Assignment 3 - Data Set Generation and First Analysis

MPP-E1180 - Hertie School of Governance, Spring 2016

Md Mujahedul Islam & Johannes Schulz-Knappe

April 27, 2016

1 Introduction

As proposed in the research proposal of Assignment02, we want to understand the factors that explain the success and rise of AfD in Germany's political landscape. At the three state elections in Bade-Württemberg, Rhineland-Palatinate, and Saxony-Anhalt, the AfD received double digit results from scratch, meaning that it has not been present in the state parliaments before. In Baden-Württemberg (BW), the party received 15.1%, in Rhineland-Palatinate (RP) 12.6%, and in Saxony-Anhalt (SA) even 24.3%, becoming the third strongest party in BW and RP, and the second strongest in SA.

While our initial attempt was to gather data on constituency level, we had to adapt to available data sources. This is why we changed the level of analysis to the administrative districts of the three German states.

Thus, the updated research question is:

"*What explains the recent electoral success of the AfD in the different administrative districts of Baden-Württemberg, Rhineland-Palatinate, and Saxony-Anhalt?"*

We want to understand which structural factors affect the electoral success of the AfD and whether the characteristics of its voters identified in post-election analyses are reflected in the district level data. Also, we want to find out whether it follows the same patterns as the electoral success of radical right populist parties in the rest of Western Europe

In the following, we present our data collection efforts and initial analysis that was undertaken with the generated dataset.

2 Data and Variables

As stated in the Assignment 2 research proposal, we combine two main data types in order to get our variables.

The first data type is the electoral data provided by the State Offices for Statistics on the last state elections in Baden-Württemberg, Rhineland-Palatinate, and Saxony-Anhalt. The variables retrieved from these datasets are:

Variable	Name in dataset	Description
Dependent Variable (DV)	**vote.AfD**	The vote share of the Alternative for Germany in the current state election
Independent Variable 1 (IV1)	**lag.turnout**	Overall voter turnout of the previous state election
IV2, IV3, IV4, IV5, IV6	**lag.CDU**, **lag.Greens**, **lag.SPD**, **lag.FDP**, **lag.Linke**	The vote share of political parties in the previous state election (CDU, SPD, Greens, FDP, and Linke)

The second data type is on the structural characteristics on the administrative districts. We retrieve this data from the Federal Statistical Office DESTATIS. As this data is only available on district level, we adapted our data collection on it. The variables retrieved from these datasets are:

Variable	Name in dataset	Description
IV7	**abitur.ratio**	Ratio of school leavers per district with general qualification for university entrance (Abitur)
IV8	**nodegree.ratio**	Ratio of school leavers per district with no school degree
IV9	**GDP.cap**	Gross domestic product (GDP) per capita
IV10	**unemp.rate**	Unemployment rate in the district
IV11	**n.refugees**	Number of asylum seekers per district

2.1 Data Gathering

Election Data

Gathering the election data was complicated due to the independence of the Statistical State Offices. Every German state office chooses its own format to provide data online. While we thought about collecting data on all state elections in Germany where the AfD has won seats, we had to limit our efforts to the three states Baden-Württemberg (BW), Rhineland-Palatinate (RP), and Saxony Anhalt (SA) in order to provide a working dataset in the given time frame. However, for the final analysis, we will try to extend our dataset to achieve a greater sample.

We retrieved district level data on the 2016 and 2011 state elections from the respective websites. While the data was available in separate files for RP and SA, district level data was only available for BW in a combined table that had to be divided in following steps. In total, we gathered five different datasets on the elections.

Structural Data

The website of the DESTATIS online database "Genesis-Online" was the source of the structural data files. As DESTATIS does not offer an API, the data had to be manually downloaded from the database. We saved four datasets as csv files that were subsequently loaded into our project. While this is only initial data, the similar structure of DESTATIS data allows us to include further datasets on new variables in case we see the need for that during analysis.

2.2 Data Cleaning and Merging

The biggest challenge during the data cleaning was the generation of a common identifier for the different dataset. Except for the SA data, the election datasets is not provided with the district key which is the main identifier for the DESTATIS data. In order to add the identifier to the election data, we converted the district names that were available in both the election and the structural DESTATIS data into the same format and structure. Then, we sorted the data frames in the same way to add the identifier to the election data.

