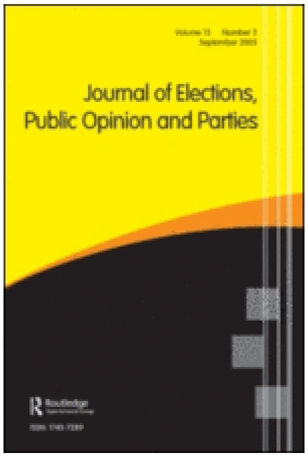


Learning VAAs: A New Method for Matching Users to Parties in Voting Advice Applications

Journal:	*Journal of Elections, Public Opinion & Parties*
Manuscript ID	JEPOP-2018-0141.R2
Manuscript Type:	Original Article
Keywords:	Voting Advice Applications, Machine Learning, Conceptual Spaces, E-Democracy, Issue Voting

Answers to the first Reviewer

We would like to thank the first reviewer for the interesting comments and suggestions. We have considered almost all of the minor and major points suggested and have made significant changes to the paper accordingly.

First, we have mentioned the hypotheses in the introduction and therefore the reader does not have to wait until the Results section. Second, we have included section 2 as subsections of the introduction: ``1.1 What are VAAs", ``1.2 Low and high dimensional matching methods" and ``1.3 Evaluating VAAs. A new Section 2 focuses on a brief explanation of the current methods of recommendation and their flaws; Section 3 introduces our proposal; Section 4 elaborates on the methodology and the data. Section 5 shows the empirical results; Section 6 discusses the results and implications of the proposed method; Section 7 summarises the conclusions. We have also included a sentence (with literature) about the growing number of VAA platforms and users in the introduction, as suggested: "Due to VAAs' momentum, language diversity and geographic span, it has become difficult to obtain the number of them currently in operation (van der Linden and Dufresne 2017)".

The Results section does not start as abruptly anymore as we have added a previous section of Methodology and Data. We have also added some lines explaining the implications of adopting the ``Learning VAAs" in the Discussion section. Moreover, we have included the number of iterations that other community-based VAAs need for stabilising their recommendation. Concerning the missing data, we have included the preprocessing of the dataset as a section in the appendix. As we have mainly considered users who have answered all the questions, we believe that the missing data has not affected the VAA recommendation significantly.

Furthermore, we have also addressed the two suggestions that we deemed most important: validating the algorithm in several subsets of the data and expanding the dataset for the experiment. We have validated our method of recommendation by running the algorithm in several random subsets instead of one and explained the nuances of it in the text. Regarding the dataset, we have included more countries in the analysis of the EU-Vox 2014 (Portugal, England, Italy, and Greece) and have extended the analysis to a VAA launched for a first-order election (Aquienvoto 2019, a VAA launched for the Spanish general election in 2019). We have displayed all of our outcomes in the Results section and the results have been similar in all countries and datasets employed.

Last, we have carried out further proofreading to improve the language and revised the phrasings that were not sufficiently clear. We hope that the quality and clarity of the paper has improved substantially from these changes.

Answers to the second Reviewer

We appreciate the efforts made by the second reviewer and want to thank all the thorough and intelligent comments. We have considered all the suggestions and believe that the paper substantially improved as a result. We agree with one of their firsts considerations: the fact that the paper ``has a wider contribution in terms of the decisional logic underpinning voter's party choice and a methodological contribution in terms of applying innovative statistical/machine learning techniques". The reviewer considers that the balance of the paper is tilted towards a computer science perspective, and we acknowledge it. To alleviate this concern, we have included part of the mathematical notation and algorithm optimisation in the appendix, giving more focus to the political science aspects.

Concerning the other points mentioned by the reviewer, we nuanced the paper's contribution. Regarding the inductive/deductive approach, the reviewer is right by pointing out our misunderstanding. As it is said by them, the deductive/inductive discussion does not properly fit the context of the paper because it does not capture the innovation of our proposal: the incorporation of empirical user-adapted saliencies and distance matrices. We have consequently changed the terminology. Concerning the low dimensionality discussion, we have included in the introduction the difference between low and high dimensional approaches in VAAs and clarified that our method is a proposal within the "high dimensional" group. As it is said by the reviewer, some inductive methods have been discussed in low dimensional mapping. We have included them in the paper consequently. Moreover, we have duly cited the GESIS data archive in the manuscript.

Regarding the interesting points of the saliency, we agree with much of what it is said by the reviewer. We now mention in the paper that the EUvox had the saliency component by assigning weights. We believe that the Learning VAA may efficiently combine in the future the possibility of assigning weights by the users and the learning component of our proposal. Under the Learning VAA, the user could assign weights to the items (i.e. by saying that an issue is very important for her) and then the algorithm would adapt the weights accordingly to what the user has said (e.g. by never giving more saliency to questions where the user did not say anything compared to questions that marked as very important for her). Another alternative would be to cluster users by their agreement in their answers and then train independent models that would assign a different saliency distribution to each cluster. This modification would include some degree of customisation of the model depending on the user preferences while adapting the model from like-minded users. However, although interesting, we feel that we cannot solve this problem in the current paper and have discussed these alternatives in the Discussion section. We still believe that our method captures better users' preferences when considering the saliency than the standard ones, since employing a collective saliency of the items is an advance with respect to static methods. Combining the Learning VAA method with other improvements to make the weight of the policy issues more customised can be the next step. With regard to removing issues from the computation of the recommendation, this feature is still available by answering "No opinion".

Concerning transparency, we agree that it may not be such a core contribution of our method and have made our statements more nuanced. However, although the computation of saliency weights and distance matrices is out of the understanding of non-experts, the user can still calculate by hand their match with a party exactly in the same way as in traditional VAAs if provided with the saliency weights and distance matrices. Regarding the community-based methods, they indeed do not need to be a ``black box" and we have modified this point to include the nuances. However, it is also true that many of the Social VAA methods in different papers can be accurately described as a ``black box" (and we have identified them in the paper). We argue in the new version of the paper that our proposal is as good as the best community-based method in terms of transparency and, in a striking difference with most of them, ours compares between users and parties' positions, and not exclusively from other users. We have included the reference that an alternative community-based method was available to the users in the EUVox2014.

Regarding the empirical analysis, we have tried to improve the clarity of our analysis. Probably because the European elections are a single constituency, the EUvox did not consider the regional differences in Spain and matched users and political parties following only their answers to the policy issues. A VAA giving an adjusted saliency can better capture the importance of some questions for most of the users, such as those about the unity of Spain and, and, in this case, not over-recommend nationalist parties. We have included a second dataset (Aquienvoto) in our analysis that corresponds to a first-order election and also four more countries in the EUVox 2014: England, Italy, Greece and Portugal.

Answers to the third Reviewer

We want to thank the third reviewer for the excellent comments which have been extremely useful to improve the paper. Regarding the main point, we have considerably changed the structure and content of the paper to better accustom a political science journal, and particularly to previous JEPOP publications which have addressed the same topic (such as Mendez [2017]). We have as well extended the analysis to other European countries and a second dataset.

Regarding the main considerations of the review, we consider that any community-based recommendation method shares the problem of users' misinformation. There is a trade-off between using empirical evidence from user data for recommending and the problem of misinformation. As far as our knowledge goes, no community-based method in the literature has addressed this problem either. The reviewer argues that misinformation is not randomly distributed and therefore using a voting intention learning algorithm will potentially learn this misguided information and spread it to other users. This is a serious concern and it is true that it has not been sufficiently addressed in the first version of the manuscript (however, note that a similar point can be made to most papers studying Social VAAs). To alleviate this concern, we have raised the standards of the slice of data on which the algorithm is trained and have discussed the problem of misinformation in the paper. In the new version, only users who are educated, politically informed and vote because of ideological considerations (issue voters) are selected for training the model. We define such users as the ones who reply ``Very" to the question ``To what extent are you interested in politics?", reply ``The ideas of the party are closest to my own" to ``What is the main reason for voting for your party of choice?", and state they have at least high-school education. Considering all the above, the current assumption of our model is more solid: the majority of those politically motivated and educated issue-voters know the positions of the parties they will vote for. In the appendix, we have explained all the data which has not been used due to our requirements in political sophistication and level of education. Therefore, as suggested, the model does not learn with ``standard VAA users". Selecting only those users that declare that ``the ideas of the party are closest to my own" may indeed be an insufficient selection criterion, but using the combination of educated, issue voters and very interested in politics type of users may alleviate these problems. In any case, we consider that further research should be done regarding the learning model and, due to space limitations, we leave further investigation for future work.

We have also included three extra countries in our analysis of the EU-Vox 2014 ---Portugal, Italy, England and Greece---, and another VAA launched in a first-order election (Aquienvoto, the VAA launched for the 2019 Spanish General Elections). This test bench should give a better picture of the flaws and benefits of the proposed method of recommendation. We have also added in the appendix more information about the cleaning process of the database to ensure a high quality of the data, as otherwise the learning method can be mislead. However, we have not been able of producing all the models that the reviewer suggested. There are many distance matrices resulting from the experiments (30 directly displayed in the paper and 134 in an online repository). We have made a preliminary analysis and did not find significant differences in the matrices of the different socio-demographic users, but a more rigorous analysis should be done (which are beyond the objectives of this paper). We consider that this can be a paper by its own, as it could shed light on the different conceptual spaces among voters. We have not been able to compare the Learning VAA with different ways of assigning the weights for standard VAAs either (although, as far as our knowledge goes, this has not been done by any other study on VAAs). This is because, unfortunately, we do not have data on VAAs in which the weights are assigned by experts, and the information about the importance that users give to the policy issues is missing for the EUVox 2014 dataset. Again, we believe that this would be a very interesting contribution that should be addressed in the future, but it is beyond the scope of this paper and could constitute another paper by its own. Due to all these interesting proposals of the reviewer, we have nuanced the contributions of our paper. Certainly, our proposed method should still be further tested before implementation, but we believe that this paper makes a valuable contribution to the debate on new methods of recommendation.

Concerning transparency, it is true that not all the recommending methods of Social VAAs necessarily lack transparency, and we have added some comments to clarify this misunderstanding. First, in this new version of the paper, our criticism is directed toward those matching methods of Social VAAs (according to our knowledge, most of them) that use a community-based recommendation without considering the positions of the parties. Second, it is true that not all methods proposed in the Social VAA literature are equally transparent. Some of them are easy to interpret and can be understood by the user (such as decision trees). Others, however, are not, since they use complex mathematical transformations with learned parameters that lack direct interpretation (such as neural networks or Support Vector Machines). In our model, all the learned parameters have a direct interpretation in the process of deciding a political party and therefore the models resulting from training are easy to explain. A user would only need to know the saliency of the policy issues and the matrix distance to calculate the party that is closest to her ideas (exactly in the same fashion as traditional VAAs) and, moreover, can interpret the saliency vector and every single distance matrix produced (provided they have enough knowledge to understand what a proximity distance matrix is). It is true that the method for training the model is out of the understanding of users, but we believe that there is a trade-off between the complexity of the method and its capabilities to attain better outcomes.

