

---

# Contributions to Modeling with Set-Valued Data: Benefitting from Undecided Respondents

Dominik Sebastian Kreiß

---



München 2023



---

# Contributions to Modeling with Set-Valued Data: Benefitting from Undecided Respondents

Dominik Sebastian Kreiß

---

Dissertation  
an der Fakultät für Mathematik, Informatik und Statistik  
der Ludwig-Maximilians-Universität  
München

vorgelegt von  
Dominik Sebastian Kreiß  
aus Frankfurt am Main

München, den 31.07.2023

Erstgutachter: Prof. Dr. Thomas Augustin

Zweitgutachter: Prof. Dr. Joseph Sakshaug

Drittgutachter: Prof. Dr. Martin Spieß

Tag der mündlichen Prüfung: 11.10.2023

## Acknowledgement

I would like to express my sincere gratitude to everyone who contributed to this dissertation and accompanied me for the past years! A special thanks goes to...

- ... Thomas Augustin for being the best supervisor one could hope for. Thank you for the support, advice, guidance, endurance, trust in me, and most importantly all the time you took. You made my happy years at the institute possible. I am deeply grateful for what I have learned from you and also for you as a person!
- ... Joseph Sakshaug and Martin Spieß for their willingness to fill the role of the external reviewers and Volker Schmid and Christian Heumann for steering the examination committee.
- ... my coauthors Thomas Augustin, Eyke Hüllermeier, Malte Nalenz, Julian Rodemann, and Georg Schollmeyer for the great collaboration.
- ... all former and current members of my working group: Thomas Augustin, Hannah Blocher, Eva Endres, Cornelia Fütterer, Gilbert Kiprotich, Christoph Jansen, Malte Nalenz, Julian Rodemann, Georg Schollmeyer, and Patrick Schwaferts. I truly enjoyed the time with the AG and I am happy that we also kept contact during the lockdowns.
- ... Christoph Jansen for proofreading the introduction of my dissertation and encouraging me to start my Ph.D. at our working group.
- ... all remaining former and current colleagues at the Department of Statistics for the excellent general atmosphere. Special thanks go to Ben, who helped me with organizational questions.
- ... Civey for the great cooperation and for implementing our research questions.
- ... the LMU Mentoring program at Faculty 16 for enabling interactions between Ph.D. candidates, support with equipment, and interesting workshops.
- ... Elke Höfner and Brigitte Maxa for their friendly administrative help.
- ... my wife Tabea for all the support, listening, and joy you bring to my life and to my daughter Lea for enriching my life ever since yours started seven months ago.

