Examination numbers

415 (UK exam), 11 (DK exam), 4 (DK exam) and 15 (DK exam)

Course title:

Social Data Science,
Winter Exam 2015

Submission date:
December 15, 2015

Number of pages including front page:

20

Preface

This reports examines a range of research questions from within the field of political science and Danish politics, using a combination of data science tools and econometric methods. Throughout the report, we use a data set containing the responses of almost every Danish candidate for the General Election 2015 (Folketingsvalg), to the most popular Voting Advice Application (VAA); a survey consisting of fifteen political questions that voters use to find the candidates with whom they agree the most. The data has been scraped from Danmarks Radio; the public, Danish, national broadcasting service who hosted the VAA. More than 25% of voters completed the test up to the general election in 2015.

The three main research questions we examine in this report are,

1. How do candidates' political positions fit with common theories of the Danish political landscape?

2. Are the Danish political parties distinguishable from one another?

3. Do candidates who are distinct relative to their peers receive more personal votes?

While an examination of any single of the above questions could easily have occupied all pages in this report, we have instead chosen to attempt an answer of each. Two main reasons have been behind this. First, we have wanted to demonstrate the broad array of research possibilities that lies in examining VAA data - including some that previous literature seem to not have dealt with. Second, and in line with the previous argument, we have wanted to employ a range of different data science methods: Question (1) deals with unsupervised learning in the form of dimensionality reduction, Question (2) uses a standard data scientific tool for supervised learning whereas Question (3) focuses on classical econometric inference.

The limited scope of this report has unfortunately not allowed us to provide a lengthy introduction to Danish politics, the electorate system and the different political parties. Hence, readers who possess some basic knowledge within this field are likely to find the report more interesting.

We hope that the contents of this report may indicate directions for further research projects within the crossing between political science, data science and econometrics.

Susanne Sundgaard Hansen, Dennis Hansen, Ann-Sofie Hansen, Oskar Harmsen

University of Copenhagen, December 2015

Introduction

The Danish political system is based on representative democracy, in which voters select representatives for four-year periods. Voters can vote for either a party or a specific candidate within the voter's constituency.[1] In recent elections, Voting Advice Applications (VAAs) have becoming an increasingly familiar part of the process. In short, a Voting Advice Application (VAAs) is a test that helps voters find candidates with whom she agrees the most. According to recent literature, VAAs have become crucial in electoral campaigns in several countries, including Denmark, and are thought by some to have a sizable impact on election outcomes (Ladner et. al., 2012).

This report is organized in three different parts, each examining a different research question. All parts use data on candidate responses to the largest Danish VAA from the 2015 General Election (Folketingsvalg) matched with election outcomes for each candidate.

Part 1 examines the Danish political landscape. By reducing the political positions of every candidate into two political dimensions, we can construct a picture of the current Danish political landscape. We make two primary contributions: First, by using candidate level data, our picture of the Danish political landscape can document both the spread within and the overlaps between political parties, contrary to methods using 'expert' interviews or coding of party programs. To the best of our knowledge, we are the first to do this on Danish VAA data, although it has been done for other countries. Second, the political dimensions we present do not coincide perfectly with standard theories of GAL - TAN dimensions (or 'fordelingspolitik' and 'værdipolitik'), suggesting that literature on Danish political dimensions might be suited for revision.

Part 2 examines the party affiliation of each candidate. We construct a supervised statistical learning model, for predicting party affiliation based on the VAA responses. Examining the model and its predictive power allows us to see what issues are most important in predicting party affiliation, and why some party affiliations may be easier to predict. We conclude that parties that have a combination of a large positional spread and overlaps with other parties are the least distinguishable.

Part 3 turns to the question of VAAs impact on election outcomes. The issue has been discussed extensively in the literature, and is of interest not just for academics, but as a general objective of obtaining a deeper understanding of how democracies work. Entangling the effects of VAAs has however proven difficult - previous literature mostly relies on experimental designs where groups of voters are asked more or less directly whether they were affected by VAA results (see e.g. Ladner et al. 2012). We present a new method that could indicate VAA impact, testing whether candidates that are more distinct receive more personal votes. We find that the opposite seems to hold true: candidates who agree the most with their peers generally receive more votes.

Dataset

Our dataset consists of the responses of candidates in the 2015 general election, to the most popular Danish Voting Advice Application - Danmarks Radio's "Kandidattest". This particular

[1] While election results is based on a rather complex interaction between personal votes, party votes, constituency borders, our analysis focuses only on personal votes. This measure can be assumed to be most strongly linked to individual candidates' performance, and is a strong predictor for whether a candidate is elected.

VAA is heavily used and thought to play an important role in the election - more than 1 million unique users completed the test, corresponding to approximately 25% of eligible voters. The candidates are asked to specify their agreement to 15 political statements on a five-point Likert scale, ranging from 'highly disagree' (DK: meget uenig) to 'highly agree' (DK: meget enig). The questions vary from tariffs on cigarettes over public sector growth to the amount of influence given to EU, and represents both issues that were specific to the time around the election as well as broader questions - a full list of questions and translations can be found in the appendix. Voters completing the test are presented with two list of candidates from their own constituency with whom they respectively agree the most and the least, and with links to more information on each candidate. In Part 3 of this report, we formally introduce the definition of agreement.

We have scraped the responses of candidates directly from DR's website. Of the 799 candidates running in the election, 724 (91%) completed the VAA, representing 93% of the personal votes given (source[2])[3]. The five candidates outside the parties (independents) are not included in the analyses in this report. None of them were elected. To this primary dataset, we have added background data for each candidate (gender, age, current position, whether they ran for parliament at the previous election and whether they were elected at this election or not, and more) from either Danmarks Statistik and from the crowdsourced website folketingsvalg-2015.dk. Links to the data set and all code used in this report can be found in the reference list.

It is important to note that data of this kind represents "stated preferences" as opposed to "revealed preferences". The candidates might have incentives to deviate from their true political views in order to obtain more votes - increasing their chances of election may well be the main motives for the candidates to take part in the VAA in the first place (in Part 3, we examine a specific such deviation). When trying to map political ideologies revealed preferences would be preferred, but we will make do with the VAA data and assume that they, to a large extent, correspond to the candidates' actual views. While scraping data from websites should generally warrant critical ethical evaluation, we find relief in the fact that the candidates on whom we are gathering data, are openly running for parliament and that the main webpage from which we have been scraping is the homepage of the public service national broadcasting corporation, DR.

