To what extent do voter age and level of education explain differences between SDP and Keskusta votership in Finland?

Miika Piiparinen

June 2023

1 Introduction

This project will concern two Finnish political parties and their votership in relation to population demographics of Finnish municipalities. The parties in question are the Social Democratic Party (SDP) and the Centre Party (KESK) (figure 1).

(a) SDP (b) Keskusta

Figure 1: Logos of both parties

The analysis will try to explain votership of SDP and Keskusta in terms of age and education. The analysis will be done on a municipality level.

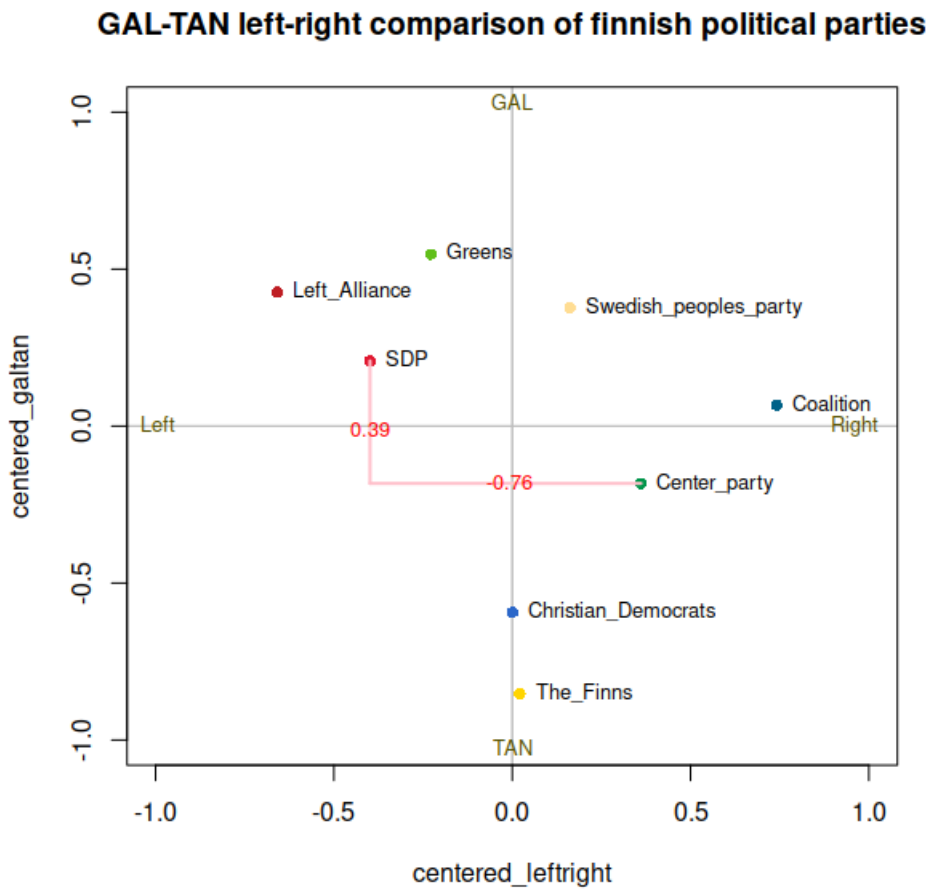

Figure 2: Δ visualized between SDP and KESK

1.1 Rationale for the research question

Let's consider the theoretical GAL-TAN framework, as well as the traditional economic left-right axis. The abbreviations stand for Green-Alternative-Liberal and Traditional-Authoritarian-Nationalist, and are a political science framework meant to distinguish between different sets of values held by political groups. Let's look at the distribution of Finnish parties between said axes [1] (figure 2). We can see that the difference in GAL-TAN is not as large as the difference in the traditional left and right axis.

What does this mean for our research question? We will consider the fact that the traditional axis of votership can be characterized, generally, through the lens of education and age as follows, as per the assumptions of the writer:

1) Universities and higher-education institutions tend to be left-leaning. Education would predict for voting left.

2) Education leads to higher wealth, higher wealth tends to vote right. Thus education would probably also predict for voting right.

3) Age tends to lead to higher wealth and more conservative values. Thus educated – wealthier – and old people would be more likely to vote right.

First assumption would seem to imply that SDP would be more popular among the educated, as it is in the left. Second assumption would imply that Center would be more popular among the educated, as it is in the right. To synthesize, the writer speculates that the age of a voterbase will correlate with a tendency to vote right, whereas education alone without wealth would predict for voting left.

Thirdly, the writer supposes that GAL-TAN axis is predicted by age. Older people tend to move towards TAN. Older people tend to also become wealthier. Thus, it seems logical to conclude that old age and education would predict for voting centre, whereas younger educated people would vote for SDP. Hence, this concludes the rationale for the research question.

Note that this analysis intently only looks at party position on the following map, and draws the rationale for the research question from my own first instincts. This is to test how accurate a pure value-based analysis is when compared to looking at precise party demographics such as KESK being a rural party or SDP being a workers, and to test my own political intuitions.

1.2 Concerning Data

The data used within this analysis are provided by Tilastokeskus. Three different files are used: Vaalitilastot [2], Koulutustilastot [3] and Väestörakenne [4], which concern electoral, educational and demographical data on each municipality. Data is aggregated on the basis of municipalities. Naturally within each municipality there can be – for example – sections of population that do not follow a general trend of education-level or age (f.e. school districts of a municipality could have clusters of young people despite a municipality being generally old). This effect is known as Simpsons Paradox [5].

Considering the prior, it can be that a deviating sub-sect of a municipality would be more inclined to vote than the people who go along with the population demographics. Elderly people tend to vote more, for example, so it would be a mistaken assumption to presume that a municipality with a low average age having certain voting preferences necessarily implies that young people in that municipality will have those voting preferences.

The paper will try to guide the reader to draw the right conclusions, and point out such issues whenever they could arise. As the raw data is not accessible by the writer, his capacity to comment on such issues is, sadly, limited.

2 Description of Data

This section is aimed at giving a rundown on the broad qualities and caveats of the data.

