Pair Assignment 3: Gathering, Cleaning, Merging and Exploring our Data

Course: Introduction to Collaborative Social Science Data Analysis

Malte Berneaud-Kötz & Jonas Markgraf

Hertie School of Governance

14 April 2016

1. Gathering the Data

1.1. *Sparkassen* Board Membership Data

We use a hand-collect a unique panel dataset containing information on board members in Bavaria's *Sparkassen*[1]. This data is available on *Bundesanzeiger* for the years from 2006 to 2015. The dataset includes detailed information on board member profiles (name, occupation, and position within board) which enables us to identify mayors on bank boards.

1.2. Municipal Election Data

A database on mayoral elections in Bavaria is available from the state statistical office upon request. It contains information on direct municipal elections between 1948 and 2014 including election date; name of all candidates and their vote shares; and party affiliation. With this database we are covering 79 of the 416 German *Sparkassen* (19%) and 2,099 municipalities (19% of all municipalities in Germany).

After obtaining the raw data, we cleaned the data set and subsequently created additional variables needed in our analysis. The individual steps taken are outline in the subsequent sections.

[1]This dataset was collected as part of a research project by Guillermo Rosas (Washington University in St. Louis; grosas@wustl.edu) and Jonas Markgraf (Hertie School of Governance; markgraf@hertie-school.org).

2. Cleaning the Data

2.1. *Sparkassen* Data

In order to make our analysis more accessible, we re-defined the variable names in the *Sparkassen* dataset. Moreover, we standardized them to follow use underscores to seperate words and use lower case and deleted redundant whitespaces.

Furthermore, we created an additional variable for top positions in *Sparkassen* boards. The raw data includes information on the position of the respective person (normal board member, vice chairman or chairman); since chairmen and vice chairmen alternate over the years in Bavarian savings banks, there is no difference between chairmen and vice chairmen in reality and they can be amalgamated into one category (*top position*).

Finally, we created four sub-data frames by subsetting the initial dataframe in order to analyze different aspects of *Sparkassen* boards in greater detail. Hence, we created a subset containing unique board member profiles, another one with unique profiles of mayors on the board, a data frame with only persons in top positions on the board, and a data frame with unique banks.

2.2. Creating Additional Variables

In addition to the variables available in the data set already, we created variables (1) distinguishing 'first time mayors', (2) identifying competitive elections where there were more than one candidate, (3) the number of times a single mayor was elected, which allowed us to identify mayor's first re-election. This last point is important because our research is interested in the first re-election of mayors specifically.

2.3. Municipal Election Data

The municipal election data as provided by the Bavarian Statistical Service was provided as an Excel worksheet, which also meant that the columns where named in a way which was not computer-readable. As a result, we had to clean the names of the data set almost entirely. Aside from containing spaces, they also frequently contained line breaks and carriage returns.

Aside from the variables included in the data set, we extracted information on PhD titles from the names of the candidates from the variable for candidate's name and created a new variable indicating whether the person has a Dr. title or not; this was necessary in order to merge the *Sparkassen* dataset and the Mayor Election dataset because name and title of board members was collected as two separate variables.

Additionally, we created a variable which indicates whether or not the election of the mayor is contested. We want to use this variable in subsetting our data set later, as it identifies mayors which could have possibly leveraged their position in a *Sparkassen* board to secure re-election. Using the total of valid votes and votes for the winning cadidate, We calculated the vote shares of the winning candidates.

Subsequently, we lagged the winner's vote shares and their sex by one time period so that we can account for mayors' sex and previous election results in estimating our model on the re-election chances.

The municipal election data set was subsetted to only include the time period for which we have data on *Sparkassen* board membership, which is from 2006 to 2016. We excluded elections which required run-offs, as leading candidate in those elections does not have to be the one winning the run-off. Moreover, elections where no candidate besides the incumbent mayor ran for office were excluded. The study data set contains 2559 observations across two municipal election rounds (2008 and 2014).

3. Merging the Data

In order to analyze whether mayors with a board seat in a public savings bank are more likely to be re-elected, we need know whether a mayor was board member in a *Sparkasse*. While we have detailed information about a mayor's profile in the Municipal Election data frame, we lack information about his or her connection to a *Sparkasse*; this information we get from our *Sparkassen* Board Membership dataset. In order to combine this information with the Election dataset, we add a variable to our Election dataset indicating whether the name of election winner is listed in our Board Membership dataset, i.e. if the mayor has a board seat[2].

However, as we are interested in whether the incumbent (and not the election winner) was holding a board seat, we had to add another variable indicating if the election winner of the previous election (i.e. the incumbent) was a board member.

