# Case Study: Predicting the outcomes of the 2017 Dutch General Elections

*Ilse van Beelen, Floor Komen*

*januari 15, 2019*

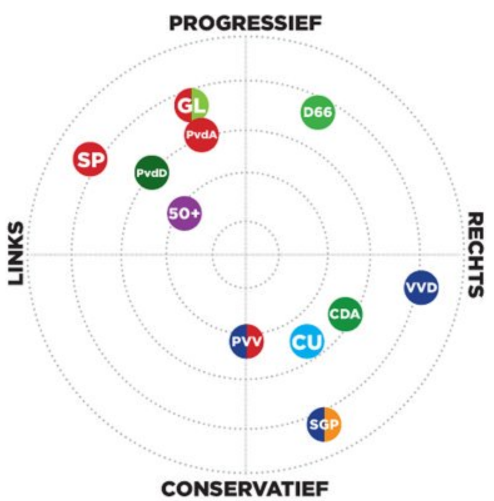
**Abstract**

Put the abstract over here

## 1. Introduction

### 1.1 Motivation

For this case study, it was decided to combine the outcome from the Dutch elections of 2017 and demographic data. Both are collected per municipality and are well maintained and reliable. This makes A lot of information is available for both in the Netherlands. This will hopefully result in observing voting trends per demographic group. The final goal is to validate the model for making futyre predictions.




**Dutch political parties**

This figure displays the diffences between the political parties in the Netherlands. The Netherlands has a total of 13 parties. This investigation focusses on only one party. This party should not be to extreme left/right/conservative/progressive and should also be one of the bigger parties. Otherwise, there is not enough data available, making the results less reliable. Therefore, party CDA is chosen.

In this research the above described demographics are chosen because of their influence on a municipality level. The expectation is that a municipality with more non-western residents for example votes different than a municipality with less non-western residents. 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 large 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 urbanity (more than 2500 addresses per $km^2$)
- Strong urbanity (1500-2500 addresses per $km^2$)
- Moderate urbanity (1000- 1500 addresses per $km^2$)
- Little urbanity (500-1000 addresses per $km^2$)
- No urbanity (less than 500 addresses per $km^2$)

Per municipality the amount of $km^2$ 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, the results of the 2017 general election are used. This is the most recent national election and is of the most important election type in the Netherlands. Furthermore, it had a turnup of 81.9%. 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 these are described in the next section.

## 1.3 Data cleaning

An extensive amount of data cleaning had to be done. Below these steps are describes and a small part of code is displayed.

**Electoral data**

**Demographical data**

The variable *non-western residents* are divided in three groups:

- Municipalities with less than 5
- Municipalities with 5-10
- Municipalities with mre than 10

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

```r
Data_CDA$Non_west <- ifelse(Data_CDA$Non_west_frac < 0.05, 1, NA)
Data_CDA$Non_west <- ifelse(Data_CDA$Non_west_frac >= 0.05 & Data_CDA$Non_west_frac <
    0.1, 2, Data_CDA$Non_west)
Data_CDA$Non_west <- ifelse(Data_CDA$Non_west_frac >= 0.1, 3, Data_CDA$Non_west)
Data_CDA$Non_west <- as.factor(Data_CDA$Non_west)
```

At last, the electoral data and demographic data are combined again. Only the municipality Boxmeer is removed, due to a mistake not all the votes are reported here[3]. The final dataset has no NAs

```r
summary(Data_CDA)
```

```
##      Muni              CDA_frac       Urban_index     High_educated_frac
##  Length:366         Min.   :0.0310   Min.   :0.0000   Min.   :0.1200
##  Class :character   1st Qu.:0.1170   1st Qu.:0.6623   1st Qu.:0.2200
##  Mode  :character   Median :0.1420   Median :1.2305   Median :0.2600
##                     Mean   :0.1528   Mean   :1.4280   Mean   :0.2662
##                     3rd Qu.:0.1820   3rd Qu.:2.1750   3rd Qu.:0.3000
##                     Max.   :0.4200   Max.   :3.7890   Max.   :0.4700
##  Mean_income     Non_west_frac         CDA_abs         Total_abs
##  Min.   :20.80   Min.   :0.01000   Min.   :  421   Min.   :  2727
##  1st Qu.:24.30   1st Qu.:0.03000   1st Qu.: 1737   1st Qu.: 11516
##  Median :25.60   Median :0.05000   Median : 2510   Median : 16915
##  Mean   :25.91   Mean   :0.06574   Mean   : 3254   Mean   : 25162
##  3rd Qu.:27.00   3rd Qu.:0.08000   3rd Qu.: 4023   3rd Qu.: 27087
##  Max.   :41.80   Max.   :0.38000   Max.   :18813   Max.   :440854
##   Frac_60plus      Non_west
##  Min.   :0.0700   1:178
##  1st Qu.:0.1200   2:111
##  Median :0.1300   3: 77
##  Mean   :0.1327
##  3rd Qu.:0.1400
##  Max.   :0.1800
```

