Mining Democracy

Vincent Etter* Julien Herzen* Matthias Grossglauser Patrick Thiran
School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
firstname.lastname@epfl.ch

ABSTRACT

Switzerland has a long tradition of direct democracy, which makes it an ideal laboratory for research on real-world politics. Similar to recent *open government* initiatives launched worldwide, the Swiss government regularly releases datasets related to state affairs and politics. In this paper, we propose an exploratory, data-driven study of the political landscape of Switzerland, in which we use opinions expressed by candidates and citizens on a web platform during the recent Swiss parliamentary elections, together with fine-grained vote results and parliament votes.

Following this purely data-driven approach, we show that it is possible to uncover interesting patterns that would otherwise require both tedious manual analysis and domain knowledge. In particular, we show that traditional cultural and/or ideological idiosyncrasies can be highlighted and quantified by looking at vote results and pre-election opinions. We propose a technique for comparing the candidates' opinions expressed before the elections with their actual votes cast in the parliament after the elections. This technique spots politicians that do not vote consistently with the opinions that they expressed during the campaign. We also observe that it is possible to predict surprisingly precisely the outcome of nationwide votes, by looking at the outcome in a single, carefully selected municipality. Our work applies to any country where similar data is available; it points to some of the avenues created by user-generated data emerging from open government initiatives, which enable new data-mining approaches to political and social sciences.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms; Experimentation

*These authors contributed equally to this work.

COSN'14, October 1–2, 2014, Dublin, Ireland.

ACM ACM 978-1-4503-3198-2/14/10$15.00.
http://dx.doi.org/10.1145/2660460.2660476.

Keywords

Political data; voting advice application; dimensionality reduction; vote results prediction

1. INTRODUCTION

In order to promote transparency and accountability, as well as to stimulate citizen awareness, an increasing number of governments across the world are adopting *open government* directives [20]. As of 2014, the website Datacatalogs [1] references more than 350 such local, regional and national datasets. These initiatives result in the release of massive amounts of structured data about multiple aspects of state affairs, politics, and governmental agencies in various countries.

In parallel to these efforts, several governments, organizations and academic groups set up *voting advice applications* (VAA's) in the form of websites that allow politicians and interested citizens to express their preferences on political issues, by answering a series of pre-determined questions spanning a variety of topics. The candidates have public profiles containing their responses (as well as various other information, such as their birthdate, interests, or Facebook profile), and the voters are matched with candidates based on their own responses. Examples of such VAA's include *Vote Compass* [6] in Canada, the USA and Australia, *Stemwijzer* [5] in the Netherlands, *Wahl-O-Mat* [7] in Germany, *Stemtest* [4] in Belgium and *smartvote* [3] in Switzerland.

In this paper, we propose an exploratory, data-mining approach that uses some of the data released by governments, together with data obtained from VAA's, to analyze a country's democracy and political trends. We consider the case of Switzerland: This country has a diversified party landscape, with frequent votes on a wide variety of topics, both at parliamentary and citizen levels. We use three different datasets:

1. The set of all vote results in each municipality, for each national vote between 1981 and 2011.
2. The set of all votes in the parliament, by all parliament members, during the current legislature (which started in 2011).
3. The set of opinions given on the smartvote VAA [3] by hundreds of thousands of citizens, as well as more than 82% of candidates for parliamentary elections.

We give more details on our datasets in Section 2.2. The initial reason for the existence of these datasets was to increase government accountability and citizen participation.

abbreviation	full name	ideology	obtained votes
SVP	Swiss People's Party	National conservatism	26.6%
SP	Social Democratic Party	Social democracy	18.7%
FDP	Free Democratic Party	Classical liberalism	15.1%
CVP	Christian Democratic People's Party	Christian democracy	12.3%
Greens	Green Party	Green politics	8.4%
BDP	Conservative Democratic Party	Conservatism, economic liberalism	5.4%
GL	Green Liberal Party	Green liberalism	5.4%

Table 1: The seven major parties after the Swiss National Council elections of 2011. The last column lists the percentage of votes each party obtained during these elections.

Yet, as a byproduct, they also provide researchers with new ways of mining and (re-)discovering patterns that are peculiar to political life, but that usually require tedious manual analysis and domain knowledge. Although the nature of this data (e.g., the individual opinions of politicians or votes in municipalities) is not new, its scale is unprecedented. It enables us to address interesting questions such as

- What are the similarities between the ideological trends of parliament candidates (the representatives) and voters?
- Considering *all* votes at the municipality level over a period of 30 years, are there some clear patterns linking geography and voting behaviors?
- Is it possible to predict nationwide vote outcomes by looking only at the outcome in a single municipality?
- How "redundant" are political parties, given the opinions expressed by their candidates in smartvote surveys?
- How should a candidate fill a VAA survey during the campaign, in order to maximize her likelihood to appear at the top of the voting recommendations?
- In order to hold a candidate accountable for her public statements, can we use the survey she filled out during the election campaign, once she has been elected at the parliament?

Our intent is to exploit the scale of the data to tackle these questions from a statistical perspective. Of course, our results are not universal, and they are valid only in the contexts for which our datasets are representative. Yet, our various procedures apply to any context where similar data is available, and our main intent in this paper is to show how systematic data mining can shed new light on questions related to political and social sciences.

The remainder of the paper is organized as follows. In Section 2, we give some background information on Swiss politics and the datasets we use. In Section 3, we study the ideological landscape and the differences between parliament candidates and voters. In Section 4, we study trends at a geographical level and observe to what extent vote results of individual municipalities can be used to predict national vote outcomes. In Section 5, we show that voting advice applications can be abused by candidates to obtain better ranks in voting recommendations. We also propose a technique to use these same VAA surveys in order to check on the consistency of the votes at the parliament. Finally, we discuss some related work in Section 6, and we give some concluding remarks in Section 7.

2. BACKGROUND

In this section, we briefly describe the Swiss political system, and its various components. We list the main political parties on which we focus, and then describe the three datasets we use in this paper. The first dataset contains vote outcomes at the municipality level, the second consists of votes of the members of the parliament, and the last dataset contains political opinions of candidates and voters, gathered on an online voting advice application.

2.1 Politics of Switzerland

The political system of Switzerland consists of a *Federal Council* (7 seats) and a bicameral parliament, which is composed of the *Council of States* (46 seats) and the *National Council* (200 seats). The Federal Council serves as head of state and executive power, and the parliament possesses the legislative power (together with citizens, as per the constitutional right for citizens to launch initiatives[1]). The Council of States represents the *cantons* (which are the states of the federal state), and each canton is attributed two seats (except six "half" cantons that have only one seat). The National Council represents the people, and each canton is attributed a number of seats proportional to its population.

The National Council and the Council of States are elected at the same time every four years, most recently in 2011. Several political parties are represented in the parliament. In this paper, we focus on the seven largest parties (in terms of votes obtained during the National Council elections in 2011) shown on Table 1.

2.2 Description of the Datasets

In this section, we describe the three datasets that we use for our analysis. We provide a summary in Table 2.

2.2.1 Municipality Votes

Our first dataset consists of the outcomes of the federal (i.e., nationwide) votes for each municipality between January 1981 and December 2011. There were 245 such votes on various topics, including military, finances, transportation, culture, integration of foreigners, public health, education, equality of rights, working conditions, energy policies, abortion, etc. The results (i.e., the proportions of "yes") are publicly available for each Swiss municipality[2]. In December 2011, there were 2,515 municipalities in Switzerland. We discard the results

[1]Initiatives, similar to propositions in California, allow any citizen or organization to gather a predetermined number of signatures to propose a new piece of legislation [36].