2.1 Votership and something about rurality

Let's look at the distribution of KESK and SDP voterships (figure 2). To note, the graphic depicts sorted municipalities in terms of the party being described. Thus the highest votership municipality of KESK (Toholampi) in the image does not coincide with highest voted municipality of SDP (Imatra).

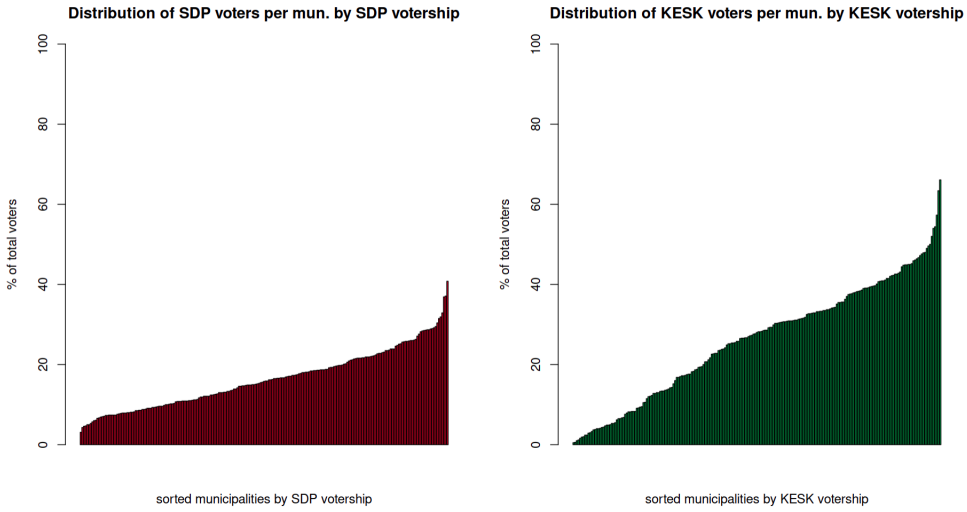

Figure 3: Share of voters per municipality sorted ascending by SDP and KESK voterbase respectively

Curiously, based on this plot and given the assumption that the municipalities in question are equally populated, we may assume KESK to be more popular than SDP. However, this assumption is mistaken, as seen in figure 4:

Here we can conclude that although KESK can compete with SDP in the municipalities with the least population, highest population municipalities are the ones where SDP manages to get the lead, and thus ultimately get most votes. The prior makes sense, as KESK strongly identifies itself as a "maalaispuolue", which loosely translates to "rural party". We can affirm this by sorting the municipalities by proportion living in rural areas (figure 5).

We can see that there is a general declining trend for SDP in rural municipalities, whereas the inverse is true for KESK. More about this in the analysis section, including precise information about the line of best fit.

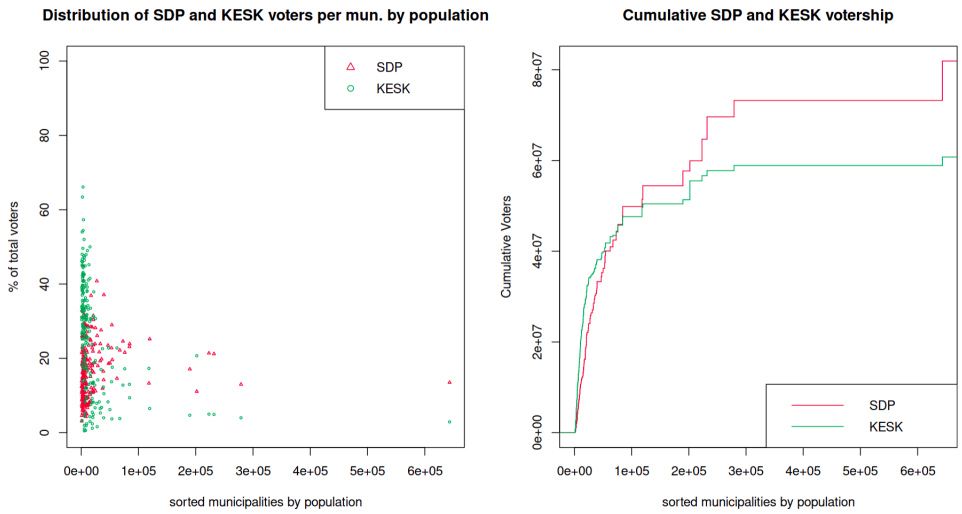

Figure 4: SDP and KESK votership proportionately (left) vs SDP and KESK votership absolute (right) with x-axis being sorted by municipality population

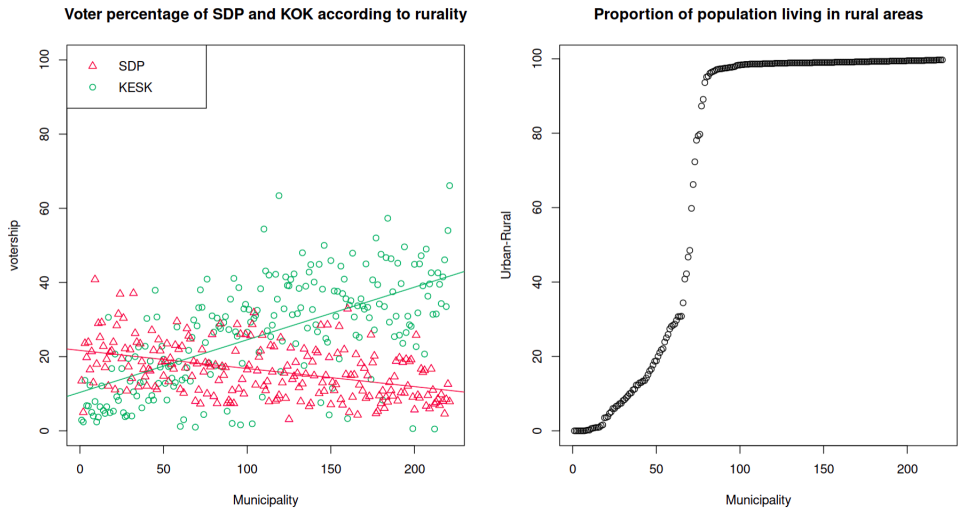

Figure 5: SDP and KESK votership proportionately (left) vs SDP and KESK votership absolute (right) with x-axis being sorted by municipality population in both images

2.2 Education Levels

The definition of education in the data occurs through seven different variables described in figure 6. Importantly, the data only concerns residents above 15 years of age, though the data also includes the percentages of municipality populations under that age bracket. This is done as not doing so would compromise certain variables, such as Vain_perusaste.