## Zusammenfassung

Diese Dissertation entwickelt einen methodischen Rahmen und Ansätze, um unentschlossene Befragte, im Speziellen in Vorwahlbefragungen, besser berücksichtigen und modellieren zu können. Nachdem Entscheidungen als Prozesse betrachtet werden können, die schrittweise Alternativen ausschließen, bis sie zu einem endgültigen Ergebnis gelangen, argumentieren wir, dass unentschlossene Teilnehmende in einschlägigen Umfragen am besten durch die Menge ihrer infrage kommenden Optionen dargestellt werden können. Im Gegensatz zu der herkömmlichen Vernachlässigung der Unentschlossenen kann diese Art der Erhebung potenziell Verweigerungen verringern und neue, wertvolle Informationen sammeln. Wir betrachten die resultierenden mengenwertigen Daten als Random Sets, die auf zwei Arten interpretiert werden können und entwickeln jeweils Modellierungsansätze. Die erste Interpretation wird als *ontisch* bezeichnet, bei der jede Position zum Zeitpunkt der Erhebung als eine Einheit an sich betrachtet wird. Dies kann als eine präzise Darstellung von etwas gesehen werden, das von Natur aus mengenwertig vorliegt. Damit ergeben sich neue Möglichkeiten für strukturelle Analysen, die insbesondere auch unentschlossene Personen berücksichtigen können. Wir zeigen, wie die zugrundeliegende kategoriale Datenstruktur bei diesem Formalisierungsprozess der Antworten für bestimmte Modelle erhalten bleibt und wie gängige Methoden weitgehend übertragen werden können. Nachdem die Menge die letztendliche Entscheidung enthält, kann sie nach der zweiten Interpretation als eine vergrößerte Version einer zugrundeliegenden Wahrheit betrachtet werden, was als *epistemische* Sichtweise bezeichnet wird. Diese ungenaue Information über etwas eigentlich Präzises kann dann zur Verbesserung von Wahlprognosen verwendet werden. Neben mehreren Modellierungsansätzen wird eine Faktorisierung der Likelihood als Grundlage für Prognosen, die mengenwertige Daten mit einbeziehen, entwickelt. In unseren Ansätzen wird explizit auf den Abwägungsprozess zwischen der Stärke von Annahmen und der mit Intervallen kommunizierte inhärenten Unsicherheit der Ergebnisse eingegangen. Um die Ansätze zu evaluieren und etablieren, wurde in Zusammenarbeit mit dem Meinungsforschungsinstitut *Civey* eine Vorwahlbefragung für die Bundestagswahl 2021 durchgeführt, bei der erstmals auch unentschlossene Wähler mengenwertig mit einbezogen wurden. Hierdurch konnten unsere theoretischen Überlegungen empirisch fundiert und weiterentwickelt werden.

## Summary

This dissertation develops a methodological framework and approaches to benefit from undecided survey participants, particularly undecided voters in pre-election polls. As choices can be seen as processes that – in stages – exclude alternatives until arriving at one final element, we argue that in pre-election polls undecided participants can most suitably be represented by the set of their viable options. This *consideration set sampling*, in contrast to the conventional neglect of the undecided, could reduce nonresponse and collect new and valuable information. We embed the resulting set-valued data in the framework of random sets, which allows for two different interpretations, and develop modeling methods for either one. The **first interpretation** is called *ontic* and views the set of options as an entity of its own that most accurately represents the position at the time of the poll, thus as a precise representation of something naturally imprecise. With this, new ways of structural analysis emerge as individuals pondering between particular parties can now be examined. We show how the underlying categorical data structure can be preserved in this formalization process for specific models and how popular methods for categorical data analysis can be broadly transferred. As the set contains the eventual choice, under the **second interpretation**, the set is seen as a coarse version of an underlying truth, which is called the *epistemic* view. This imprecise information of something actually precise can then be used to improve predictions or election forecasting. We developed several approaches and a framework of a factorized likelihood to utilize the set-valued information for forecasting. Amongst others, we developed methods addressing the complex uncertainty induced by the undecided, weighting the justifiability of assumptions with the conciseness of the results. To evaluate and apply our approaches, we conducted a pre-election poll for the German federal election of 2021 in cooperation with the polling institute *Civey*, for the first time regarding undecided voters in a set-valued manner. This provides us with the unique opportunity to demonstrate the advantages of the new approaches based on a state-of-the-art survey.

**This cumulative dissertation is based on the following six chronologically ordered contributions:**

**Contribution 1** focuses on the forecasting potential of this newly obtained information. We introduce a point-valued approach reliant on a homogeneity assumption between the decided and undecided, given the covariates. The approach is compared to the very cautious so-called Dempster Bounds and the conventional approach of neglecting the undecided entirely, arguing with Manki’s law of decreasing credibility. We applied the methods on artificially constructed consideration sets to establish initial methodology for the consideration set sampling as no real dataset was available yet.

**Contribution 2** discusses one machine learning application for either view of the random set. The ontic view enables insights with clustering, while random forests are used to improve forecasting once again with a homogeneity assumption for the epistemic one.

**Contribution 3** takes a Bayesian perspective. Hereby, the main goal is to find compromises between, on the one hand, cautious approaches, which tend to produce broad probability intervals as results, and on the other hand, possibly too strong assumptions leading to spurious, seemingly precise results. We utilized distribution assumptions to combine sparse and broad forecasts in a Bayesian manner.

