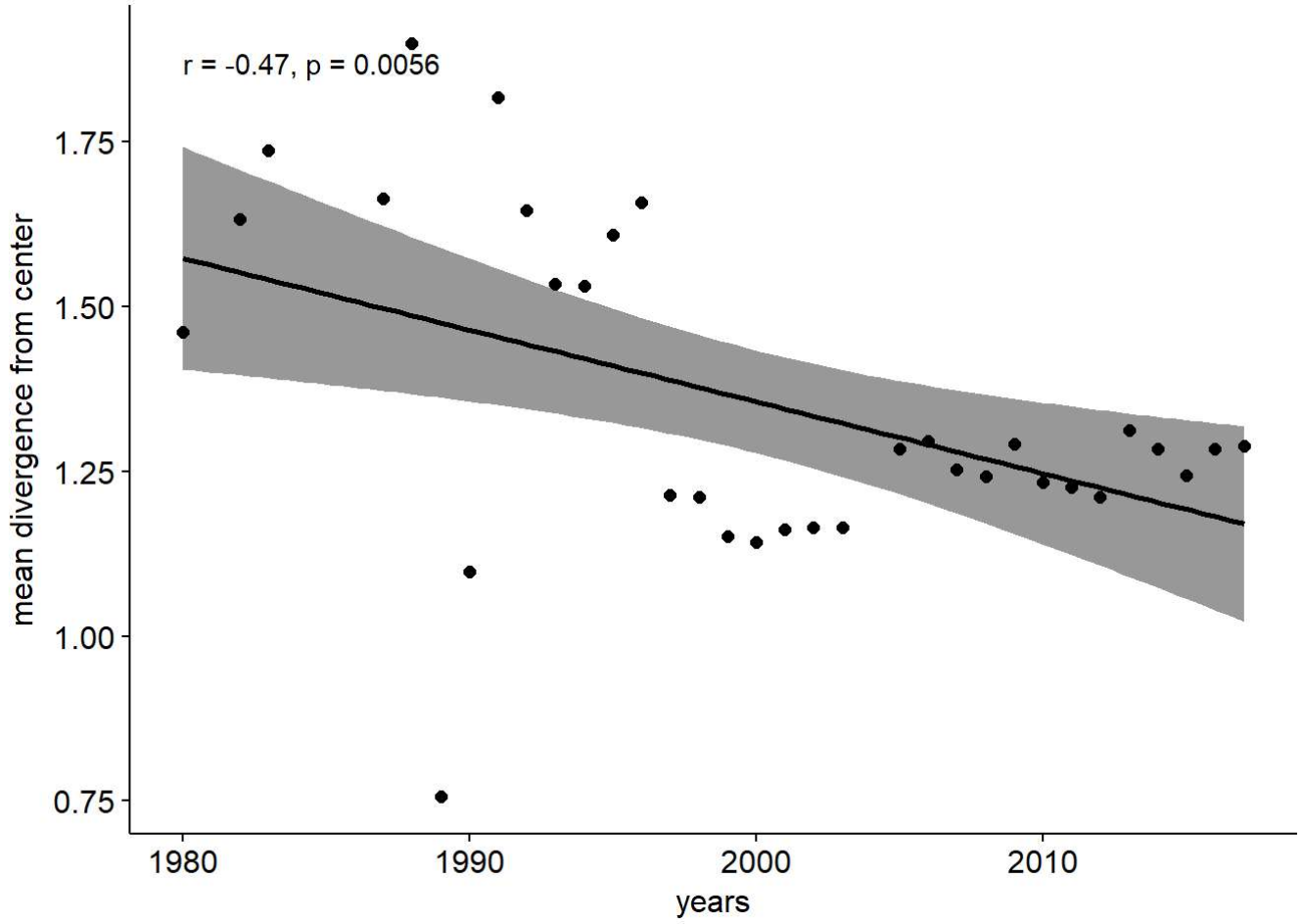
2. The general trend of left-wing vs center vs right-wing association across time.