Regarding the last part of the main considerations, we agree with the reviewer that the model should not ``update the parameters in real time" and we have modified this point. We now propose to train the model only at the beginning, after reaching a minimum number of users and keep this model static (with possible health checks at specific points later on to verify that there were no significant changes in users' positions). This goes in line with the results of previous work on VAAs. For example, Germann et al (2017) found that ``drawing on the first 2.000–5.000 entries tends to lead to good results", while Katakis et al. (2014) claim that the first few hundreds of users already provide well-tuned models.

On the formal comments, we have changed the structure of the paper to early clarify the position of the proposed method in the VAA literature. Now the introduction contains a discussion on the difference between low and high dimensionality and also rapidly clarifies what our proposal of VAA is. We have abandoned the inductive/deductive labels because they lead to misunderstanding, as it was suggested. In general, we believe that now it is easier for the reader to understand ``what the paper is actually about" early on. We followed the suggestions from the reviewer and mention that the paper is a hybrid VAA based on expert coding and user data.

Regarding the mathematical notation, we have performed major simplifications and hope that are now easier to understand.

Last, we have also followed the suggestions made by the reviewer and included the section ``What are VAAs" in the introduction, we have linked the discussion on directionality and proximity theory to the literature, and have improved the language via proofreading

Learning VAA: A New Method for Matching Users to Parties in Voting Advice Applications

Voting Advice Applications (VAAs) aim to increase voters' political competence by providing them with the closest political party according to their preferences. To do this, VAAs usually compare and aggregate the positions of users and political parties on a set of policy issues by defining a conceptual space and some distance metric on it. In this paper, we argue that the main method for performing the comparison adapts to users' preferences unsatisfactorily because 1) they use unjustified *a priori* decisions for weighting policy issues and 2) they employ the same issue-voting space on all policy issues. Some exceptional cases address these issues by providing a community-based recommendation, but often come with lack of interpretability. To fill these gaps, we propose an adaptive algorithm that learns the configuration of the conceptual space from users' answers. We employ a hybrid VAA that uses expert coding for the party positions and users' data to adjust the calculation of the distance between users and parties. This new matching method, the Learning VAA, innovates by adjusting the saliency and issue-voting space for every policy issue. We argue and empirically demonstrate that our model fits better the users' preferences while providing a higher degree of interpretability.

Keywords: Voting Advice Applications, Machine Learning, Spatial Model, E-democracy, Election Studies, Vote Choice, Issue Voting.

1 Introduction

The main aim of this paper is to contribute to the existing literature on the methods used by Voting Advice Applications (VAAs) to recommend a political party to each user based on their preferences. In the last years, VAAs have spread through different European countries and been used by a considerable part of their population (Garzia and Marschall 2014; Garzia, Trechsel and De Angelis 2017). Due to VAAs' momentum, language diversity and geographic span, it has become difficult to obtain the number of them currently in operation (van der Linden and Dufresne 2017). As a result of their popularity, there is an extensive ongoing discussion on the effects and implications that these applications have on the final voting decisions made by their users (Garzia and Marschall 2012; Anderson et al. 2014; Marschall and Garzia 2014; Garzia, Trechsel and De Angelis 2017).

The discussion on VAAs is crucial as they are not a mere entertainment artefact but have real effects in the political world (Anderson et al. 2014). VAAs are considered to play a role in the electoral process in many countries, and therefore it seems necessary to evaluate the empirical validity and the normative implications of the methodology that they use for providing recommendations (Fossen and Anderson 2014). Due to VAAs' recent growth, several papers and books have been published trying to understand the potential and limitations of this web application (Marschall and Garzia 2014; Anderson et al. 2014). As a result, new statistical and normative questions regarding VAAs have been raised by researchers from different fields (Garzia and Marschall 2014). There is a branch of literature researching methodological questions of VAAs, such as the design of the algorithms matching users and political parties (Mendez 2012; Katakis et al. 2014;

Mendez 2017), the positions of the political parties (Gemenis 2013) or the political dimensions used on VAAs to cluster the policy issues (Germann et al. 2015).

This paper proposes a new method of recommendation and thus focuses on the methodological aspect of the application, although it also addresses some of the normative concerns that current VAAs face. Our goal is to propose a VAA method of aggregating and comparing users' and parties' stances on the policy issues, reaching similar or better performance than previous VAA methods. We argue that standard methods of recommendation have flaws that can be partially improved by using learning algorithms that incorporate empirical user-adapted information for adjusting the saliency and distance metric of the policy issues. Therefore, we propose a hybrid VAA, the Learning VAA, based on expert coding of party positions and users' data. We hypothesize that adding these features will improve VAAs' performance compared to standard methods of recommendation while being more interpretable than community-based methods.

1.1 What are VAAs?

VAAs are online tools that are designed to help users to decide which party to vote for (Garzia and Marschall 2014). For doing so, users need to fill in a questionnaire with their preferences on a set of policy issues previously chosen by the VAA designers. According to their answers to the questions, users receive matching scores to each party based on an aggregation of the user's preferences and the positions of the political parties on these issues. By offering users a recommended party, the policy issues at stake, and the positions of the political parties within them, VAAs are expected to increase voters' political competence (Fossen and van den Brink 2015).

Following spatial theories of voting (Downs 1957), VAAs assume that voters compare their positions with those of the political parties in the set of policy issues that are essential to determine their vote. Once she is situated in this conceptual space (Gärdenfors 2004), she can determine which is the best party to vote for according to her preferences. Therefore, competent voters should 1) compare their policy preferences with those of the parties in a set of relevant policy issues, and 2) aggregate the disagreements in the policy issues in a meaningful way. As this process of comparison and aggregation involves many difficulties, VAAs are supposed to facilitate the task by helping users to see which party better fits their policy preferences. It is essential to note that VAAs are only able to increase voters' political competence by assuming a world in which only the preferences in the policy issues determine which party is preferable to be chosen. This is a strong assumption because it has been empirically tested that many other important factors influence voting behaviour beyond ideology, such as leadership, social group, and parties perceived political competence, and VAAs are unable to treat them fairly (Mendez 2012).

1.2 Low and high dimensional matching methods

VAAs can provide users with recommendations by following a high- or a low-dimensionality matching method (Mendez 2017). While the former considers every policy issue separately in the aggregation algorithm, the latter assumes that there are latent dimensions that explain the positions of the users on a set of intertwined policy issues (Benoit and Laver 2012). Some researchers have focused on the advantages and disadvantages of using low and high dimensional approaches (Mendez 2017; van der Linden and Dufresne 2017; Louwerse and Rosema 2014). High dimensional approaches in VAAs consider that each policy issue constitutes by itself a dimension. They typically use ex-ante chosen matrices to compare the positions of the users and parties and then aggregate them into an agreement score (Mendez 2017).

Meanwhile, low dimensional approaches aggregate the policy issues in some latent dimensions and often offer a score based on the Euclidean distance of the low-dimensional space, with the additional benefit of being able to include an understandable diagram of two (or more) dimensions. Although the discussion on low-dimensional approaches has also involved the possibility of using users' data to inductively adapt the methods (via scale validation) (van der Linden and Dufresne 2017), low-dimensional methods as a proxy for computing recommendation scores have proven to be less reliable than high-dimensional approaches (Louwerse and Rosema 2014; Mendez 2017). Their main contribution comes from the diagrams they produce as an additional source of information for users. Many VAAs, such as *EUvox 2014*, *Horizon 2019* and *Aquienvoto*, use high-dimensional approaches to make their recommendation, even when most VAAs often display a two-dimensional graph with the positions of the user and the parties at stake.

1.3 Evaluating VAAs

VAA designers must unavoidably make some critical decisions which will determine whether the application increases the users' political competence. At a minimum, the designers must consider what the policy issues at stake are, the positions of the political parties within them (Gemenis 2013), how the disagreement between users and parties is computed by a distance function (Mendez 2017) and the saliency of the questions (Lefevere and Walgrave 2014). Therefore, even the best imaginable VAA will not merely reflect "what is at stake in the election by neutrally passing along information" (Fossen and van den Brink 2015, 341). Instead, political information is always structured by developers' decisions.

These key decisions will determine the degree to which the VAA is adapted to the users' preferences; and therefore, how likely it is to increase their political competence. In this paper, we focus on two of them: the distance function and the saliency (weights) of the questions. We also believe that VAAs must be accountable to users and researchers alike since decisions can only be trusted when the process of making them is open and interpretable. Openness addresses the fact that the assumptions made by VAA developers are public, exposing methods that may hide developers' interest or that are just not reliable. Interpretability rejects methods that lack a plausible explanation of the outcomes, which currently is a substantial concern in the general development of algorithms that make decisions on behalf of users (Goodman and Flaxman 2017).

The rest of the paper is structured as follows. In Section 2 we describe VAAs' current methods of recommendation and their flaws; in Section 3 we present our proposed recommendation method: the *Learning VAA* ; Section 4 is focused on the methodology and data utilized for testing our approach; in Section 5 we show the results obtained in simulated experiments, and in Section 6 we discuss the implications, potentialities, and limitations of the Learning VAA.

2 State-of-the-art: current high-dimensional methods of recommendation

In this section, we first introduce the main high-dimensional method of recommendation that is currently used, which we refer to as the *standard method*, and discuss the problems that it incurs (Section 2.1). We then present and discuss the problems of the only marginal alternative method, *community-based methods* (Section 2.2).

2.1 Standard high-dimensional methods of recommendation

In most VAAs that use high-dimensional methods, the distance function that compares the positions of users and parties within a specific issue is chosen ex-ante according to proximity or directional voting logic. There is a vast literature comparing these theories, but no clear winner has emerged (Downs 1957; Rabinowitz and Macdonald 1989; Westholm 1997; Lewis and King 1999; Bølstad and Dinas 2017; Gallati and Giger 2019). While proximity theory states that, all other things being equal, a person will vote for the candidate with the least distance position to theirs (Downs 1957), in directional theory, a citizen will vote for the party that is on the same “side of the fence", and has the most visible position on that side (Rabinowitz and Macdonald 1989). Therefore, directional theories assume that any policy issue has two antagonist sides, and users only aim to advance in their side of the fence. Some VAAs also use a voting logic related to the unified model proposed by Merrill and Grofman (1999) as they incorporate elements from both proximity and directional logic. Since VAAs’ answers are typically coded as a Likert scale, these theories can be materialized into *distance matrices* that associate specific values to each user-party combination of stances on the issue. Some examples of such matrices, as defined in Mendez (2017), can be seen in Figure 1.

Regarding the aggregation of the distance scores for all policy issues, high-dimensional methods typically assume a *Manhattan* distance, where the scores are directly summed up. The saliency of policy issues can be incorporated into this method by performing a weighted sum instead, with each of the weights representing the issue saliency. Most VAAs include this saliency factor by providing users with the opportunity to declare whether a question is more or less important (Marschall and Garzia 2014; Wagner and Ruusuvirta 2012). This action doubles or halves the weight of the question. Meanwhile, other VAAs perform the unweighted aggregation (Wagner and Ruusuvirta 2012).