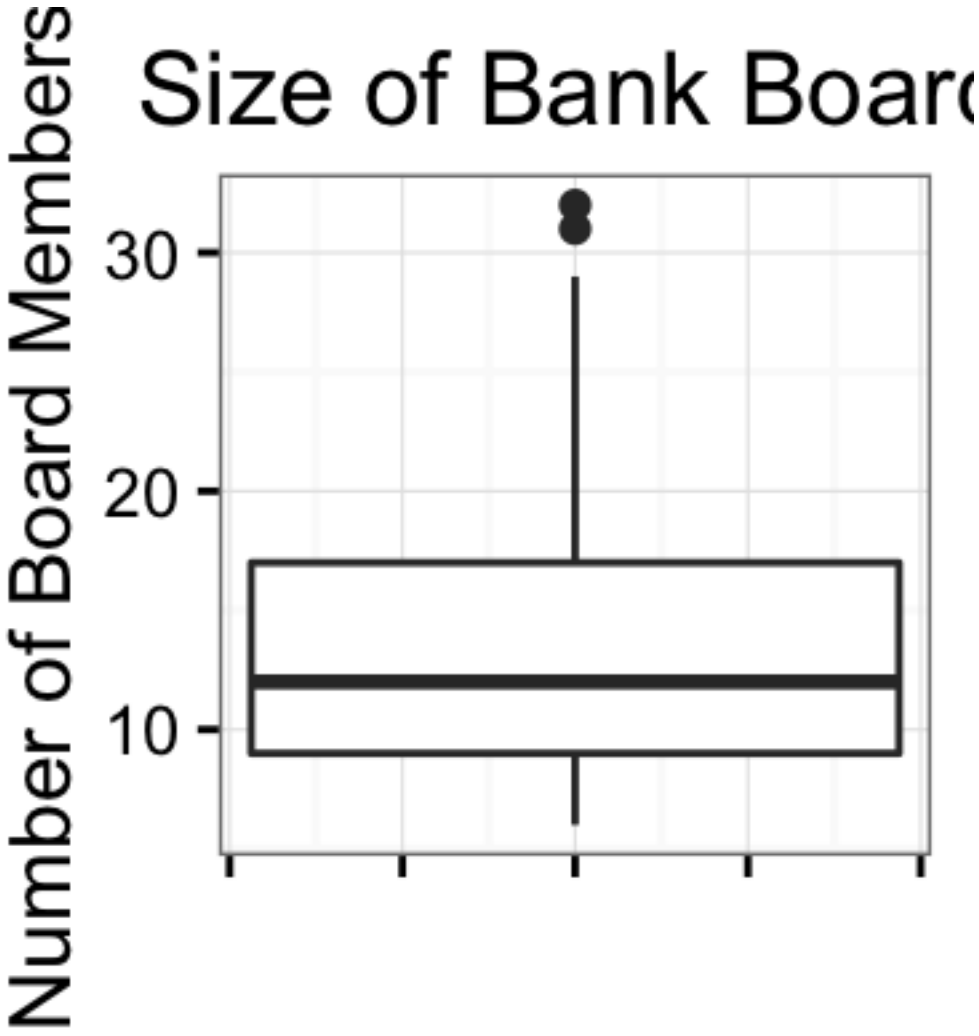
[2]As the focus of this paper is the analysis of the electoral effect of board membership, we focus on the re-election chances of mayors with board membership in a *Sparkasse* compared to those without a board seat. A closer examination of the patterns of politicization of banks' boards, such as partisanship within boards, is therefore not conducted and goes beyond the scope of this paper. Future steps of the PhD research project will, however, analyze those patterns.