Descriptive statistics and visualizations

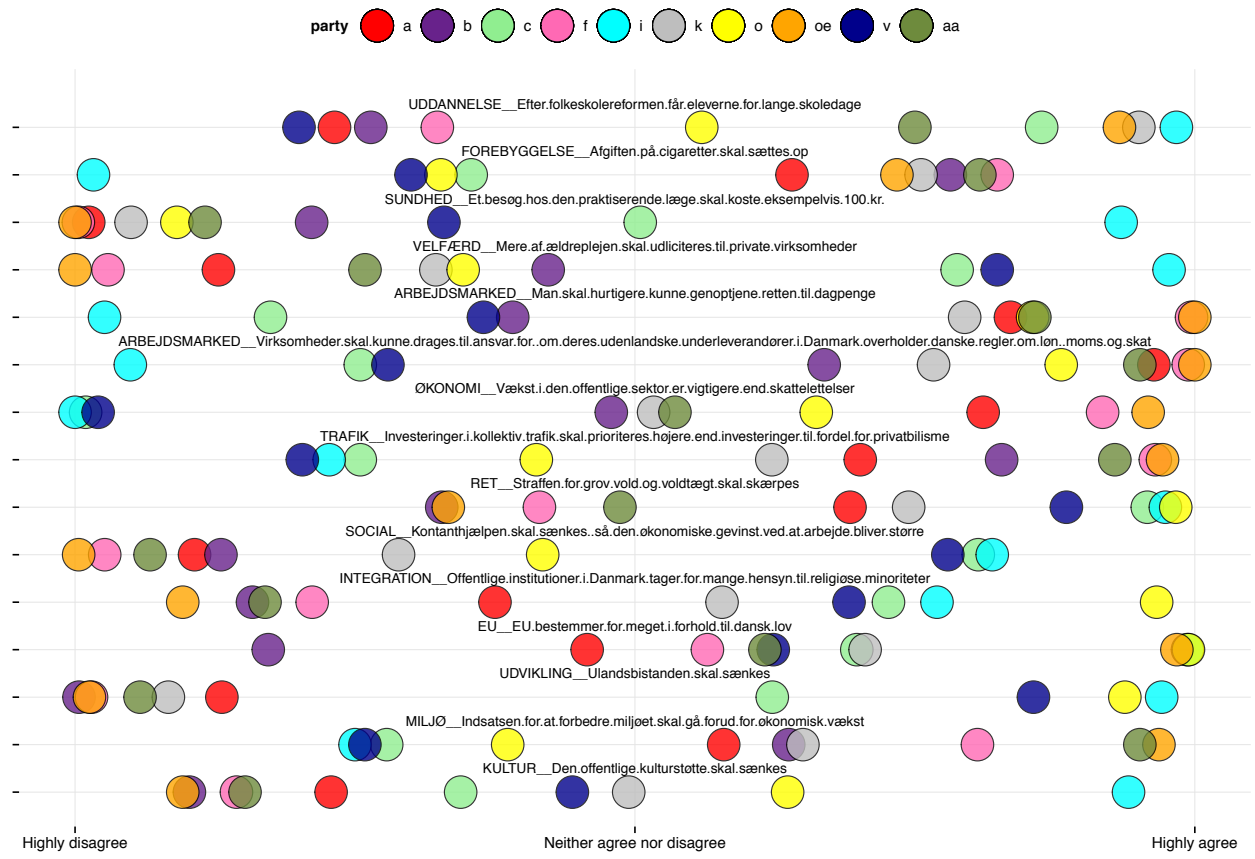
In this section we provide overview of the primary dataset. Figure 1 plots the mean response to all questions by party. This can give a very detailed overview, but with some limitations: we leave out some data variation by collapsing within-party variation, and we do not overcome the complexity of having to compare every party's average response across each of the fifteen questions. Some insight do however emerge: First, we see that the mean responses to all questions seem rather spread out in the sense that there is no immediate clusterings around one end of any questions. Second, we see a strong tendency for a pattern at least for the Enhedslisten (oe - orange) and Liberal Alliance (i - turqoise) that seem to linger around the

[2]http://www.dr.dk/nyheder/politik/valg2015/kandidat-testen

[3]To a certain extent, the external validity of our results hinge on the selection into VAA responses among candidates. When comparing the groups of candidates who have answered the VAA and who have not, it does not seem that selection is completely random. Notably, neither of the two primary candidates for the position of Prime Minister, Lars Løkke Rasmussen (V) and Helle Thorning-Schmidt (A), have completed the VAA. However, in excess of these two, no candidate that did not respond to the VAA were elected. Due to the overwhelming majority of personal votes accounted for in the dataset matched with VAA, we believe our dataset will be well suited to assess questions related to personal votes.

(opposing) edges of the mean response spectrum and rarely around the middle. There are also some dots that are almost completely, or to a large extent, overlapping - as is often the case for Enhedslisten (oe - orange) and SF (f - pink). However we also see parties that hold opposite views in some questions joining views in matters of others, e.g. Enhedslisten (oe - orange) and DF (o - yellow) on the questions of EU, and Enhedslisten (oe) and Liberal Alliance (i) on public school reform. This provides an early indication that a one-dimensional representation of the political landscape may be insufficient to capture the Danish political parties.

Figure 1: Mean response to survey questions, by party



These observations are validated even further when we compute the average distance from the "neither agree nor disagree" for each party (figure 2), where Enhedslisten and Liberal Alliance, not surprisingly cf. figure 1, are distinctively more extreme in their opinions. The average distance for the most extreme (Enhedslisten) is almost twice as large as the least extreme (Kristendemokraterne).

1 The political landscape: Theory and empirics

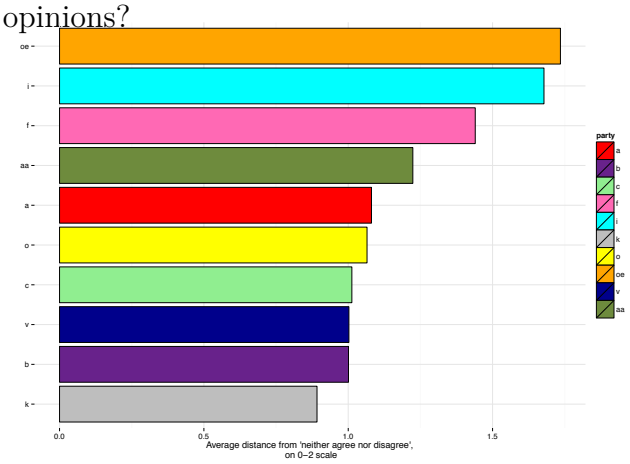
In this section we explore the Danish political landscape. By landscape, we mean a simplified exposition of political positions, using a few 'policy dimensions'. Using dimensions to represent political positions is essentially an exercise in simplication: using a simple framework that can collapse positions on a large number of issues into a few easily understandable constructs. In political theory, policy positions was originally viewed along a single left/right-dimension: left represents a socialist ideology, right represents a capitalist ideology (Wheatley,

2012). However, as was alluded to in the previous section, some parties may agree on one issue but disagree on another, lending support for an idea that political positions should not be viewed as one-dimensional. A general suggestion is to assess positions on another dimension, including 'newer' political position on e.g. environmental protection and crime along. A well-known theory of this dimension is the GAL-TAN (GAL: Green/Alternative/Libertarian, TAN:Traditionalism/Authority/Nationalism) [ibid]. In a Danish context, a roughly corresponding way to depict policy, rests on dimensions of distributional policy (DK: fordelingspolitik) and values policy (DK: værdipolitik) (See Appendix Figure 7 for a illustration of this)

1.1 Dimensionality reduction through PCA

A well-known data science method parallel to the exercise of reducing political position in to a few dimensions, is known as dimensionality reduction. One such method, is Principal Component Analysis (PCA) - a multivariate analysis method within unsupervised statistical learning for reducing dimensions in a parameter space. Our VAA-data is a 15-dimensional parameter space, making it difficult to visualize all dimensions in a meaningful way. Using PCA, allows us to collapse these 15 questions into 'synthetic' dimensions capturing a decreasing part of the original variance in the data. Intuitively, PCA creates linear combinations of existing dimensions, that capture as much variance from the original data set as possible, while ensuring uncorrelated dimensions.