Name of variable	**Summary**	**Finnish mean %**
Vain_perusaste	Elementary education	23.27
Toinen.aste	High school, vocational training	43.47
Alin_korkea.aste	2-3 year long vocational degrees	7.84
Alempi_korkea.aste	3-4 years of vocational or a bachelors degree	9.59
Ylempi_korkea.aste	University education, usually masters	8.22
Tutkija.aste	Researcher, doctorate	0.89
VKTM.indeksi	Index describing education	309.68

Figure 6: Education variables

Understanding the VKTM index is going to be critical to understand this paper, so a brief description is in order. The VKTM index describes the time spent in education of every 20 year old by the mean duration of completion of their highest level of education [6].

2.3 Age Demographics

Age is split into three categories, below 15 years of age, working age and above 65 years of age. Figure 7 Describes the broad strokes of Finnish age demographics on a municipality level. We additionally gain a perspective towards distributions of people; top 13 municipalities contain 52.1% of people. Additionally, the highest population municipality (Helsinki) contains some 14% of the Finnish population.

Further we can inspect middle age and it's distribution between municipalities through figure 8. We notice that there tends to be a mild negative relationship between population of a municipality and it's middle age. Furthermore, middle age seems to be spread quite evenly among municipalities, ranging from 30 to 56 years of age, and only partially being determined by size of municipality.

This concludes the data exploration section.

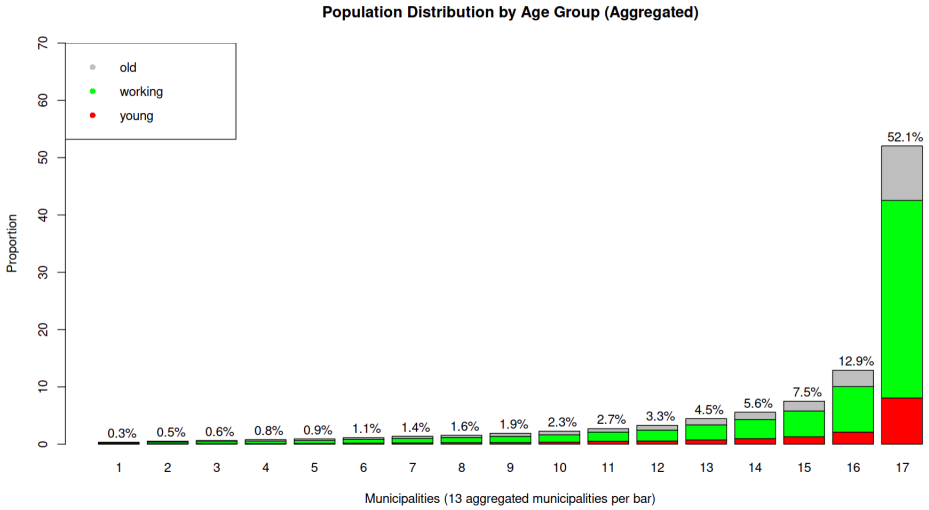

Figure 7: Population distribution and age distribution of 13 municipality clusters from low population to high. Old means larger than 65, working means between 15 and 65, young means below 15.

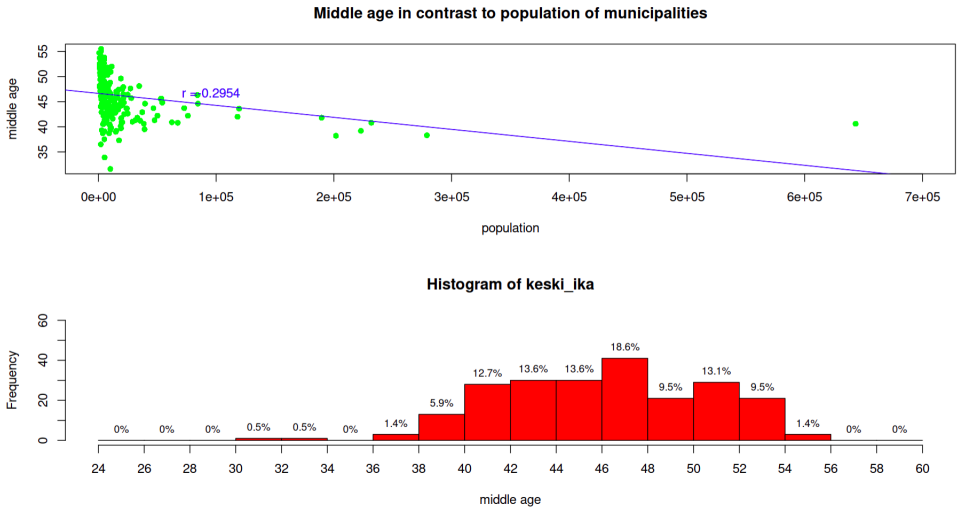

Figure 8: The relationship between middle age and size of municipalities (above) spread of middle age (below)

3 Methodology

The question we are analysing is as follows: To what extent do voter age and level of education explain differences between SDP and Keskusta votership in Finland?

3.1 Hypotheses

The hypotheses we are investigating are:

H1: As voter education level increases, SDP will be voted for
H2: As voter age increases, KESK will be voted for

And the respective null hypotheses are:

H1:0: There is no significant relationship between SDP votership and education
H2:0: There is no significant relationship between KESK votership and age

3.2 Statistical tools used

This analysis will primarily use linear regression. The hypotheses will be answered through simple linear models, and a more robust answer to the research question will be considered via a multiple regression model.