[2]`http://www.bfs.admin.ch/bfs/portal/de/index/themen/17/03/blank/data/01.html`

dataset	content
Municipality votes	Outcomes (percentage of "yes") for 245 nationwide votes in 2,389 municipalities between 1981 and 2011. This amounts to 585,305 outcomes.
Votes in the parliament	2,494 votes (yes/no/abstention) by 181 of the 200 national councilors, since the beginning of the current legislature in 2011. This amounts to 451,414 votes.
Smartvote pre-electoral opinions	Responses given by 2,985 candidates (82.4% of all candidates) and 229,133 citizens (∼9% of total turnout) on the smartvote VAA [3] during the campaign of the 2011 parliamentary elections.

Table 2: Specifications of our three datasets.

smartvote 2011 statistics	
Number of questions in the short survey	32
Number of questions in the long survey	75
Number of candidates who took the long survey	2985
Approximate number of unique voters that requested recommendations	436,726
Number of voters who completed all questions of the short survey	229,133
Number of voters who completed all questions of the long survey	80,067

Table 3: Statistics about the political opinions dataset, that contains the responses given on the smartvote VAA by the citizens and candidates during the Swiss parliamentary elections of 2011.

for all the municipalities that have merged during the period 1981-2011, and our final dataset contains the vote results for 2,389 municipalities.

2.2.2 Votes in the Parliament

Our second dataset consists of all the votes of the members of the National Council since the beginning of the current legislature, between December 2011 and December 2013. There were 2,494 votes by the 200 national councilors during this period[3]. In order to compare the opinions given on smartvote with the votes in the National Council, we discard the votes of the councilors that did not reply to the smartvote survey, hence our final dataset contains the votes (or abstentions) of 181 national councilors.

2.2.3 Opinions Expressed on Smartvote

Our third dataset consists of the responses given on the smartvote VAA [3] by the citizens and candidates during the Swiss parliamentary elections of 2011[4]. Smartvote proposes a long and a short survey. The short survey is composed of 32 questions and the long survey is composed of 75 questions, that include the 32 questions from the short survey. The *voters* (i.e., the visitors of the website) had the freedom to choose which survey to answer, but the candidates had to answer all the questions of the long survey. The questions address various topics ranging from society to economy and finance, and they were carefully selected to cover topics as representative as possible of current political issues. An answer consists in selecting one of the following options: *strongly agree - agree - disagree - strongly disagree*. An additional set of "budget questions" require selecting one of the options: *less - no change - more*. Finally, the voters can also select "no answer" (an option not available to the candidates). Each possible answer is mapped internally by smartvote to a number in the set $\{0, 0.25, 0.75, 1\}$ for regular questions, and in the set $\{0, 0.5, 1\}$ for budget questions. The final recommendation given to each voter is a list of candidates, in decreasing order of distance (using the l_2-norm) to this voter [32].

2,985 candidates filled out the survey, which represents about 82.4% of all the candidates. Unless otherwise specified, we consider the responses given by voters who participated in the short survey (which was the most popular survey). This amounts to about 229,000 voters[5], which corresponds to 9.3% of the total voter turnout of 2011. Detailed statistics about this dataset are summarized in Table 3.

3. IDEOLOGICAL SPACE

In this section, we provide a first analysis of the political landscape of Switzerland. We observe that simple dimensionality-reduction techniques can produce useful visual representations of political positions. We then analyze the difference of distribution and polarization between voters and candidates (before and after the elections), in such political spaces. Finally, we compute pairwise similarities between political parties, as measured by the opinions expressed by their members.

3.1 Dimensionality Reduction

In this section, we consider the dataset of opinions expressed on smartvote. Each candidate who took the short survey can be represented as a point in a space of 32 dimensions. Because it is likely that some politicians tend to think similarly on several questions, we can expect that some of these dimensions are strongly correlated. For instance, it could be the case that two persons who answer similarly to the question "Should access to naturalization be made more difficult?" also answer similarly to the question "Are you in favor of legalizing the status of illegal immigrants?". Therefore, one of the first questions that we could ask concerns the intrinsic dimensionality of this dataset. In the following, we use principal component analysis (PCA) [28, 30] in order to compute the sets of questions that best capture these correlations.

Denote by $\mathcal{A}$ the set of possible responses to any question on smartvote[6]. Let n be the number of questions and C the number of candidates. Using this notation, we define $\mathbf{C}$ as the $C \times n$ matrix of candidates' responses, whose $(i,j)^{\text{th}}$ entry $\mathbf{C}_{i,j} \in \mathcal{A}$ is the response of the i-th candidate to the

[3]The data is publicly available via a dedicated web-service: `http://ws.parlament.ch/votes`.

[4]The smartvote dataset can be obtained on demand for research purposes, by sending a request to `contact@smartvote.ch`.

[5]Note that obtaining a precise figure for the number of unique voters is difficult, as one voter can ask several recommendations on the website. This number is an estimate, obtained after filtering out identical web sessions.

[6]We merge budget and regular questions, and take $\mathcal{A} = \{0, 0.25, 0.5, 0.75, 1\}$ for all questions.

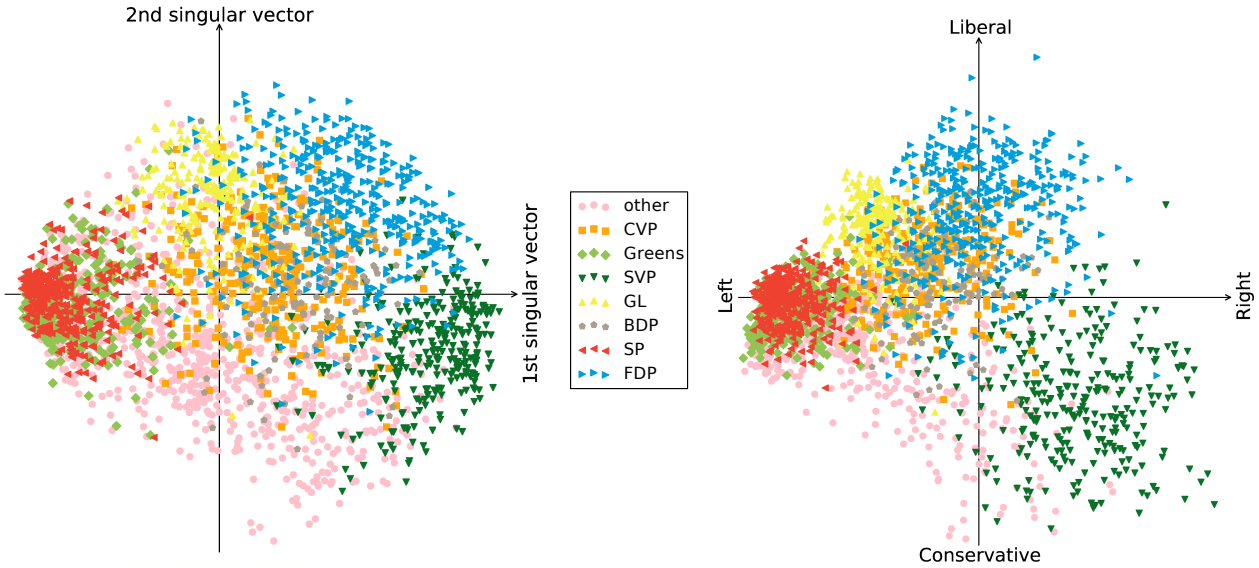

Figure 1: Left: 2-D projection of candidates onto the first two singular vectors of the matrix of their smartvote responses. Right: projection obtained by the smartvote *Smartmap*, with qualitative axes referring to traditional ideological separations.

