Undecided Voters as Set-Valued Information – Machine Learning Approaches under Complex Uncertainty

Dominik Kreiss[1], Malte Nalenz[2], and Thomas Augustin[3]

[1] LMU Munich, Department of Statistics, Ludwigstr. 33, Munich
dominik.kreiss@stat-uni.muenchen.de
[2] LMU Munich, Department of Statistics, Ludwigstr. 33, Munich
malte.nalenz@stat-uni.muenchen.de
[3] LMU Munich, Department of Statistics, Ludwigstr. 33, Munich
thomas.augustin@stat-uni.muenchen.de

Abstract. Undecided voters in pre-election polls, even though an increasing phenomenon and issue in electoral research, have mostly been neglected in conventional analysis so far. We argue to include this inherent form of uncertainty in a set-valued manner, in order to make the most of the valuable information, not improperly reducing voters' response to either an spuriously precise answer or to drop outs. The resulting consideration set consists of all elements the individual is still pondering between and can be interpreted in two ways, depending on the question at hand. First, for the sake of forecasting, it can be seen as a coarse version of the yet unknown element the individual ends up choosing, using the information for so-called epistemic modeling. Second, from an so-called ontic view, it can be seen as entity of its own, representing the individual's current position accurately and thus allowing to examine structural properties within the population. Both views provide good opportunities for machine learning. In this paper we introduce one exemplary approach based on each view, analysing structural properties using spectral clustering and forecasting using random forests, providing initial methodology for this type of complex, non-stochastic uncertainty. The theory is applied with constructed consideration sets to the most recent German federal election of 2017, using data from the *German Longitudinal Election Study*. The results are promising, laying the groundwork for further machine learning approaches concerning this natural type of inherent uncertainty.

Keywords: Epistemic imprecision · Ontic imprecision · Set-Valued Data · Consideration Sets · Random Forests · Spectral Clustering · Election

1 Introduction

Increasing numbers of undecided voters before an election[4] urge us to find new ways to deal with these individuals in statistical analysis and empirical election

[4] see for example [19,4]

research. Conventionally, the undecided voters are either forced by the questionnaire to give a precise answer or neglected in further analysis reliant on possibly unjustified assumptions (e.g. [17,15]). This leaves the undecided with the options to either over-simplify their position conveying incorrect information, or to drop out. Hence, recently in [17,16,15,12,13] the authors argue to include set-valued response options in surveys. Several arguments are put forward, like the reduction of nonresponse, the natural procedure or the more accurate representation of uncertainty. Despite these advantages, set-valued response options are regrettably not yet included in most surveys, also because methodology handling this type of information is in the beginning stages only. Thus, with this paper we contribute to a solution of the resulting "chicken-egg dilemma" [9, p. 7], providing approaches and ideas for such data.

Human choice generally, as argued by [16, p. 256], can be seen as a process in stages, excluding possibilities until arriving at one final element. Thus, at a given point in time before an election, which resembles a choice of N individuals amongst a finite set of alternatives $\{1, \cdots, s\} = S$, not every individual's position can be determined by only one element of the choice set. As several individuals are still pondering between options, the most accurate representation of their position is a set, excluding all options of S they will definitively not choose. This set, consisting in the case of a decided voter of one and a still undecided voter of several elements, determines naturally and accurately their position and will from now on be called *consideration set* following [16].

Indecision amongst voters is hereby a natural and very interesting example with practical relevance for the theoretical groundwork laid by Couso and Dubois (e.g. [8,7]). Following them, the resulting set-valued information can be interpreted in two ways, dependent on the question at hand. First, considering the election outcome, it can be seen as a coarse version of one true but at the time unknown element contained in the set, providing incomplete information. This is the so-called *epistemic* or *disjunctive* view. Second, focussing on the time point of the survey, the set represents the positions as a non-reducible entity of its own. This so-called *ontic* or *conjunctive* view regards a decided or undecided alike as a viable position with its own characteristics. Both views, even though very different, are justified, dealing with complementary issues.

In this paper we develop initial methodology for either view, providing first approaches and opportunities for machine learning to incorporate this set-valued information. With the ontic approach, regarding the undecided between specific parties as positions of their own, new structural properties concerning the political landscape can be examined. We generate socioeconomic clusters (using *spectral clustering*) and assess structural properties within the undecided and decided before the German federal election of 2017. For the epistemic view, we develop a forecasting approach incorporating the otherwise wasted information of the undecided. We hereby estimate *transition probabilities* of the undecided with *random forests* based on the decided individuals and provide an overall forecasting approach, reliant on simulation and assumptions, that is able to take the information of the consideration set into account. Both approaches are ap-

plied to data of the most recent German federal election of 2017, provided by the *German Longitudinal Election Study* [10] with constructed consideration sets.