In order to find out the trend of political affiliation regardless of party association, the self reported stance of political leaning in the “left-right-continuum” was looked at. More specifically, the item of interest was the percentage of people associating with left wing ideologies, right wing ideologies or centrist points of view. This set of values was calculated for each year where data was available and compared.



The graph shows a stacked percentage barplot where the colors represent the mean political tendencies across the span of 7 years. While the amount of center-associated positions seems to be expanding across time, the left-wing positions seems to not only increase but also get more intense. A similar increase in intensity can be seen for right-wing opinions although the general percentage of people holding right wing positions seems to be receding towards 1993.

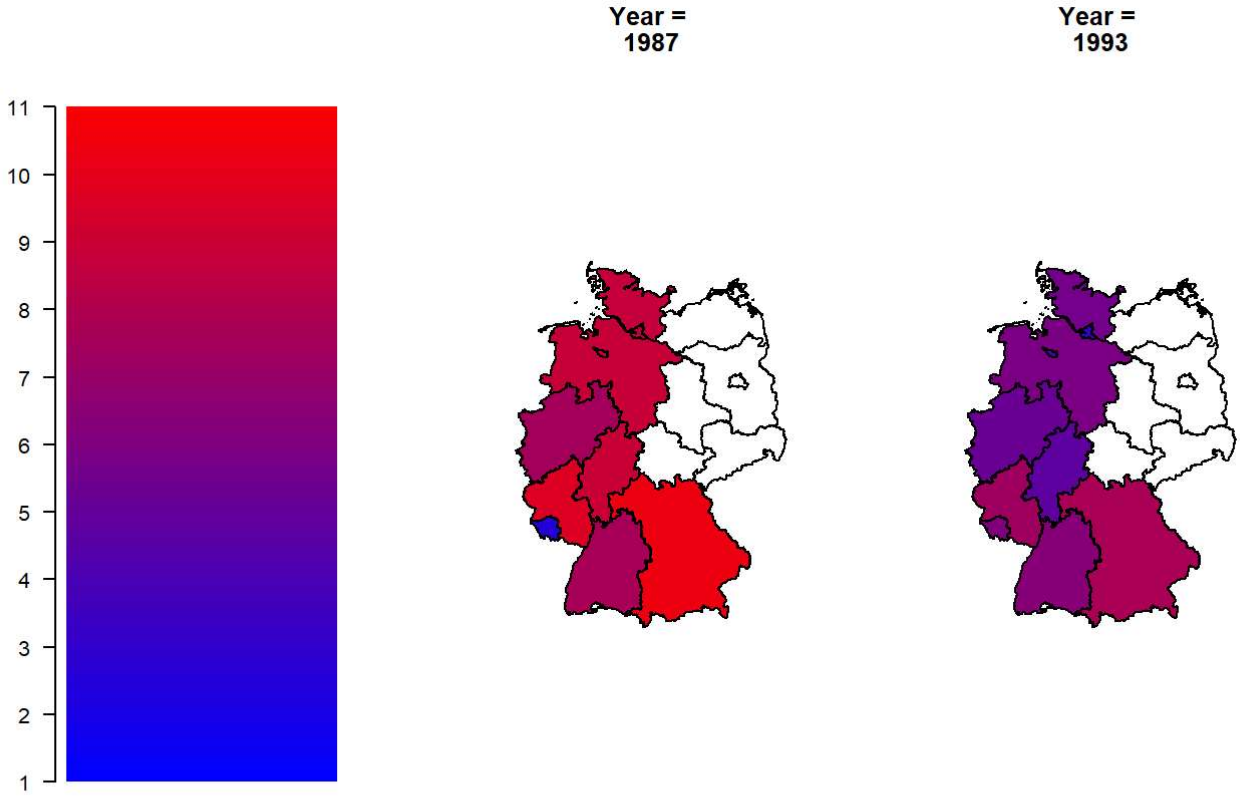
In order to analyse this pattern over a longer timeframe, an additional analysis was done to observe how the deviation from the centrist point of view in public opinion over time. In order to quantify this change, Spearman's correlation coefficient (Spearman (1904)) was estimated and plotted below.