3.3 Why regression?

Linear regression is a simple and elegant method of investigating whether there exists a directional relationship between variables. The method is limited, as there is an expectation of linearity, but even in the case where perfect linear relationship is not achieved, with proper visualization the analysis will explain some tendency of the underlying data.

3.4 Assumptions of linear models

Linear regression in the case of multiple independent variables expects that the variables are not correlated with each other heavily. This is prominent in 4.4, where the variables must correlate with each other. This is taken into account in the analysis as far as possible, but with the research question in hand it's arguable that this problem is unavoidable. The scope of this analysis failed to account for homoscelasticity and other more advanced limitations due to time constraints.

4 Analysis

Now we will concern ourselves with the research question at hand. This analysis will constrain itself to primarily a few variables, those being the ones that were given attention in section 2.

4.1 Linear regression analysis

Linear regression analysis is about forming a line of best fit through data points. This occurs in the form of an equation, as follows:

$$Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ...x_{in}\beta_n$$

Generally, linear models include dependent and independent variables. In this paper, the dependent variable Y is always either SDP or KESK. Linear regression has a few key statistics that we should understand.

1) Coefficient of variable, or the β terms for each respective x_i. This describes the slope of the line drawn through data. It is important to note that scaling of data in a similar way is vital to be able to do 1:1 comparisons of coefficients. In this analysis the slope magnitude is only looked at in specific scenarios. The sign of the coefficient is potentially more important.

2) Standard error. Estimates the variance of the distances of each point from the regression line. The smaller this value, the better the fit.

3) T-statistic. Describes the ratio of the coefficient to the standard error. Very roughly: higher the t-value, the more significant the model. We will roughly state that a null will be rejected if the absolute value of the t-statistic is higher than 2.

4) P-value: The probability, given that there is no true relationship between the independent and depenedent variables, of receiving a sample at least as extreme as this one. Generally speaking, a p-value of less than 0.05 is said to be an indicator of being able to reject the null hypothesis.

5) $\mathbf{R^2}$: Once again very roughly: what percentage of the variation in the results is explained by the independent variables? I will use the term "robust" to describe a model with a high $\mathbf{R^2}$ value.

We will initially look at simple linear models with only one independent variable in order to neatly think about the prior hypotheses, but in 4.4. we'll also look at a multiple regression model.

4.2 SDP Votership and Education

We can consider the first pair of hypotheses through linear models.

H1: As voter education level increases, SDP will be voted for
H1:0: There is no significant relationship between SDP votership and education

Let's first plot SDP votership per municipality and contrast this to the respective VTKM index (figure 9). We can interpret some slight relationship with this metric. However, we have more variables to consider when looking at education. Figure 9 contains a linear model of each of them.

Clearly we can see that while those with elementary education tend to vote for SDP less, those with a professional degrees are the primary voterbase for the party when looking at education. Higher education levels tend to have a positive relationship too.

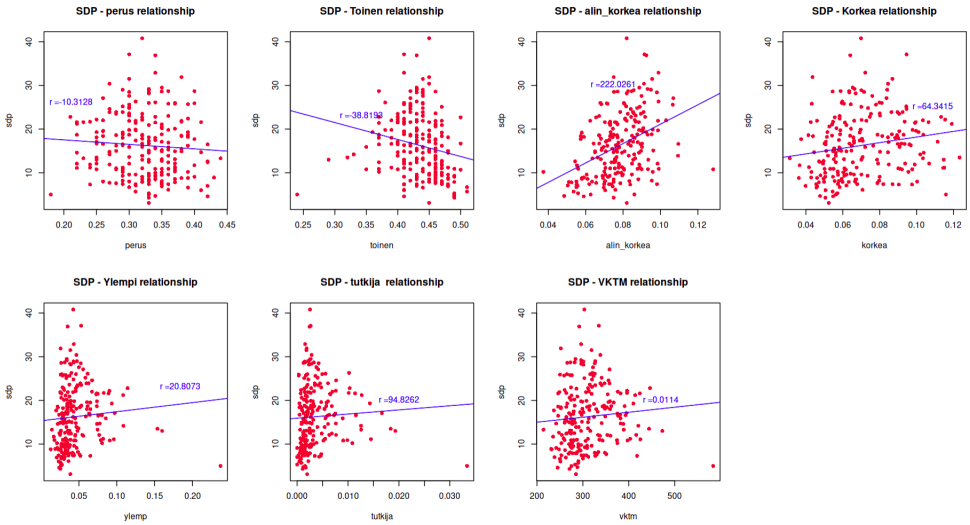

Figure 9: SDP linear relationship with education variables

We can also describe further qualities of each model. They're present figure 10. Their significance is explained in section 4.1. It is important to note that all others are somewhat comparable par from VKTM, as it is not on a percentage-based scale.

The standard error in most cases is telling us that there isn't that much of a linear trend going on, especially when combined with the information from the plots. The best performer here is Vain_perusaste, but even then the standard error leaves a lot to be desired.

Looking at the t-statistic, we have more problems. The standard critical 95% significance region is above the absolute value of 2. We can see that only

Variable	Coefficient	Standard error	t statistic	p-value	R^2
Vain_perusaste	-10.313	9.086	-1.052	0.294	0.005
Toinen.aste	-38.819	12.865	-3.017	0.00285	0.03
Alin_korkea.aste	222.026	34.708	6.397	9.47e-10	0.15
Alempi_korkea.aste	64.34	24.59	2.617	0.0095	0.003
Ylempi_korkea.aste	20.8073	19.0568	1.092	0.276	0.005
Tutkija.aste	94.8262	125.3436	0.757	0.45	0.002
VKTM.indeksi	0.011444	0.009468	1.209	0.228	0.006

Figure 10: Key Statistics in variables as explainers for SDP votership

Toinen.aste, Alin_korkea.aste and Alempi_korkea.aste are significant enough to be rejected.