Standard methods of recommendation have some methodological and normative flaws. Here, we focus on two important issues that undermine the effectiveness of standard high-dimensional methods of recommendation: the weighting (saliency) of the questions and the distance matrices employed. Concerning the weighting of the issues, the VAAs that perform an unweighted aggregation of the dimensions assume that voters give the same importance to every policy issue. This assumption does not go in line with any previous spatial voting theory. Having the weights chosen by experts would improve the situation, although it makes the VAA prone to biases from the developers due to a lack of reliable data that could be used for evaluating them. In any case, and to the best of our knowledge, current VAAs neither explore this possibility nor have considered using the parties' manifestos as a proxy for adjusting the saliency of each policy issue. Regarding the VAAs that allow users to state the importance of the questions, the decision of doubling or halving the weights according to the user's declaration is also chosen ex-ante. Introducing this feature does not solve the problems, as multiplying/dividing the outcome score of the issue by two is still arbitrary and can only marginally improve the results[1]. It is especially hard to find a sound explanation for choosing two and not any other number, as it is challenging to know what the users mean when they declare that a question is more/less important. Furthermore, there is another crucial problem that arises from the weighting used by the above-mentioned methods: the recommendation outcomes can be significantly affected by the composition of questions included in the VAA (Walgrave, Nuytemans and Pepermans 2009; van der Linden and Dufresne 2017). For instance, if two or more questions belong to issues that are closely related, their multiplicity will amplify the importance of the topic when giving the recommendation (which would effectively override the user signalling an unrelated issue as relevant).

The second important issue is related to the distance matrices, typically selected from proximity, directional or hybrid logic. This is problematic since it is not clear that proximity and directional models are liable to explain voting behaviour. For example, empirical studies have found that “key statistical assumptions [on proximity and directional models] have not been empirically tested and, indeed, turn out to be effectively untestable with existing methods and data" (Lewis and King 1999, 21). The vast literature on proximity and directional theory has not declared an ultimate winner

[1] The other options that have been used, such as giving the users a scale from 1 to 10 to evaluate the saliency of the questions, have only been scarcely utilized and convey the new problem of demanding too much from the users (Wagner and Ruusuvirta 2012).

and, instead, it is discussed whether contextual factors affect the fit of these theories (Pardos-Prado and Dinas 2010; Singh 2010; Fazekas and Méder 2013; Gallati and Giger 2019). Therefore, we can conclude that the current matrices are empirically contestable and are not necessarily sound for VAAs. Different matrices could have been tested following theories on spatial voting other than proximity or directionality. In particular, the unified model proposed by Merrill and Grofman (1999), incorporating some elements of each proximity and directional logic, could be used to build a sound matrix. Although the usage of the hybrid matrix in Mendez (2012, 2017) resembles the unified model in that it is a mix of both theories, it is an arbitrary combination of them that was not linked to any voting logic theory in the paper where it was suggested (Mendez 2012). This situation brings suspicion about why other distance matrices have not been extensively tested, e.g. following other combinations of values for the parameters of the unified model (Merrill and Grofman 1999). Moreover, even assuming that these theories are sound enough, there is another problem with the distance matrices: the usage of the same distance matrix for every policy issue (Mendez 2012, 2017). It is assumed that voters follow the same logic across all the political dimensions. On the contrary, we may expect that voters use different logic in distinct policy issues, as they potentially respond to different patterns of behaviour that are challenging to disentangle ex-ante. Previous research has found that the inner logic of any political dimension is affected by the degree of polarisation and contextual factors and may change over time (Pardos-Prado and Dinas 2010). Previous VAA research has never explored the possibility of different dimensions using different theories according to features such as their polarisation or saliency.

From a normative perspective, the a priori decisions taken on standard methods regarding the saliency and the distance functions are problematic because they may be inconsistent with users' real preferences. Previous research has found that, in current VAAs, "some parties may find themselves in the beneficial position of receiving more voting recommendations because of artefacts in the VAA construction" (Gemenis 2013, 281). In the studies where some empirical measures have been used to evaluate whether the method of recommendation is reliable, standard VAA results have not been too promising (Katakis et al. 2014; Mendez 2017). Furthermore, van der Linden and Dufresne (2017) argue that "training algorithms using pilot data [should be] a fundamental prerequisite in VAA design" and "to rely on ex-ante estimates is not only methodologically problematic but practically irresponsible" while indicating that some of the developers' decisions should be based on "empirics".

2.2 Alternative matching methods: community-based recommendations

As an alternative to the standard model, some researchers have proposed community-based recommendation methods (Katakis et al. 2014; Tsapatsoulis et al. 2015; Agathokleous and Tsapatsoulis 2016). However, their implementation has been quite limited (for example, *Choose4Greece* or in *EUvox 2014* as an alternative method offered to the user for obtaining the recommendation), with standard methods remaining as the preferred option for almost all cases. Community-based recommendation dispenses with expert party positioning; recommendations are purely based on other users' answers (Tsapatsoulis et al. 2015). Before filling up the questionnaire, users are requested to indicate which party they plan to vote for, and this information can be used for inferring what voters of a specific party value most. Recommendations are then based on the parties that are prevalent in the user's vicinity, i.e. the parties that like-minded users (the ones with similar answers) already declared they planned to vote for (Katakis et al. 2014).

Community-based recommendation methods are more successful than standard methods in terms of performance, as measured by indicators such as accuracy and mean rank (Katakis et al. 2014). Moreover, in an experiment by Tsapatsoulis et al. (2015) in which users did not know what kind of recommendation method was employed, a higher percentage of users declared that community-based recommendations were more useful than standard ones. Probably, the better consideration of these recommendations is correlated with a) the higher percentages of accuracy and mean rank (Tsapatsoulis et al. 2015); b) an indirect capture of the saliency of the issues within the models (Agathokleous and Tsapatsoulis 2016). It is worth remarking that these methods of recommendation are based on other users' answers as a proxy to the parties' position, so there is an underlying assumption that most users should reflect their preferred party's position on the issues they value most. This proxy helps to overcome possible biases introduced when placing the party positions on the issues by VAA developers, but may also blur the function of VAAs as the proximity of a user to a party is not directly reflected.

However, a significant problem of community-based recommendation methods is their lack of transparency. There are a wide variety of techniques that can be employed in community-based recommendation methods, but most of them implement complicated mechanisms of matching users and political parties that can act like a *black box*. The few techniques that do not face such problems, such as decision trees, are hard to link to the actual decision processes taken by the users, and therefore lack interpretability.

Furthermore, complex models with many parameters and high predictive power can potentially learn spurious rules from correlations in the data that may not be justifiable. Their opaqueness may prevent researchers from detecting such spurious rules of classification and hinders the accountability of the application to the users, who may lose their trust in the methods employed. This lack of transparency along with the technical complexity of the method may be among the reasons behind its marginal adoption. Furthermore, they impose a higher technical complexity on the development of the web platform that web providers may not have.

3 Our alternative: Learning VAAs

In order to overcome some of the issues in state-of-the-art VAAs identified in the previous section, we propose a recommendation method that lies between standard and community-based methods. While mimicking standard recommendation methods in the way distances between users and parties are computed, it employs users' data for empirically adjusting the parameters that rule the saliency weights and distance functions. Like previous community-based recommendation methods, the adjustment is made by an algorithm that trains on users' answers to the policy issues and learns model parameters that align its recommendations to users' voting intentions. Unlike community-based methods, agreement scores are always computed using both the user and parties' answers, and all the parameters in the model can be easily interpreted from a theoretical political science perspective. A formal introduction to the model and the training process can be seen in Appendix A.

The adjustment of the model parameters is carried out at an early stage of the VAA when a critical mass of users' data is achieved (with possible health checks at specific points later on to verify that there have not been significant changes in users' positions). Although early users may have altered results, the algorithm should stabilize quickly and work realistically for the vast majority, as shown by other empirical approaches Katakis et al. 2014, van der Linden and Dufresne 2017). While Katakis et al. (2014) report that their community-based method already stabilizes with a few hundreds of users (after filtering), van der Linden and Dufresne (2017) and Germann and Mendez (2016) show that there is not much difference in the loadings of dynamic scale validation using data from a few thousands of early users (1444 and 3000, respectively). Importantly, the algorithm only learns from those users who are educated, politically informed and vote because of ideological considerations (issue voters). We define such users as the ones who reply "Very" to the question "To what extent are you interested in politics?", reply "The ideas of the party are closest to my own" to "What is the main reason for voting for your party of choice?", and state they have at least high-school education.

Few recommendation algorithms in the literature resemble our approach. The work of Katakis et al. (2014) presents a weighted standard model as a baseline against which other community-based methods are compared. This model employs users' data to adapt the saliencies of the issues according to their information gain. However, the model keeps a static single distance matrix for all questions and is dismissed as its performance scores were not as high as other community-based methods. Mendez (2017) also presents a model that incorporates statistical learning to a model primarily based on standard recommendation methods. However, differently from ours, this model fully preserves the standard model and only adds a multinomial regression layer after the final recommendation, so the input to the regression would be the agreement scores produced by the standard method plus some demographic data of the user. Still, this model uses a single distance matrix (hybrid, proximity or directional) for all issues instead of allowing different distance matrix for each policy issue, and the saliency remains equal for all issues. Furthermore, the regression layer performs a mapping of the agreement score space to itself, which lacks interpretation from the perspective of the cognitive decision process made by users. We also consider that the inclusion of demographic data as additional information on which to base the recommendation should not be done in VAAs. This is the case because the goal of VAAs should be to inform of proximity on the policy preferences and other factors such as political identity are not within their scope. Lastly, although dynamic-scale validation (Germann et al. 2015; van der Linden and Dufresne 2017) is both interpretable and based on empirical data, its use is reduced to generating spatial maps, as they have been proved to be less suitable for building recommendations (Mendez 2017).

4 Methodology and Data

We perform an empirical demonstration of our proposed learning matching method on data from the *EUvox 2014*, a VAA launched for the European Parliamentary Election of 2014, and *aquienvoto.org* (AQV), the biggest VAA for the Spanish General Election of April 2019. In the case of the *EUvox*, we have taken a subset of the data from Greece, Italy, Portugal, Spain and the UK, countries that have either the most user entries and or that are have been used in previous studies (Mendez 2017; Djouvas, Mendez and

Tsapatsoulis 2016). The selection of VAAs employed for our experiments cover both *second-order* elections (the European elections [Reif and Schmitt 1980]), and a national *first-order* election (AQV).

The *EUvox 2014* VAA is a survey conducted across 27 countries of the EU in the context of the European parliamentary elections in 2014. It has 30 policy-issue questions of which 25 were common across every country, and the remainder were national-specific. The AQV survey was specific for Spain and had 44 questions on different political issues. The total tally of observations for the algorithms is almost half a million users, and the minimum sample for any of the selected country survey has more than 25,000 observations. Although both VAAs gave users the opportunity of declaring which issues were more relevant for them, this information is not available in the *EUvox* dataset. While the *EUvox 2014* data is publicly available on the GESIS data archive (Mendez and Manavopoulos 2018), the data from AQV was shared with us by the creators of the survey and is expected to be available to the general public shortly. The treatment of the missing data and a more detailed explanation of the datasets can be found in Appendix B.