This paper is structured as follows: First, in Section 2 we consolidate the ontic and epistemic methodology and introduce possible approaches for either view. We later apply the approaches to the most recent German federal election in Section 3. The concluding remarks in Section 4 reflect on the possibilities and challenges of this new way of incorporating undecided voters.

2 Methods

2.1 The Ontic and Epistemic Views

Dependent on the question at hand, a set consisting of the same elements can be interpreted in two different ways. To take a meanwhile classical example (e.g. [8]), if we are interested in the languages an individual is capable to speak, the set {English, French, German} is a precise representation of the truth, while if we are interested in the language he or she feels the most comfortable with, the same set contains only incomplete information. Equally, in the case of an undecided voter before an election, we can either focus on the indecision itself, which is accurately represented by the set as a whole, or focus on the choice outcome, in which case only incomplete information is provided. Thus, set-valued information obtained by a pre-election survey can be used in two different ways. Reflecting uncertainty in electoral analysis in a set-valued manner is a natural and especially interesting application for the theoretical groundwork laid by Couso and Dubois, presented for example in [8,7,3]. The state space of the consideration sets consist of all possible combinations of the original options, which can naturally be represented by the power set $P(S)$ of the set of the original options. Hence, in the case of an undecided, we are provided with a set ℓ that can be described as the realization of a measurable mapping $\mathcal{Y} : \Omega \to P(S)$ from some underlying space Ω into the set of all combinations. This set-valued representation can now be interpreted under ontic or epistemic imprecision.

Starting with the set as entity of its own, also called ontic or conjunctive interpretation, we consider undecided voters between specific parties as a further position. In this case, the consideration set is a precise representation of something naturally imprecise. Hence, it cannot be reduced or improved in any way. As the original choice set consists of finite elements measured on a nominal scale, the power set does as well, satisfying the same basic mathematical properties. Hence, methodology based on conventional approaches can broadly be transferred. Quite naturally, but most importantly, this protruding trait of ontic approaches opens up a wide range of options to apply state of the art machine learning approaches to data with this type of complex non-stochastic uncertainty. By this, the ontic view of undecided voters prior to the election enables new ways to examine structural properties within the political landscape.

The epistemic view, in contrast, focuses on the election outcome. Hereby, the set at the time point of the poll, accurately representing the position of an

undecided individual, is a coarse version of the one true element the individual ends up choosing. In other words, the set-valued information is an imprecise version of something precise. Thus, only incomplete information about the phenomena of interest (the eventual choice) is provided within the consideration set. To obtain statements about the precise value of interest, next to incorporating further information, one can make rather rigorous assumptions or reflect the uncertainty within interval-valued results. After all, we are only provided with incomplete information in the sense that $\forall \omega \in \Omega$ only $Y(\omega) \in \ell = \mathcal{Y}(\omega)$ is observable, with $\mathcal{Y}$ again as a mapping $\Omega \to P(S)$ now representing the set of mappings $\{Y : \Omega \to S, \forall \omega, Y(\omega) \in \mathcal{Y}(\omega)\}$, where we assume one of each is the true underlying mapping (e.g. [7, p. 1504]). As a consequence, reducing the set or assigning probabilities to each of its elements is usually strived for, in order to retrieve as precise information as possible about the variable of interest.

The following two sections reflect on possible applications of ontic as well as epistemic imprecision conducted with data from pre-election polls.

2.2 More on the Ontic Approaches

While in conventional pre-election voter analysis the undecided are neglected, we try to show in this section how including those individuals in a set-valued manner can open up new perspectives and findings about structural properties. The common procedure to monitor each month and regular before elections political orientations and developments in the political landscape of a country[5] could be enriched by these approaches, including further positions of interest. As the consideration sets are, as described in Section 2.1, the most accurate representation of the undecided, ontic approaches not only enable new findings, but also represent the current structural properties of the political landscape in the most accurate way. Several approaches are possible, examining different aspects of the political landscape concerning the undecided. Recently, as one example, we [12] extended discrete choice models with the undecided's consideration sets, providing new findings about the undecided in Germany.