The plot shows mean deviation regardless of left- or right-wing association over the span of 41 years. The plot shows that while initially extremist positions persisted before and after unification, the trend is negative going further towards centrist positions continuing onwards from 1997. The correlation coefficient of *-0.47* indicates that their is a medium negative correlation between extremists position and the timeframe.

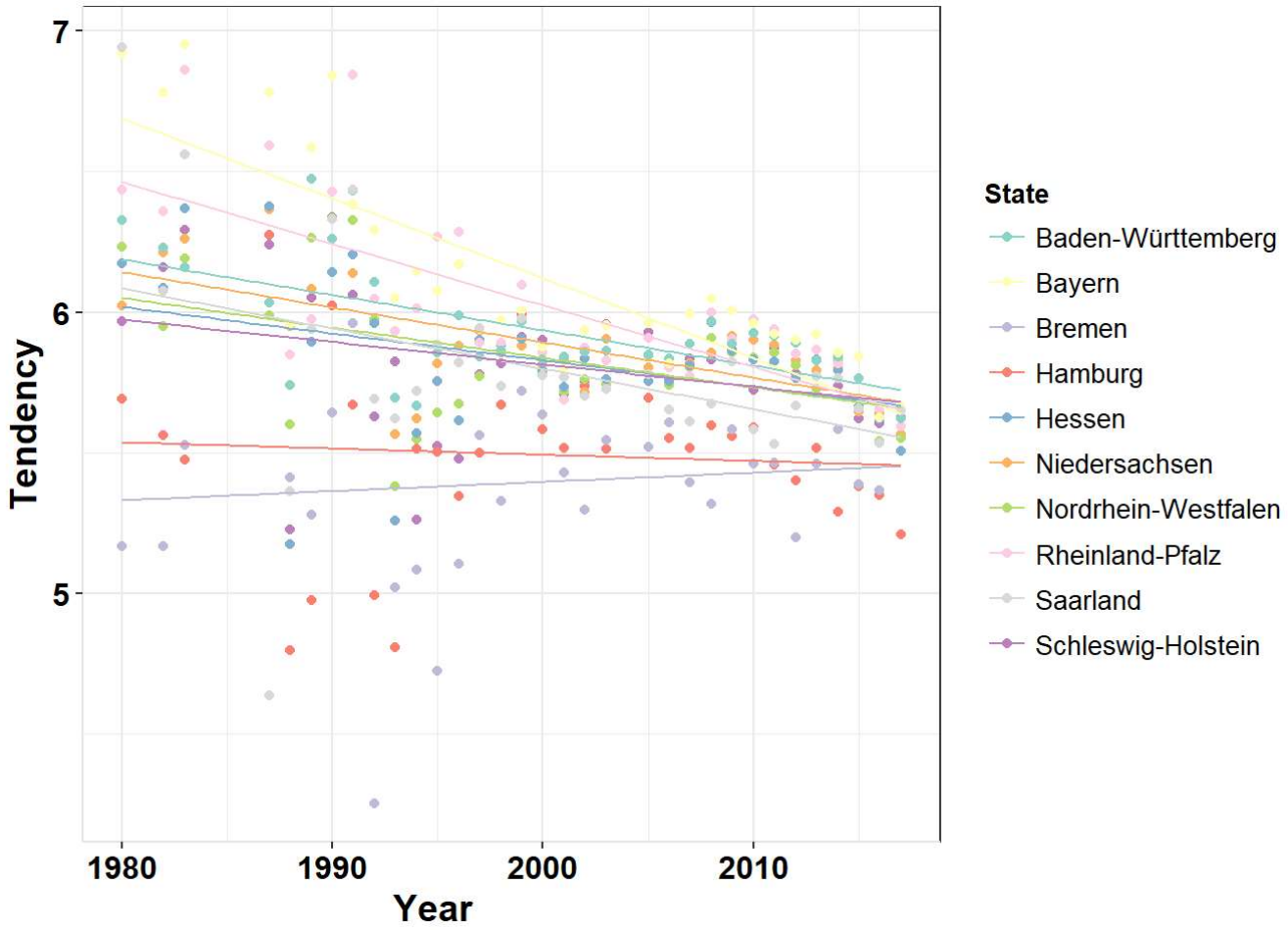
3. Regional differences of political tendencies.