Regarding the evaluation of a VAA, choosing an appropriate performance metric is not straight-forward, since the information of what the correct recommendation should be is not accessible. The voting intention can be utilized as a proxy for it, as it is done to train learning and community-based methods. Performance metrics in the literature typically compare the voting intention with the given recommendation for building metrics such as accuracy, mean-rank, or F-score (van der Linden and Dufresne 2017; Mendez 2017; Katakis et al. 2014). To ensure the results can be generalized, each data-set has been randomly partitioned into two subsets: the training set and the test set. While the model parameters are adjusted based on data from the training sets, all performance metrics are measured on data from the test set. The random partitioning is performed ten times, with each partition resulting in a different model, and results are averaged over them (this technique is commonly known as *cross-validation*). A more detailed explanation of the performance metrics employed here (accuracy, F-score) can be found in Appendix C.

5 Results

We compare the performance of the Learning VAA to both standard and community-based methods. For the standard method, we use the hybrid voting logic as the distance matrix, as it was the one applied initially on the *EUvox 2014* VAA since it was marked as the best performing one in Mendez (2012). For the community-based method, we employ a Support-Vector Machine (SVM) architecture, as it has been claimed to outperform other architectures in various studies on community-based VAAs (Katakis et al. 2014; Djouvas, Antoniou and Tsapatsoulis 2018). The performance metrics of the three methods are shown in Table 1. For each metric, the values shown are the mean and standard deviation over the ten random partitions of the datasets (see Section 4). We find that our method always performs better than the standard method and similarly to community-based methods. A more detailed figure showing the F-scores that every political party obtained with each method can be found in Appendix D.

Figure 2 and Figure 3 give some insights into the model that has been learned by the algorithm. Figure 2 shows the distance matrices of five sample questions for the *EUvox 2014* VAA (distance matrices for all data-sets can be found in Appendix E).

We find that different questions follow a different voting logic, supporting our hypothesis that applying a unique distance matrix to all policy issues is unrealistic (see Section 2). The two rightmost matrices deserve further discussion. The one labelled as "Other possible paradigms" exhibits an interesting behaviour: on one side, it joins together voters and parties that both completely agree or both completely disagree with the statements and, on the other side, it groups voters and parties that agree, disagree, or are neutral. It seems as if the voting logic was that moderate voters are close to moderate parties —even when they are at the other side of the fence— and they are far from extreme parties —even when being in the same direction. In fact, the shown matrix corresponds to the question "The possibility of independence for an Autonomous Community should be recognized by the Constitution" for the *EUvox 2014* VAA in Spain, which is an issue with extreme polarisation at both sides of the spectrum [2]. This situation also goes in line with the discussion about the *region of acceptability* in directional voting logic, that modifies it to penalize parties that are on the same side of the fence but have a too extreme position (Rabinowitz and Macdonald 1989).

The rightmost matrix has a pattern that does not seem to follow any reasonable logic as a distance metric. It corresponds to the question "Same sex couples should enjoy

[2] As a matter of fact, the question regarding the independence of Autonomous Communities was the most important policy issue according to our learning algorithm. This is quite plausible, as the discussion regarding the Independence of Catalonia was already a salient topic in 2014. By giving to this issue the same weight as others, the *EUvox 2014* was penalising the Spanish voters that, instead of having a non-nationalist position, share their positions with the Catalan nationalist parties in other policy issues.

the same rights as heterosexual couples to marry”. We believe that a reason for such a resulting matrix is a misplacement of some party answers by the VAA developers, making the algorithm learn this spurious rule from the mismatch between a VAA developers’ decision and the behaviour of the users.

Figure 3 displays the saliency weights [3] of the questions of three sample countries: England, Greece, and Spain. For all countries, nearly half of the questions have remarkably low importance, while a small number of other questions stand out. These especially important questions are related to Brexit in the case of England, leaving the euro in the case of Greece, and Catalonia in the case of Spain. The low importance of some questions can be due to: a) an effective low saliency of the issue by the voters, b) the saliency is divided between a few questions that are highly correlated (both users and parties have analogous opinions on them), c) the coding of the parties’ positions is not accurate, d) users do not believe that the party is able to execute the policy as they campaign to (van der Linden and Dufresne 2017).

6 Discussion

We have shown that the Learning VAA performs better than standard ones and similarly to community-based ones, as measured with accuracy and F-score. Therefore, our results suggest that the Learning VAA can capture the political space of the users in an

[3] Each weight is calculated by multiplying the weight parameter of the corresponding question by the Frobenius norm of its distance matrix, as the magnitude of the distance matrix indirectly affects the importance given to the question.

acceptable way. Moreover, the fact that we can disaggregate the saliency and distance matrix of each of the policy issues indicates that it can be more interpretable than previous community-based VAAs. In any case, it is worth noting that there are limitations regarding what can be considered an ideal performance for VAAs. This is the case because these applications assume a world in which the preferences in the policy issues determine what the voting party closest to the user is. This is a strong assumption because it has been tested that there are many other important factors that influence voting behaviour beyond ideology, such as leadership and parties perceived political competence, and VAAs are unable to treat them fairly (Mendez 2012; van der Linden and Dufresne 2017). Further research should try to shed light on this issue. We also need to take into consideration that there is a substantial overlap between some parties in the ideological space. An ideal method would also need to recommend the competing parties equally and not favour one of the parties of the confronted area over the others. For all these reasons, performance results should be taken with a grain of salt.

Despite its limitations, there are several normative and empirical implications of using our methods. The Learning VAA can discover the importance of a policy issue and assign different distance matrices to each of them. Therefore, it can help to undercover the underlying political space of a significant share of voters who are issue voters and politically competent. As an example, we have shown that some questions seem to follow proximity or directional logic, while others may be subject to different paradigms that can be further analysed. These unexpected matrices can be a result of either an indeed different voting logic paradigm —unveiling new, interesting facts for researchers—, or just spurious rules learned by the system. In a striking difference with most community-based recommendation methods, such spurious rules can be spotted due to the parameters of the system having direct interpretation. Further restrictions can be imposed on the parameter space to limit them to reasonable combinations —e.g. forcing monotony: given a specific user position on an issue, parties that are further than others cannot have a better score than them on the issue.

We also argue that our learning method is as good as the best community-based method in terms of transparency. Although our method is indeed more complicated than standard VAAs (we cannot expect most users to understand anything about the learning mechanisms), users can still calculate by hand their match with a party precisely in the same way as in standard methods if provided with the saliency weights and matrices computed after the learning phase. Moreover, users can know precisely the reasons for their matching (i.e. the saliency of any question or how it is compared) and make sense of it. This is the case because every learned parameter has a direct interpretation in the process of deciding a political party, and therefore, the models resulting from the training process are easy to explain. Furthermore, since in our proposed method the computation of the model parameters occurs offline and only once, the implementation of the method on a web platform is as simple as standard methods. We also attempt to facilitate the application of our method with the public release the code for building our model and reproducing the experiments shown here. It can be found at the public code repository *GitHub* [4].

[4] It can be accessed at https://github.com/Juillermo/VAA-data.

As VAAs are used by many citizens looking for which party to vote for, and they expect to obtain political information with these applications, it is crucial to analyse the problems of learning from their voting intention. Any VAA relying on other users to make a recommendation (in our case, to adapt saliencies and distance functions) may have the problem of misinformation, which appears when most users that may consider voting for a party are misled concerning its position on some of the issues. There is currently no published work considering the problem of community-based VAAs and employing data from these mislead users. From our perspective, there is a trade-off between using empirical evidence from users' data for improving recommendations and the problem of misinformation. As we can expect that mistakes among users will not be evenly distributed (users are prone to bias as a result of heavy political campaigns), any recommendation method that uses other users' information may have the risk of learning incorrect information and therefore providing wrong recommendations [5].

This would be the case if the algorithm learns from all users, especially considering the influence of fake news in the political debate. In our model, in which we only take educated and politically active issue voters for training, we assume the following: the majority of those politically motivated and educated issue-voters know to a reasonable degree the positions of the parties they will vote for in the election. This is a precondition of our model to work correctly and further research should be done in this area, in particular, to study whether these conditions are enough for a learning scheme. In any case, considering exclusively the educated and politically motivated users might also

[5] We would like to thank one of the anonymous reviewers for pointing out this problem

carry some drawbacks as these users could have different preferences than the rest and not be representative. For that reason, in further research two models should be compared and evaluated, one trained with educated and politically motivated users only and another trained with all users, to know whether there are significant differences between the models produced with either group. It is also worth remarking that users may have intertwined reasons to vote for a party, and a single supporting question may be insufficient to correctly classify users between issue and non-issue voters, as it has been done. This fact adds some noise to the data and further limits the performance upper bound. However, statistical learning techniques are prepared to cope with the noise and capture the trends out of it.

Our method, as here presented, adapts to the users globally, i.e. the weights and distance matrices are the same for all users. It could be argued that different users do not need to agree on the importance they give to each question, and their voting logic could be different as well. We argue that users could be clustered based, for instance, on their declarations about which questions they consider more important. A different model could be trained for each of these clusters, providing that they contain a minimum number of users for the model to learn from the data correctly. A second alternative can be to cluster users by the agreement in their answers and then train independent models that would assign a different saliency distribution to each cluster. These modifications would include some degree of customisation of the model depending on the user profile, although it would require a more significant amount of user data to work properly. We leave these issues to be addressed in future work.

7 Conclusion

In this paper, we have argued that standard high-dimensional methods of recommendations for VAAs fail to offer appropriate outcomes since they are not able to correctly map users' preferences. While there are alternative community-based methods that base their recommendations on empirical data, most remain opaque: they cannot successfully explain the reasons behind the recommended party and researchers are not able to validate and understand the process of recommendation given by them. We have presented an alternative method that learns to adapt its saliency weights and distance matrices based on users voting intentions and answers to the questions. We have argued that Learning VAAs can adapt themselves to users' way of aggregating and comparing policy preferences. Furthermore, we believe that the Learning VAA can provide unprecedented evidence of voting behaviours from VAA users and even provide new paradigms regarding issue-voting in spatial voting theories. However, there are still several improvements to be made to our approach, such as giving all parties an equal treatment or training different models for different clusters of users as users need not share the same saliencies and voting logic. We expect that we can tackle some of these drawbacks in the future, as well as implement our method into real VAAs for coming elections to measure user satisfaction with different methods of recommendation.

References

Agathokleous, M., and Tsapatsoulis, N. 2016. "Applying Hidden Markov Models to Voting Advice Applications". *EPJ Data Science*, 5(1): 34. doi:10.1140/epjds/s13688-016-0095-z.

Anderson, J., and Fossen, T. 2014. "Voting Advice Applications and Political Theory: Citizenship, Participation and Representation." In *Matching Voters with Parties* *and Candidates: Voting Advice Applications in Comparative Perspective*, edited by Garzia D., and Marschall, S., 217-226. ECPR Press.

Andreadis, I. 2012. "To Clean or not to Clean? Improving the Quality of VAA Data". Paper presented at the XXII World Congress of Political Science (IPSA), Madrid, July 8-12.

Andreadis, I. 2014. "Data Quality and Data Cleaning". In *Matching Voters with Parties and Candidates. Voting Advice Applications in Comparative Perspective*, edited by Garzia. D., and Marschall, S., 79–91. ECPR Press.

Benoit, K., and Laver, M. 2012. "The Dimensionality of Political Space: Epistemological and Methodological Considerations". *European Union Politics*, 13(2):194–218. doi:10.1177/1465116511434618.