For the ontic approach, we focus on the connection between socioeconomic clusters within the population and the undecided. Hereby, trends of indecisiveness could be located and assigned towards specific clusters. Thus, we cluster our data according to socioeconomic variables and examine structural differences of decided and undecided within the resulting socioeconomic groups. Conclusions from the composition of the clusters can then be interpreted from a political science perspective. We use spectral clustering (e.g. [18]) as a common machine learning approach for dividing our population in characteristics based on similarity in their covariate values. Hereby, we make use of the spectrum of a similarity matrix in order to perform dimensionality reduction and natural scaling on the data before clustering in fewer dimensions. The eventual clustering on this new data is usually performed by a simple algorithm like k-means.

[5] like for example in Germany the *Politbarometer* `https://www.forschungsgruppe.de/Aktuelles/Politbarometer/` last visited: 28.07.2020

The approach introduced in this paper is only meant to exemplify the opportunities of machine learning to describe this new type of data under ontic imprecision. It goes without saying, that there are numerous possibilities for straightforward applications of machine learning approaches, examining structural properties concerning the undecided, while already this rather simple one can initiate new ways to think about the political landscape.

2.3 More on the Epistemic Approaches

The epistemic approach, like sketched in Section 2.1, concerns itself with the yet unknown element in the consideration set the individual ends up voting for. Hence, in contrast to the ontic approaches addressing diverse questions, the epistemic ones try to improve forecasting, using the potentially valuable information of the undecided. As there is no information about the final choice of the undecided provided, either rather strong assumptions have to be made, or the uncertainty is manifested in the results using interval-valued identification. Thus, several approaches are possible, weighting the justifiability of assumptions with the precision of the results.[6] In a recent paper [13], we discuss this question, considering different approaches to incorporate the set-valued information into election forecasting, resulting in three different suggestions. Here, we pick up on the second one, achieving point-valued estimation by assuming that, given the covariates, the undecided choose identical to the decided with the consideration set as restriction of the possible outcomes.

Each individual holds a consideration set $\ell \in P(S)$ and covariates $X = x$ in some space $\mathcal{X}$. The consideration set is written as an event $\{\mathcal{Y} = \ell\}$ with $\ell \in P(S)$ and his or her possibly unknown choice on election day as $\{Y = l\}$ with $l \in S$. In order to estimate transition probabilities, the approach uses the distribution of the decided $P(Y = l|X = x, I_d = 1)$, which can be estimated from the data, with I_d as the indicator function for being decided. In order to incorporate the information of the consideration sets, all options not in ℓ are excluded. Therefore, scaling the estimates from the decided to comply with the multinomial distribution results in:

$$\underbrace{\hat{P}(Y = l|\mathcal{Y} = \ell, X = x)}_{Transition\ Probabilities} = \frac{\hat{P}(Y = l|X = x, I_d = 1)}{\sum_{a \in \ell} \hat{P}(Y = a|X = x, I_d = 1)} \tag{1}$$

leading to point-valued estimation of every parameter. Hence, to ensure point valued estimation, some implicit assumption of independent coarsening in the sense that undecided behave identical to the undecided is made. This resembles a random coarsening process, but satisfies mathematical properties different from the common CAR assumption of [11].

We utilize random forests [5] to estimate the conditional distributions for each undecided individual in Equation (1). Random forests grow a sequence of independent decision trees on bootstrap samples of the original data. At each

[6] also see Manski's Law of Decreasing Probability [14, p. 1]

node, only a subset of the covariates is used for splitting, efficiently reducing the correlation between the individual trees. These decorrelated, individually weak, trees are subsequently combined into an ensemble, typically through voting or by averaging the probability estimates. The resulting ensemble classifier was generally shown to significantly improve generalization performance and stability. As random forests are based on a set of decision trees, they posses several properties that are desirable in epistemic forecasting:

- They can naturally capture interaction effects between variables, without the need of prespecification.
- Non-linear effects can be appoximated. While single decision trees struggle to capture linear relationships, random forests can approximate them reasonably well.
- Both numeric and categorical covariates are natively supported without the need of any preprocessing.

Another reason to chose random forests over other popular ensemble methods, such as gradient boosting, is their stability towards a large grid of reasonable parameter choices [1].