Finally, we merged all datasets into a single one which contains the variables mentioned above, the election year, and a state indicator on 94 observations.

3 Empirical Findings

With the generated dataset we were able to conduct initial analysis of correlations and distributions. While our dependent variable is a [0, 1] bounded variable (meaning, the vote share can take values between 0% and 100%), we will use beta regression for our final analysis. For the beginning, we use OLS to get a basic idea of the relationships of the selected variables.

3.1 Descriptive Statistics

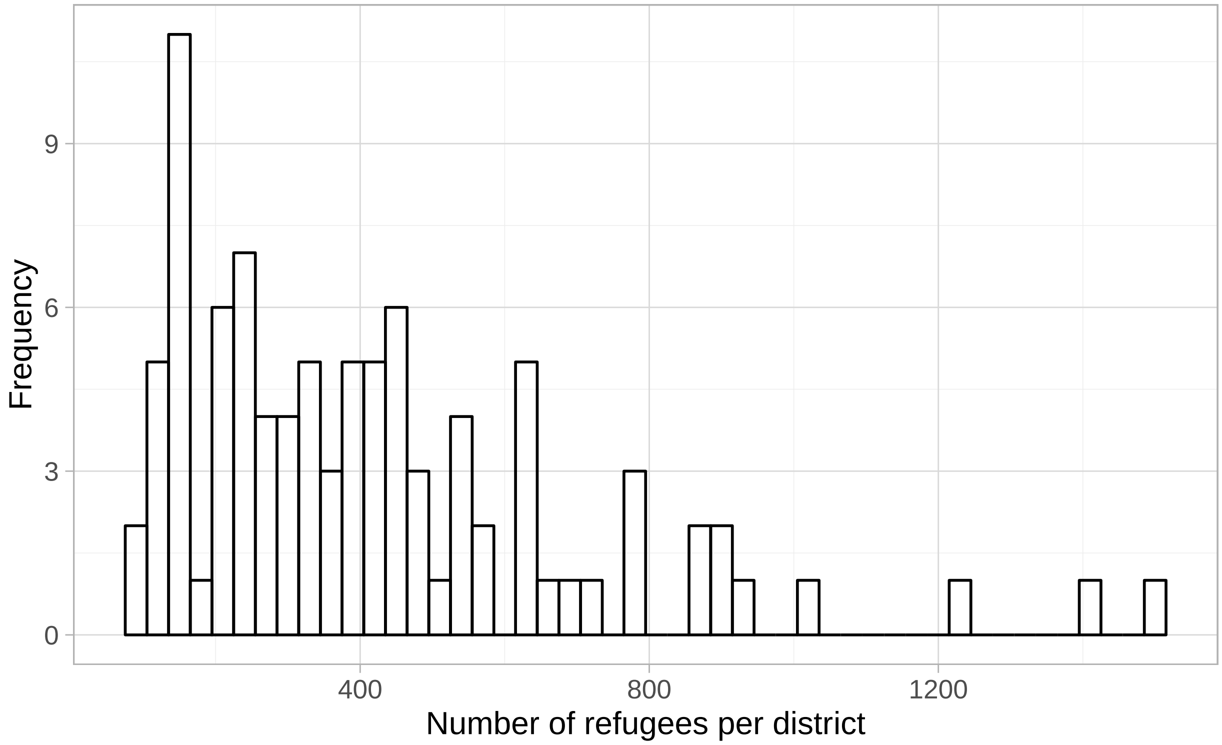
Table 3 shows general descriptive statistics of our variables. Except for the "GDP per capita" and "number of refugees" variables they are all in percent.

The standard deviations for "GDP per capita" and "vote share of Linke" are particularly large in comparison to the respective means. GDP per capita differs substantially between rural districts and heavily urbanized ones and while the Linke is very strong in Saxony Anhalt, it is barely present in BW and RP.