Each Western German federal state was looked at separately and related with eachother in order to account for regional differences. The idea was to find out whether the geographic proximity to their Eastern German counterparts would influence the way public opinion is formed. The following shows the regional difference in voting behavior for the years 1987 and 1993, three years before and after German reunification. The map data was taken from GADM ((2019a)).



The scale on the left shows the left-right barometer where 1 represents a hard-left stance, 6 represents a centrist position and 11 represents a hard-right stance. The Red-Blue shift clearly shows that almost all states with the exception of Saarland turned more left-leaning between these two years.

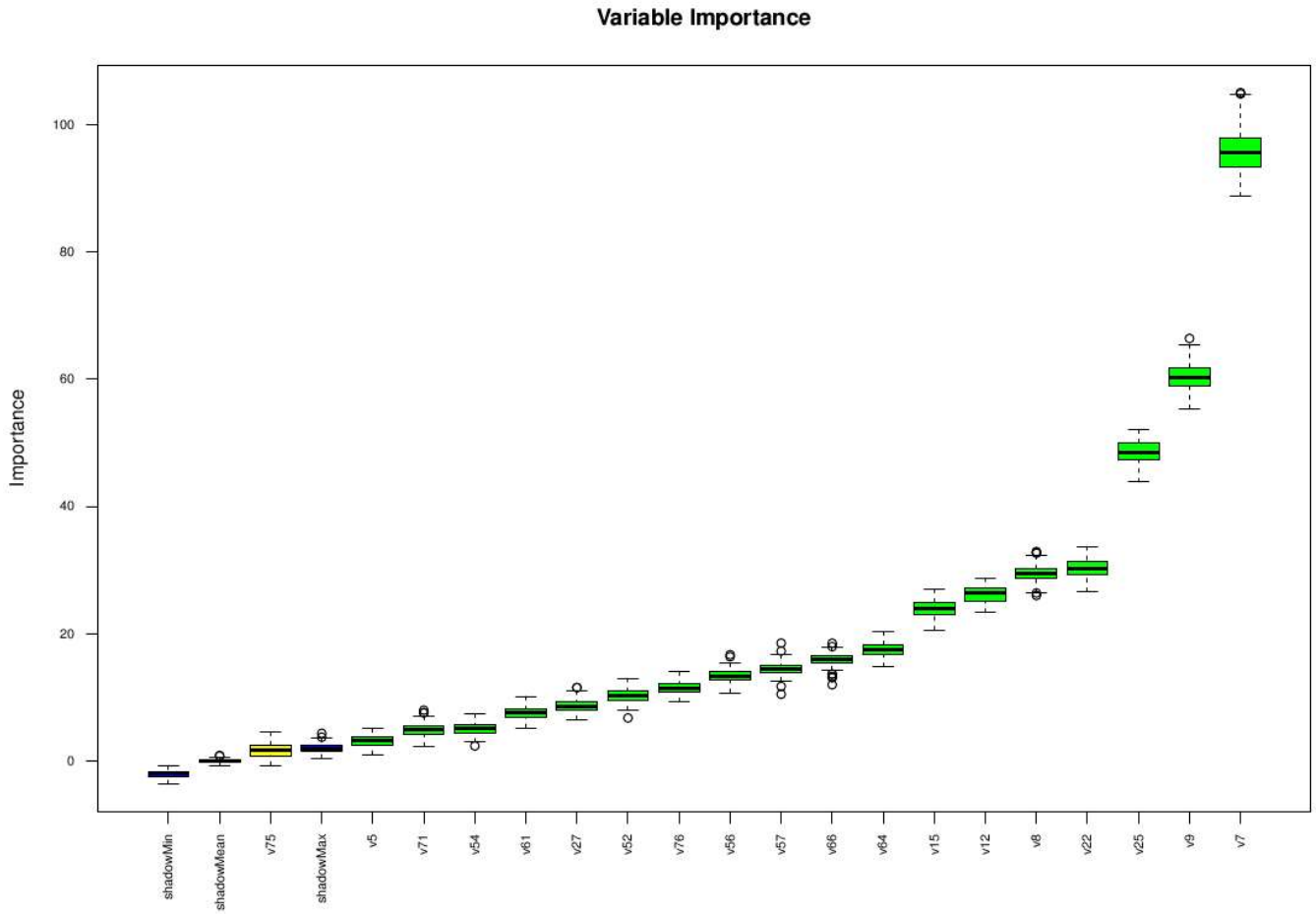
In order to correclty quantify this change over time, a linear regression line was fitted using the *easyGgplot2* package for all federal states separately, the results can be seen in the plot below.