Figure 2: Which parties have the most extreme opinions?



We use the princomp to do our PCA, which progresses in several stages - firstly, a correlation matrix Σ of the data matrix X is constructed. For each eigenvector α, corresponding to this correlation matrix, we want to find the eigenvectors that maximize the variance of the data, i.e. $\alpha'\Sigma\alpha$. This can be seen as a Lagrange optimization, subject to $\alpha'\alpha = 1$ due to normalization.

$$\alpha'\Sigma\alpha - \lambda(\alpha'\alpha - 1), \text{ differentiation wrs. } \alpha_i \text{ gives } \Sigma\alpha_i - \lambda\alpha_i$$

Hence, λ can be seen as an eigenvalue to Σ and α_i is the corresponding eigenvector. The (in our case 15) eigenvectors, called "loadings" in princomp, are then constructed so that the corresponding eigenvalues are maximized and in decreasing order throughout the components. Lastly the "scores" are computed by multiplying the original, standardized values from X with our loadings. This provides us with transformed variable values for each component suitable for plotting. Thus, the "loadings" can be seen as a measure of "importance" from the different parameters - i.e. the questions provide the most explanatory power.

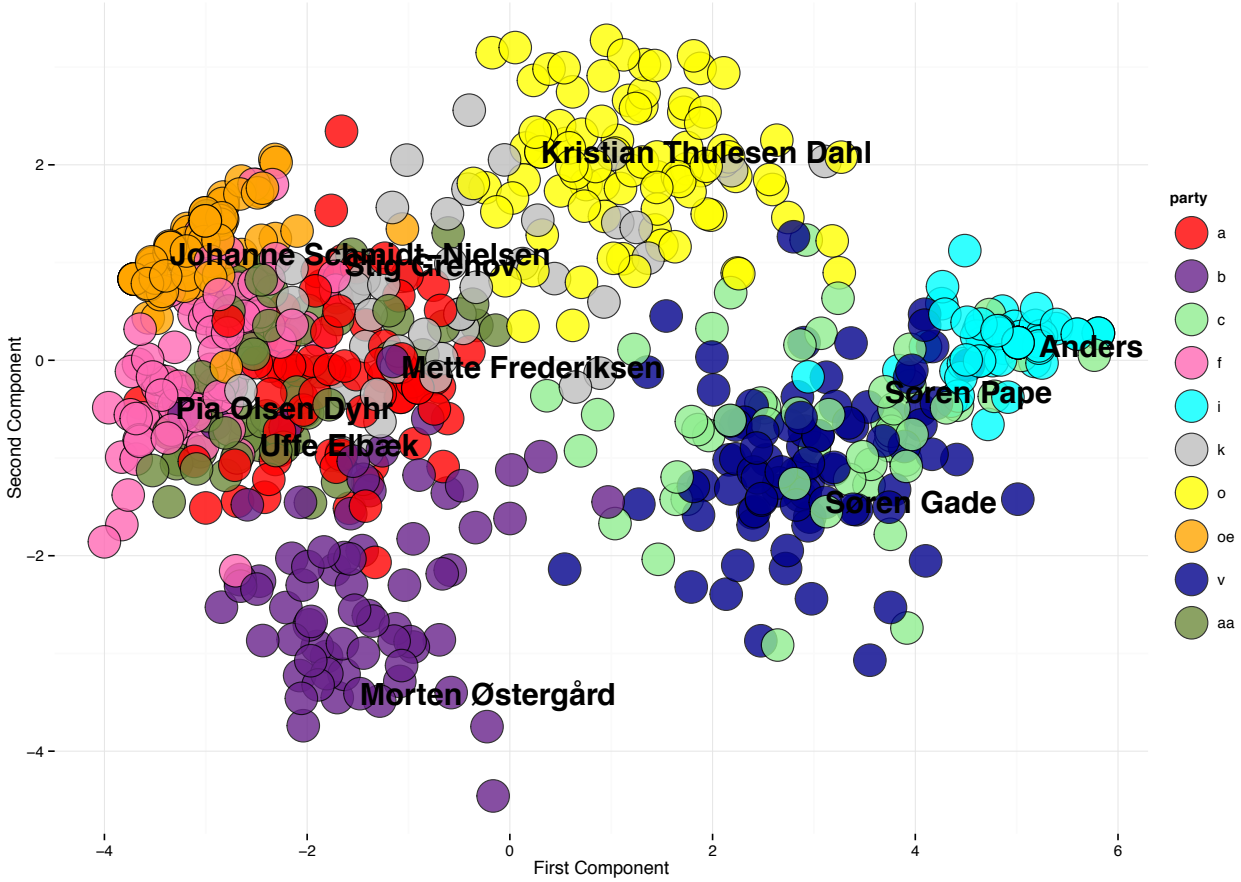
1.2 Results

The first synthetic dimension created by our PCA captures 54% of the variance in the original data set. Adding the second component explains 12%, implies that we can retain a cumulative

proportion of 66% of the variance of the responses to the VAA by collapsing the data from 15 to two dimensions.

The PCA is plotted in Figure 3, with the first component on the horizontal axis and the second component on the vertical axis. A few immediate insights emerge, before we turn to the axes: (1) There seems to be a fairly distinct grouping of the parties although some have significant overlaps: Especially the horizontal axis, which captures the most variance, serves to distinguish parties from one another, but some parties, e.g. Venstre (v) and Konservative (c) can not be distinguished along these axes. (2) The spread of candidates varies highly across parties: seen in relation to the view of what parties use the most extreme ends of the scale, its seems that these parties (e.g. i and oe) are also those where candidates tend to agree the most in their VAA responses.

Figure 3: Principal Component Analysis



The figure is a plot of the two first components from the PCA. Positions of the party leaders are shown in the plot - the specific position is by the first letter of their first name. Since party leaders of Socialdemokraterne and Venstre are missing from the VAA data, the two candidates that recieved most personal votes, Mette Frederiksen and Søren Gade, are representing those parties.