Bølstad, J., and Dinas, E. 2017. "A Categorization Theory of Spatial Voting: How the Center Divides the Political Space". *British Journal of Political Science*, 47(4):829–850.

Djouvas, C., Antoniou, A., and Tsapatsoulis, N. 2018. "Improving Social Vote Recommendation in VAAs: The Effects of Political Profile Augmentation and Classification Method". Paper presented at the 2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Zaragoza, September 6-7. doi:10.1109/SMAP.2018.8501885.

Djouvas, C., Mendez, F., and Tsapatsoulis, N. (2016). "Mining Online Political Opinion Surveys for Suspect Entries: An Interdisciplinary Comparison". *Journal of Innovation in Digital Ecosystems*, 3(2):172–182. doi:10.1016/j.jides.2016.11.003.

Downs, A. 1957. "An Economic Theory of Political Action in a Democracy". *Journal of political economy*, 65(2):135–150.

Fazekas, Z., and Méder, Z. Z. 2013. "Proximity and Directional Theory Compared: Taking Discriminant Positions Seriously in Multi-Party Systems". *Electoral Studies*, 32(4):693–707.

Fossen, T., and Anderson, J. 2014. "What's the Point of Voting Advice Applications? Competing Perspectives on Democracy and Citizenship". *Electoral Studies*, 36:244–251. doi:10.1016/j.electstud.2014.04.001.

Fossen, T., and van den Brink, B. 2015. “Electoral Dioramas: On the Problem of Representation in Voting Advice Applications”. *Representation*, 51(3):341–358. doi:10.1080/00344893.2015.1090473.

Gallati, L., and Giger, N. 2019. “Proximity and Directional Voting: Testing for the Region of Acceptability”. *Electoral Studies*. doi:10.1016/j.electstud.2019.02.015.

Gärdenfors, P. 2004. “Conceptual Spaces: The Geometry of Thought”. MIT Press.

Garzia, D., and Marschall, S. 2012. “Voting Advice Applications Under Review: The State of Research”. *International Journal of Electronic Governance*, 5(3-4):203–222. doi:10.1504/IJEG.2012.051309.

Garzia, D., and Marschall, S. 2014. *Matching Voters with Parties and Candidates: Voting Advice Applications in a Comparative Perspective*. ECPR Press.

Garzia, D., Trechsel, A. H., and De Angelis, A. 2017. “Voting Advice Applications and Electoral Participation: A Multi-Method Study”. *Political Communication*, 34(3):424–443.

Gemenis, K. 2013. “Estimating Parties’ Policy Positions Through Voting Advice Applications: Some Methodological Considerations”. *Acta Politica*, 48(3):268–295. doi:10.1057/ap.2012.36.

Germann, M., and Mendez, F. 2016. “Dynamic Scale Validation Reloaded”. *Quality & Quantity*, 50(3):981–1007. doi:10.1007/s11135-015-0186-0.

Germann, M., Mendez, F., Wheatley, J., and Serdült, U. 2015. “Spatial Maps in Voting Advice Applications: The Case for Dynamic Scale Validation”. *Acta Politica*, 50(2):214–238. doi:10.1057/ap.2014.3.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. “Deep Learning”. MIT Press. http://www.deeplearningbook.org.

Goodman, B., and Flaxman, S. 2017. “European Union Regulations on Algorithmic Decision-Making and a ‘Right to Explanation’”. *AI Magazine*, 38(3):50–57. doi:10.1609/aimag.v38i3.2741.

Katakis, I., Tsapatsoulis, N., Mendez, F., Triga, V., and Djouvas, C. 2014. “Social Voting Advice Applications—Definitions, Challenges, Datasets and Evaluation”. *IEEE Transactions on Cybernetics*, 44(7):1039–1052. doi:10.1109/TCYB.2013.2279019.

Lefevere, J., and Walgrave, S. 2014. “A Perfect Match? The Impact of Statement Selection on Voting Advice Applications’ Ability to Match Voters and Parties”. *Electoral Studies*, 36:252–262. doi:10.1016/J.ELECTSTUD.2014.04.002.

Lewis, J. B., and King, G. 1999. “No Evidence on Directional Vs. Proximity Voting”. *Political analysis*, 8(1):21–33.

Louwerse, T., and Rosema, M. 2014. “The Design Effects of Voting Advice Applications: Comparing Methods of Calculating Matches”. *Acta politica*, 49(3):286–312. doi:10.1057/ap.2013.30.

Mendez, F., and Manavopoulos, V. 2018. “Euvox2014: Voting Advice Application Data for the 2014 European Parliament Elections”. doi:10.7802/1750.

Marschall, S., and Garzia, D. 2014. “Voting Advice Applications in a Comparative Perspective: An Introduction”. In *Matching Voters with Parties and Candidates. Voting Advice Applications in Comparative Perspective*, edited by Garzia. D., and Marschall, S., 1–10. ECPR Press.

Mendez, F. 2012. “Matching Voters with Political Parties and Candidates: An Empirical Test of Four Algorithms”. *International Journal of Electronic Governance*, 5(3/4):264. doi:10.1504/IJEG.2012.051316.

Mendez, F. 2017. “Modeling Proximity and Directional Decisional Logic: What Can We Learn from Applying Statistical Learning Techniques to VAA-Generated Data?”. *Journal of Elections, Public Opinion and Parties*, 27(1):31–55. doi:10.1080/17457289.2016.1269113.

Merrill, S., and Grofman, B. 1999. *A Unified Theory of Voting: Directional and Proximity Spatial Models*. Cambridge University Press.

Pardos-Prado, S., and Dinas, E. 2010. “Systemic Polarisation and Spatial Voting”. *European Journal of Political Research*, 49(6):759–786. doi:10.1111/j.1475-6765.2010.01918.x.

Rabinowitz, G., and Macdonald, S. E. 1989. “A Directional Theory of Issue Voting”. *American Political Science Review*, 83(1):93–121.

Reif, K., and Schmitt, H. 1980. “Nine Second-Order National Elections–A Conceptual Framework for the Analysis of European Election Results”. *European journal of political research*, 8(1):3–44.

Singh, S. P. 2010. “Contextual Influences on the Decision Calculus: A Cross-National Examination of Proximity Voting”. *Electoral Studies*, 29(3):425–434.

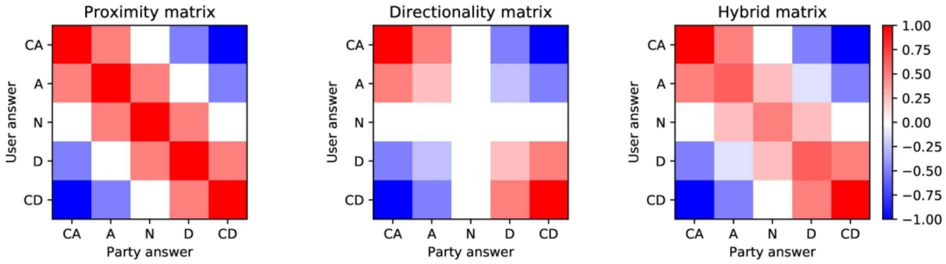
Tsapatsoulis, N., Agathokleous, M., Djouvas, C., and Mendez, F. 2015. “On the Design of Social Voting Recommendation Applications”. *International Journal on Artificial Intelligence Tools*, 24(03):1550009. doi:10.1142/s0218213015500098.

van der Linden, C., and Dufresne, Y. 2017. “The Curse of Dimensionality in Voting Advice Applications: Reliability and Validity in Algorithm Design”. *Journal of Elections, Public Opinion and Parties*, 27(1):9–30. doi:10.1080/17457289.2016.1268144.

Wagner, M., and Ruusuvirta, O. 2012. “Matching Voters to Parties: Voting Advice Applications and Models of Party Choice”. *Acta Politica*, 47(4):400–422. doi:10.1057/ap.2011.29.

Walgrave, S., Nuytemans, M., and Pepermans, K. 2009. “Voting Aid Applications and the Effect of Statement Selection”. *West European Politics*, 32(6):1161–1180. doi:10.1080/01402380903230637.

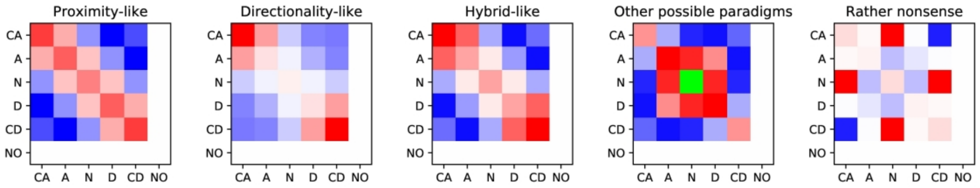
Westholm, A. 1997. “Distance Versus Direction: The Illusory Defeat of the Proximity Theory of Electoral Choice”. *American Political Science Review*, 91(4):865– 883.



Common distance matrices in standard high-dimensional VAA recommendation methods. Higher scores mean that the users and the party are more coincident on the policy issue and thus have more chances to be matched. Abbreviations have the following meanings: Completely Agree (CA), Agree (A), Neutral (N), Disagree (D), Completely Disagree (CD). The hybrid matrix is the average of the other two and relates to a unified model.

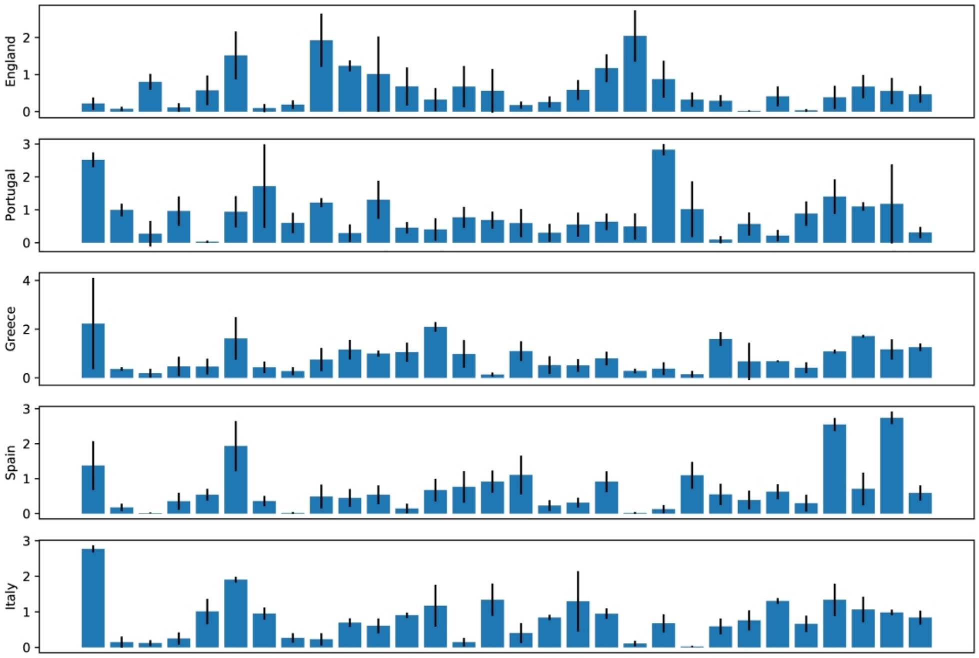
266x76mm (300 x 300 DPI)

		Learning	Community-
	Standard VAA	**VAA**	based (SVM)
Accuracy	0.597 ± 0.013	**0.683 ± 0.009**	0.659 ± 0.014
F-score	0.579 ± 0.013	**0.673 ± 0.009**	0.646 ± 0.016
Accuracy	0.440 ± 0.015	0.555 ± 0.017	0.562 ± 0.014
F-score	0.447 ± 0.017	0.502 ± 0.018	**0.541 ± 0.017**
Accuracy	0.503 ± 0.019	0.578 ± 0.013	0.593 ± 0.019
F-score	0.510 ± 0.017	0.513 ± 0.015	**0.559 ± 0.022**
Accuracy	0.282 ± 0.004	0.403 ± 0.006	0.397 ± 0.009
F-score	0.273 ± 0.005	0.335 ± 0.007	**0.353 ± 0.016**
Accuracy	0.638 ± 0.013	0.701 ± 0.018	0.702 ± 0.021
F-score	0.631 ± 0.013	0.672 ± 0.019	0.690 ± 0.021
Accuracy	0.485 ± 0.004	0.582 ± 0.009	—
F-score	0.479 ± 0.004	0.568 ± 0.010	—



Five sample distance matrices produced by the learning model. Red tiles have positive values and blue tiles are negative. Their intensity is directly proportional to the absolute value. Green tiles represent combinations that do not happen in the database. It can be appreciated that some matrices resemble the previously proposed distance matrices (the first three samples), while others move away from them (the last two).