### 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.

[3] https://www.gelderlander.nl/boxmeer/7-600-stemmen-in-boxmeer-niet-meegenomen-in-uitslag-verkiezingen~a063ee9e/

Table 1: Data summary

| CDA_frac | Urban_index | High_educated_frac | Mean_income | Non_west_frac | CDA_abs | Total_abs | Frac_60plus | Non_west |
|---|---|---|---|---|---|---|---|---|
| Min. :0.031 | Min. :0.00 | Min. :0.12 | Min. :21 | Min. :0.010 | Min. : 421 | Min. : 2727 | Min. :0.07 | 1:178 |
| 1st Qu.:0.117 | 1st Qu.:0.66 | 1st Qu.:0.22 | 1st Qu.:24 | 1st Qu.:0.030 | 1st Qu.: 1737 | 1st Qu.: 11516 | 1st Qu.:0.12 | 2:111 |
| Median :0.142 | Median :1.23 | Median :0.26 | Median :26 | Median :0.050 | Median : 2510 | Median : 16915 | Median :0.13 | 3: 77 |
| Mean :0.153 | Mean :1.43 | Mean :0.27 | Mean :26 | Mean :0.066 | Mean : 3254 | Mean : 25162 | Mean :0.13 | NA |
| 3rd Qu.:0.182 | 3rd Qu.:2.17 | 3rd Qu.:0.30 | 3rd Qu.:27 | 3rd Qu.:0.080 | 3rd Qu.: 4023 | 3rd Qu.: 27087 | 3rd Qu.:0.14 | NA |
| Max. :0.420 | Max. :3.79 | Max. :0.47 | Max. :42 | Max. :0.380 | Max. :18813 | Max. :440854 | Max. :0.18 | NA |