The plot shows political tendencies again, where 6 represents a centrist position,values above represent a right-leaning position and values below represent a left-leaning position with the same scaling as in the plot before. As was already hinted at by the last graph, there is a clear trend for all federal states to move towards a left-leaning stance. What also can be seen is that the political affiliation becomes more homogeneous over time across all states.

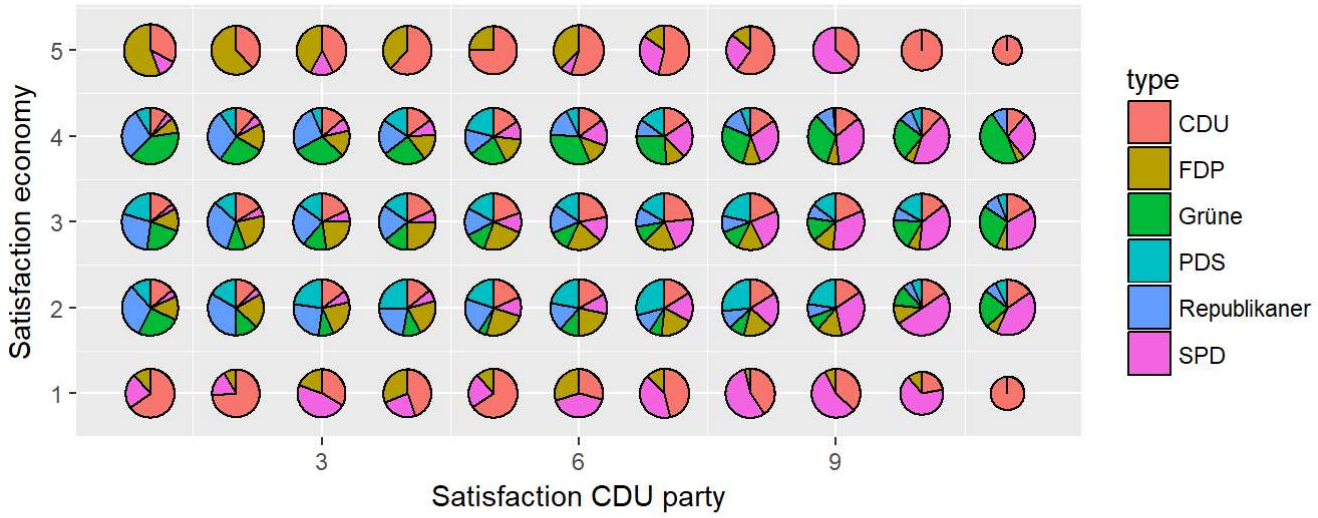
4. Predictiveness of voting behavior based on other factors.