j-th question. We start by centering $\mathbf{C}$ so that it has zero mean. We then compute the SVD factorization of $\mathbf{C}$ as

$$\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^T,$$

where $\mathbf{U}$ is the $C \times C$ matrix whose columns are the left-singular vectors of $\mathbf{C}$, $\mathbf{\Sigma}$ is a $C \times n$ diagonal matrix, whose n non-zero entries are given by the singular values of $\mathbf{C}$, and $\mathbf{W}$ is the $n \times n$ matrix whose columns are the right-singular vectors of $\mathbf{C}$. We adopt the usual convention, according to which the columns of $\mathbf{U}$ and $\mathbf{W}$, and the diagonal elements of $\mathbf{\Sigma}$ are ordered by decreasing amplitude of the corresponding singular values. The projection of $\mathbf{C}$ onto the basis constituted by its singular vectors is given by $\mathbf{C}' := \mathbf{C}\mathbf{W}$. The matrix $\mathbf{C}'$ has a diagonal covariance matrix, i.e., all its dimensions are uncorrelated. Furthermore, if we denote s_i the singular value associated with the i-th singular vector, the variance of the data along the i-th dimension of $\mathbf{C}'$ is proportional to s_i^2. It follows that, for any $k \leq n$, the first k dimensions of $\mathbf{C}'$ are the k dimensions that capture most of the variance of the data.

We use this property in Figure 1 (left) to obtain a graphical representation of the candidates on the plane, by showing the first two columns of $\mathbf{C}'$, i.e., the projection of $\mathbf{C}$ onto its first two singular vectors. In Figure 1 (right), we also show the representation of the same candidates using the *Smartmap* provided by smartvote [3]. The Smartmap employs a similar dimensionality-reduction technique based on correspondence analysis, and it has been manually validated in order to obtain the correspondence with traditional left/right and liberal/conservative directions.

The relative positions of candidates and political parties are qualitatively similar in both cases, which confirms that our dimensionality-reduction approach is consistent with traditional ideological representations.

Interestingly, PCA easily recovers the usual left/right and liberal/conservative divisions, by looking *only* at the responses (and not at the questions themselves). In Table 4, we show the two most important questions corresponding to the first three singular vectors (i.e., the two questions with the largest absolute weights for each axis). It very clearly appears that the first two axes refer, broadly speaking, to openness and integration of foreigners, and to economic liberalism. Interestingly, the third axis (not used for the 2-D representation in Figure 1) seems dominated by "ethical" issues, such as drug consumption and adoption by same-sex couples.

Singular vector	**First two questions**
1st	1. Would you support foreigners who have lived for at least ten years in Switzerland being given voting and electoral rights at municipal level? 2. Are you in favour of legalizing the status of illegal immigrants?
2nd	1. Are you in favour of the complete liberalization of shop opening times? 2. Should Switzerland conclude an agricultural free trade agreement with the EU?
3rd	1. Should Switzerland legalize the consumption of hard and soft drugs? 2. Should same-sex couples who have registered their partnership be able to adopt children?

Table 4: Two most important questions of the first three singular vectors, for the dataset of candidates' responses to the smartvote survey. These questions are those that contribute the most, in absolute value, to each of the singular vectors. They can be used to interpret the different themes on which the candidates tend to disagree the most.

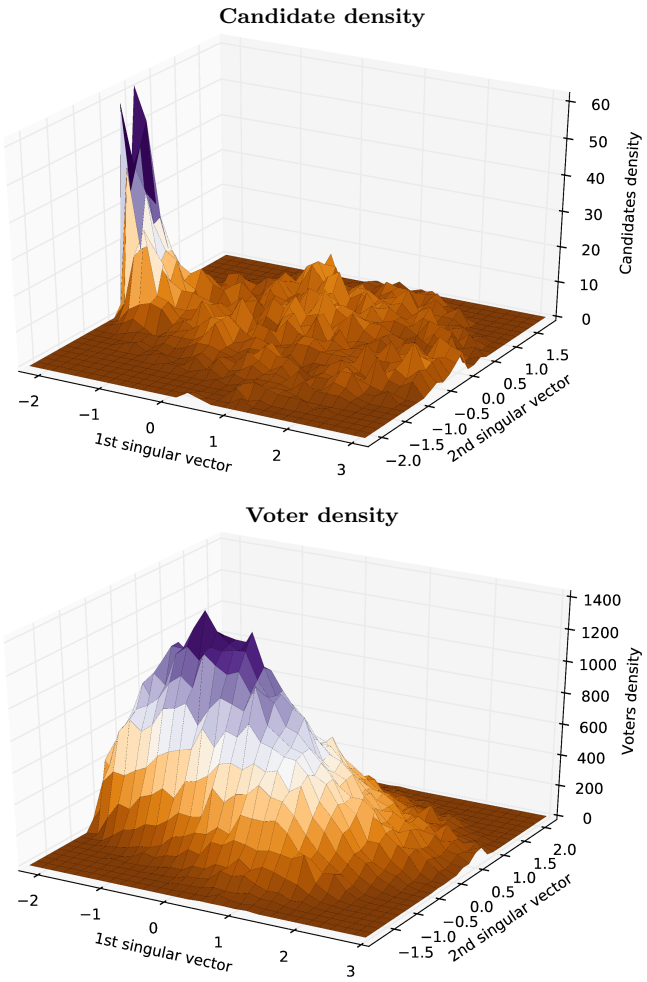

Figure 2: Density of candidates and voters in the ideological space, computed from their smartvote responses. The distributions are very different, and the candidates at the "left" of the space exhibit especially low variance.

3.2 Candidates, Voters, and Polarization

Using the dimensionality-reduction approach presented in the previous section, we can compare the distribution of candidates with that of voters in the ideological space. To this end, we divide the 2-D region of Figure 1 in a 30×30 grid and compute the candidate density as the number of candidates falling into each cell. We follow the same procedure for voters and show both densities in Figure 2. Perhaps the most striking feature of this figure is the comparatively large density of candidates residing on the "left" of the political space. As has already been observed [13], left-wing candidates appear to be very consistent in their responses and exhibit little variance. It seems to be that these candidates, more than the others, tend to strongly agree on the issues raised in the first two singular vectors. It is also possible that this is partly an artifact due to the (publicly admitted [17]) existence of "guidelines" provided by some parties and used by their candidates to answer smartvote questions.

The difference between the two densities of Figure 2 also suggests that politicians are somewhat more polarized than citizens. This fact has often been observed by political scientists, in particular in Switzerland [21]. It is confirmed by the first two plots in Figure 4 that show the proportion of total variance that is captured by each of the first three singular vectors (as well as the remaining variance, captured by the remaining singular vectors). We see that the first three singular vectors capture about a third of the variance in the voters responses, while candidates have 58% of their variance captured in these first three dimensions.

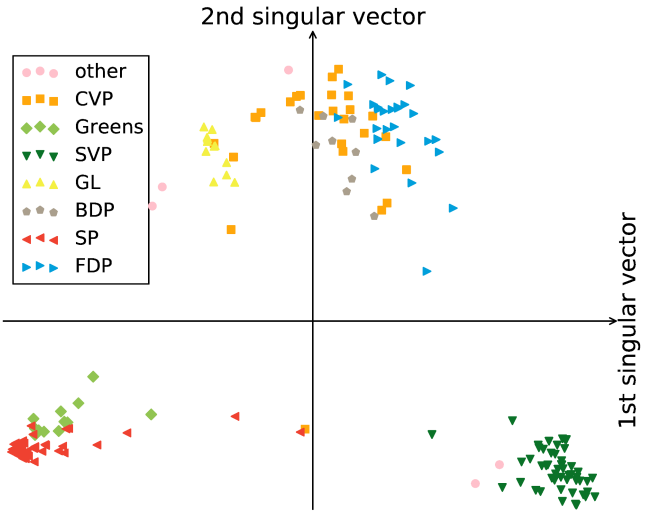

Figure 3: 2-D representation of parliament members, obtained from the dataset of votes in the parliament. The votes in the parliament are more clustered and polarized than the pre-electoral opinions given by candidates and represented in Figure 1.