4. Descriptive Statistics

4.1. Sparkassen Dataset

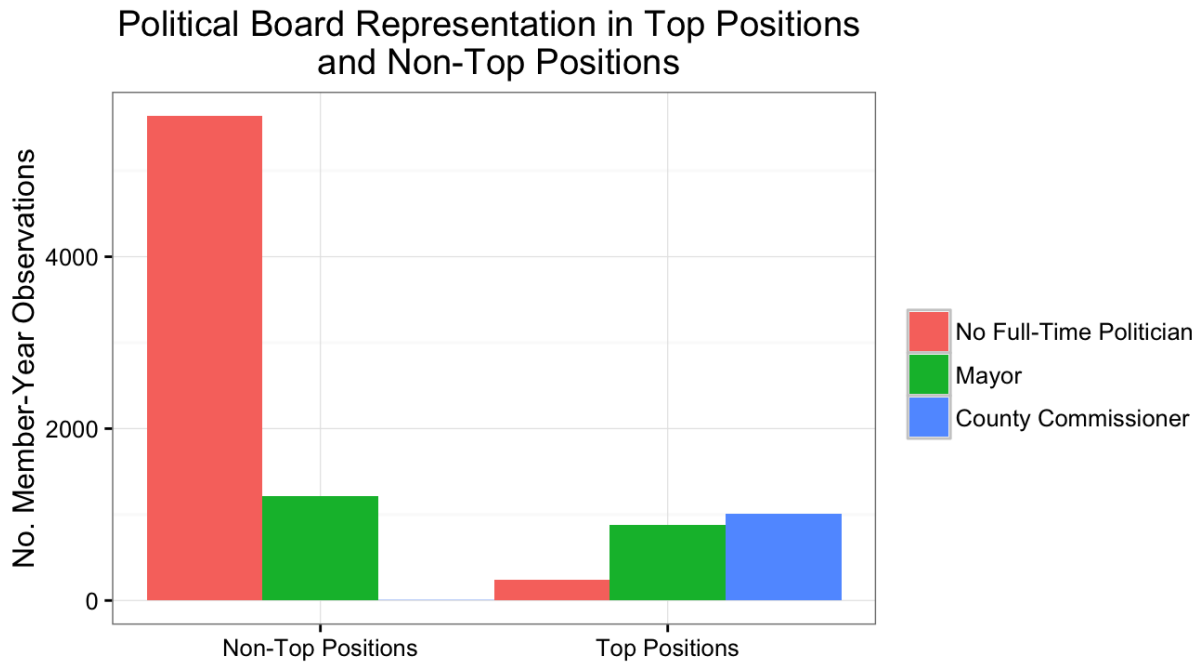
As outlined above, the *Sparkassen* dataset contains information on names, political position (no full-time politician; mayor; county commissioner) and position within the board (top position; non-top position). This allows us the estimate the degree of politicization of boards and the patterns of politicization.

The Board Membership dataset contains 79 banks. Board size varies between a maximum of 32 board seats and a minimum of six seats; on average (median), Bavarian savings banks have 12 persons on the supervisory board.



For those 79 banks, we have overall almost 9,000 board member-year observations, about 1,600 unique board member profiles of which 410 are mayors (23%) and 175 are county commissioners (11%).

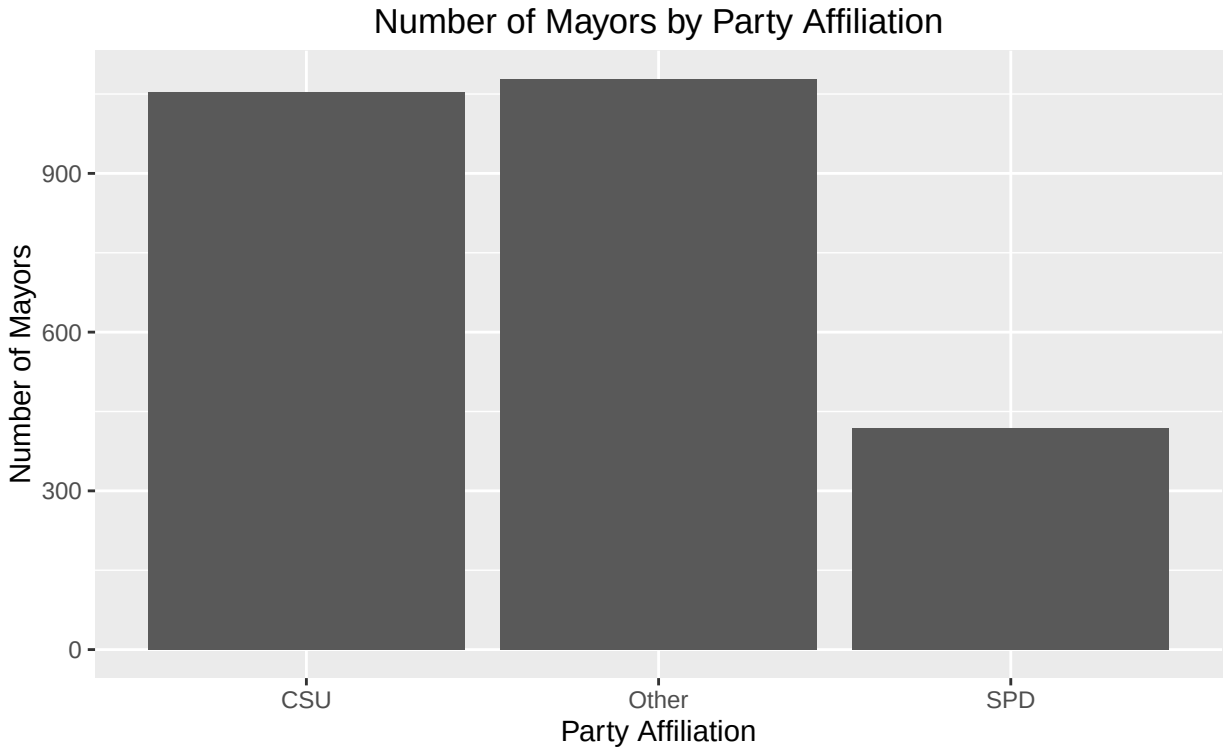
Of the 9,000 board member-year observations, about one-fourth (2,115 observations) are member-year observations for board members in top positions. Of those 2,115 observations, 875 are mayors and 1,006 are county commissioners; only 238 top-position observations belong to persons that are not full-time politicians.