As for the decided voters both the outcome Y and the covariate values X are known, random forests are applied directly, using the decided as training data. This implicitly presupposes, in accordance with above, that the conditional distributions of Y given the covariates are equal for decided and undecided voters, hence $P(Y = l|X = x, I_d = 1) = P(Y = l|X = x, I_d = 0)$. For easier reference in the discussion, we call this *structural similarity assumption*. Thus, for the undecided voters we can estimate the conditional multinomial distribution over all possible parties for each individual, using the structural similarity assumption. Note, however, that the random forest output is only a first level prediction, that is subsequently refined by taking into account the information given by the consideration sets, using Equation (1). This combines the predicitve power of random forests with the additional information given by the consideration sets. [7]

Provided with the estimated transition probabilities resulting from Equation (1), hence the probability an undecided chooses a particular party from their consideration set, we want to estimate the overall distribution together with the decided individuals. To this end, we use a Monte Carlo simulation approach: For the undecided we simulate precise decisions, drawing from the restricted multinomial distribution of each individual. Thus, the decided and the simulated data from the undecided can be used together for straightforward estimation of the overall distribution. In order to minimize the variance of the results, we repeat the process, averaging over the different estimates. The resulting point-valued estimates can be directly used for forecasting. Nevertheless, one should explicitly mention that the underlying assumptions are disputable.

[7] We do not use the undecided in the first level of estimation with some kind of simulation, in order to avoid strong assumptions about the final outcome in the consideration sets.

Thus, this approach can be seen as only a first example of how to integrate state of the art machine learning reliant on set-valued information of the undecided.

3 Application

3.1 The Data from The GLES

The ideas developed in Section 2.2 and 2.3 are applied for the most recent German federal election of 2017, using the state of the art pre-election poll conducted by the *GLES*[8]. Set-valued answer options are regrettably not included in this survey, but the assessment of the parties by the individuals and their statement about the certainty of their choice are, enabling construction of a consideration set as already conducted by [17, p. 261].

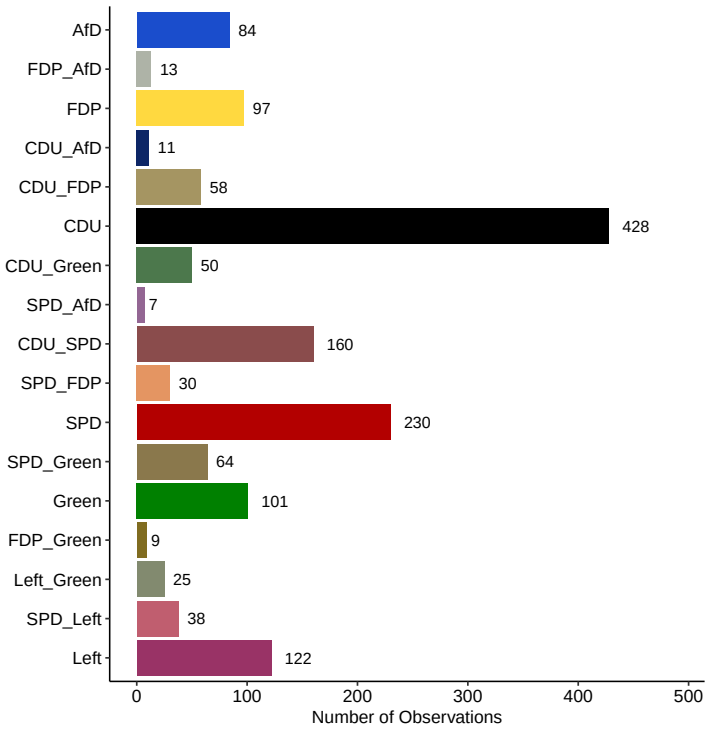

Fig. 1. The plot illustrates the distribution of the positions in our dataset, including decided and undecided individuals between exactly two parties. On the x-axis the numbers of observations and on the y-axis the corresponding position are shown.

[8] German Longitudinal Election Study: Pre- and post- election cross-section available under `https://www.gesis.org/wahlen/gles/daten`; last visited: 27.07.20

For our analysis, we use the so-called second vote[9] for the six main parties[10] anticipated to reach at least one seat in the parliament, in addition not including non-voters. As always in our illustrative example, structures of nonresponse in the dataset are not explicitly adjusted for. Moreover, we only focus on the most common case of indifference between exactly two parties.