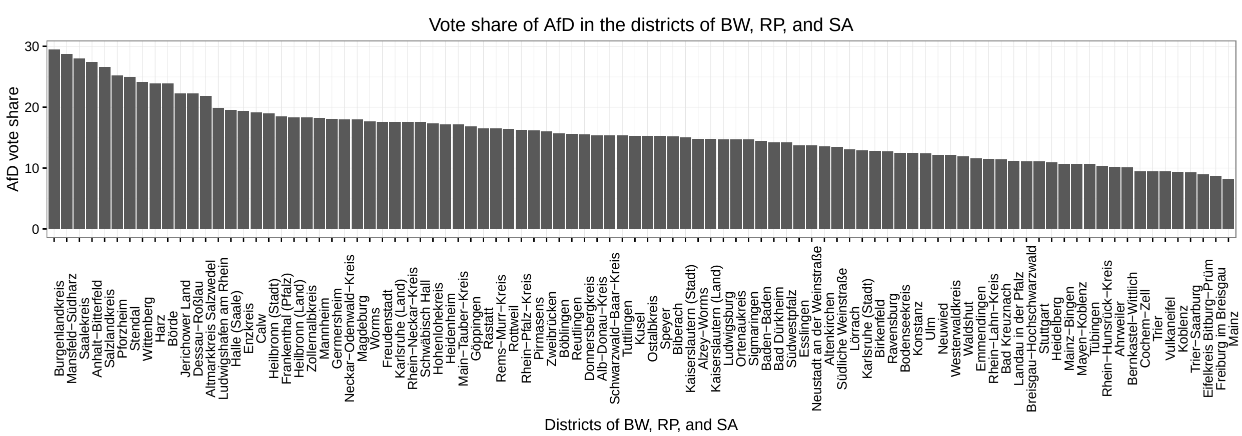
Table 3: Summary statistics of all variables

Statistic	N	Mean	St. Dev.	Min	Max
Vote share of AfD in 2016	94	15.825	4.856	8.200	29.440
Vote turnout in 2011	94	61.587	6.176	47.060	73.400
Vote share of CDU in 2011	94	36.675	6.350	21.500	51.200
Vote share of Greens in 2011	94	17.303	8.288	2.830	43.000
Vote share of SPD in 2011	94	27.731	7.606	16.600	46.300
Vote share of FDP in 2011	94	4.107	1.987	0.090	8.400
Vote share of Linke in 2011	94	5.986	7.283	2.000	25.180
High school ratio	94	29.524	8.957	13.900	55.200
No degree ratio	94	5.798	2.594	1.600	15.900
GDP per capita	94	31,543.490	10,577.930	14,998	70,052
Unemployment rate	94	6.246	3.032	2.900	15.100
Number of refugees	94	423.266	283.213	97	1,487

The following graph shows the frequency distribution of “n.refugees” for the observed districts. The distribution shows that there are some outliers when it comes to the number of asylum seekers. We would estimate that these outlier districts are urban districts with higher population density. For further analysis it would be valuable to calculate a “refugees per capita” variable.



When lining up the districts according to the vote share of the AfD, it becomes clear that the party achieved an overall stable and even vote share. The only real outliers are a couple of districts in Saxony Anhalt where the AfD achieved its by far greatest successes.



3.2 Correlations Among Variables

Table 4 provides a correlation matrix of a selection of our variables.

It is important to note that the AfD vote share is negatively correlated with the SPD vote share, but only slightly with the CDU vote share. GDP per capita and high school degree ratio are negatively correlated with AfD vote share, while the unemployment rate and the number of refugees are positively correlated with it.