When we look at the P-values of each variable, we can see that all of the p-values are below the accepted treshhold of 0.05, though Tutkija.aste remains pretty close to it. Appears that based on these we can reject the null hypothesis. However, when combined with the information from the t-statistic, we can see that the ones that pass the t-test are also the ones with a definitively low p-value.

Finally, R-squared is quite small in all other models except from Alin_korkea.aste, in the case of which we actually see that the variable explains some 15% of the total variance. No other variable comes close to this.

So, based on these models, the variables that do good prediction appear to be Toinen.aste, Alin_korkea.aste and Alempi_korkea.aste. Even these have a somewhat non-linear trend and a high standard error. It is also important to consider that our hypothesis requires us to operationalize education-level as a concept, which I think would include the weaker variables. It appears that there isn't a holistic relationship, as the VKTM.indeksi remains insignificant according to standard t-statistic conventions.

In conclusion, H1 was not specific enough. However, if we interpret it as is, so we mainly look at VKTM.indeksi and the other variables in aggregate, we fail to reject the null-hypothesis, and we do not get a very robust prediction. If we allow freedom of interpretation, some types of education, even some higher education, are positively related to SDP votership. If we'd only look at Alin_korkea.aste, the prediction is actually both significant and robust. Only having a secondary education is negatively related, and also explains about 3% of the variance. A more specific formulation of H1 that is only looking at these variables would reject the null.

4.3 KESK votership and age

Now we will consider the second hypothesis. Let's make some regression models where we conceptualize KESK votership through the four aforementioned age-related variables (figure 11). Keep in mind that keski-ika is on a different scale than nuori, tyoika and vanhoja.

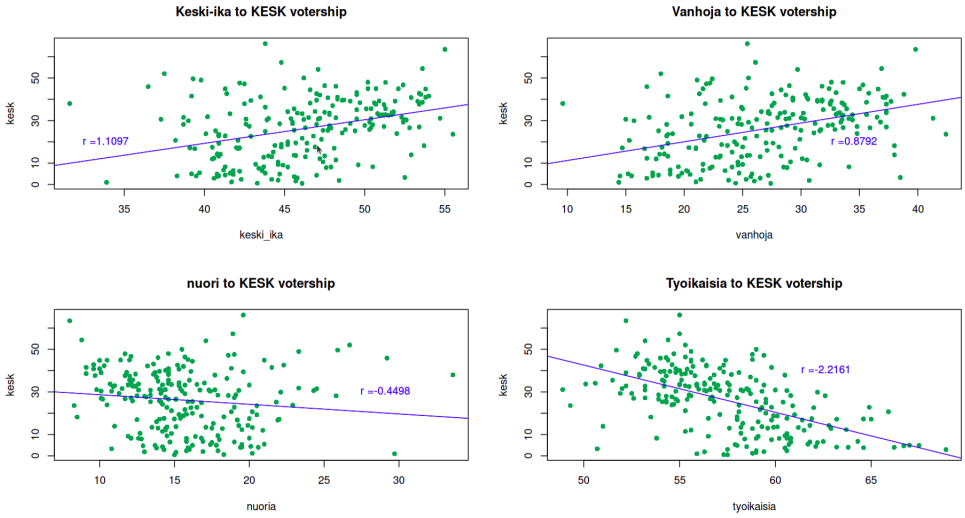

Figure 11: KESK votership relationship to age

We should also look at the significance of the relationships, presented in figure 12.

Variable	Coefficient	Standard error	t statistic	p-value	R^2
keski-ika	1.110	0.203	5.467	1.24e-07	0.1201
vanhoja	0.8792	0.1394	6.305	1.57e-09	0.1536
nuori	-0.4498	0.2384	-1.887	0.0605	0.016
tyoika	-2.2161	0.2239	-9.90	¡2e-16	0.3092

Figure 12: Key Statistics in variables as explainers for KESK votership

Age seems to generally a pretty robust predictor for KESK votership. All of the variables solidly pass the p-value test, with nuori being the only one that fails at the used boundaries of the t-statistic. We can see that the R squared for all other par from nuori are also very formidable, suggesting that a lot of variance is explained between the variables.

Due to the formulation of H2 we can say that old age does predict for KESK votership in a reasonable manner. Thus we fail to reject the null hypothesis.

4.4 Is KESK votership positively predicted for more by age and education than SDP?

Finally, we would like to look at both of the variables being explained by a similar model. Let's create a multiple regression model!

After a lot of tinkering a model with the variables depicted seems quite handy at answering our research question. Let's look at the following models:

```
Call:
lm(formula = sdp ~ keski_ika + vanhoja + vktm + alin_korkea +
    korkea + toinen + tyoikaisia)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4137  -4.6649  -0.3203   4.3688  22.5392

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.67317   19.91609   0.435 0.663651
keski_ika    -1.22089    0.83479  -1.463 0.145076
vanhoja       0.98506    0.69512   1.417 0.157911
vktm         -0.08807    0.02634  -3.343 0.000979 ***
alin_korkea 230.68453   44.26808   5.211 4.43e-07 ***
korkea       96.64650   66.71249   1.449 0.148891
toinen      -28.57240   15.46350  -1.848 0.066028 .
tyoikaisia    0.91117    0.37855   2.407 0.016937 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.439 on 213 degrees of freedom
Multiple R-squared:  0.2364,    Adjusted R-squared:  0.2113
F-statistic: 9.421 on 7 and 213 DF,  p-value: 3.598e-10
```

Figure 13: KESK votership relationship to age

We can see that both models are significant, and that both contain the same independent variables but explain for a different dependent variable. We can

```
Call:
lm(formula = kesk ~ keski_ika + vanhoja + vktm + alin_korkea +
    korkea + toinen + tyoikaisia)

Residuals:
    Min      1Q  Median      3Q     Max
-33.271  -5.885   0.316   4.435  31.821

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   22.52271   28.08481   0.802  0.42347
keski_ika      3.80307    1.17719   3.231  0.00143 **
vanhoja       -2.79423    0.98023  -2.851  0.00479 **
vktm           0.05859    0.03715   1.577  0.11625
alin_korkea -426.41583   62.42495  -6.831 8.70e-11 ***
korkea       -52.62073   94.07510  -0.559  0.57651
toinen       144.31723   21.80596   6.618 2.90e-10 ***
tyoikaisia    -2.44845    0.53381  -4.587 7.68e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.08 on 213 degrees of freedom
Multiple R-squared:  0.6171,    Adjusted R-squared:  0.6046
F-statistic: 49.05 on 7 and 213 DF,  p-value: < 2.2e-16
```