The distribution of the positions in our data is illustrated in Figure 1. As one can see, the decided make up the major positions within this dataset, but 546 of the overall 1558 individuals are undecided, constituting one third of the population. A big proportion of the undecided is pondering between the two biggest and currently governing parties CDU and SPD with 160 observations, while there are few voters undecided between (combinations with) smaller parties in our dataset. These first descriptive results already hint towards a structural difference between the decided and undecided.

3.2 Clustering to Examine Ontic Structures

The approach sketched in Section 2.2 can be divided into two parts. First, we use spectral clustering with the three variables *age*, *household size* and *household income* to identify three separate socioeconomic groups within our population. The results are shown in Figure 2. While the first cluster mostly represents rather

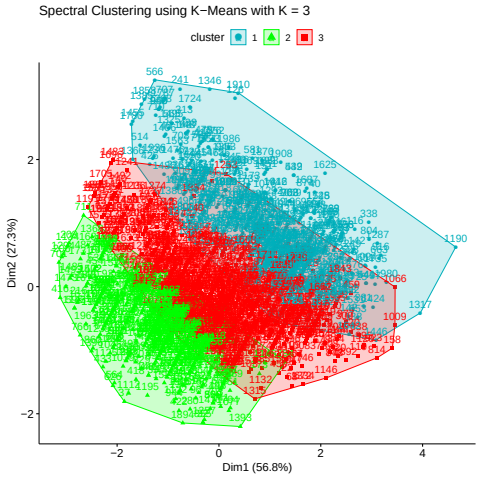

Fig. 2. This figure visualises the resulting three clusters using spectral clustering with the three variables *age*, *household size* and *household income* and k-means.

[9] The second vote basically determines the distribution of the seats among the parties, and thus is usually used for forecasting. For more information see: `https://www.bundeswahlleiter.de/en/bundestagswahlen/2021/informationen-waehler/wahlsystem.html`, last visited 27.07.20

[10] The parties are: AfD, FDP, CDU (including CSU), SPD, Green, Left

young and well earning individuals, living in a household with in average almost three individuals and the second one consist predominantly of pensioners, the third one is more intermixed. Considering we used three variables, the separation visualised in Figure 2 is proficient for our purposes.

Second, we examine the distribution of the consideration sets amongst the clusters as viable positions of their own. Thus, Figure 3 visualises the distribution of the positions, on the left side for the decided only and on the right side for the consideration sets, separate for the three clusters. As we can see, the

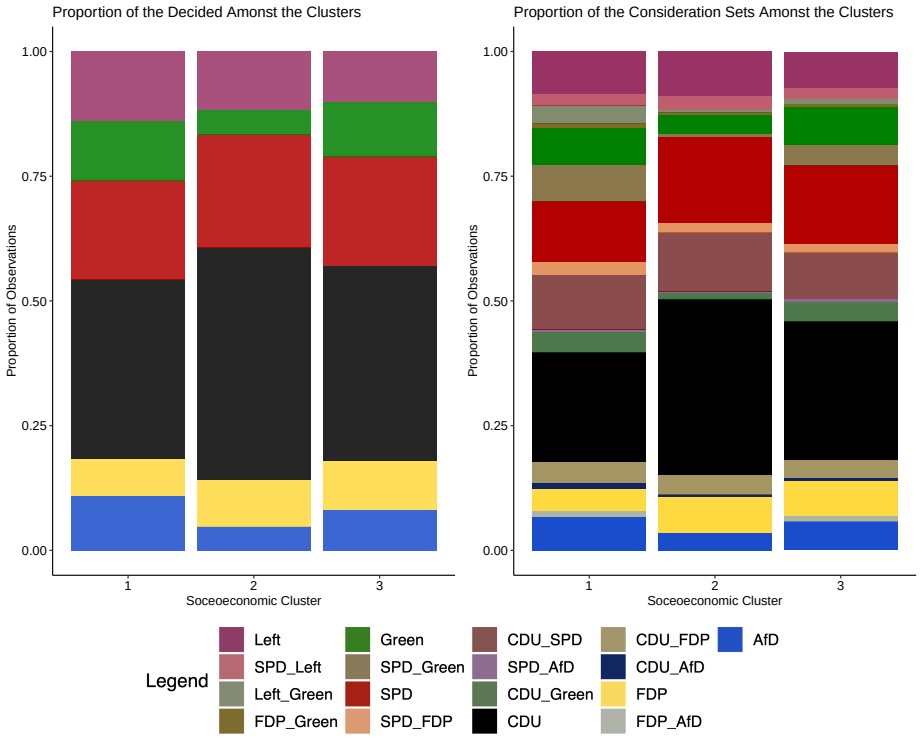

Fig. 3. This figure illustrates the composition of the three socioeconomic clusters, on the left for the decided only and on the right for the consideration sets.