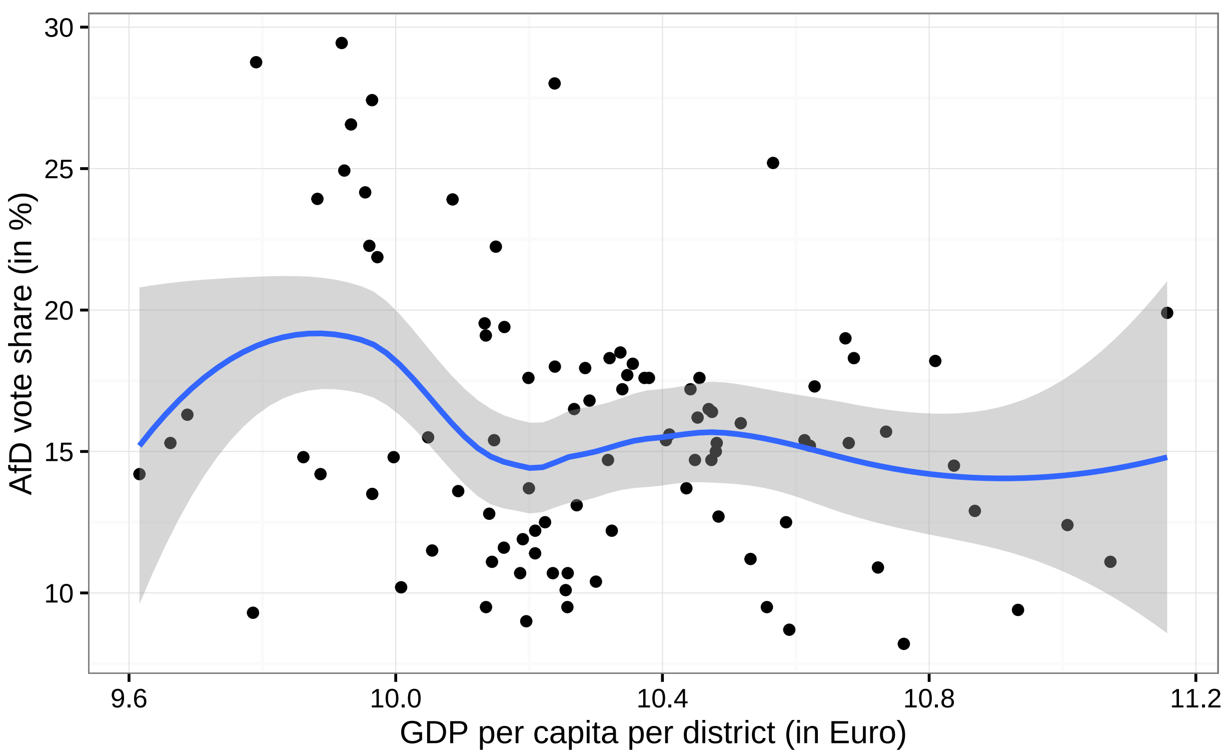
When it comes to IV/IV correlations, the large positive correlation between GDP per capita and the high school degree ratio is not surprising. However, it is interesting to see that the high school degree ratio is negatively correlated with the previous vote share of the CDU, and positively with the unemployment rate.

Table 4: Correlation Matrix of different variables

	vote.AfD	lag.CDU	lag.SPD	GDP.capita	unempl.rate	n.refugees	abitur.ratio
vote.AfD	1	-0.079	-0.337	-0.232	0.615	0.020	-0.298
lag.CDU	-0.079	1	-0.409	-0.019	-0.527	-0.037	-0.481
lag.SPD	-0.337	-0.409	1	-0.119	0.045	-0.400	0.227
GDP.capita	-0.232	-0.019	-0.119	1	-0.119	0.185	0.468
unempl.rate	0.615	-0.527	0.045	-0.119	1	-0.079	0.236
n.refugees	0.020	-0.037	-0.400	0.185	-0.079	1	-0.021
abitur.ratio	-0.298	-0.481	0.227	0.468	0.236	-0.021	1