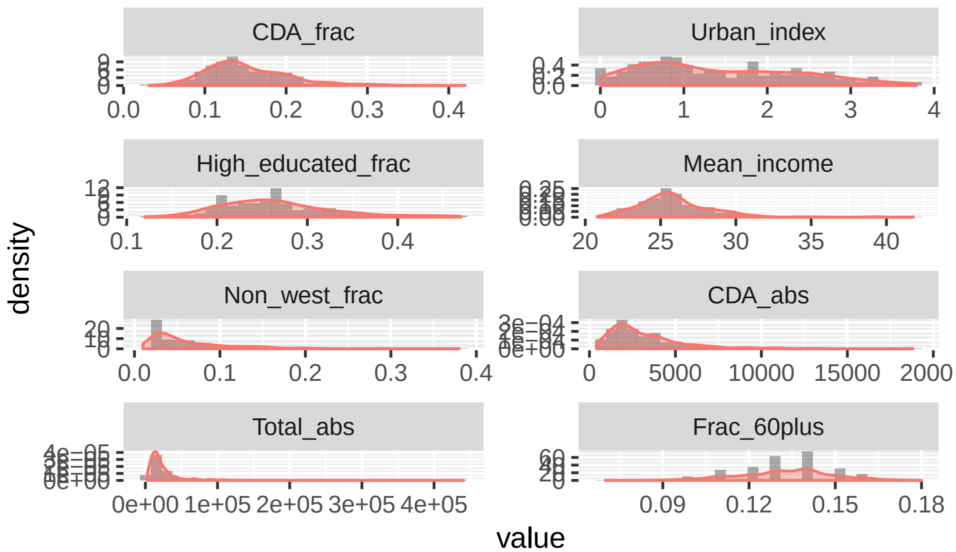



Figure 1: Density plot

**Correlation heatmap** In this heatmap (figure 2) 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.

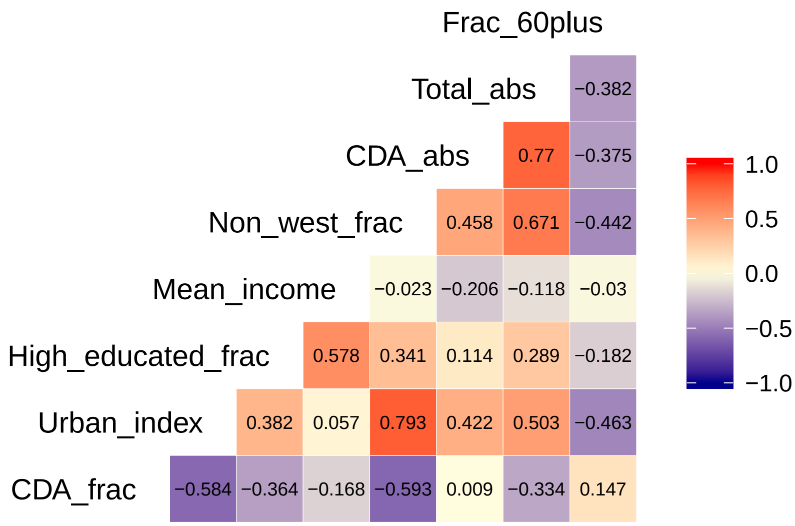



Figure 2: Correlation between explanatory and respons variables

**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.

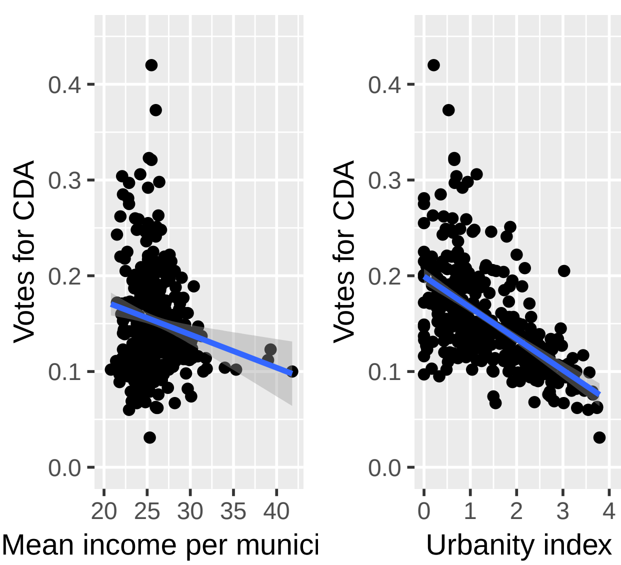



Figure 3: Scatterplots CDA

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

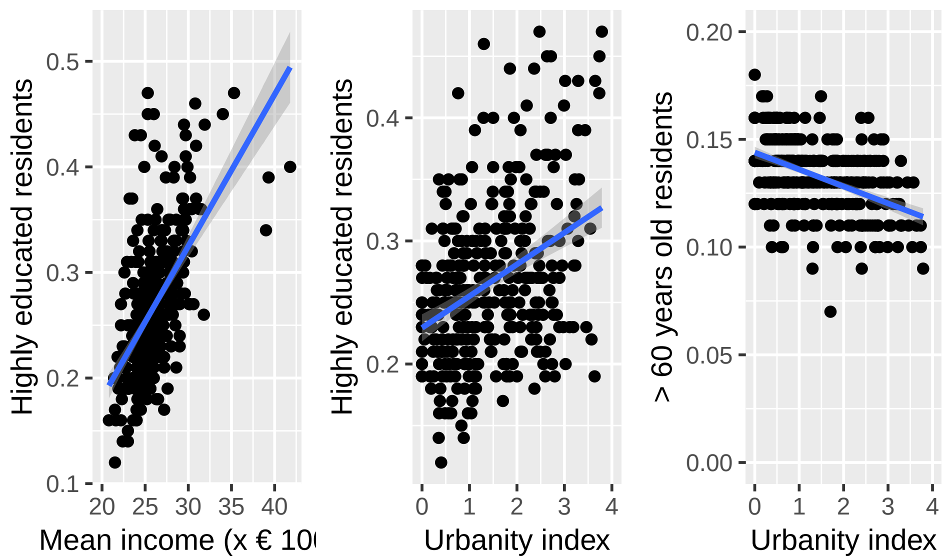



Figure 4: Scatterplot explanatory variables

**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.

## 2. Multiple linear regression

### First model

### Final model

### Cross validation

## 3. Logistic regression

The raw respons variable is the absolute amount of residents per municipality that voted for CDA. For linear regresssion, we transformed this variable to a fraction. However, we also know the total amount of votes per municipality. Therefore, a better fit to our data would be a binomial model. We use the logit as link function to transform the range of the respons. The choice for the logit was easily made. Because the inverse of the logit is directly interpretable as the log-odds ratio. This link displays the underlaying pattern of our data best. Below, the formula for our link function:

$\eta = log(\frac{\theta}{1-\theta})$

Where $\theta$ is the absolute amount of votes for CDA.

In GLM, we still have to make diagnostics plots to visualise if any of the assumptions for the error-term are violated. The assumptions made for the error-term in a binomial model are slightly different than for the linear model. Below we state the assumptions:

> Check deviance residuals
>
> Leverages are no longer just a function of X and now depend on the response through the weights W.

### First model

### Second model

### Final model

### Cross validation

## 4. Discussion

### Limitations