In order to find which factor contribute towards the voting for a given party, a feature selection using random forests was done. The method used for this was Boruta feature selection (Kursa, Rudnicki, and others (2010)) using the *Boruta* library ((2019b)). The Boruta algorithm tries to estimate the importance of a given variable in estimating a predictor by iteratively generating random forests using a randomized subset of variables with replacement. This way, the contribution of a given variable can be estimated given enough permutation, and uninformative variables can be discarded during the process. The results can be found in the plot below.



The plot shows all remaining variables sorted by importance in estimating *v6*, i.e. the prognosis of who a subject is voting for in the next election. Unsurprisingly, the best indicator seems to be *v7*, i.e. the party a subject has voted for in the last election. More interesting are the next two variables. Variable *v9* contains the answers to the question of a subjects general opinion of the *CDU* party, while *v25* contains the answers to the question what a subject thinks of the current state of the German economy.

In order to display their relative importance in predicting *v6*, an analysis was made of how each possible combination of answers influences the relative likelihood to vote for a certain party. The results can be seen in the plot below.



The plot was made using the *easyGgplot2* package. Each point in the scatterplot contains a piechart, which compares the percentage of voters who held specific views and voted for a certain party over all votes for a certain party. The parties that were compared reflect a big spectrum of the political landscape in Germany. *SPD* is a social democratic party, *Grüne* is an ecologic party, *CDU* and *Republicaner* are conservative parties, *FDP* is a economically liberal party and *PDS* is a party with a socialist agenda. Both variables go from the lowest meaning very low satisfaction to the highest value, meaning totally satisfied. This means that a higher percentage of voters vote comparatively for *CDU* when they have a high satisfaction with the party. This point of view is both seen when the economy is judged very well or very poorly, which might be interpreted as the *CDU* being seen as a bringer of good times for these people. On the contrary side, if you are very unsatisfied with the *CDU* and hold the economy in high regards, a great percentage of people tends to vote for the liberal party *FDP* which is a frequent coalition partner of the *CDU* with similar economic policies.

Conclusion

The notion that political opinion after the fall of the Berlin wall has become more extreme is disproven insofar as political extremist tendencies have decreased over time, especially in the last decade. However, as the analysis shows, there was still a strong sentiment of dissatisfaction during and directly after the events of 1989 and 1990. The plots show that not only did West Germany recover from the aftereffects of the reunion, the federal states are actually homogenising in their political opinions.

Deployment

All code used in this analysis was made publicly available on https://github.com/FabianWalocha/ElectionPolling (https://github.com/FabianWalocha/ElectionPolling). Furthermore, a link to all datasets has additionally been provided in the *readme.md*, all relevant citations are found in the bibtex file.

References

[Chord, and Vincent Zoonekynd. 2019. “Chord Diagram in R.” *Stack Overflow*. https://stackoverflow.com/questions/14599150/chord-diagram-in-r?rq=1 (https://stackoverflow.com/questions/14599150/chord-diagram-in-r?rq=1).

Forschungsgruppe Wahlen, Mannheim. 2019. “Partial Cumulation of Politbarometers 1977-2017.” Data file. GESIS Data Archive, Cologne. http://dx.doi.org/10.4232/1.13243 (http://dx.doi.org/10.4232/1.13243).

Kursa, Miron B, Witold R Rudnicki, and others. 2010. “Feature Selection with the Boruta Package.” *J Stat Softw* 36 (11): 1–13.

Spearman, Charles. 1904. “The Proof and Measurement of Association Between Two Things.” *American Journal of Psychology* 15 (1): 72–101.

2019a. *Gadm.org*. https://gadm.org/data.html (https://gadm.org/data.html).

2019b. *DataCamp Community*. https://www.datacamp.com/community/tutorials/feature-selection-R-boruta (https://www.datacamp.com/community/tutorials/feature-selection-R-boruta).