To further investigate the polarization of politicians, we apply the same dimensionality-reduction approach to the dataset of parliament votes. The resulting 2-D representation of the members of the parliament is shown in Figure 3. Once elected, politicians are much more clustered (which is essentially explained by the existence of coalitions in the parliament). We also show in the last plot of Figure 4 the variance captured by the singular vectors of the dataset of the parliament votes. It confirms that votes in the parliament are strongly polarized, with 66% of the total variance explained by only the first three axes. The candidates, in contrast, are somewhat less polarized during the pre-electoral campaign, but still significantly more than the voters.

3.2.1 Party Overlaps

Figure 1 shows that some subsets of the political parties significantly overlap with each other. In order to check whether such overlaps still exist in the original 32-dimensional space, we compute, for each party, the proportion of candidates of this party who are closer to the median answer of the candidates of *at least* one other party, than to the median of their own party. These proportions are shown in Figure 5. It appears that several of the main parties have a large proportion of their candidates who are closer to at least one other party. This concerns more than 20% of the candidates of four of the seven parties. The FDP, CVP and BDP show exceptionally large figures; more than 35% of FDP, 45% of CVP and 50% of BDP candidates are closer to the median

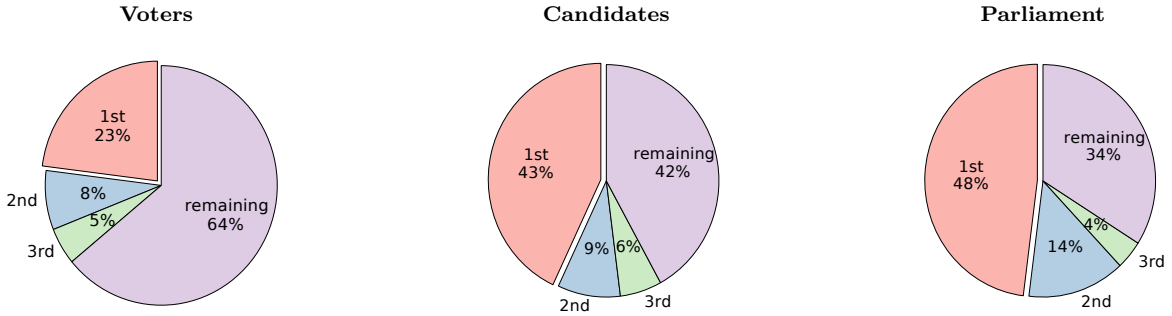

Figure 4: Proportion of the variance captured by the first three singular vectors of candidates, voters, and parliament members. Votes at the parliament are more polarized than the opinions given on smartvote. In turn, the opinions given by the candidates are more polarized than the opinions given by the voters.

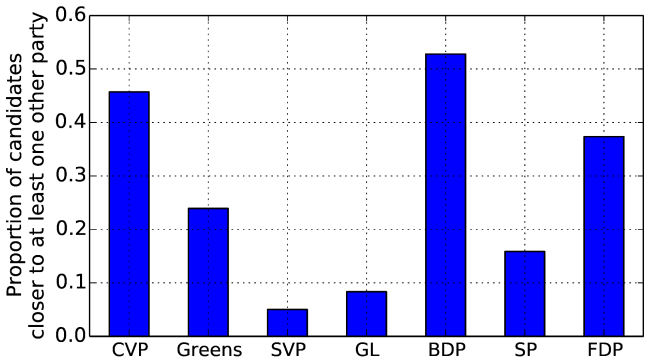

Figure 5: Proportions of candidates of each party that are closer to the median of *at least* one other party than to the median of their own party.

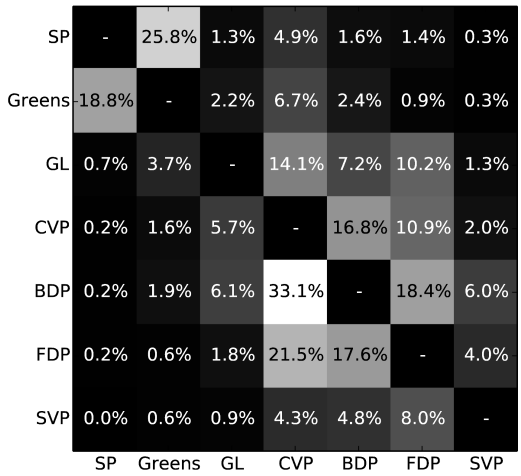

Figure 6: Inter-party overlaps. The number in row i and column j indicates the percentage of candidates of party i that are closer to the median position of party j.

answer of at least one other party. These parties do not belong to political extremities, but rather share a region near the center of the political space, which partly explains why they largely overlap. In practice, this means that using smartvote questions, it is hard to determine which party suits best a person with centrist opinions.

In order to gain more insights into which parties are actually closer, we look at detailed pairwise overlaps. Specifically, for each pair of parties (i, j), we show in Figure 6 the proportion of candidates of party i who are closer to the median of party j than to the median of their own party i. Note that here too, these proportions are computed in the original space, and are thus not subject to distortion due to dimensionality reduction. It is surprising that even opposite parties (such as SVP and SP, or SVP and Greens) have a few overlapping candidates.

4. MUNICIPALITIES

In this section, we take a closer look at the voting patterns of municipalities, with respect to both their main language and their geographical location. Then, we compare the outcome of votes in municipalities with the outcome at the federal (national) level, and show that it is possible to identify unique municipalities that have a great predictive power.

Throughout this section, we use the dataset of votes at the municipality level, described in Section 2.2. It contains the results of federal votes in Swiss municipalities, as well as the result at the federal level (whether the object was accepted or not). We denote by $\mathbf{M}$ the $M \times V$ matrix containing, for each municipality $m \in \{1, \ldots, M\}$ and federal vote $v \in \{1, \ldots, V\}$, the proportion $\mathbf{M}_{mv}$ of *yes* obtained in the municipality for this vote. Finally, we write $\mathbf{o} = \{o_v, v = 1, \ldots, V\}$ for the outcome of these votes at the federal level. We have $M = 2,389$ municipalities and $V = 245$ votes.

4.1 The Infamous "Röstigraben"

Following the dimensionality reduction procedure described in Section 3.1, we project each line of $\mathbf{M}$, corresponding to the results of all votes in each municipality, onto the first two singular vectors of $\mathbf{M}$, and show the result in Figure 8. In this figure, each municipality is represented by a point whose shape indicates the language spoken by the majority.

The figure shows two clear clusters, corresponding to the French-speaking municipalities on one side, and the remaining municipalities on the other, separated by what Swiss people humorously call the *Röstigraben*[7] (literally *hashbrown curtain*). The gap between the two clusters reflects the difference in votes that often arise during federal elections in

[7]This term describes the cultural difference between the German-speaking Switzerland, on one side, and the French-speaking part (sometimes together with the Italian-speaking part) on the other.

Switzerland, where the results from French-speaking cantons are different from those of German-speaking cantons. It is interesting to note that while the Italian-speaking municipalities are culturally closer to the French-speaking ones (and are usually placed on the same side of the Röstigraben), their voting patterns seem to be closer to those of German-speaking municipalities in this projection.