positions are very unevenly distributed amongst the clusters. Notable, for example, is the high proportion of undecided between the Green and other parties within the first cluster, as mentioned above mostly consisting of young voters with comparable high income. The proportion of overall undecided is the highest within this first cluster in our data as well. Next to the insights into the political landscape, Figure 3 also shows structural differences between the decided and undecided. This underlines the importance of including undecided voters in electoral forecasting in order to avoid bias. The results of this first analysis are

therefore twofold. First, we examined structural properties, analysing predominate affiliation of specific undecided voters towards specific clusters. Second, we established structural differences between the decided and undecided.

3.3 Epistemic Forecasting

As described in Section 2.3, a random forest was applied using all available covariates, consisting of sociodemographic variables and several batteries of opinion questions. For training only the decided voters were used, as argued above. Using 10-fold cross validation on the decided voters led us to an estimated error rate of 25.4 %. This suggests that some of the covariates are clearly predictive. Furthermore, restricting the outcome space via the consideration sets adds important information. The Monte Carlo simulation to obtain overall estimates as explained in Section 2.2 is repeated 1000 times, leading to results illustrated in Figure 4 next to the ones only based on the decided.

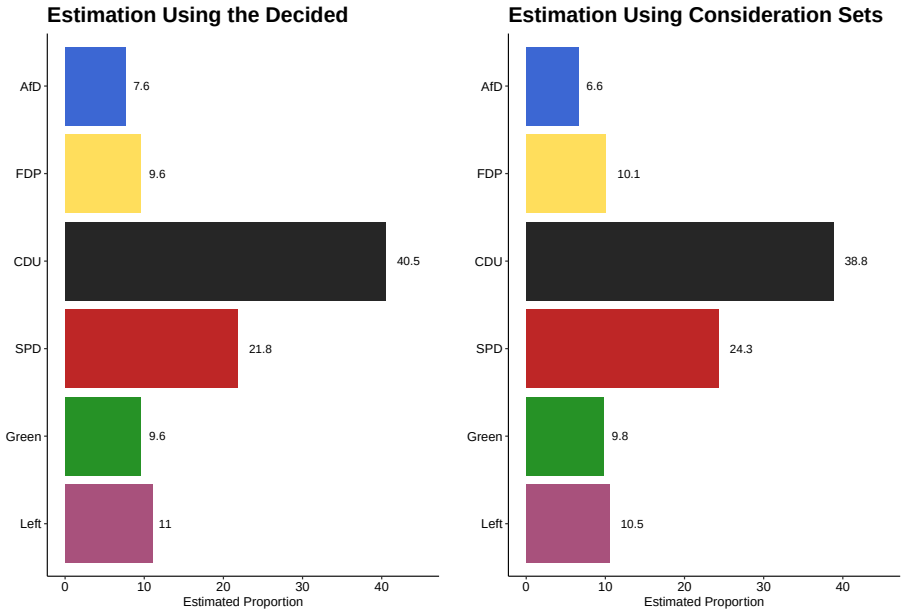

Fig. 4. The plot illustrates the forecasts of the overall distribution for the six main parties. On the left side based only on the decided and on the right incorporating undecided voters using random forest and simulation. The y-axis shows the six main parties while the x-axis shows the corresponding estimated proportion.

There are notable differences, stressing the impact of including the undecided. The biggest party CDU is less strongly represented including the undecided, while the SPD has a higher proportion. While the Green Party and FDP have

slightly higher estimates including the undecided, the wing parties AfD and Left Party have lower ones.

When drawing conclusion on political issues, one has to be cautious not to overinterpret our results, as the nonresponse structures are not adjusted for and the consideration sets had to be constructed. Nevertheless, including the undecided using random forests with the structural similarity assumption is straightforward applicable, providing first sound methodology which could be improved by further research.

4 Concluding Remarks

In this paper we proposed new ways to include the otherwise wasted information of undecided voters by making use of their consideration sets. For the ontic view, common methodology can broadly be transferred as the power set satisfies the same basic mathematical properties of the original data, while for the epistemic view, rather strong and untestable assumptions are necessary in order to obtain more concise forecasting. Thus, numerous approaches are possible, integrating machine learning into this natural type of uncertainty. While the ontic view focuses on new findings in structural properties, the epistemic one may improve election forecasting by including this valuable information.

