Case Study Elections Netherlands

Ilse van Beelen and Floor Komen

december 21, 2018

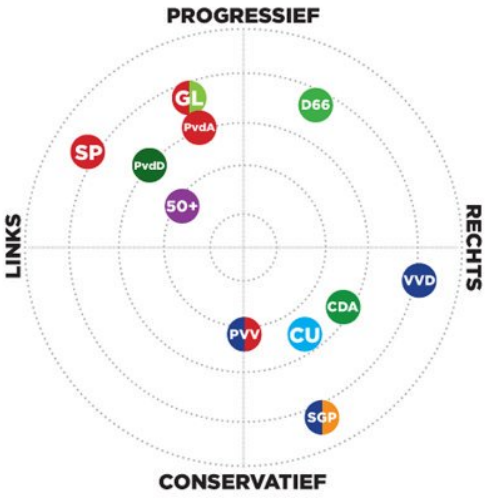
Abstract

Put the abstract over here

1. Introduction

In this report the election

1.1 Motivation



Dutch political parties

In this landscape the diffence between the parties is graphically displayed in this figure. In this research, two parties are chosen to investigate. These two parties had to be different, so that some comparisons could be made. The parties should not be to extreme left/right/conservative/progressive, so that the model will be proven to work on less-extreme parties. Therefore, the chosen parties are: CDA and GroenLinks. After the data-cleaning the CDA will be researched first, afterwards GroenLinks will be researched.

In this research the above described demographics are chosen because of there influence on a municipality level. The thought is that a more non-western municipality for example votes different than a less non-western municipality. This is the same for the other two demographics. Other demographics are also researched, for example gender, but on a municipality level there is no big difference between the amount of men and women per municipality. So that is a more interesting demographic to research on an individual level. *The standardized income per municipality* are given in thousands. *the urban index of a municipality* is a database with five categories per municipality. These five categories are; really strong urban (more than 2500 addresses per km2), strong urban (1500-2500 addresses per km2), moderate urban (1000- 1500 addresses per km2),

little urban (500-1000 addresses per km2) and not urban (less than 500 addresses per km2). Per municipality the amount of km2 per category is given. The *non-west residents per municipality* is given in an amount per municipality, also the total amount of residents is given per municipality.

1.2 Data sources

Electoral data For the electoral data we used the results of the 2017 general election. This is the most recent national election and is of the most important election type. Therefore, it seems plausible that the data for this election is representative of the political makeup of different municipalities. We downloaded the raw data directly from the official government source.[1] This contained a .csv file with the raw number of votes for every party in every municipality.

Demographical data
We got our demographical data from the CBS, the official Dutch statistical agency.[2] From the wealth of demographical information available we picked a handful of attributes that we suspected (based on prior research and some gut feeling) to be useful as predictor variables. We landed on five demographical attributes: education grade, average income, age, urbanization and the amount of people with a non-western background. Note that the data we downloaded from the CBS site usually had to be transformed to get it in a useful predictor variable format. The specifics of this are described in the next section.

1.3 Data cleaning

The data cleaning is the process in which we used

Electoral data

Demographical data

The variable *non-western residents* are divided in three groups. Municipalities with less than 5 % non-western residents, 5-10 % non-western residents and municipalities with more dan 10 % non-western residents.

```
Data <- read.csv("1_clean_data/voting_and_demographics.csv", stringsAsFactors = F,
    header = T)
Data <- Data[, -15]
colnames(Data) <- c("Muni", "VVD", "CDA", "PVV", "D66", "SP", "GL", "PvdA",
    "CU", "50PLUS", "PvdD", "SGP", "FvD", "DENK", "Urban_index", "High_edu_perc",
    "Mean_income", "Dutch_perc", "West_perc", "Non_west_perc")
Data$Non_west <- ifelse(Data$Non_west_perc < 0.05, 1, NA)
Data$Non_west <- ifelse(Data$Non_west_perc >= 0.05 & Data$Non_west_perc < 0.1,
    2, Data$Non_west)
Data$Non_west <- ifelse(Data$Non_west_perc >= 0.1, 3, Data$Non_west)
Data$Non_west <- as.factor(Data$Non_west)
Dat_cda <- Data[, c(1, 3, 15, 16, 17, 21)]
Dat_cda <- Dat_cda[complete.cases(Dat_cda), ]
```

[1] https://data.overheid.nl/data/dataset/verkiezingsuitslag-tweede-kamer-2017
[2] https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072ned/table?ts=1544803364892

Table 1: Data summary

CDA	GL	Urban_index	High_edu_perc	Mean_income	Non_west	Perc_60plus
Min. :0.031	Min. :0.0025	Min. :0.00	Min. :0.12	Min. :21	Min. :1.0	Min. :0.067
1st Qu.:0.116	1st Qu.:0.0539	1st Qu.:0.66	1st Qu.:0.22	1st Qu.:24	1st Qu.:1.0	1st Qu.:0.123
Median :0.142	Median :0.0657	Median :1.22	Median :0.26	Median :26	Median :2.0	Median :0.134
Mean :0.152	Mean :0.0714	Mean :1.43	Mean :0.27	Mean :26	Mean :1.7	Mean :0.133
3rd Qu.:0.182	3rd Qu.:0.0846	3rd Qu.:2.18	3rd Qu.:0.30	3rd Qu.:27	3rd Qu.:2.0	3rd Qu.:0.142
Max. :0.420	Max. :0.2055	Max. :3.79	Max. :0.53	Max. :42	Max. :3.0	Max. :0.178