To investigate the relationship between the geographical location of a municipality and its voting pattern, we map each point of the two-dimensional space represented in Figure 8 to a color, illustrated by the gradient in the upper-right corner of Figure 9. We then draw the map of Switzerland in Figure 9, where each municipality is shown with the color corresponding to its location in Figure 8. Thus, two municipalities having similar voting patterns have a similar color on the map. Lakes and municipalities for which some vote results are missing (e.g., due to a merging of municipalities) are shown in white[8].

Again, the separation between the French and German-speaking parts is clearly visible. Moreover, it is possible to identify different types of municipalities: urban centers, such as the greater areas of Geneva, Lausanne, Bern and the Zürich area have relatively similar tints of green, indicating that they share similar voting patterns, whereas rural areas in the German-speaking part share a deep purple color. It is interesting that the French-speaking part of the mountainous canton of *Valais*, located in the southwestern part of Switzerland, has its own unique voting pattern, shown in light blue.

4.2 Vote Outcome Prediction

We have seen in Section 4.1 that municipalities vary substantially in their voting patterns. One question that arises from this observation is whether it is possible to find one municipality whose voting behavior is representative of the global national outcomes. To answer this question, we study in this subsection the predictability of the outcome of votes at the federal level, using the outcome in a single municipality as unique feature.

We therefore define the following learning problem: Given the outcome $\mathbf{M}_{mv} \in [0,1]$ of vote v in a municipality m, can we predict its outcome $o_v \in \{\text{yes}, \text{no}\}$ at the federal level? We split our dataset of 245 votes by taking the first 80% (196 votes) as a training set, and the remaining 20% (49 votes) as a test set. We train one binary classifier[9] for each municipality $m \in \{1, \ldots, M\}$. The parameters of the classifiers are selected using a 10-fold cross-validation on the training set. Figure 7 shows the cumulative distribution function of the accuracy of these M classifiers, averaged over the 10 validation sets (i.e., over the 10 cross-validation folds).

The results are quite surprising: about 10% of municipalities correspond to an accuracy higher than 90%, which means that knowing their results allows us to predict the outcome at the federal level with less than 10% of mistakes. Moreover, some municipalities reach accuracies of more than 96% on the validation set. The municipality reaching the highest average prediction accuracy on the validation sets is *Ebikon*, a town of 12,000 inhabitants in the canton of Lucerne. We evaluated the accuracy of the predictor which uses the vote outcome of Ebikon as feature, and we found that it obtains

[8]A more detailed map can be found online [2].

[9]We use a Gradient Boosted Decision Tree, implemented in Python with `scikit-learn` [23].

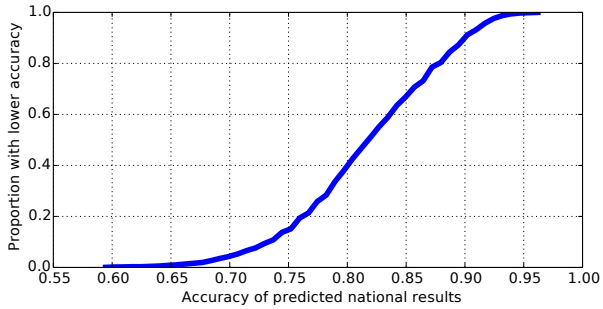

Figure 7: Cumulative distribution function of the accuracy of the prediction of the outcome of votes at the federal level, given the outcome of a single municipality. The accuracies are averaged over 10 cross-validation folds. 10% of municipalities allow to predict vote results at the national level with an accuracy higher than 90%.

a prediction accuracy of 95.9% on the test set. This means that out of the 49 votes of our test set, only 2 are incorrectly predicted by the classifier of Ebikon.

Although surprising[10], these results can be partly explained by the characteristics of Ebikon: located in the heart of Switzerland, it is quite representative of the overall diversity of the country. During the 2007 National Council elections, it had overall demographic features similar to that of Switzerland, and the proportion of votes for the different parties related relatively closely to the proportion of seats obtained. Moreover, as pointed in Figure 8, its voting pattern falls in the bulk of the German-speaking cluster.

Having such a representative sample would be extremely useful to many: Polling institutes, political parties and even news agencies would be able to target this municipality instead of sampling the population at random, thus maximizing the utility of their opinion surveys.

5. OTHER USES OF VAA'S

We showed in Section 3 that the data obtained from VAA's can be used to dress a fairly interesting portrait of the political landscape of the country. In this section, we show that an unscrupulous candidate could turn the public opinions of her adversaries to her advantage, by crafting a specific profile that would gather more voting recommendation than any other candidate. Moreover, we also show how a concerned citizen could turn these public profiles against potential cheaters, by comparing their votes after their election with their advertised opinion in order to spot changes of opinion.

5.1 Crafting the Ideal Opinion

As explained in Section 2.2, the smartvote VAA emits voting recommendations to each visitor by first computing the l_2-distance between her responses and those of each candidate, and then recommending the candidates who are the closest. This means that the responses a candidate gives to each question in the smartvote survey influences directly the number of voting recommendations she gets. Hence, it

[10]A similar effect was observed in the USA for the presidential elections, with the state of Ohio.

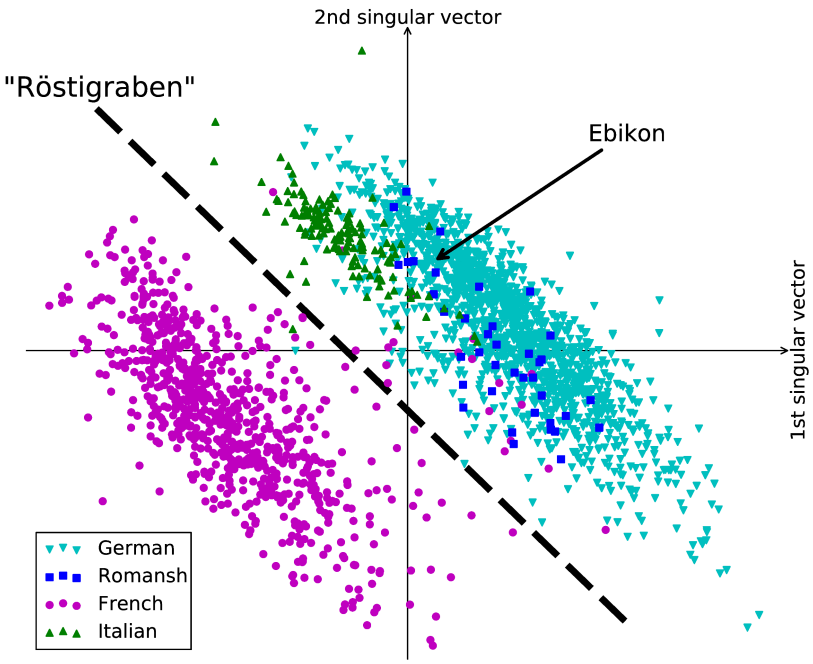

Figure 8: Projection of the vote results in each municipality onto the first two singular vectors of the municipality votes matrix M. The shape of each point indicates the language spoken by the majority in the municipality. A clear separation is visible between the French-speaking municipalities and the remaining municipalities.