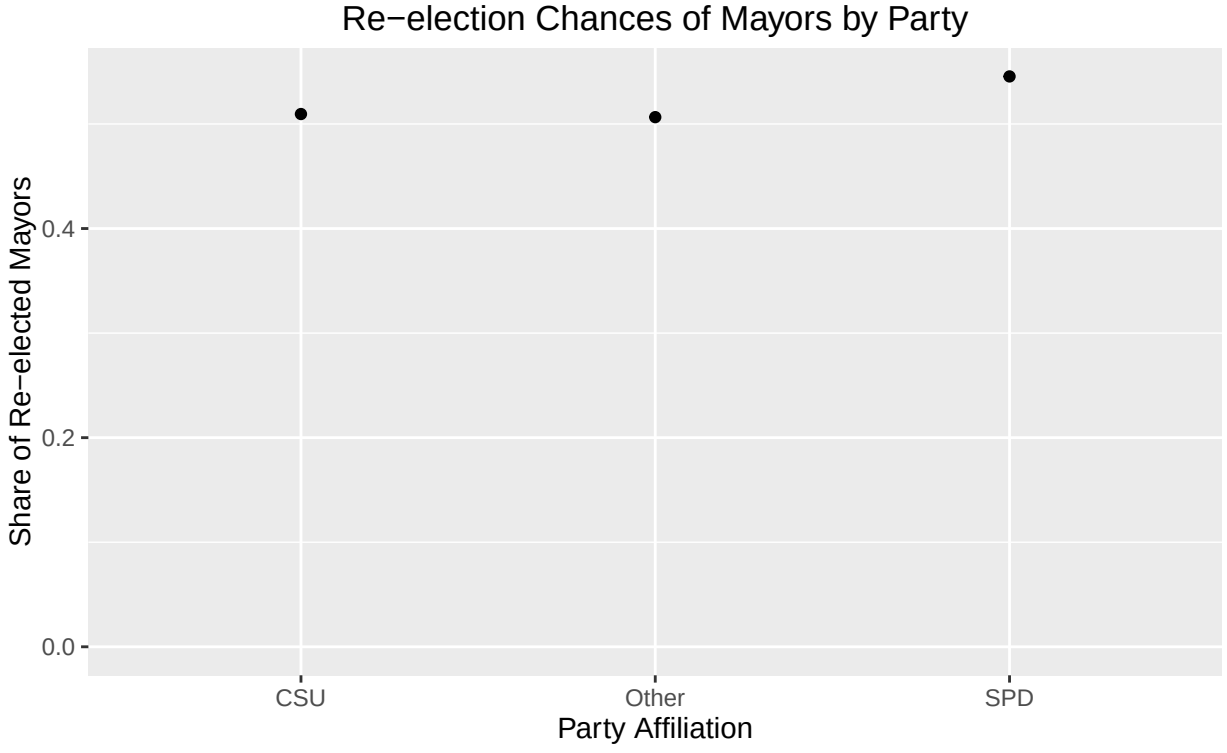
Overall, this shows that a considerable share of bank boards is hold by full-time politicians (34%). Moreover, political representation is pronounced in top positions of banks' boards (county commissioners are only in top positions); as board seat holders of top positions have much higher influence on banks' strategic decisions (e.g. as the chairman is always member of the credit committee), this finding confirms our suspicion that German *Sparkassen* are subject to potentially high political control and shows that politicization of public savings banks mainly works through top positions.