304x63mm (300 x 300 DPI)



Saliency weights of our model for the questions of three sample countries. The bars are the average of the saliencies over ten random training-test partitions, with error bars denoting the standard deviation across partitions.

279x190mm (300 x 300 DPI)

Appendix A: Formalisation of the algorithm

We will now formally define our method and how it is adaptive. A user answers on a L-points Likert scale the N questions offered in the VAA. Each answer is processed as a one-hot vector (a vector of length L with value one at the point of the Likert scale chosen by the user and zero otherwise), and all vectors are stacked onto a single $N \times L$ matrix, U, called *profile*. We have M users, each with a user profile, U^i, and a voting intention, v_i,$i \in \{1,...,M\}$, and K parties with their corresponding profiles, P^k,$k \in \{1,...,K\}$. The agreement score for issue j between user i and party k is computed with a distance matrix, as in the standard method:

$$Q_j^{ik} = U_{jl}^i \cdot D_{lm}^j \cdot P_{jm}^k. \quad (1)$$

The main difference concerning standard methods is the inclusion of N distance matrices, D^j,$j \in \{1,...,N\}$ (one per policy issue) instead of a single one. We impose symmetry in both diagonals, as it is typical in spatial voting theory. The aggregated score for each party k is a weighted sum of the N issue scores (Manhattan distance), in the same way as in high-dimensional standard methods:

$$s_k^i = \sum_{j=1}^{N} w_j \cdot Q_j^{ik}. \quad (2)$$

A diagram explaining the process described so far is shown in Figure A1.

Finally, to render the output more interpretable, the agreement scores are passed through a *softmax* function that places them in the range [0,1] and make them sum up to one:

$$t_k^i = \frac{e^{s_k^i}}{\sum_{l=1}^K e^{s_l^i}}.$$

(3)

The weights for the aggregation of the issues and the values in the distance matrices are flexible in our model; they will gradually decrease or increase as the model learns from users' answers. The goal of the algorithm is to find which parameters — which values for the distance matrices and squashing weights — lead to the most realistic outcomes. This is achieved when the party that the user intends to vote for, $v_i \in \{1,...,K\}$, gets the maximum agreement score. How well this is met can be quantified by the *log loss* function[1], A common performance measure for multi-class classification tasks in the field of *machine learning*. It outputs the classification error per user as $err_i = -log\left(t_{v_i}^i\right)$, with $t_{v_i}^i$ being the score given to the party the user declared in their voting intention, v_i.

The algorithm minimizes the total error on a training sample of user answers by adjusting its parameters. A new regularisation term should be added to the error function to prevent parameters from growing arbitrarily. The *L2 regularisation* term fulfils this purpose by penalising too large parameter values and includes a parameter, λ, that expresses the strength of this penalisation, which is a hyperparameter chosen by the developer. The final function to be minimized has the form

$$\min_{D,w} \sum_{i=1}^M err_i + \lambda\left(\|D\|_2^2 + \|w\|_2^2\right),$$

(4)

[1] Also known as the *cross-entropy* function.

where the operation $\| \cdot \|_2$ is the l^2-norm or the Frobenius norm for vectors and matrices, respectively.

The above optimisation problem could be solved by computing the partial derivatives in the parameter space and finding the critical points[2]. Nevertheless, the inclusion of complex non-linear functions, such as the softmax, prevents us from efficiently computing a closed-form solution and numerical methods must be utilized. A conventional method that performs a local search in the parameter space is called *gradient descent*. This method relies on the information carried by the derivatives. The gradient of a function at a certain point shows the direction in the parameter space in which the error diminishes the fastest. By taking a step (changing the parameters) in that direction, the new parameter set should produce a lower error[3]. By iterative steps, the algorithm eventually reaches a local minimum. The gradient of the function determines the size of the step at the point and the learning rate, μ, —chosen by the developer among a wide number of different techniques allow for an efficient selection (Goodfellow, Bengio and Courville 2016, 84)—:

$$D^j_{t+1} = D^j_t - \mu \sum_{i=1}^{M} \frac{\partial err_i}{\partial D^j_t}, w_j^{t+1} = w^t_j - \mu \sum_{i=1}^{M} \frac{\partial err_i}{\partial w^t_j}, \tag{5}$$

[2] Critical points are the ones whose partial derivatives are equal to zero.

[3] Steps should be small to ensure that local properties are conserved. Too large steps could lead to divergence.

where t and $t + 1$ represent subsequent iteration steps. The update steps are restricted to enforce all saliency weights and the main diagonal of the distance matrices to remain positive.

For the experiments shown in the paper, our algorithm has been trained for 300 iterations with a learning rate of $\mu = 1$, a regularisation term of $\lambda = 0.01$, and weights randomly initialized from a standard normal distribution with zero mean and variance of one.

Appendix B: Datasets and data cleaning

The *EUvox 2014* data was retrieved from the official repository before it ceased to be available online. From the original dataset, a subset constituted by the Greek, Italian, English, Portuguese, and Spanish users was selected. Table B1 shows how many parties were included for recommendations in each country. In order to only evaluate the method on observations with complete responses, we only considered those users who answered all policy questions in the survey.

Following the recommendations forwarded by (Andreadis 2012, 2014), further filters were imposed upon the resulting subset of observations. This removal was done to diminish the potential impact of rogue users who did not complete the survey promptly or that showed patterns of disregard when answering the survey questions. To this end, the following filters were imposed on the original retrieval of the VAA data.

(1) Users whose time to complete the items section took under 120 seconds were removed

(2) Users with response time for an item under 1 second were removed

(3) Users who had responded 3 or more item questions under 2 seconds were also removed

(4) Users whose age was stated to be below 17 were removed

(5) Users who completed the survey using either the Android or the iPhone platform were removed

Filters 1-3 intend to ensure that enough care and attention has been put into the survey. Filter 4 intends to remove users who could not have been able to vote in the EU 2014 elections. Lastly, Filter 5 is devised to ignore entries affected by an error on the platform that did not present the questions appropriately in the mobile device version of the survey.

Then, as it has been done in Mendez (2012, 2017), we have split the users between issue and non-issue voters since users who do not base their voting on the parties' policy programs cannot be taken as reliable for training the algorithm nor evaluating its performance. This separation was done by picking only users who replied "The ideas of the party/bloc are closest to my own" to the supplementary question "What is the main reason for voting for your party/bloc of choice?". As explained in Section 3, further restrictions were applied to the users to improve the quality of the data: we only take educated and politically active voters, defined as those who reply "Very" to the question "To what extent are you interested in politics?", reply "The ideas of the party are closest to my own" to "What is the main reason for voting for your party of choice?", and state they have at least high-school education.

In the case of the AQV survey, we retrieved a sample from the first 217175 users that responded to the survey on the 2019 Spanish general election. Since in this survey no data about the amount of time employed per question or along with the survey was gathered, we only applied the age removal filter while also removing users who completed less than 25 policy questions. The questions for detecting issue, educated, politically motivated voters, were not available and hence these distinctions could not be done for this database.

Since the voting intention is a requisite for training and evaluating the algorithms, users without this information were removed from both datasets. The impact of each filter in the original sample in each country and dataset can be seen in Table B2. The attrition suffered by the imposition of the filters amounts to significant losses in some countries but are necessary to ensure the integrity of the data. The removal of voters that are not issue-voters or politically active incurs in the highest impact to the original data. All subsets from *EUvox 2014* were reduced to around a 10% of their original size.

Statistical learning methods impose an additional requirement in the dataset: they must be trained with a subset and evaluated in a different one, typically labelled as *training set* and *test set*, respectively. This splitting is done because the algorithm could try to learn all the small details of the training set, including contingent random variations, thus losing generality. The evaluation of the algorithm in such a set would then be misleading; we should evaluate it on new, unseen data to prove that it learned the abstract features, so the test set is kept aside and only used at the very end for evaluating. For our experiments, we allocated 90% of the data at random to the training set and the

10% remaining for the test set, although maintaining the same party proportions as in the original dataset.

Appendix C: Evaluation metrics

The results of the recommendation process are high-dimensional: each user receives ten scores (one per party) and has a voting intention. There are many ways of squashing these into a single dimension (i.e. performance score) for all the users. There is a trade-off in the way of executing this reduction since making performance metrics more intuitive comes with the loss of valuable information. At the most intuitive side of the spectrum, *accuracy* states the percentage of users to whom their voting intention is given as the first recommendation.

Next, the *F-score* is the harmonic average between *precision* —from the users that were recommended a certain party first, which percentage did intend to vote for it— and *recall* —how many of the users that intended to vote for a certain party had it recommended—. Unlike accuracy, F-score treats all parties equally and hence can better reflect a systematical mis- or over-representation of some parties. Unluckily, it is not as intuitive: bounded in the interval [0,1], the reader can only know that values close to zero are bad and close to one are good, while it allows comparisons between F-scores. It can be computed as

$$F\text{−}score = \frac{1}{K}\sum_{k=1}^{K} 2\left(\frac{1}{precision_k} + \frac{1}{recall_k}\right)^{-1}. \qquad (6)$$

Both accuracy and F-score do not differentiate between recommending a voting intention party at the second place close to the first recommended party (fair recommendation) and giving it at the eighth position with a low score (bad recommendation); accounting for a failure of the same magnitude in both cases. A performance metric that captures this information is the *logloss* (as explained in Appendix A), but all intuition about the metric is lost apart from knowing that lower values are comparatively better.