We introduced one approach each, analysing structural properties with spectral clustering and extending forecasting reliant on the structural similarity assumption and random forests. Both approaches, even though not yet perfected, yield promising results. Thus, we provided initial methodology which must be further developed and improved. Concerning forecasting, new sources of information could be incorporated like decisions in previous elections or expert knowledge in a (generalised) Bayesian way. Furthermore, set-valued approaches are promising. This includes cautious data completion explicitly [2] (see also, e.g. for classifiers, [6]) as well as working in the spirit of partial identification following [14], permitting to weaken assumptions resulting in more credible results. For ontic approaches, discrete choice models are of particular interest, examining connections between attributes and indecision between specific parties. Hereby, highlighting attributes of individuals determined to vote for the right-wing party AfD compared to those only considering it, might provide essential insights into the trend towards nationalistic parties.

With this paper, we open up this complex uncertainty structure towards exciting applications for a broad spectrum of machine learning methodology.

Acknowledgement. We sincerely thank the anonymous reviewers for their helpful remarks. Further we thank the LMU mentoring, supporting young researchers, and the GLES for providing the dataset.

References

1. Aggarwal, C.C.: Outlier analysis. In: Data mining. pp. 237–263. Springer (2015)
2. Augustin, T., Walter, G., Coolen, F.: Statistical inference. In: Augustin, T., Coolen, F., de Cooman, G., Troffaes, M. (eds.) Introduction to Imprecise Probabilities, pp. 135–189. Wiley (2014)
3. Augustin, T., Coolen, F., De Cooman, G., Troffaes, M., (Eds.): Introduction to imprecise probabilities. Wiley (2014)
4. BBC: Why has the UK become a nation of political swingers? BBC News (2017), `https://www.bbc.com/news/uk-politics-39103972`, (Last visited 28.07.2020)
5. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
6. Corani, G., Zaffalon, M.: Learning reliable classifiers from small or incomplete data sets: the naive credal classifier 2. Journal of Machine Learning Research **9**, 581–621 (2008)
7. Couso, I., Dubois, D.: Statistical reasoning with set-valued information: Ontic vs. epistemic views. International Journal of Approximate Reasoning **55**(7), 1502–1518 (2014)
8. Couso, I., Dubois, D., Sánchez, L.: Random sets and random fuzzy sets as lll-perceived random variables. Springer (2014)
9. Fink, P.: Contributions to reasoning on imprecise data. Ph.D. thesis, LMU Munich, Faculty of Mathematics, Computer Science and Statistics (2018), `https://edoc.ub.uni-muenchen.de/22547/`
10. GLES: German longitudinal election study (2019), `https://www.gesis.org/wahlen/gles/`, (Last visited 28.07.2020)
11. Heitjan, D., Rubin, D.: Ignorability and coarse data. The Annals of Statistics pp. 2244–2253 (1991)
12. Kreiss, D.: Examining Undecided Voters in Multiparty Systems. Master's thesis, LMU Munich, Department of Statistics (2019), `https://epub.ub.uni-muenchen.de/70668/`
13. Kreiss, D., Augustin, T.: Undecided voters as set-valued information, towards forecasts under epistemic imprecision. In: Davis, J., Tabia, K. (eds.) SUM 2020. Springer (2020)
14. Manski, C.: Partial identification of probability distributions. Springer (2003)
15. Oscarsson, H., Oskarson, M.: Sequential vote choice: Applying a consideration set model of heterogeneous decision processes. Electoral Studies **57**, 275–283 (2019)
16. Oscarsson, H., Rosema, M.: Consideration set models of electoral choice: Theory, method, and application. Electoral Studies **57**, 256–262 (2019)
17. Plass, J., Fink, P., Schöning, N., Augustin, T.: Statistical modelling in surveys without neglecting 'The undecided'. In: Augustin, T., Doria, S., Miranda, E., Quaeghebeur, E. (eds.) ISIPTA 15, pp. 257–266. SIPTA (2015)
18. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**(4), 395–416 (2007)
19. Zeit: Die Hälfte der Wähler hat sich noch nicht entschieden. Die Zeit: Online Newspaper (2017), `https://www.zeit.de/politik/deutschland/2017-08/bundestagswahl-umfrage-waehler-unentschlossen`, (Last visited 28.07.2020)