Figure 14: KESK votership relationship to age

also see some interesting and counter-intuitive relationships with the variables. The most significant predictors for SDP votership are now vktm and alin_korkea, only the latter of which was initially analyzed to be significant. Age as a variable group seems to be insignificant in the terms of the model, as only tyoikaisia passes our tests of significance. 23% of SDP variance is explained with the variables in the model. It seems that adding old age variables to the model increases the robustness somewhat, but generally the model is only significant when it comes to the tyoikaisia variable in this context. This makes sense, as the SDP party is very much a workers' party. However it also seems that as middle age decreases, SDP is voted for more, but at the same time the proportion of old people seems to be related to SDP votership positively. This effect cannot be explained in the context of this analysis, but the variables in it are not the most significant ones. Another interesting note is that vktm is negatively related to SDP votership. This goes against my initial speculations as well as the results presented in 4.1. However, generally in the model education seems to positively predict for SDP votership.

As we look at Keskusta, we can immediately see that the R^2 is almost triple that of SDP. This is interesting. It would also appear that education on all fronts except on the secondary is negatively related to Keskusta votership. Further analysis – as done with SDP and education – could be interesting to see how Keskusta and education relate to eachother. We can also see that many variables in the model are significant, with the exception of vktm and korkea. This is somewhat odd, as I would have expected a negative significant relationship. However, even odder still is the fact that tyoikaisia and vanhoja are both negative trends, completely contrary to the age-based analysis we performed in 4.2. Yet, keski_ika is positively related. This could be due to interrelation between the variables that the multiple regression model expresses like this. The most interesting thing is that alin_korkea is also a very significant predictor here. It appears that my theory in the beginning was wrong. Age and education do not predict for KESK votership. Education is actually very much negatively related to Keskusta votership.

5 Discussion

The general ideas uncovered during this study are the following:

1) SDP is a worker's party, such that having some professional education or a bachelors' predicts robustly and significantly for SDP votership, that being about 15% of variance explained.

2) Keskusta votership is predicted for by older age. This could be because of the fact that middle age tends to be higher in high population municipalities (illustrated in figure 8), which tend to be less rural; Keskusta is a rural party, after all.

3) Considering a multiple regression model in the case of SDP, looking at only age and education, we can get to at least 23% of variance explained. It appears that more important factors are at play here.

4) Considering a multiple regression model in the case of Keskusta, looking at only age and education, we can get to at least 61% of variance explained. This is a very robust model, with many of the variables being significant.

5) The stand-out variable in the analysis is alin_korkea, predicting with great significancy in both multiple regression models.

We can notice that my initial theorization was off in terms of Keskusta. Although old age predicts for Keskusta, it in tandem with education seems to not given an unanomous positive trend. Education on almost all fronts seems to indicate lsser likelihood of keskusta votership. This could be because older generations are less educated as a whole, especially when looking at farmers, who would be the main demographic of the party.

When looking at SDP, my original analysis was somewhat more accurate. Education, though not especially high education, predicts for SDP votership, and at least in the multiple regression model there were mixed associations between age and SDP votership, with keski_ika being a negative predictor, while vanhoja and tyoikaisia had a positive one. However, only the last is statistically significant, implying that my analysis failed to consider that SDP is a worker's party. Thus, ultimately the analysis failed.

6 Conclusion

The initial rationale of the research question was – expectedly – misguided. Simply looking at party positions on a GAL-TAN left-right graphic does not explain all the context required to draw conclusions about votership (rural party, workers' party). Furthermore, some of my speculations could have been faulty to begin with. However, when looked at as a jump-off point toward proper analysis, one can say that these findings have been fruitful. The models robustly and significantly predict for SDP and Keskusta votership, especially with the latter. Furthermore, the analysis has clarified the characteristics of each voterbase to a reasonable degree.

7 sources

1: Åbo akademi, Voting and public opinion 2019, https:/
/www.abo.fi/wp-content/uploads/2019/10/
Voting_and_Public_Opinion_2019_second_edition_Digi.pdf
2: vaali (given in assignment)
3: koulutus (given in assignment)

4: väestö (given in assignment)
5: Wikipedia, Simpsons Paradox https://en.wikipedia.org/wiki/Simpson%27s_paradox
6: MayorsIndicators, Hyvä koulutus https://www.mayorsindicators.com/
index.cfm?area=indicator&g_id=4&i_id=93&charttype=line