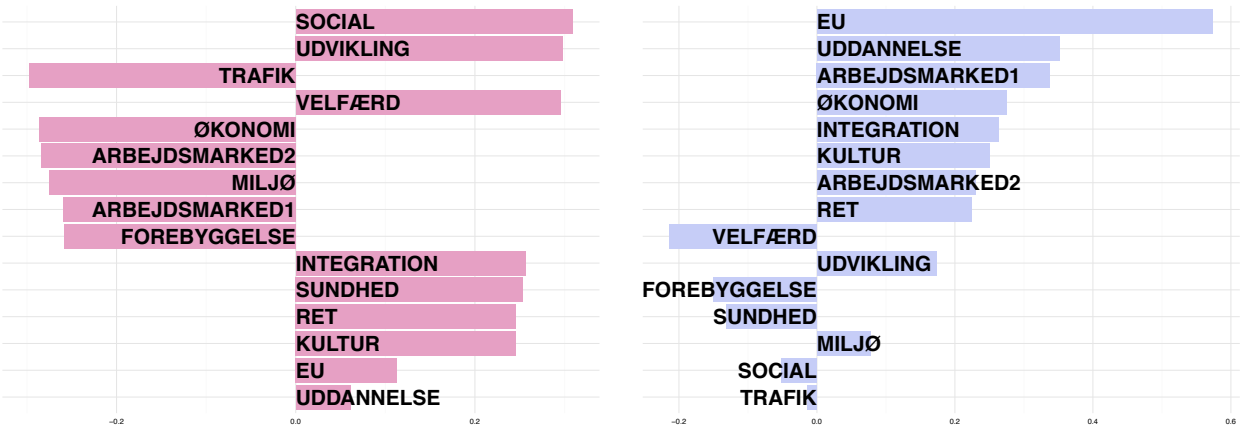
At first sight, the horizontal axis seems to capture a fairly classical left-right scale. The proclaimed left wing parties generally group together on the left, with Enhedslisten clustered closely together and Socialdemokratiet, SF and Alternativet slightly more spread out. In the opposite end, candidates from Liberal Alliance are also clustered neatly together whereas both Venstre and Konservative are vastly spread out. However, when we take a closer look at the factor loadings this picture changes slightly. The left panel of figure 4 plots the linear combination of the responses that make up the first dimension, in decreasing order of factor loading. We observe that a wide range of questions have a high loading into the dimension,

and that these questions span both typical distributional issues (Social: size of welfare benefits, Økonomi:public sector growth) and value issues (Udvikling: foreign aid levels and Miljø: trade-off between). This explains that Radikale Venstre (b) has a position that is to the left of Dansk Folkeparti (o), while RV would generally be thought to be more right-leaning on distributional issues than DF.

Turning to the vertical axis (second component) we see that it captures less variance, and mainly serves to distinguish the Radikale Venstre (b) and Dansk Folkeparti (o) from the rest. Standard political theory would predict the secondary axis to be driven by values-based issues which makes sense having Radikale and DF opposed to each other. However, having the parties Dansk Folkeparti (o) and Enhedslisten (oe) lean towards the same side on classical values-based issues seems contrary to popular belief. Looking into the factor loadings, we see that the second component is driven mainly by the question on EU's role and influence, but also by questions on the public school reform and unemployment benefits.

Conclusively, we can see that there exist a clear dimension that separates most Danish parties from each another, which does not only capture positions on classical distributional issues, but rather a much wider range of political questions. Second, the next-most important dimension in Danish politics seems to not be the classical 'values-based' dimension, but rather a dimension that capture openness to internationalisation and national sovereignty, especially in relation to the European Union.

Figure 4: Loadings



(a) Loadings, 1st component

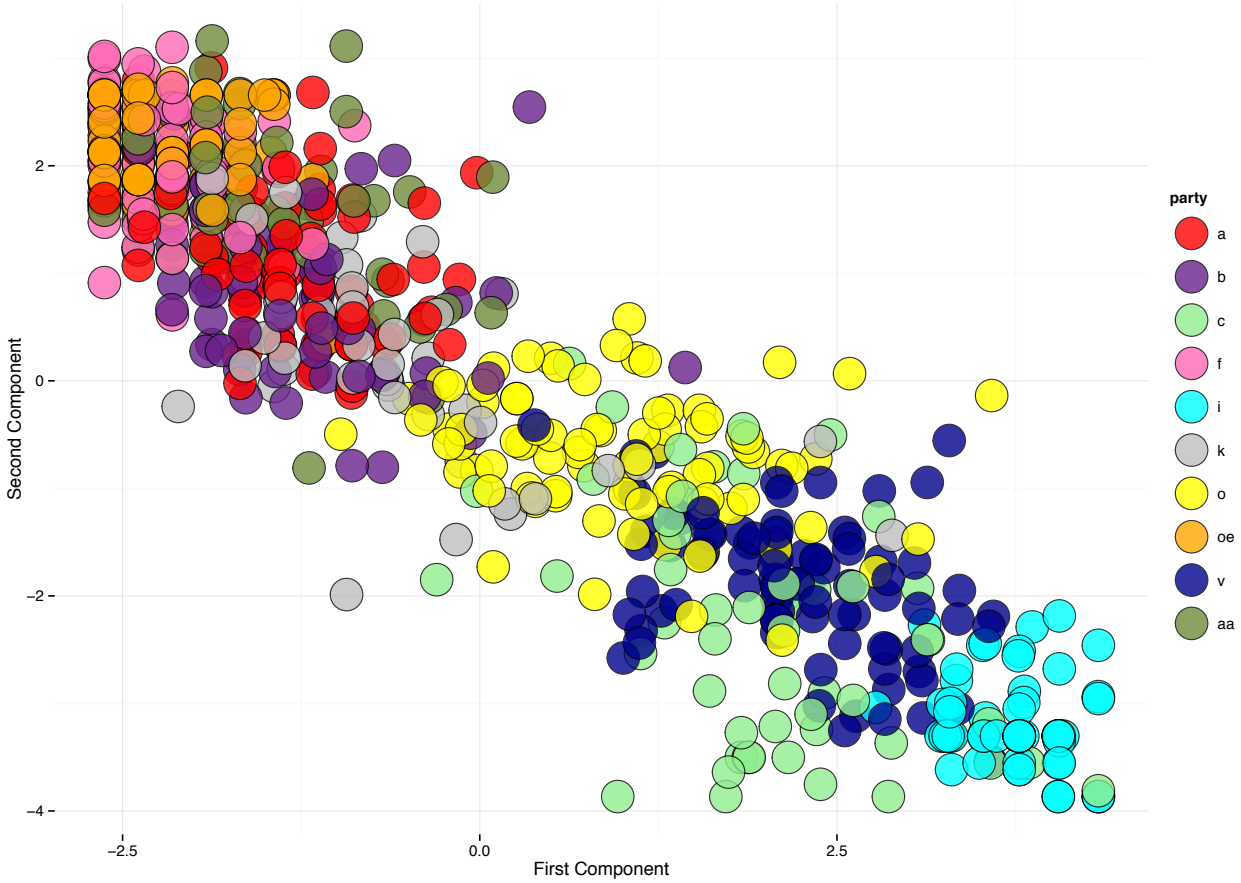
(b) Loadings, 2nd component

A reason why we may not see a crystal clear political spectrum aligned with the political science literature could also be the weighting of the different questions - or rather, the lack thereof. One could argue that a candidate's views on criminal justice and integration should not weigh as heavily as views on public sector size and levels of unemployment benefits when trying to depict the candidates on a distributional left-right-scale. Vice versa we would want issues on integration and environmental politics to weigh more when plotting the value-based political spectrum. The candidates are not asked to weigh the questions individually (as is for instance seen in Talonen and Sulkava, 2011).