Appendix D: Per-party F-scores

Figure D1 compares the F-scores of the three methods for the political parties in each of the countries. Small variations can be seen between the three methods. For many of the parties across different countries, the outcome of the three methods is similar. The standard method has a poorer performance with the *Labour* (England), *CDU* (Greece), and *IU* (Spain), while doing better for *DIMAR* (Greece). The community-based method (Social VAA) performs badly for *BE* (Portugal), *ANEL* (Greece), *Equo* (Spain), and *NCD UDC* and *FDI* (Italy) and does better for *ERC* (Spain). The learning method, however, does not perform much worse than any of the other two methods for any party, while having an improved performance with *Potami* (Greece), *PNV_CiU* (Spain). All methods perform badly for *MPT, PCTP* and *PAN* (Portugal), and *DIMAR* and *Greens* (Greece). Most of these parties had 100 or fewer users intending to vote for them.

Appendix F: Distance matrices of the *EUvox 2014* for Spain

Figures F1, F2, F3, F4 and F5 contain the distance matrices of the thirty questions for the *EUvox 2014* in Spain. Distance matrices resulting from the other countries can be found in the online repository where the code is stored: https://github.com/Juillermo/VAA-data. Five colormaps are associated to each question in the following order (left to right):

(1) the distance matrix learned by our algorithm, with red cells as positive numbers, blue cells as negative ones (the intensity corresponds to the magnitude), and green cells as combinations that do not happen in the dataset,

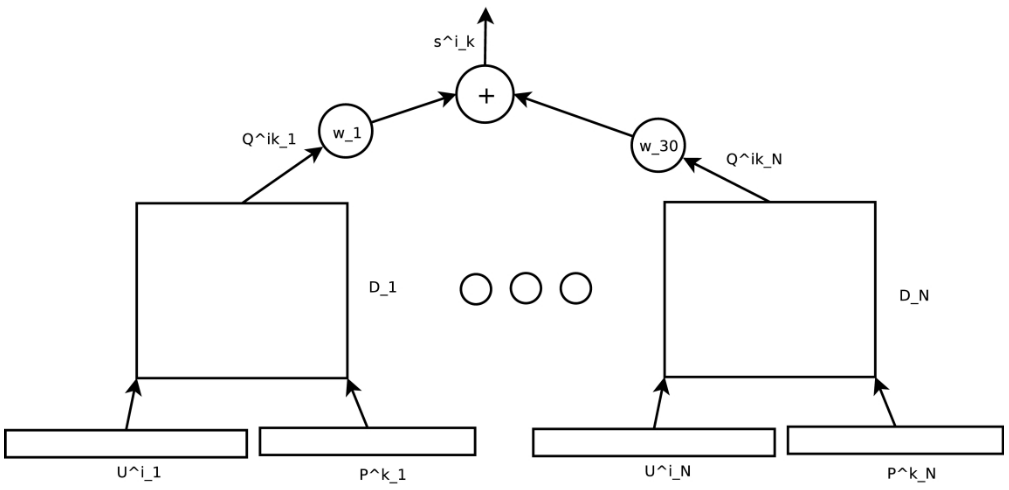
(6) the frequencies of answer combinations, which reflect which cells of the distance matrix are being used,

(7) the frequencies of per-party users' answers, showing which are the most common answers from the VAA users,

(8) the frequencies of per-party users' answers normalized per party, which reflect better what the prevalent answers from the users of each party are,

(9) and the parties' positions on the questions, decided by the VAA developers.

Distance matrix plots are all on the same scale, so stronger distance matrices (i.e. questions with higher weight) have more intense colours.



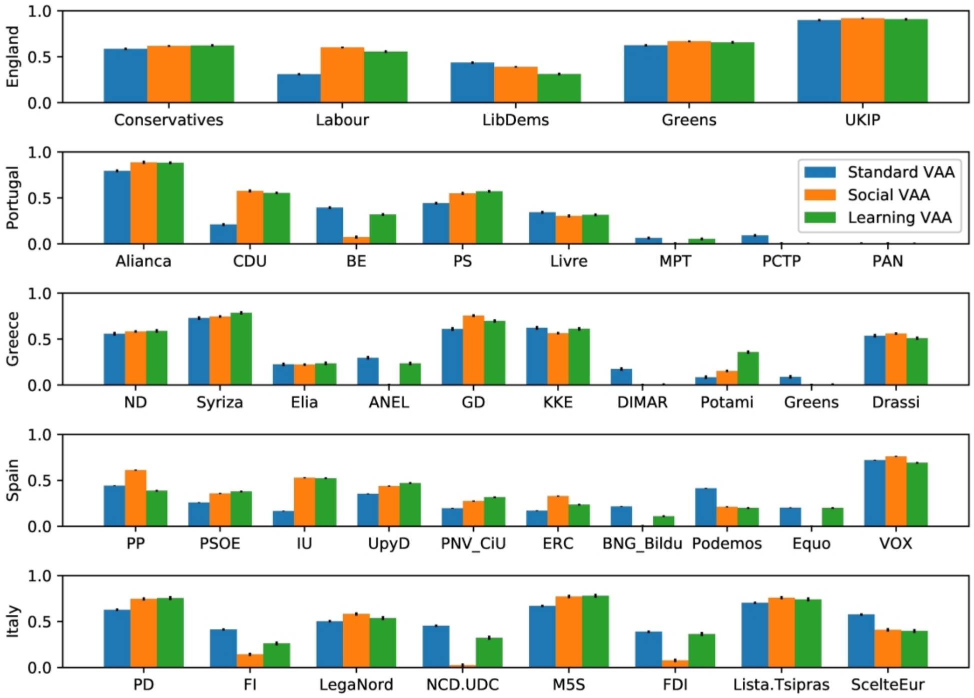
Transformation of a user profile and a party profile for getting the conjoined score, s^i_k. Our method uses different weights and distance matrices instead of the same ones for every policy issue.

422x202mm (300 x 300 DPI)

	EUvox 2014					**Aquienvoto**
	Greece	Italy	Portugal	Spain	England	Spain
N. of parties	10	8	8	10	5	9

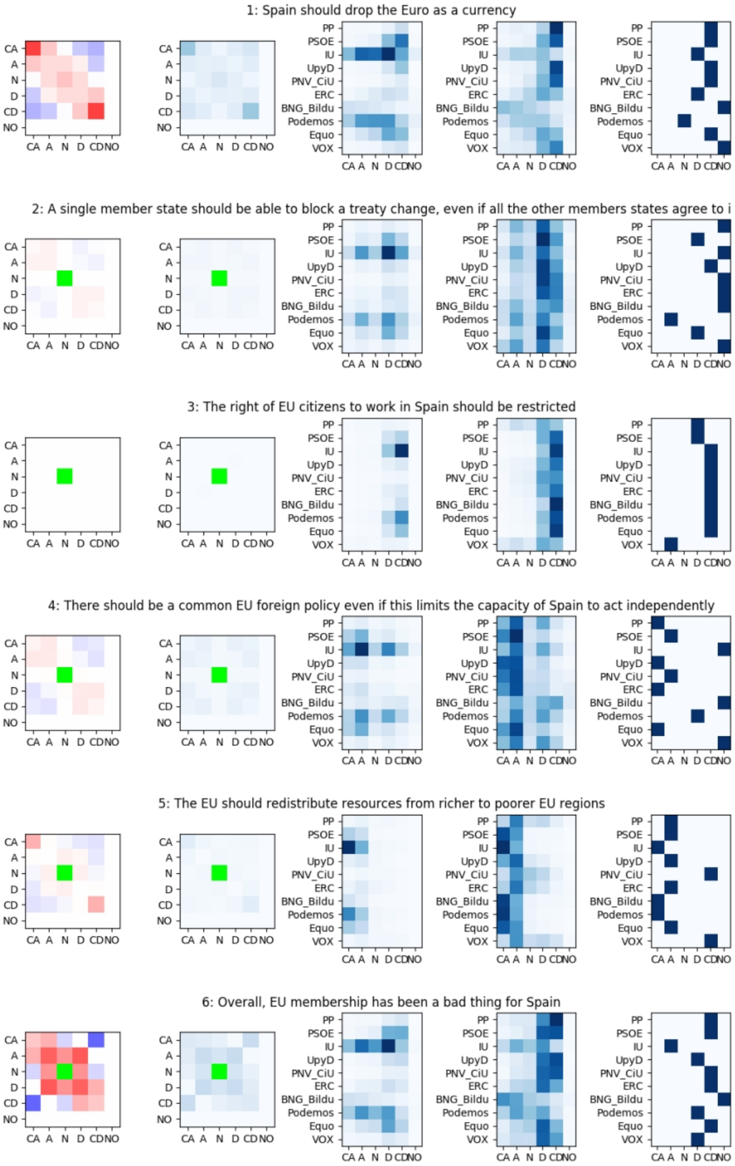
Filter Effects

Survey	Country	Condition	Original	Cumulative Attrition
EUvox 2014	Greece	Total	63923	0%
		All answers	61051	4.49%
		Filters	45274	29.17%
		Demographic	7697	87.96%
		Voting intention	5406	91.54%
	Italy	Total	36943	0%
		All answers	35422	4.12%
		Filters	25160	31.90%
		Demographic	3617	90.21%
		Voting intention	3206	91.32%
	Portugal	Total	55263	0%
		All answers	53119	3.88%
		Filters	43414	21.44%
		Demographic	5585	89.89%
		Voting intention	5007	90.94%
	Spain	Total	281559	0%
		All answers	268689	4.57%
		Filters	172434	38.76%
		Demographic	36998	86.86%
		Voting intention	30808	89.06%
	England	Total	122630	0%
		All answers	113915	7.11%
		Filters	59291	51.65%
		Demographic	11951	90.25%
		Voting intention	10385	91.53%
AQV 2019	Spain	Total	217145	0%
		>= 25 answers	111558	48.63%
		Voting intention	109252	49.69%



F-scores of standard, community-based and Learning VAAs for each party in the EUvox 2014.