1.3 Data visualisation

In this part the cleaned data is visualized, so that a good picture can be obtained of the current data. First of all some demographics of data will be showed. In figure 1 of the *parties*, *the urban index*, *the percentage of highly educated residents*, *the mean income*, *The non west residents factor* and * the percentage 60 plus* are plotted. As you can see in the plot, they are normal distributed. Because of the low values at the x-axis, the CDA, GroenLinks, 60 plus percentage and the highly educated densities are above 1. The area beneath the curve sums to 1, so it is correct.

Figure 1: Density plot

Correlation heatmap In this heatmap the correlation between explanatory and respons variable are showed. The red color means a positive relation, the purple color means a negative relation. The non_west variable is not taken into account, because it is a factor and the other variables are continous. VERDER UITLEG

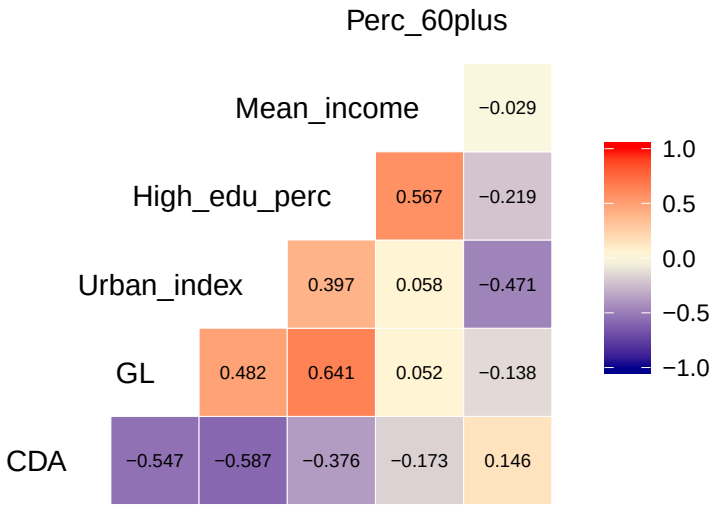


Figure 2: Correlation between explanatory and respons variables

Multilineair plots CDA In these two plots you can see a scatterplot with on the y-axis the votes for CDA in percentages and on the x-axis on the left graph the mean income per municipality in 1000 euro. The right plot has the urbanity index as x-axis. As you can see, the trend is that when the mean income goes up, the votes for CDA goes down. Same with the urbanity index. In the model formulation graph these trends are checked.

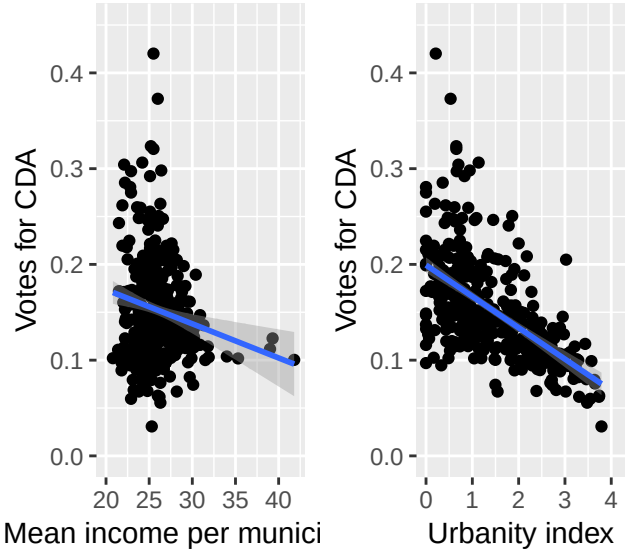


Figure 3: Scatterplots CDA

Multilineair plots Groenlinks

In these plots the percentage of votes for GL are compared with urbanity index and percentage of highly educated residents. When municipalities get more crowded, the trend is that votes for GL

goes up. The other visible trend is that when more residents in a municipality are highly educated, GL got more votes.