8 R-code

Code for figure 2:

```
galtan <- list(Greens=3.53,
           Left_Alliance=3.41,
           Swedish_peoples_party=3.36,
           SDP=3.19,
           Coalition=3.05,
           Center_party=2.80,
           Christian_Democrats=2.39,
           The_Finns=2.13)
leftright <- list(Greens=1.92,
               Left_Alliance=1.49,
               Swedish_peoples_party=2.31,
               SDP=1.75,
               Coalition=2.89,
               Center_party=2.51,
               Christian_Democrats=2.15,
               The_Finns=2.17)
coldict = c(
  '#61BF1A',
  '#BF1E24',
  '#FFDD93',
  '#E11931',
  '#006288',
  '#01954B',
  '#2B67C9',
  '#FFD500'
)
mean_galtan <- mean(unlist(galtan))
centered_galtan <- lapply(galtan, function(x) x - mean_galtan)
mean_leftright <- mean(unlist(leftright))
centered_leftright <- lapply(leftright, function(x) x - mean_leftright)
pos <- plot(centered_leftright, centered_galtan, col=coldict, pch=16,
     main = "GAL-TAN left-right comparison of finnish political parties",
     xlim = c(-1,1),
     ylim = c(-1,1),
     panel.first = c(abline(v = c(0,0), lwd = 1, lty = 1, col='grey'),
                     abline(h = c(0,0), lwd = 1, lty = 1, col='grey')))
text(centered_leftright, centered_galtan,
```

```
labels = names(centered_galtan), pos = 4, cex = 0.8)
sdp_x_diff <- centered_galtan$SDP - centered_galtan$Center_party
sdp_y_diff <- centered_leftright$SDP - centered_leftright$Center_party
segments(centered_leftright$SDP, centered_galtan$SDP,
         centered_leftright$SDP, centered_galtan$Center_party,
         col = "pink", lwd = 2)
segments(centered_leftright$SDP, centered_galtan$Center_party,
         centered_leftright$Center_party, centered_galtan$Center_party,
         col = "pink", lwd = 2)
text(centered_leftright$SDP, centered_galtan$Center_party,
     labels = round(sdp_x_diff, 2), pos = 3, col = "red", offset = 1.5,cex=0.8)
text(centered_leftright$Center_party, centered_galtan$Center_party,
     labels = round(sdp_y_diff, 2), pos = 2, col = "red", offset = 3.5,cex=0.8)
text(-0.9, 0, "Left", pos = 2, col = "#555000", cex = 0.8)
text(0.85, 0, "Right", pos = 4, col = "#555000", cex = 0.8)
text(0, 1.1, "GAL", pos = 1, col = "#555000", cex = 0.8)
text(0, -1.1, "TAN", pos = 3, col = "#555000", cex = 0.8)
```

Code for figure 3

```
sdp <- merged_df$SDP
order_sdp <- order(sdp)
sdp_ordered <- sdp[order_sdp]
kesk <- merged_df$KESK
order_kesk <- order(keskusta)
kesk_ordered <- kesk[order_kesk]
par(mfrow=c(1,2))
barplot(sdp_ordered, col="#E11931",
        main='Distribution of SDP voters per mun. by SDP votership',
        ylim=c(0,100),
        ylab='% of total voters',
        xlab='sorted municipalities by SDP votership')
barplot(kesk_ordered, col='#01954B',
        main='Distribution of KESK voters per mun. by KESK votership',
        ylim=c(0,100),
        ylab='% of total voters',
        xlab='sorted municipalities by KESK votership')
```

Code for figure 4

```
par(mfrow=c(1,2))
plot(vaesto_ordered, sdp_ordered, col="#E11931",
     main='Distribution of SDP and KESK voters per mun. by population',
     ylim=c(0,100),
     ylab='% of total voters',
     xlab='sorted municipalities by population',
     pch=2,
```

```
     cex=0.4)
points(vaesto_ordered, kesk_ordered, col='#01954B',cex=0.4)
legend("topright", legend=c("SDP", "KESK"), col=c("#E11931", "#01954B"), pch=c(2,1))
order_vaesto <- order(merged_df$Vaesto)
vaesto_ordered <- merged_df$Vaesto[order_vaesto]
sdp_cumulative <- c(0,
cumsum(merged_df$SDP[order_vaesto]*merged_df$Vaesto[order_vaesto]))
kesk_cumulative <- c(0,
cumsum(merged_df$KESK[order_vaesto]*merged_df$Vaesto[order_vaesto]))
sdp_stepfun <- stepfun(vaesto_ordered, sdp_cumulative)
kesk_stepfun <- stepfun(vaesto_ordered, kesk_cumulative)
plot(sdp_stepfun, col = "#E11931", lty = 1, ylim = c(0, max(sdp_cumulative)),
     main = "Cumulative SDP and KESK votership",
     xlim = c(0, max(merged_df$Vaesto)),
     ylab = "Cumulative Voters", xlab = "sorted municipalities by population",
     do.points=FALSE)
lines(kesk_stepfun, col = "#01954B", do.points=FALSE)
legend("bottomright", legend = c("SDP", "KESK"), col = c("#E11931", "#01954B"), lty = 1)
```

Code for figure 5

```
par(mfrow = c(1, 2))
ordered <- order(rural)
rural_sorted <- rural[ordered]
kesk_sorted <- merged_df$KESK[ordered]
sdp_sorted <- merged_df$SDP[ordered]
plot(kesk_sorted, col = '#01954B',
     main = 'Voter percentage of SDP and KOK according to rurality',
     ylab='votership',
     xlab='Municipality',
     ylim = c(0, 100))
points(sdp_sorted, col = "#E11931", pch = 2)
lm_sdp <- lm(sdp_sorted ~ seq_along(sdp_sorted))
abline(lm_sdp, col = "#E11931", lty = 1)
lm_kesk <- lm(kesk_sorted ~ seq_along(kesk_sorted))
abline(lm_kesk, col = "#01954B", lty = 1)
legend("topleft", legend = c('SDP', 'KESK'), col = c("#E11931", "#01954B"), pch = c(2, 1))
plot(rural_sorted,
     main = 'Proportion of population living in rural areas',
     ylab = 'Urban-Rural',
     xlab = 'Municipality')
```

Code for figure 7

```
sorted_df <- merged_df[order(merged_df$Vaesto),]
sorted_df$Alue
length(sorted_df$Alue)
```

```
nuori <- (sorted_df$Alle15Osuus * sorted_df$Vaesto) / sum(sorted_df$Vaesto)
tyo <- (sorted_df$TyoikaistenOsuus * sorted_df$Vaesto) / sum(sorted_df$Vaesto)
vanh <- (sorted_df$Yli65Osuus * sorted_df$Vaesto) / sum(sorted_df$Vaesto)
nuori_aggregated <- tapply(nuori, ceiling(seq_along(nuori)/13), sum)
tyo_aggregated <- tapply(tyo, ceiling(seq_along(tyo)/13), sum)
vanh_aggregated <- tapply(vanh, ceiling(seq_along(vanh)/13), sum)
par(mfrow = c(1, 1))
x <- 1:length(nuori_aggregated)
stacked_heights <- rbind(nuori_aggregated, tyo_aggregated, vanh_aggregated)
total_percentages <- unname(nuori_aggregated + tyo_aggregated + vanh_aggregated)
locs <- barplot(stacked_heights, col = c('red', 'green', 'grey'),
        main = 'Population Distribution by Age Group (Aggregated)',
        xlab = 'Municipalities (13 aggregated municipalities per bar)',
        ylab = 'Proportion',
        ylim = c(0,70),
        beside = FALSE)
text(x = locs+0.075, y = stacked_heights[1, ] + stacked_heights[2, ] + stacked_heights[3, ],
     labels = paste0(round(total_percentages, 1), "%"), pos = 3, cex = 1, offset = 0.3)
legend("topleft", c('old','working','young'), pch=c(16,16,16),col=c('grey','green','red'))
```