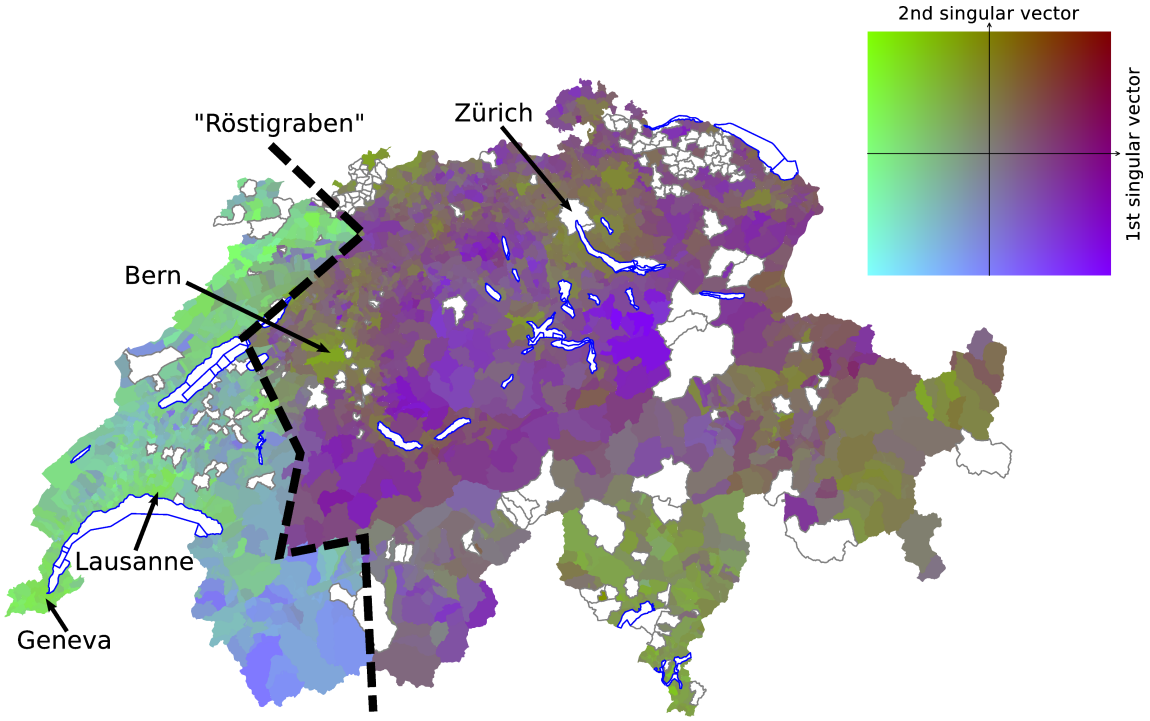

Figure 9: Voting patterns of Swiss municipalities. The color of a municipality is assigned using its location in Figure 8 and the color gradient shown in the upper right corner. Two municipalities with similar colors have similar voting patterns. The *Röstigraben*, corresponding to the cultural difference between French-speaking municipalities and German-speaking ones, is clearly visible from the difference in voting patterns. Regions shown in white are lakes or municipalities for which some vote results are missing (due to a merging of municipalities, for example). A more detailed map can be found online [2].

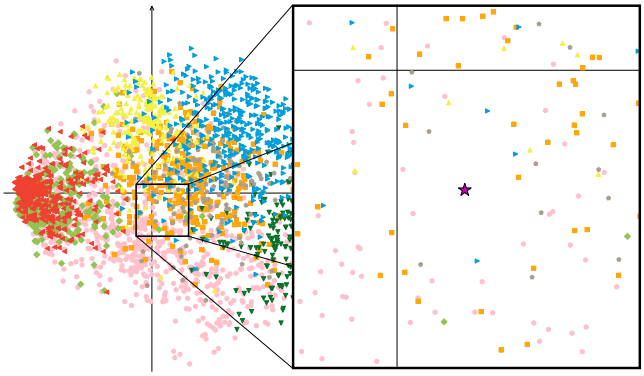

Figure 10: Zoom in the 2-D representation of candidates shown in Figure 1. We clearly see an area with no candidates, which we choose for the location of our crafted opinion, represented as a star.

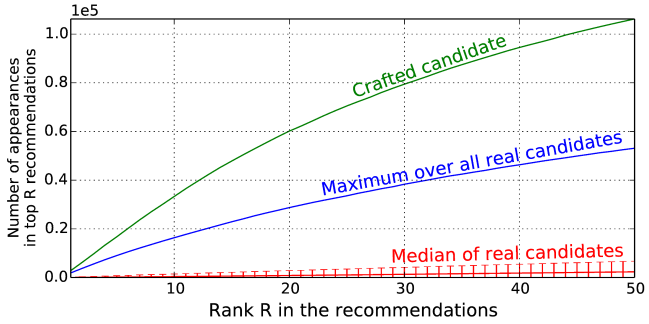

Figure 11: Number of appearances of candidates in the top recommendations. The curves show how many times the median, best, and crafted candidates appear in the top R recommendations for voters. The crafted candidate uses responses corresponding to the star shown in Figure 10. It gets more recommendations than any other candidates, appearing in the top 50 recommendations for close to half of the 230,000 voters.

is interesting to see if it is possible to create an "optimized" profile, in order to get as many recommendations as possible.

Computing the optimal set of answers that maximize the likelihood of a candidate to appear on top of recommendations would require to know both the answers of candidates and voters. However, at the time of completing the survey, a candidate can only access the answers given by her fellow candidates (which are publicly listed on the website). Furthermore, even if the set of answers given by voters were known in advance, the computation of an optimal profile is of combinatorial complexity; if there are n questions with k possible choices, an exhaustive search requires $O(k^n)$ computations. More efficient techniques (e.g., based on geometric approximation [16]) could be used to solve this problem. We leave a more formal study of this optimization problem for future work.

Instead, we propose a simple but efficient heuristic to craft a new candidate profile, by looking only at the answers of the other candidates. Our method consists in inspecting the distribution of candidates in the two-dimensional ideological space depicted in Figure 1. We see that there are several spots where the density of candidates is quite low. However, from Figure 2, we know that voters tend to have a more uniform distribution, thus suggesting that these spots might correspond to "unrepresented" citizens. Thus, we choose to place our crafted candidate in one of those spots, filling a gap in the ideological space but staying far from the extremes.

Such an "optimal" positioning problem has been studied from a game-theoretical point of view in simpler settings [10, 22], and it has been shown that choosing the median position leads to the best results. However, selecting the median answer to each question as our crafted profile did not give satisfactory results in this setting.

To compute the actual responses this crafted candidate should give to the smartvote survey, we proceed as follows: First, we get the coordinates of an empty spot in the ideological space, represented in Figure 1, that is still close to the center of the space. The intuition behind this choice is that we want to be as far as possible from any other candidate, and still be close to the majority of voters. Such a location is illustrated in Figure 10. Then, we perform the inverse operation of the projection explained in Section 3.1, to project a 2-dimensional point back onto the 32-dimensional space of smartvote responses. Because the responses can only take values in $\mathcal{A}$, we round each component of the resulting projected answer to the closest value.

Finally, we add this crafted set of responses (obtained from the point shown in Figure 10) to the list of candidates and compute recommendations for each voter. We count, for each candidate, the number of times she appears in the top R recommendations of a voter, for $R \in \{1, \ldots, 50\}$ and show the results in Figure 11. The lower curve shows how many times the median candidate appears in the top R recommendations, and the error bars indicate the standard deviation. The middle curve shows the maximum number of times a real candidate appears in the top R recommendations. The upper curve shows how many times our crafted profile appears in the top R recommendations.

We see that our crafted profile appears significantly more often in the top recommendations than any other candidate. For example, it appears more than 100,000 times in the top 50 recommendations, about twice as much as the best real candidate. As our dataset consists of around 230,000 voters, this means that our crafted profile is recommended to almost half of the voters. Although the effect of these recommendations on direct votes has not been clearly determined [34], Ladner et al. indicate that 67% of smartvote users state that smartvote had an influence on their choice of party [18]. This influence is even more significant for swing voters [19]. Thus, both parties and individual candidates would benefit from an increased number of recommendations.