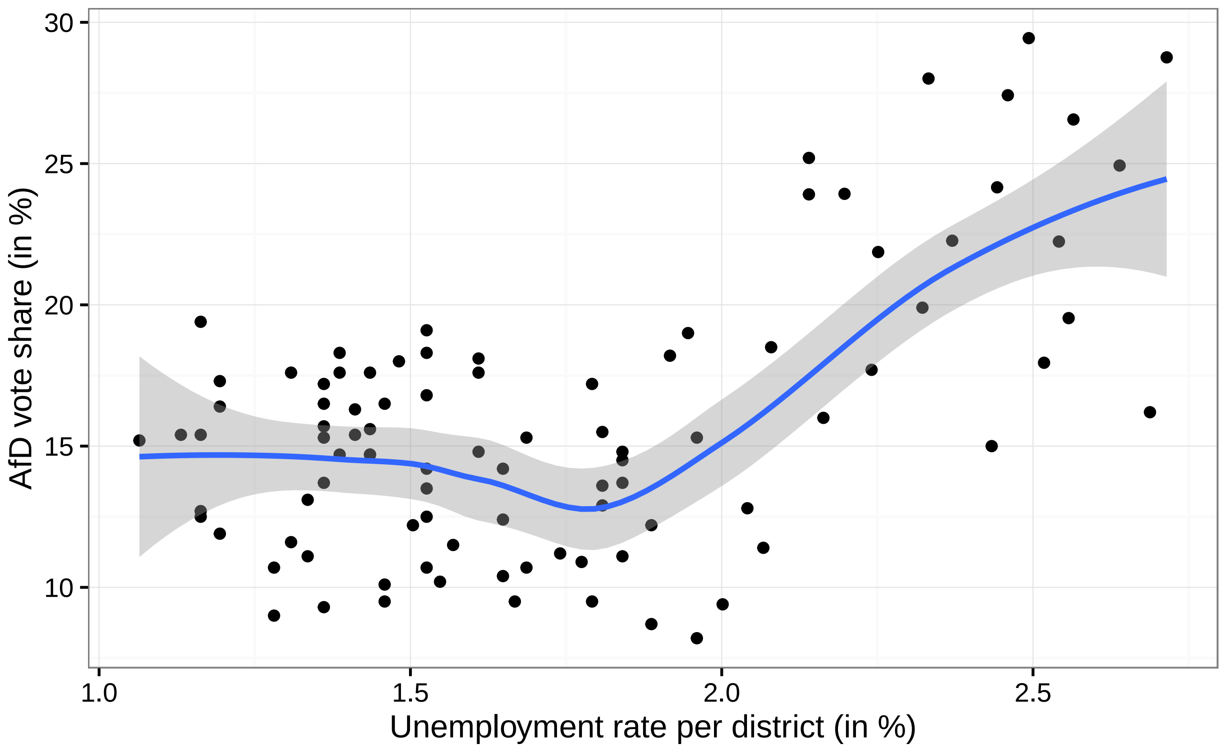
3.2.1 Correlation of AfD vote share and GDP per capita

The following graph shows a slightly statistically significant (t value is -2.04) negative correlation of GDP per capita with AfD vote share. The effect seems to be mostly influenced by a small group of outliers (high vote share, low GDP per capita) and the variance is not equal across the independent variable (posing a problem with heteroscedasticity).



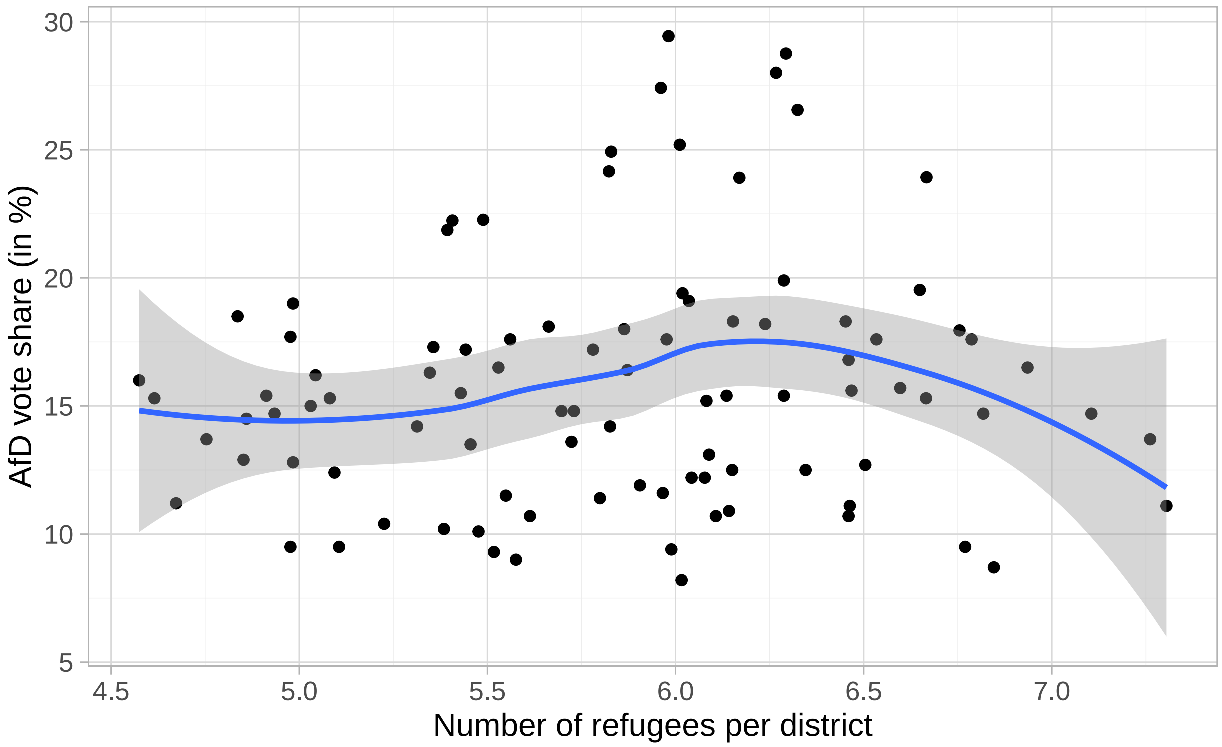
3.2.2 Correlation of AfD vote share and unemployment rate