1.3 Discussion

The method for dimensionality reduction used above did not seem to confirm popular theories of the Danish political spectrum and dimensions. To delve further into why this is the case, and how the two 'standard' dimensions (distributional and values-based) come to show in the dataset, we attempt another method of assisted dimensionality reduction. First, we separate the questions in two sets of distributional and value-based politics respectively (see Appendix for the exact grouping). We perform two separate PCAs on these segregated datasets and plot the first components from each set of questions in the same plot, in order to see if the standard political spectrum (see Appendix) is more visible now. In this way we force the PCA to only choose questions higher from an ex ante dimension in each set. The resulting plot is shown in figure 5. The distributional political views are on the horizontal axis and the value-based issues are on the vertical axis. Obviously, the resulting dimensions now capture less variance than our PCA analysis above, and the parties are much less distinguishable. More interestingly, the two dimensions now appear highly correlated: parties do not stray far from a straight downward sloping line through the dimensions, and the ordering of the parties is the same along the two dimensions. This implies a concretisation of our conclusion above: in Danish politics, positions on classical values-based issues and distributional issued are highly correlated, and do therefore not represent to distinct 'dimensions' that the political space can be reduced to. Rather, both distributional and some values-based issues seem to make up a primary dimension of Danish politics, whereas a second (less important) dimension consists of issues on the European Union, education and labor policies.

Figure 5: Principal Component Analysis - separated data

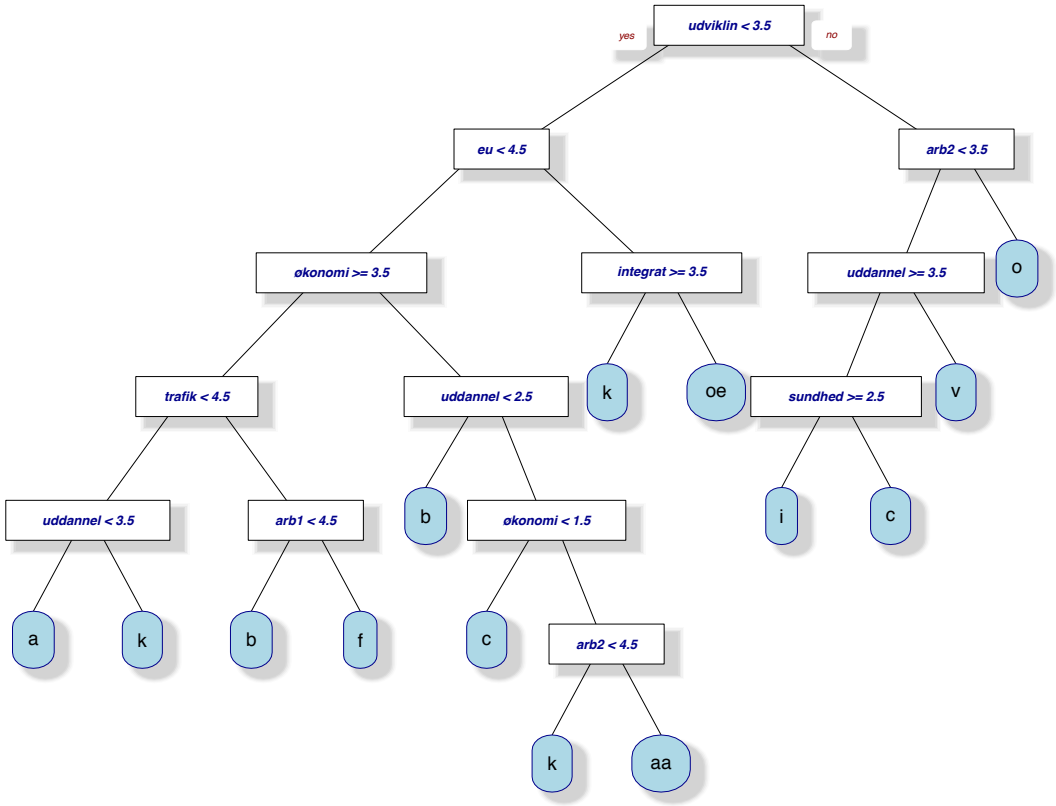


2 Can we predict the party outcome?

In the following section we'll make a model to classify the candidates into parties. In order to do so a simple classification decision tree will be applied to our scraped VAA dataset. The approach of running this type of model will leave some insights about: (1) how well suited the questions from the VAA is to classify the party of the voters, (2) an implication of the key determinants of which the different parties differentiates themselves and (3) get a hint of to what extent the homogeneity of the party (in terms of the prediction outcomes of our model) do affect the mean of personal votes.

To actually make the model we first divide our VAA dataset into a training dataset by a random sample containing a third of the observations and a testing dataset containing the rest of the observations. The objective for the model is to make predictions, which is the reason why we train the model on only a subset of the dataset. Else, we would run into problems with overfitting in the sense that the model to a high degree would replicate the actual data. This is in contrast to the standard econometric approach where the whole dataset usually is evaluated to make inference on the relationships among variables. Next, we run the recursive binary splitting algorithm to actually make the classification tree (using the rpart-package in R). Even though we divided our data into a training set, we still have a possible danger with overfitting, which will lower the predictive power of the model when applying it to the rest of the data. To avoid that, we'll grow the full tree and prune it down afterwards to a size that minimizes the cross-validation error. The end result is a tree with 15 nodes including only 8 out of 15 VAA questions:

Figure 6: Classification Tree



The visualized decision tree points out how the algorithm categorizes the candidates into parties. Overall the model performs with a classification error rate of 23.5% when applied to the testing dataset which simply reflects the fraction of incorrect predictions produced by the model. Two effects is in play for the classification error rate: (i) weak predictive power of the model and (ii) existence of candidates which actually has a policy that may belong to another party than the one he/she is actually enrolled into. For these reasons it can be difficult to evaluate how well the model performs but in general decision trees is not the best model in terms of predictive powerness but is easy to implement on data. One of the insights given by the model is the variable importance. The question "økonomi" is the most important one which may not a surprise in context of the question formulation which deals with the preferences of the voters in respect to the tradeoff between tax cuts versus public sector growth. During the election campaign, this was actually one of the greatest topics separating the left and right flank parties. Generally the model attaches more weight to questions that could be classified as "core political values" such as the size of the public sector and labour market and less weight to "soft political values" like culture and environment. If we treat our model as predicting the underlying truth, then we can identify the misplaced candidates in terms of party (i.e the classification errors of the model). One interesting finding of this approach is the difference in mean personal votes. We finds that the mean of personal votes for candidates not in sync with the policy of their party 1,408 in contrast to 2,525 for those in sync. That is quite a significant difference and implies that homogeneity within a party could be one of the determinants for the personal votes outcome. Another finding is simply a support of the first section, where in general it's the same parties covering a great area of the political landscape that also has the largest classification error rates, i.e. a lot of non-homogeneous candidates[4]. But with the strong assumption imposed - that the model speaks the underlying truth - we'll leave this insight simply as an implication for further exploration.

3 The political landscape: VAAs and Election outcomes

In this section, we present a new test for the impact of VAAs on election outcomes, that uses only the candidates' responses to the VAA (as well as some background data). This research should be taken as a preliminary indication of a specific type of impact test, namely testing for the possibility of 'gaming the system', by taking a position that is 'unoccupied' by others, rather than a general test for its impact. Nevertheless, we hope our approach can be a valuable addition to the literature in this account. The main hypothesis we attempt to test is the following; *Hypothesis: Candidates that make themselves distinct will receive more personal votes.*

We propose the following reasoning behind the hypothesis: if voters' preferences map into responses in a sufficiently 'noisy' way, a candidate who has a higher 'distance' to other candidates, would end up as the top candidate with a higher number of test takers, than candidates who lie very close to others. If VAAs have a discernible impact on voting behavior, candidates who are more likely to end up on top of the results with test takers should receive more personal votes. However, an opposite mechanism could be expected to exist: when candidates are observed to cluster around a specific position, we might hypothesize that it such positions would be popular among voters. As above, this would lead candidates in such positions to end

[4]See table 2 in Appendix

up as the top candidate with more test takers, if voters and candidates cluster around the same positions.[5] Assume that three candidates, A, B and C are fairly close on political opinions, but there is higher agreement between B and C, than between Candidate A and the two others. If voters responses to a VAA are spread noisily over the depicted policy space, Candidate A will have a larger area to 'herself', and will therefore be shown as the top result with more voters. However, if Candidates B and C are clustered around an opinion that is also popular with test takers (and voters) they are likely to receive more personal votes.

3.1 Measures of distance and distinctiveness of a candidate

Our identification strategy rests on two primary elements: First, we need a precise definition of the distance between two candidates. Second, we need an appropriate method of capturing the distance measure for each candidate, which requires answering the question: for any one candidate, who are the relevant peers? While the data science toolbox provides a wide array of options for assessing distance between points in multi-dimensional, potentially correlated space (e.g. Euclidean distance, Mahanolobi's distance) we have chosen the same defition used in computing the results of DR's VAA. As described in the introduction, users are shown a 'percentage agreement' with candidates based on their responses to the VAA. The percentage agreement between Candidate i and Candidate j is defined as,

$$agree_{i,j} = \sum_{k=1}^{l} \left(\frac{4 - |r_{ik} - r_{jk}|}{4 \cdot l} \right) \tag{1}$$

where r_{ik} is Candidate i's scored response to question k on a 1-5 Likert scale, and l is the number of questions in the VAA. Intuitively, the measure $agree_{i,j}$ assigns 4 points to every question that is answered identically, and a minimum of 0 points to questions at opposite ends of the scale. These points are summed across all questions and rescaled to a percentage scale; $agree_{i,j} \in [0, 1]$. To measure distinctiveness of a candidate's position we need a proper operationalisation of the definition of agreement above. We have chosen to use the average agreement with closest party members within the same voting district (storkreds). Technically, we measure for each candidate i, the average agreement with the candidates in the set P_i. The variable $agree.mean_i$ is then defined as,

$$agree.mean_i = \frac{1}{n_{P_i}} \cdot \sum_{P_i} agree_{i,p} \tag{2}$$

where n_{P_i} is the number of candidates in the set P_i. Our base definition of P_i is the set of the three candidates from the same party and same voting district, with whom candidate i agrees the most. $agree.mean_i$ has a mean value of $\mu_{agree.mean} = 0.86$ and a standard deviation of $\sigma_{agree.mean} = 0.078$, and its distribution is fairly close to normal, but truncated at $agree.mean_i = 1$. Twelve candidates in the dataset have values of $agree.mean_i$ exactly equal to 1.[6]

[5] It is important to note, that our identification method does not allow us to explicitly distinguish these two mechanism from one another. A highly robust method of doing so, would likely require data on the test takers' responses to the VAA. We were unfortunately not able to obtain this data from DR. One way of alleviating this issue is to measure distinctiveness for any candidate, not from the entire set of remaining candidates, but only for those within the same party and the same voting district (storkreds).

[6] To a certain extent, this choice of P_i is arbitrary - to test its robustness, we will present varying definitions of P_i in the results section. Our choice of the average is to avoid a measure that would record two highly outlying candidates as a 'cluster'. The choice of the three nearest candidates is based on the fact that some parties do not have more than four candidates in a voting district.

3.2 Identification strategy

In Part 2, we used a typical data science approach with prediction as the primary goal - for this, low prediction variance and out-of-sample performance of the model is more important than non-biased estimates. In this Part however, we are interested in inference, and therefore need unbiasedness. We assume an underlying data generating process, implicitly defined in our baseline regression model, and test whether our data disagrees strongly with hypotheses made, and use the entire dataset for model training. Our dependant variable is the personal votes for a candidate, determined by the following baseline model:

$$votes_i = \beta_0 + \beta_1 agree_{i,j} + \boldsymbol{\beta}'\mathbf{X}_i + \varepsilon_i \tag{3}$$

Here, β_1 measures the impact of agreeing closely with peers. A negative β_1 implies a positive effect of being 'distinct' from other candidates. Variables in X_i represent controls such as demographic background and previous election history. They are included in case there is any covariance with $agree_{i,j}$, which would otherwise imply a biased estimate of β_1.[7]