5.2 Detecting Opinion Shifts

We showed above how an unscrupulous candidate could craft a profile that would gather more recommendations than any other. This could result in the election of this candidate, who would then have to vote daily in the parliament. However, in this case, the votes she would cast in the parliament may not be in accordance with the opinion expressed by her crafted smartvote responses. As all votes of the members of the parliament are publicly disclosed, a concerned citizen could monitor legislators in order to detect *flip-floppers*, i.e., candidates changing their opinion after they are elected. We

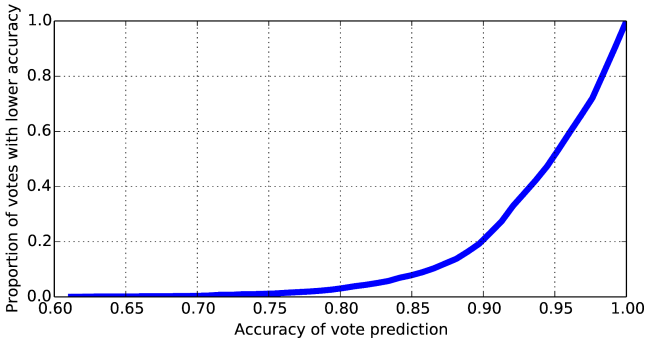

Figure 12: Cumulative distribution function of the predictability of parliament votes from smartvote responses. For each issue in the parliament dataset, we use the smartvote profile of candidates to predict their votes, and report the average accuracy over 10 folds, where 90% of the candidates are used for training and 10% for evaluation. We see that close to 50% of votes can be predicted with an accuracy of 95% or more, using only the smartvote profiles of legislators.

propose here a method to measure the shift in opinion of candidates, between the profile they advertised on smartvote (or any other VAA) during an electoral campaign, and their voting patterns in the parliament once they are elected. Note that our method aims only at quantifying opinion shifts. Of course, there are many contexts where politicians can be reasonably expected to change opinions with time, and moderate opinions shifts need not always be interpreted as bad signals.

5.2.1 Predicting Parliament Votes

The first step towards detecting changes of opinion is to map a set of smartvote responses to votes in the parliament. To do so, we identify parliament votes that can be predicted by smartvote responses. Indeed, the intuition is that, as smartvote responses are a good indicator of a candidate's political opinion, some votes can be accurately predicted from a set of smartvote responses. Therefore, we define the following learning problem: Given the smartvote profile of all candidates $\mathbf{C}$ and their votes $\mathbf{v}$ at the parliament on a given issue, learn a model that predicts the vote $\mathbf{v}_c \in \{\text{yes}, \text{no}\}$ of a candidate $c \in \{1, \ldots, C\}$ on this issue, from her smartvote opinion $\mathbf{C}_{c\cdot}$.

We train a linear classifier[11] for each of the 2,494 votes in the parliament dataset. For each vote, we filter candidates to keep those that actually voted (some are sometimes absent, or abstain) as learning samples. We evaluate the predictability of each vote by computing the prediction accuracy of our linear classifier using on 10 folds, where, for each fold, the classifier is trained on 90% of the candidates and evaluated on the remaining 10%. We then compute the average accuracy on these 10 folds, and report the results in Figure 12.

Figure 12 shows the cumulative distribution function of the prediction accuracy for each vote. We observe that the vast majority of votes at the parliament can be predicted with a high accuracy from smartvote profiles; more than 90% of votes can be predicted with an accuracy higher than 85%, and close to 50% of the votes can be predicted with an accuracy higher than 95%.

[11] We use Logistic Regression, implemented in Python with `scikit-learn` [23].

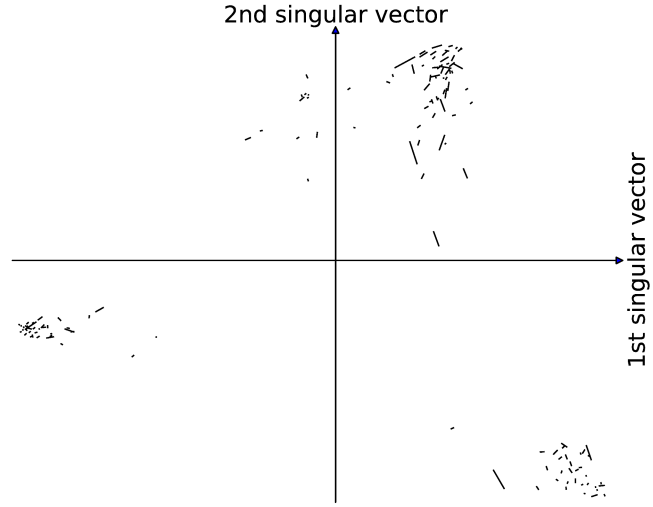

Figure 13: Opinion shifts of parliament members. Each segment represents a legislator, and goes from her expected votes (according to her smartvote profile) to her actual votes. The median legislator has only 0.3% of votes that are different than what is expected from her advertised opinion. The largest difference is 3.75%. Interestingly, the magnitudes of the shifts seem to be different for the three coalitions.

5.2.2 Computing Opinion Shifts

Now that we have a way to map smartvote opinions to parliament votes, we can compute the expected votes of legislators, based on their smartvote profile, and compare them with their actual votes. To do so, we first choose the 1,000 most predictable votes, in order to maximize the confidence in our predicted votes. This corresponds to the top 40% of votes, meaning that each of them can be predicted with an accuracy higher than about 96% (see Figure 12).

We then use the predictors trained in Section 5.2.1 to predict the expected votes of each candidate on these 1,000 issues. This means that, for each candidate, we use her smartvote profile to compute her expected votes on these issues, and we compare them with her actual votes. We compute the proportion of actual votes that differ from the expected votes. This proportion corresponds to the shift in opinion of the candidate, between her smartvote profile and her actual voting behavior in the parliament.

The 181 legislators voted on a median number of 906 issues. The median discrepancy between the votes predicted from smartvote profiles and the actual votes is only 0.3%. This means that the median candidate votes coherently with her advertised smartvote opinion 99.7% of the time. The candidate with the largest discrepancy has 3.75% of her votes in opposition to her advertised opinion. While this distance is an order of magnitude larger than the median distance, it still means that 96 votes out of 100 are coherent with what she advertised, which is a somewhat reassuring observation. A larger distance could mean that she falsely advertised her opinion on smartvote, or that she "flip-flopped", i.e.,

she changed her opinion significantly after being elected. However, it can also be expected that legislators sometimes divert from their advertised positions, for example to follow their party on a specific issue. Thus, one should be careful when interpreting such differences between expected and actual votes.

To visualize these opinion shifts, we show in Figure 13 the 2-D representation[12] of the expected and actual votes of each councilor, computed as explained in Section 3.1. Each candidate is represented as a segment, with one end corresponding to her expected votes, and the other to her actual votes. The longer a segment, the more significant the shift in opinion between her smartvote profile and her votes in the parliament. Interestingly, the magnitudes of the shifts seem to be different for the three coalitions.

6. RELATED WORK

Spatial approaches are often used to represent politicians or parties, most often using one or two dimensions. Some papers use dimensionality reduction techniques similar to ours [13, 31, 35]. However, to the best of our knowledge, we are the first to apply it on datasets of this scale. Furthermore, we show how it can be used to craft ideal VAA profiles, and put it in contrast with parliamentary and municipal votes.

Some researchers studied roll calls at the U.S. Congress [26, 25, 9]. For instance, Poole et al. study voting patterns at the Congress [26], and find that legislators can be described in a space of low dimensionality. Based on spatial voting theory [11], Enelow et al. propose a method to predict congressional votes. Their method relies only on past congressional votes to make predictions. While we study the predictability of votes, we do not use our predictors to predict future votes. Instead, we propose a method that permits to map one space (the opinions expressed on a VAA) onto another (the votes in the parliament), in order quantify opinion shifts.