Code for figure 8

```
par(mfrow = c(2, 1))
keski_ika <- merged_df$Keski.ika[order(merged_df$Vaesto)]
vaesto <- merged_df$Vaesto[order(merged_df$Vaesto)]
plot(vaesto, keski_ika, xlim = c(0, 700000),
     main = 'Middle age in contrast to population of municipalities', col = 'green',
     pch = 16, xlab='population', ylab='middle age')
fit <- lm(keski_ika ~ vaesto)
abline(fit, col = 'blue')
cor_value <- cor(keski_ika, vaesto)
text(x = 0.1 * max(vaesto), y = max(keski_ika)-9,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
hist(keski_ika, breaks = seq(24, 60, 2),
col = 'red', xaxt = 'n', ylim = c(0, 60), xlab='middle age')
axis(1, at = seq(24, 60, 2), labels = seq(24, 60, 2))
bin_counts <- hist(keski_ika,
                   breaks = seq(24, 60, 2),
                   plot = FALSE)$counts
total_count <- sum(bin_counts)
bin_percentages <- round(bin_counts / total_count * 100, 1)
text(x = seq(25, 59, 2), y = bin_counts,
labels = paste0(bin_percentages, "%"), pos = 3, cex = 0.8)
```

Code for figure 9

```
par(mfrow=c(2,4))
fit = lm(sdp~perus)
plot(perus , sdp,
     main='SDP - perus relationship',
     xlab='perus',
     ylab='SDP',
     col='#E11931', pch=16)
abline(fit, col = 'blue')
cor_value <- unname(fit$coefficients["perus"])
text(x = 0.17, y =26,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
summary(fit)
fit = lm(sdp~toinen)
plot(toinen , sdp,
     main='SDP - Toinen relationship',
     col='#E11931', pch=16)
abline(fit, col = 'blue')
cor_value <- unname(fit$coefficients["toinen"])
text(x = 0.3, y =23,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
summary(fit)
fit = lm(sdp~alin_korkea)
plot(alin_korkea , sdp,
     main='SDP - alin_korkea relationship',
     col='#E11931', pch=16)
abline(fit, col = 'blue')
cor_value <- unname(fit$coefficients["alin_korkea"])
text(x = 0.06, y =30,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
summary(fit)
fit = lm(sdp~korkea)
plot(korkea , sdp,
     main='SDP - Korkea relationship',
     col='#E11931', pch=16)
abline(fit, col = 'blue')
cor_value <- unname(fit$coefficients["korkea"])
text(x = 0.095, y =25,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
summary(fit)
fit = lm(sdp~ylemp)
plot(ylemp , sdp,
     main='SDP - Ylempi relationship',
```

```
     col='#E11931', pch=16)
abline(fit, col = 'blue')
cor_value <- unname(fit$coefficients["ylemp"])
text(x = 0.15, y =23,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
summary(fit)
fit = lm(sdp~tutkija)
plot(tutkija , sdp,
     main='SDP - tutkija  relationship',
     col='#E11931', pch=16)
abline(fit, col = 'blue')
cor_value <- unname(fit$coefficients["tutkija"])
text(x = 0.015, y=20,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
summary(fit)
fit = lm(sdp~vktm)
plot(vktm , sdp,
     main='SDP - VKTM relationship',
     col='#E11931', pch=16)
abline(fit, col = 'blue')
cor_value <- unname(fit$coefficients["vktm"])
text(x = 420, y =20,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
summary(fit)
```

Code for figure 11

```
kesk <- merged_df$KESK
keski_ika <- merged_df$Keski.ika
nuoria <- merged_df$Alle15Osuus
tyoikaisia <- merged_df$TyoikaistenOsuus
vanhoja <- merged_df$Yli65Osuus
fit <- lm(kesk~keski_ika)
plot(keski_ika, kesk, col="#01954B", pch=16, main="Keski-ika to KESK votership")
abline(fit, col='blue')
cor_value <- unname(fit$coefficients["keski_ika"])
text(x = 32, y=20,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
summary(fit)
fit <- lm(kesk~vanhoja)
plot(vanhoja, kesk, col="#01954B", pch=16, main="Vanhoja to KESK votership")
abline(fit, col='blue')
```

```
cor_value <- unname(fit$coefficients["vanhoja"])
text(x = 32, y=20,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
summary(fit)
fit <- lm(kesk~nuoria)
plot(nuoria, kesk, col="#01954B", pch=16, main="nuori to KESK votership")
abline(fit, col='blue')
cor_value <- unname(fit$coefficients["nuoria"])
text(x = 27, y=30,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
summary(fit)
fit <- lm(kesk~tyoikaisia)
plot(tyoikaisia, kesk, col="#01954B", pch=16, main="Tyoikaisia to KESK votership")
abline(fit, col='blue')
cor_value <- unname(fit$coefficients["tyoikaisia"])
text(x = 61, y=40,
     labels = paste0("r =", round(cor_value, 4)),
     pos = 4, col = 'blue')
summary(fit)
```

Multiple regression model

```
fit_sdp <- lm(sdp~keski_ika+vanhoja+vktm+alin_korkea+korkea+toinen+tyoikaisia+rural)
fit_kesk <- lm(kesk~keski_ika+vanhoja+vktm+alin_korkea+korkea+toinen+tyoikaisia+rural)

summary(fit_sdp)
summary(fit_kesk)
```