228x165mm (300 x 300 DPI)



1: Spain should drop the Euro as a currency

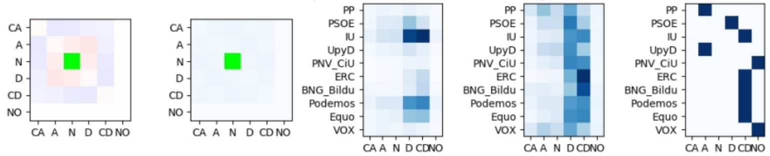
2: A single member state should be able to block a treaty change, even if all the other members states agree to i

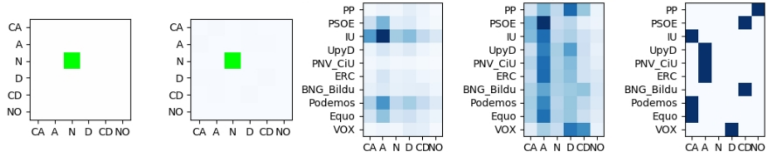
3: The right of EU citizens to work in Spain should be restricted

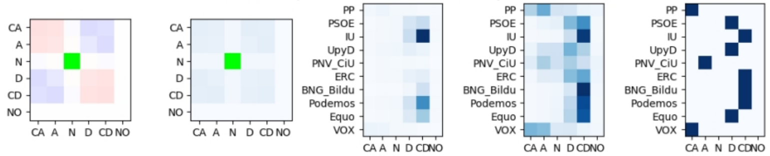
4: There should be a common EU foreign policy even if this limits the capacity of Spain to act independently

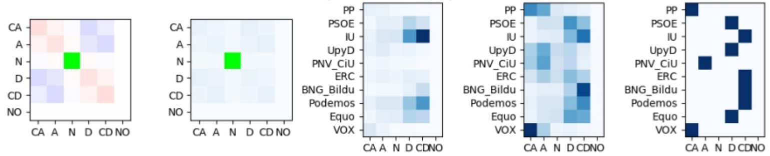
5: The EU should redistribute resources from richer to poorer EU regions

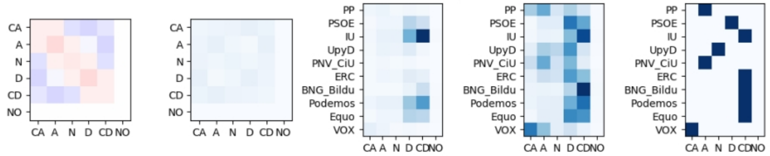
6: Overall, EU membership has been a bad thing for Spain

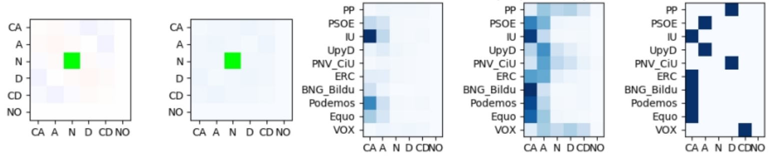

7: EU treaties should be decided by the Cortes Generales rather than by citizens in a referendum.

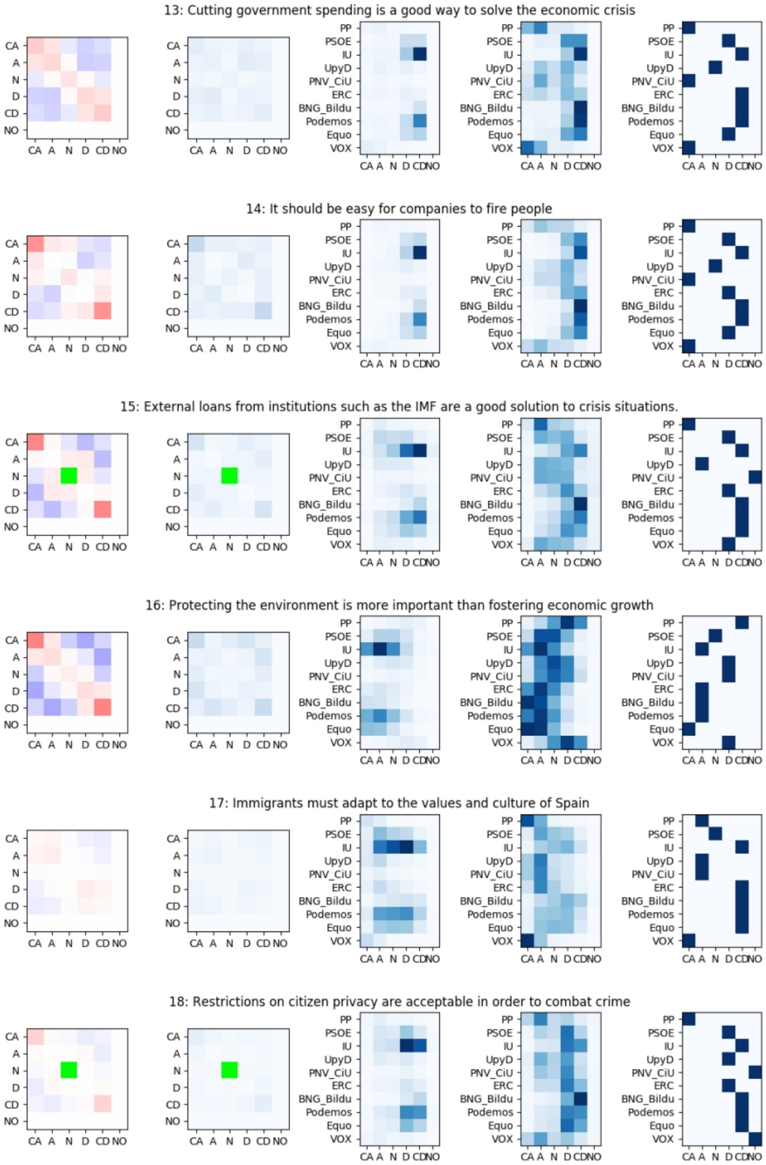

8: To address financial crises, the EU should be able to borrow money just like states can


9: Free market competition makes the health care system function better


10: The number of public sector employees should be reduced


11: The state should intervene as little as possible in the economy


12: Wealth should be redistributed from the richest people to the poorest



13: Cutting government spending is a good way to solve the economic crisis

14: It should be easy for companies to fire people

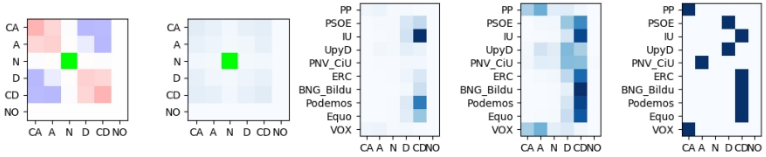
15: External loans from institutions such as the IMF are a good solution to crisis situations.

16: Protecting the environment is more important than fostering economic growth

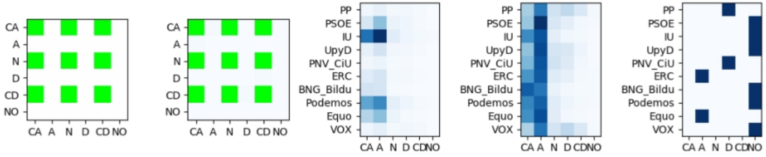
17: Immigrants must adapt to the values and culture of Spain

18: Restrictions on citizen privacy are acceptable in order to combat crime

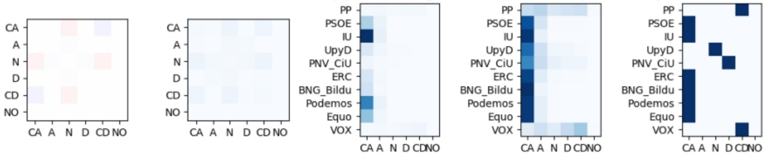
19: To maintain public order, governments should be able to restrict demonstrations



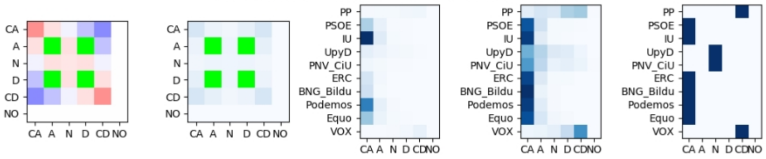
20: Less serious crimes should be punished with community service, not imprisonment



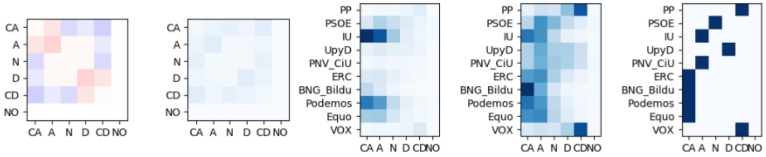
21: Same sex couples should enjoy the same rights as heterosexual couples to marry



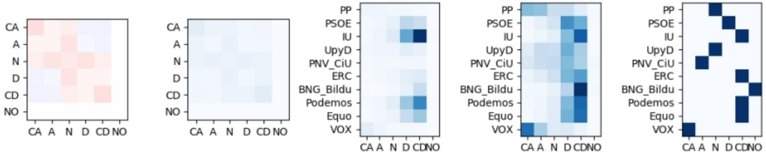
22: Women should be free to decide on matters of abortion

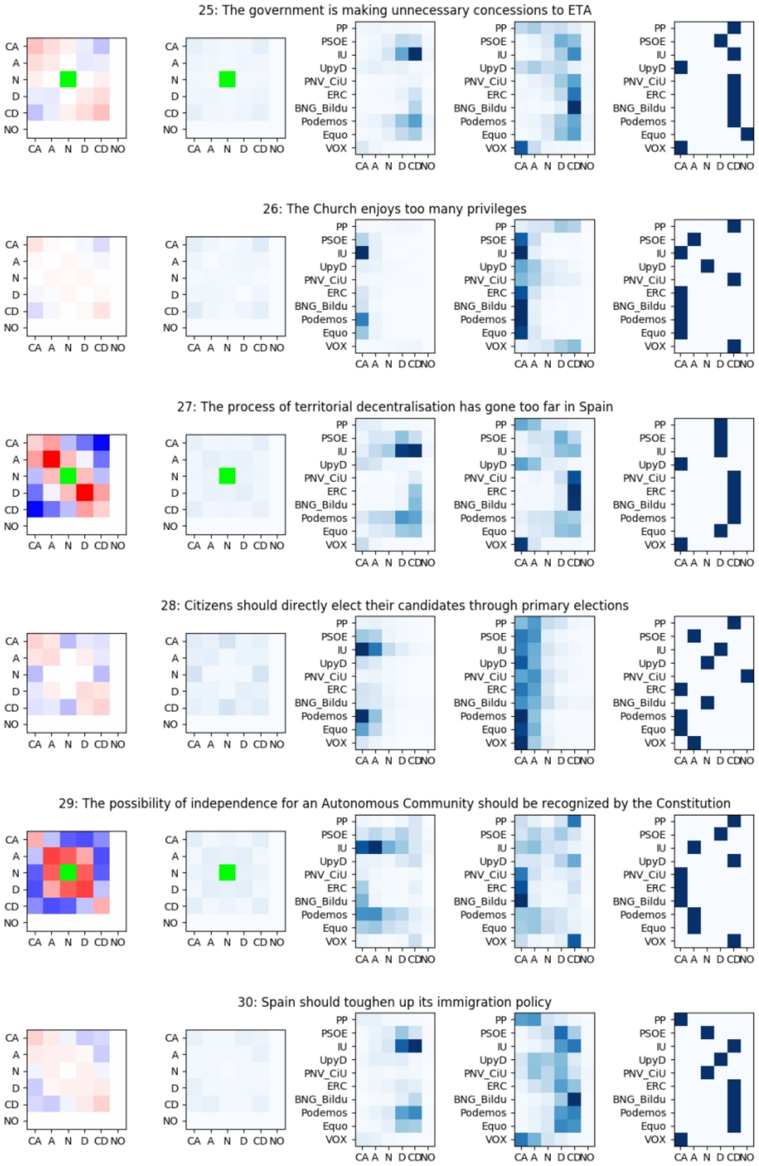


23: The recreational use of cannabis should be legal



24: Islam is a threat to the values of Spain





25: The government is making unnecessary concessions to ETA

26: The Church enjoys too many privileges

27: The process of territorial decentralisation has gone too far in Spain

28: Citizens should directly elect their candidates through primary elections

29: The possibility of independence for an Autonomous Community should be recognized by the Constitution

30: Spain should toughen up its immigration policy