Hansen et al. [15] explore the cohesiveness of political parties using VAA data, by measuring the agreement among party members. We propose a different approach, which allows us to measure the overlap between each pair of parties.

While we focus on predicting the results of national *issue voting* in Section 4.2, there is a large body of work that focuses on predicting the results of *elections*. For instance, Armstrong et al. [8] use biographical information about candidates to predict U.S. election results. In addition, several studies focus on Twitter data to predict the outcome of elections, from Germany [33] to the Netherlands [27] and Singapore [29]. However, some researchers (see e.g., Gayo [12]) have warned against relying only on tweets to predict election results, arguing that the data is inherently biased and that missing signals could be more important than observed ones.

Related to the votes prediction and the opinion shifts measurement that we propose in Section 5, Gerrish et al. [14] study the prediction of lawmakers' position on a bill, using the text of the bill. The authors use the resulting model to explore how lawmakers deviate from their expected voting patterns. Finally, Poole studies members of the U.S. Congress [24], and finds that they "adopt a consistent ideological position and maintain it over time".

[12]We restrict the parliament dataset to the 1,000 most predictable votes, instead of all 2,494 votes, resulting in a projection slightly different than that shown in Figure 3.

7. CONCLUSION

We proposed a data-mining approach towards using massive open government and VAA datasets to study different aspects of a country's politics. We considered the case of Switzerland, as this country has a strong democratic culture with a diversified political landscape. We observed that the scale of the data enables statistical approaches to uncover patterns that usually require manual investigations by domain experts.

We compared the polarization of voters with that of politicians, before and after the elections. We found out that some parties have more than 40% of their candidates that are closer to at least one other party. We showed that it is possible to learn models that predict vote outcomes at the national level with an accuracy higher than 95%, by looking at the outcome in a single municipality.

We described how an unscrupulous candidate could craft a synthetic VAA profile, in order to gather a very large number of voting recommendations. However, we also proposed a technique to hold a legislator accountable for her opinions expressed on a VAA, by mapping VAA responses to votes in the parliament and comparing her expected vote with her actual votes. Our technique can be used to spot legislators that vote in contradiction to the opinions that they expressed on a VAA.

Overall, our work applies to any country where similar data is available, and it points to some avenues created by open government initiatives that enable new data-mining approaches to political and social sciences.

8. ACKNOWLEDGMENTS

We would like to thank Jan Fivaz, Julien Fiechter and the rest of the smartvote team for providing us with their dataset, and answering the many questions we had about its format and use. We would also like to thank Olivier Dousse for the several insightful initial discussions on this topic.

9. REFERENCES

[1] Datacatalogs.org. `http://datacatalogs.org/`.
[2] Predikon eigenmap. `http://www.predikon.ch/eigenmap`.
[3] Smartvote. `http://www.smartvote.ch`.
[4] Stemtest. `http://www.stemtest2014.be/`.
[5] Stemwijzer. `http://www.stemwijzer.nl/`.
[6] Vote compass. `http://votecompass.com/`.
[7] Wahl-o-mat. `http://www.bpb.de/politik/wahlen/wahl-o-mat/`.
[8] J Scott Armstrong and Andreas Graefe. Predicting elections from biographical information about candidates: A test of the index method. *Journal of Business Research*, 64(7):699–706, 2011.
[9] Joshua Clinton, Simon Jackman, and Douglas Rivers. The statistical analysis of roll call data. *American Political Science Review*, 98(02):355–370, 2004.
[10] Anthony Downs. An economic theory of political action in a democracy. *The Journal of Political Economy*, pages 135–150, 1957.
[11] James M Enelow and Melvin J Hinich. *The spatial theory of voting: An introduction*. CUP Archive, 1984.

[12] Daniel Gayo-Avello. Don't turn social media into another 'literary digest' poll. *Communications of the ACM*, 54(10):121–128, 2011.

[13] Micha Germann, Fernando Mendez, Uwe Serdult, and Jonathan Wheatley. Exploiting smartvote data for the ideological mapping of swiss political parties. In *XXVI Congress of the Italian Political Science Association*, pages 13–15, 2012.

[14] Sean Gerrish and David M. Blei. How they vote: Issue-adjusted models of legislative behavior. In *NIPS*, pages 2762–2770, 2012.

[15] Martin Ejnar Hansen and Niels Erik Kaaber Rasmussen. Does running for the same party imply similar policy preferences? evidence from voting advice applications. *Representation*, 49(2):189–205, 2013.

[16] Sariel Har-Peled. *Geometric approximation algorithms*. Number 173. American Mathematical Soc., 2011.

[17] 24 heures. Sur smartvote, certains partis aident les candidats à répondre, August 2011. `http://goo.gl/Uru3Gt`.

[18] Andreas Ladner, Gabriela Felder, and Jan Fivaz. More than toys? a first assessment of voting advice applications in switzerland. *Voting Advice Applications in Europe. The State of the Art*, pages 91–123, 2010.

[19] Andreas Ladner, Jan Fivaz, and Joëlle Pianzola. Voting advice applications and party choice: evidence from smartvote users in switzerland. *International Journal of Electronic Governance*, 5(3):367–387, 2012.

[20] Daniel Lathrop and Laurel Ruma. *Open government: Collaboration, transparency, and participation in practice*. O'Reilly Media, Inc., 2010.

[21] Philipp Leimgruber, Dominik Hangartner, and Lucas Leemann. Comparing candidates and citizens in the ideological space. *Swiss Political Science Review*, 2010.

[22] Roger B Myerson and Robert J Weber. A theory of voting equilibria. *American Political Science Review*, 87(01):102–114, 1993.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] Keith T Poole. Changing minds? not in congress! *Public Choice*, 131(3-4):435–451, 2007.

[25] Keith T Poole and Howard Rosenthal. A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384, 1985.

[26] Keith T Poole and Howard Rosenthal. Patterns of congressional voting. *American Journal of Political Science*, pages 228–278, 1991.

[27] Erik Tjong Kim Sang and Johan Bos. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of SASN 2012*, pages 53–60. EACL, 2012.

[28] Jonathon Shlens. A tutorial on principal component analysis. *Systems Neurobiology Laboratory, University of California at San Diego*, 2005.

[29] Marko Skoric, Nathaniel Poor, Palakorn Achananuparp, Ee-Peng Lim, and Jing Jiang. Tweets and votes: A study of the 2011 singapore general election. In *HICSS 2012*, pages 2583–2591. IEEE, 2012.

[30] Lindsay I Smith. A tutorial on principal components analysis. *Cornell University, USA*, 51:52, 2002.

[31] Jaakko Talonen and Mika Sulkava. Analyzing parliamentary elections based on voting advice application data. In *IDA X*, pages 340–351. Springer, 2011.

[32] Luis Terán, Jan Fivaz, and Stefani Gerber. Using a fuzzy-based cluster algorithm for recommending candidates in e-elections. *Fuzzy Methods for Customer Relationship Management and Marketing*, page 115, 2012.

[33] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*, 2010.

[34] Kristjan Vassil. *Voting Smarter? The Impact of Voting Advice Applications on Political Behavior*. PhD thesis, European University Institute, 2011.

[35] Jonathan Wheatley, Christopher Carman, Fernando Mendez, and James Mitchell. The dimensionality of the scottish political space: Results from an experiment on the 2011 holyrood elections. *Party Politics*, 2012.

[36] Wikipedia. Ballot measure, 2013. `https://en.wikipedia.org/wiki/Ballot_measure`.