4.2. Election Data Set

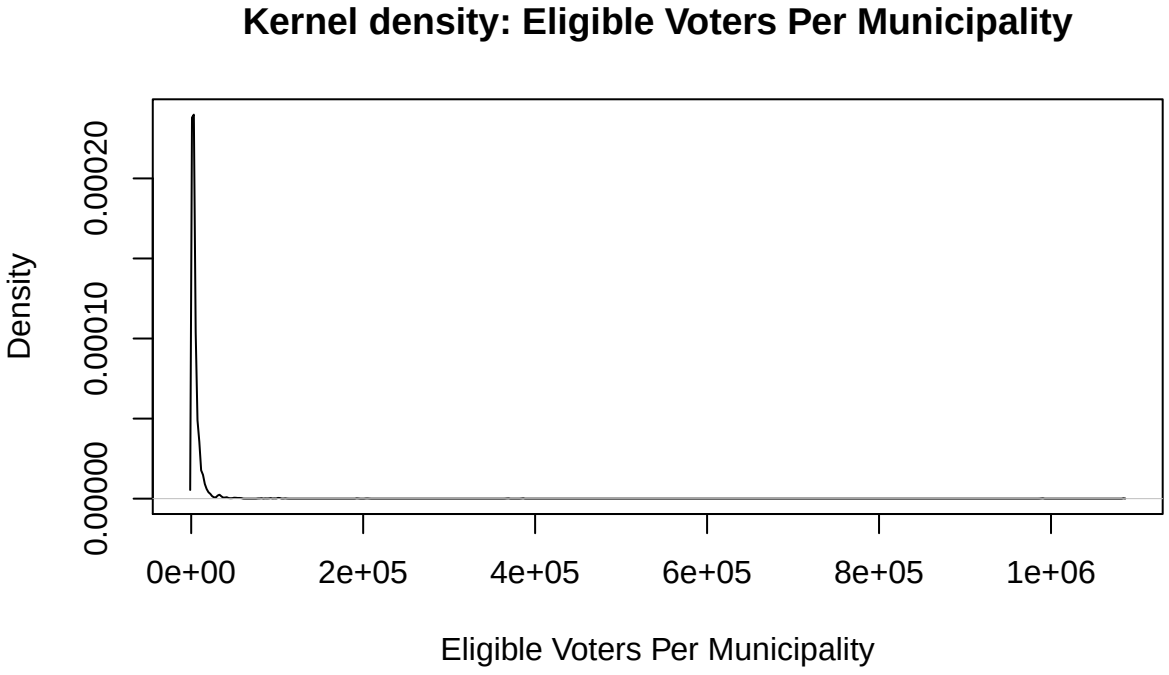
Out of all the mayors elected the time period under study, 8.871% of the mayors who were elected were female.



The biggest share of mayors are affiliated with the CSU, the second largest belong to other parties, which include independent voter lists. The SPD, a popular party in the entirety of Germany does worse than in the remainder of Germany. Out of 2559 mayors which were elected over the study period, only 418 mayors were affiliated with the SPD.



Comparing the chances of re-election for mayors of different parties, we find that mayors from all parties except the SPD are re-elected in about 18% of the cases, whose mayors can reclaim office in roughly 24% of the cases.



The number of eligible voters per municipality, which we intend to use to control for municipality size is severly right-skewed, as expected, so we will log it in our analysis.

5. First Inferences

Finally, we run first regressions in order to test the effect of board membership for re-election chances of mayors. In a first, very simple equation (1), we regress re-election (binary variable) on board membership (binary variable); in the second specification we add municipality size proxied by the number of eligible voters in a municipality to the regression. As the distribution of municipality size is strongly positively skewed, we include the variable in log format (Figure X). In regression 3, we furthermore add the previous vote share of the incumbent to the regression assuming that the previous vote share should be positively correlated with re-election chance of the incumbent. Finally, we add a dummy for the gender of the incumbent in specification 4. As our response variable is categorical, we use a logistic regression model.

Statistical models

Model 1

Model 2

Model 3

Model 4

(Intercept)

0.02

1.49***

0.82*

0.72

(0.04)

(0.15)

(0.41)

(0.41)

IncumbentSparkassenMemberTRUE

0.25*

0.25*

0.24*

0.24*

(0.11)

(0.11)

(0.11)

(0.11)

L.VoteShareWinner

-2.17***

-2.12***

-2.05***

(0.22)

(0.23)

(0.23)

log(NumberEligVoter)

0.08

0.08

(0.04)

(0.04)

L.Geschlecht1w

0.47*

(0.18)

AIC

3530.27

3372.54

3371.39

3366.55

BIC

3541.96

3390.02

3394.70

3395.68

Log Likelihood

-1763.14

-1683.27

-1681.70

-1678.27

Deviance

3526.27

3366.54

3363.39

3356.55

Num. obs.

2549

2506

2506

2506

p < 0.001, *p < 0.01,* p < 0.05