The following graph shows a statistically significant (the t value from the Pearson's product-moment correlation test is 6.27) positive correlation of the logged unemployment rate in a district with the AfD vote share. However, the correlation seems to be negative for districts with lower unemployment rates but becomes strongly positive for districts with high unemployment. The positive correlation of the high school degree ratio with the unemployment rate and its negative correlation with the AfD vote share visible in the correlation matrix could offer a possible explanation for this. We will test this in future analysis with interaction variables.



3.2.3 Correlation of AfD vote share and number of refugees

The result shows a positive correlation of the number of refugees per district with the AfD vote share. But it is not statistically significant since the t value is 0.13 (from the Pearson's product-moment correlation test). Since the data on the refugees per district is from 2013, it is highly outdated. We will have to get more recent data (for 2014 or even for 2015) to be able to conduct a meaningful analysis here.



3.3 Inferential Statistics

Testing some initial regression models, we see that the unemployment rate, the vote share of the SPD in previous elections, and the ratio of people leaving school without a formal degree are statistically significant across all models. In almost each model a one percentage point increase of the unemployment rate leads to a one percentage point increase of the AfD vote share. On the other hand, the number of refugees per district and the vote share of the CDU are insignificant in every model.

With just a small initial selection of variables, the models have surprisingly high adjusted R^2. Between 40% and 64% of the variation of AfD vote share could be explained. However, since our current type of analysis (OLS) is not as suited for the analysis of a [0, 1] bounded dependent variable as for example beta regression, these findings are likely to change in our final analysis.

Table 5: Determinants of electoral success for the Alternative for Germany

	Dependent variable:			
	Vote share of AfD			
	(1)	(2)	(3)	(4)
GDP per capita	-0.0001^{**}	-0.0001^{**}	0.00004	0.00001
	(0.00004)	(0.00003)	(0.00004)	(0.00004)
Unemployment rate	0.963^{***}	0.682^{***}	1.103^{***}	0.973^{***}
	(0.130)	(0.222)	(0.181)	(0.233)
Number of refugees	0.002	0.001	0.001	−0.0003
	(0.001)	(0.001)	(0.001)	(0.001)
Vote share of CDU in 2011 election		0.091		−0.033
		(0.074)		(0.073)
Vote share of SPD in 2011 election		-0.205^{***}		-0.201^{***}
		(0.058)		(0.055)
High school ratio		-0.241^{**}		−0.136
		(0.105)		(0.099)
No degree ratio			-0.267^{***}	-0.224^{***}
			(0.048)	(0.050)
(Intercept)			0.148	−0.118
			(0.208)	(0.203)
Constant	11.675^{***}	30.800^{***}	14.225^{***}	32.175^{***}
	(1.618)	(8.649)	(1.731)	(8.277)
District FE?	NO	NO	NO	NO
Observations	94	94	94	94
R^2	0.413	0.588	0.597	0.669
Adjusted R^2	0.394	0.560	0.574	0.638
Note:				*p<0.1; **p<0.05; ***p<0.01