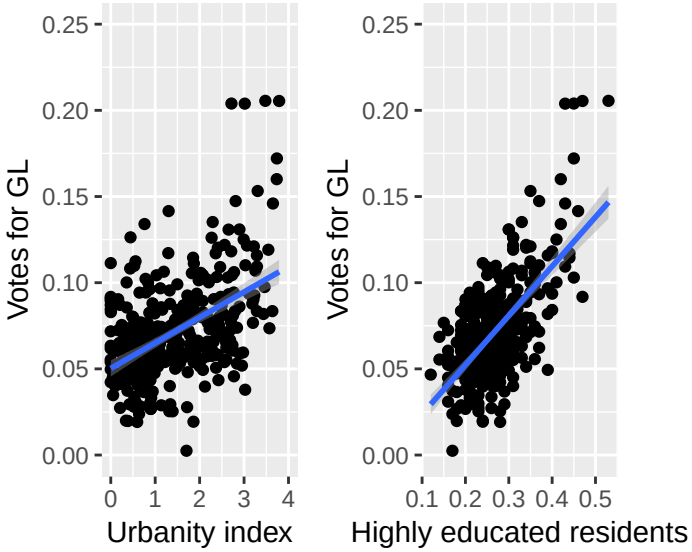


Figure 4: Scatterplot GroenLinks

Multilinear plots explanatory variables These three plots are scatterplots about explanatory variables.

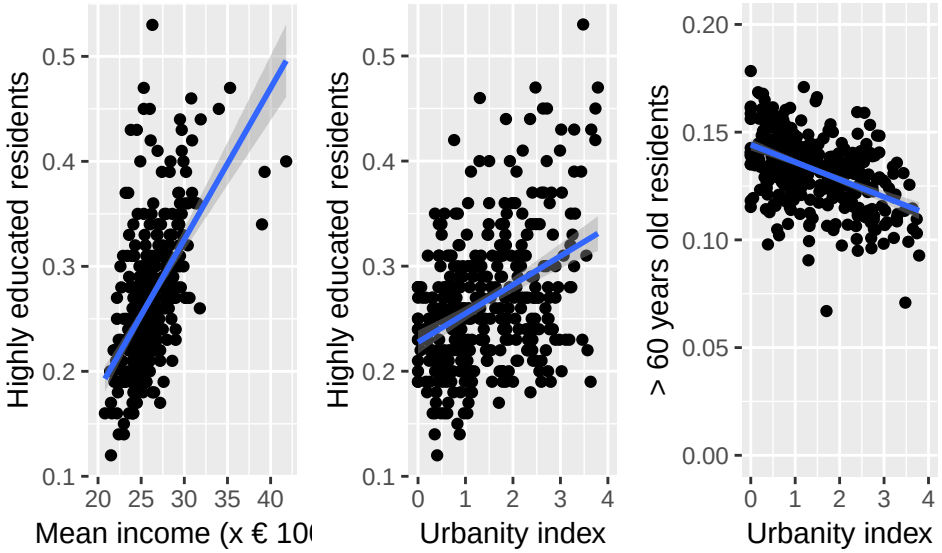


Figure 5: Scatterplot explanatory variables

Multiple boxplots In this graph boxplots are made, to compare some variables. A boxplot is a standardized way to display the distribution of data. It gives the minimum, first quartile, median,

third quartile and the maximum. If there are any outliers, the boxplot is extended with those. The line within the box is the median, the first and third quartile are the down- and upside of the box. The length of the box is the Inter Quartile Range (IQR). The minimum and maximum are 1.5XIQR distance. Outliers are thus further away than 1.5XIQR.

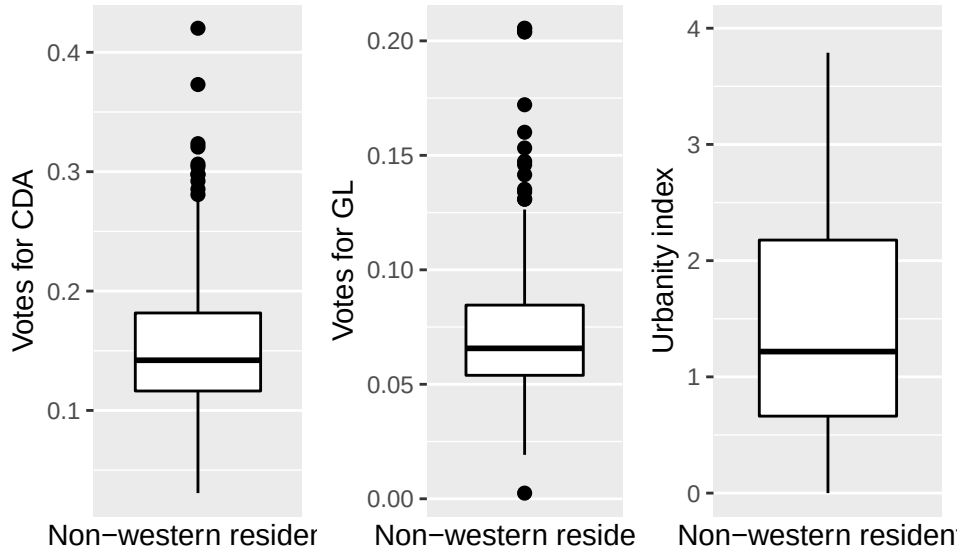


Figure 6: Three boxplots: Votes for CDA, Votes for GroenLinks and Urbanity index

2. Formulate model

Model formuleren

CDA

GroenLinks

3 Final model

CDA

Groenlinks

4 Analyse output

5 Discussion

5.1 Limitations