3.3 Results

Table 3 in Appendix presents the results from our regressions, with each column representing a different model specification. Generally, we find no benefit to being 'distinct' from peers - on the contrary, it seems that the more distant a candidate is from party members, the fewer personal votes the candidate is likely to obtain. Model (1) regresses personal votes on $agree_i$ only - we obtain a positive and statistically significant estimate of $\beta_1 = 6,222$. Including background variables in Model (2) leads to a slightly lower estimate at $\beta_1 = 5,051$, but still positive and statistically significant. The estimate implies that on average, a one standard deviation increase in average agreement with three nearest party members in the same voting district is associated with approximately 400 more personal votes. Taking into account that 1 in 5 elected candidates received less than 3000 votes, and ten percent less than 2000 votes, these results are not economically insignificant. To uncover the driving mechanism behind these results, we ran Model (3) which has interaction terms for party and $agree.mean_i$. Socialdemokratiet (A) is the reference group in Model (3), so the estimate of β_1 corresponds to the marginal effect of $agree.mean_i$ on personal votes for Social Democrat candidates. An estimate about 5-7 times larger than our baseline model suggests that the effect found in Models (1) and (2) is in large part driven by candidates from Socialdemokratiet. The interaction coefficients for each other party show that the effect of being distinct is significantly larger for candidates from Socialdemokratiet (A) than for each of the other parties, with the exception of Dansk Folkeparti (O) and Enhedslisten (Ø). However, the numerical value of each interaction coefficient estimate is smaller than β_1, which implies that the effect of being distinct is not positive for any party. Model (4) repeats Model (2), but excludes candidates from Socialdemokratiet (A), and finds that the effect of being distinct diminishes, and fails to be significant on five percent level, but remains significant at a ten percent level. Models (5) and (6) are meant as robustness checks. Model (5) uses the distance to the average party position instead of $agree.mean_i$ as the regressand, and finds a slightly significant (on a ten percent level) positive effect, corresponding to the results from Models (1) through (4). Models (6) uses the average distance to the three

[7] Other relevant background variables that would be relevant for providing robustness checks of the results would be other good predictors of personal votes. Among variables we have considered, but not tested, are position on the voting list presented to voters, social media activity, media presence, and more.

nearest non-party members instead of $agree.mean_i$. This measure is high for candidates who are close to members of other parties, and therefore typically on the periphery of their own party. This estimate is negative, and therefore consistent with the above intuition, but insignificant so we should be wary of attaching too much importance to it.

3.4 Discussion

The results above do not provide any support for our hypothesis: it does not seem that being more distinct in VAA responses translates into more personal votes. This however, does not allow us to conclude that the opposite mechanism described above - that candidates would tend to cluster around 'popular' positions - is in fact driving our results. It would be an interesting research question, to look into whether a relatively unknown candidate would be able to increase her personal votes by copying the VAA responses of more popular candidates, but our research design above is not well suited to identify such mechanisms. As a visualization of our results above we provide a scatterplot of the positions of every candidate using our PCA analysis from part 1 in Figure 8 in the appendix. Every point is scaled corresponding to number of personal votes (diameter of circle). Finally, each point i is colored according to the value of $agree.mean_i$ - redder colors indicate higher agreement with adjacent party members. Two reflections can be taken from the plot: First, the reader may note that the distance coloring does not directly correspond to the apparent Euclidean distances in the shown dimensions - this is due to the fact that the PCA collapses many dimensions into only few, so assessing Euclidean distances in the PCA dimensions essentially assigns less importance to distances in highly correlated dimensions. Second, the reader will observe that the candidates who receive the most votes (large circles) are typically fairly red, implying a low distinction to adjacent party members.

Conclusion

This report presents three ways in which the responses of candidates in the Danish General Election 2015 to a popular Voting Advice Application, can be used to examine questions within political science, through data science and econometrics. First, we have shown that the political positions of Danish candidates do not seem to be well represented by the theory of a distributional and a values-based political dimension - positions on these dimensions are highly correlated. On the contrary, we indicate that a new dimension capturing e.g. positions on the European Union serves to better separate the political parties. Second, we have shown that a large majority of candidates can be classified in their parties using only data on their VAA responses, and that classification works better within parties who have little variance in political positions. Finally, we have shown that there seems to be no benefit, in terms of votes, from making oneself distinct to peers in the VAA - rather, the opposite might be the case.

References

- BAKKER, R., Edwards, E., Jolly, S., Polk, J. and J. Rovny (2010), "The Dimensionality of Party Politics in Europe"
- CEDRONI, L. and D. Garzia (2010), "Voting Advice Applications in Europe: The State of the Art", ScriptaWeb
- FOLKETINGET (2012), "Folketinget efter valget 2011", Folketinget
- GARZIA, D. and Marschall, S. (2012), "Voting Advice Applications under review: the state of research", Int. J. Electronic Governance, Vol. 5, Nos 3/4, pp. 203-222
- LADNER, A., Fivaz, J. and Pianzola, J. (2012), "Voting advice applications and party choice: evidence from smartvote users in Switzerland", Int. J. Electronic Go2vernance, Vol. 5, Nos 3/4, pp. 367-387
- PIANZOLA, Joëlle and Andreas Ladner (2011), "Tackling Self-Selection into Treatment and Self-Selection into the Sample Biases in VAA Research", IDHEAP Lausanne, University of Lausanne
- SLOTHUUS, Rune, Rune Stubager, Kasper Møller Hansen, Michael Bang Petersen and Morten Pettersson (2010), "Måling af politiske værdier og informationsbearbejdning. Nye indeks for fordelingspolitik, værdipolitik og "Need to Evaluate" blandt danske vælgere", Aarhus Universitet and Københavns Universitet
- TALONEN, Jaakkoo and Mika Sulkava (2011), "Analyzing Parliamentary Elections Based on Voting Advice Application Data", Aalto University School of Science
- WHEATLEY, J. (2012), "Using VAAs to explore the dimensionality of the policy space: experiments from Brazil, Peru, Scotland and Cyprus", Int. J. Electronic Governance, Vol. 5, Nos 3/4, pp. 318-348
- WHEATLEY, Jonathan (2015), "Identifying Latent Policy Dimensions from Public Opinion Data: An Inductive Approach", Journal of Elections, Public Opinion and Parties, 25:2, 215-233
- The data is available at Github: `https://raw.githubusercontent.com/oskarharmsen/Assignment-2/master/Exam%20Project/dk_ft15_politician_responses.csv`

Appendix

The questions from the VAA

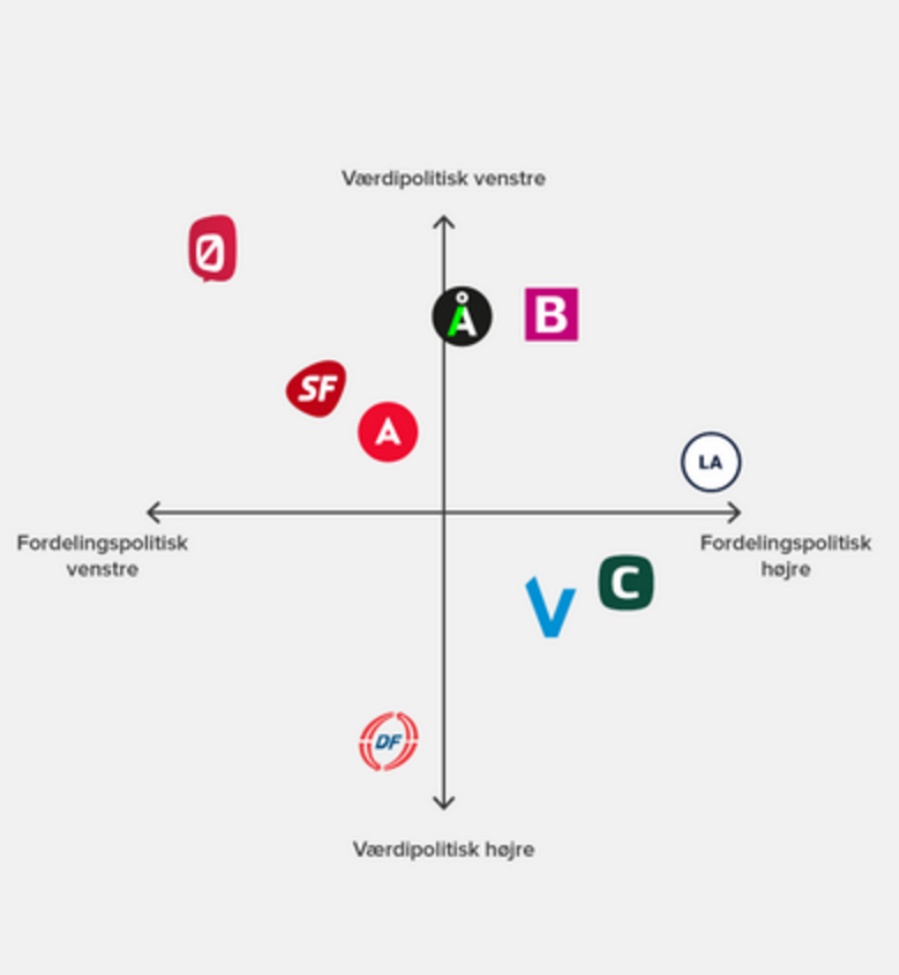
1. UDDANNELSE - Efter folkeskolereformen får eleverne for lange skoledage
2. FOREBYGGELSE - Afgiften på cigaretter skal sættes op
3. SUNDHED - Et besøg hos den praktiserende læge skal koste eksempelvis 100 kr.
4. VELFÆRD - Mere af ældreplejen skal udliciteres til private virksomheder
5. ARBEJDSMARKED1 - Man skal hurtigere kunne genoptjene retten til dagpenge
6. ARBEJDSMARKED2 - Virksomheder skal kunne drages til ansvar for, om deres udenlandske underleverandører i Danmark overholder danske regler om løn, moms og skat
7. ØKONOMI - Vækst i den offentlige sektor er vigtigere end skattelettelser
8. TRAFIK - Investeringer i kollektiv trafik skal prioriteres højere end investeringer til fordel for privatbilisme
9. RET - Straffen for grov vold og voldtægt skal skærpes
10. SOCIAL - Kontanthjælpen skal sænkes, så den økonomiske gevinst ved at arbejde bliver større
11. INTEGRATION - Offentlige institutioner i Danmark tager for mange hensyn til religiøse minoriteter
12. EU - EU bestemmer for meget i forhold til dansk lov
13. UDVIKLING - Ulandsbistanden skal sænkes
14. MILJØ - Indsatsen for at forbedre miljøet skal gå forud for økonomisk vækst
15. KULTUR - Den offentlige kulturstøtte skal sænkes

Subsetting questions attributed to distributional and GAL-TAN political views

Distributional: *FOREBYGGELSE, SUNDHED, ØKONOMI, TRAFIK, SOCIAL, UDVIKLING, KULTUR*

Value-based: *UDDANNELSE, VELFÆRD, RET, INTEGRATION, EU, MILJØ*

Figure 7: Political spectrum



source: http://www.clioonline.dk/samfundsfaget/emner/politik/det-politiske-landskab/det-nye-politiske-landskab/

Table 1: Variables of importance

økonomi	*arb1*	*velfærd*	*eu*	*social*
84.943	84.586	74.697	65.859	63.683
arb2	*uddannelse*	*sundhed*	*trafik*	*udvikling*
62.683	59.968	59.251	54.552	44.060
ret	*integration*	*forebyggelse*	*kultur*	*miljø*
41.990	39.973	28.157	27.918	16.634

Table 2: Tree - Prediction summary

party	**mean votes**	**fraction of correct predictions**
oe	1351.1	94.87
v	4553.4	94.1
i	1331.1	89.5
o	3269.4	88.6
a	5504.3	78.0
f	828.9	74.1
k	366.8	68.9
b	1181.0	66.2
aa	977.9	48.2
c	1365.4	43.4

Table 3: Regression results

	Dependent variable: votes, Method: OLS					
	(1)	(2)	(3)	(4)	(5)	(6)
	Only *agree.mean*	With background vars	Interaction with party ('A' is reference)	Without party 'A'	Dist from party	Dist. from oth. party
agree.mean	6,222**	5,051***	38,209***	3,237*		
	(2,024)	(1,901)	(8,318)	(1,849)		
agree.party.center					4,129*	
					(2,268)	
agree.other.party						−3,843
						(2,402)
agree.mean · party$_k$						
$k = aa$	-	-	−31,191***	-	-	-
			(11,186)			
$k = b$	-	-	−29,342**	-	-	-
			(11,598)			
$k = c$	-	-	−35,785***	-	-	-
			(12,438)			
$k = f$	-	-	−37,834**	-	-	-
			(16,934)			
$k = i$	-	-	−32,596***	-	-	-
			(11,816)			
$k = k$	-	-	−34,076***	-	-	-
			(11,531)			
$k = o$	-	-	−10,379	-	-	-
			(11,194)			
$k = oe$	-	-	−8,915	-	-	-
			(16,372)			
$k = v$	-	-	−29,824***	-	-	-
			(10,121)			
Background variables						
is.male	-	−243	−399	−204	−151	−227
		(310)	(290)	(313)	(321)	(310)
ran.last.election	-	3,125***	2,327***	2,610***	3,088***	3,144***
		(299)	(298)	(306)	(308)	(299)
age	-	−22.914**	−17.802	−15.092	−22.450*	−28.437**
		(11.609)	(10.915)	(11.632)	(12.026)	(11.560)
Observations	715	715	715	633	719	717
R^2	0.129	0.148	0.296	0.111	0.129	0.142
Adjusted R^2	0.124	0.143	0.273	0.106	0.124	0.137
Residual Std. Error	3,994	3,851	3,545	3,666	3,994	3,860
F Statistic	26.496***	30.729***	13.215***	19.642***	26.496***	29.359***

All regressions contain a constant. The regression method is Ordinary Least Squares.
Numbers in brackets are standard errors. P-values from T-tests:p<0.1*; p<0.05**; p<0.01***

Figure 8: Candidate distinction, votes and political position