**Contribution 4** is – without peer-review – published on arXiv one week after we obtained the data from Civey and one day before the election of 2021 to put our methodology to the test before knowing the election outcome. With the new survey data, we could evaluate our methodological considerations and deploy the developed approaches with promising results. We further introduced techniques focussing on coalition forecasting, providing more added value from regarding the undecided set-valued.

**Contribution 5** builds on the machine learning framework of superset learning in the epistemic context. This approach seeks to disambiguate the set-valued data to a certain extent. In our paper, we also address the tradeoff between caution and conciseness, for which we propose a way to construct a hierarchical family of subsets within the set-valued categorical observation. The practitioner is hereby enabled to choose the level of the coarseness of the results context-dependent.

**Contribution 6** develops a framework for structural analysis under the ontic view. It is argued in which case the new state space satisfies mathematical properties to transfer well-known methodology. Then, regression-based, interpretable machine learning, and unsupervised learning approaches are suggested as different possibilities to obtain new insights into the political landscape and applied to our data about the 2021 German federal election.



# Contents

Acknowledgement

Summary

Contributions of the thesis	i
Declaration of the author's specific contributions	iii
<b>1 Introduction</b>	<b>1</b>
<b>2 Background, literature, and aim of this work</b>	<b>5</b>
2.1 Background and General Literature Review . . . . .	5
2.1.1 Undecided Survey Participants and Consideration Set Sampling . . . . .	5
2.1.2 Set-valued Data under Ontic and Epistemic Imprecision . . . . .	8
2.1.3 Classical Election Research and Pre-Election Polls in Germany . . .	12
2.2 Aim of this Work . . . . .	14
<b>3 About the contributing material: Relations, summaries, and outlooks</b>	<b>15</b>
3.1 Forecasting with Set-Valued Data . . . . .	15
3.1.1 The Framework and Initial Approaches: Contribution 1 . . . . .	16
3.1.2 One Pseudo Bayesian Approach: Contribution 3 . . . . .	18
3.1.3 Cautious Superset-Learning: Contribution 5 . . . . .	20
3.2 Applying the Ontic and Epistemic Approaches . . . . .	22
3.2.1 Initial Ontic and Epistemic Ideas with Machine Learning: Contribu- tion 2 . . . . .	22
3.2.2 Application to the 2021 German Federal Election: Contribution 4 .	23
3.3 Structural Analysis with the Ontic View: Contribution 6 . . . . .	25
<b>4 Concluding remarks</b>	<b>27</b>
<b>Further references</b>	<b>29</b>
<b>Attached contributions</b>	<b>33</b>

---

# Contributions of the thesis

The Ph.D. project is composed of the following six chronologically ordered contributions that are referred to as *Contribution 1* to *Contribution 6* throughout the rest of this work:

1. Kreiss, D.; Augustin, T.: Undecided Voters as Set-Valued Information – Towards Forecasts Under Epistemic Imprecision. In: J. Davis and K. Tabia, editors, *International Conference on Scalable Uncertainty Management*, pp. 242-250. Springer Lecture Notes in Artificial Intelligence (2020) [https://doi.org/10.1007/978-3-030-58449-8\\_18](https://doi.org/10.1007/978-3-030-58449-8_18)
2. Kreiss, D.; Nalenz, M.; Augustin, T.: Undecided Voters as Set-Valued Information - Machine Learning Approaches under Complex Uncertainty. In: E. Huellermeier and S. Destercke, editors, *ECML/PKDD 2020 Tutorial and Workshop on Uncertainty in Machine Learning* (2020) <https://sites.google.com/view/wuml-2020/program?authuser=0> last access: July. 21, 2023
3. Kreiss, D; Schollmeyer, G. and Augustin, T.. Towards Improving Electoral Forecasting by Including Undecided Voters and Interval-Valued Prior Knowledge. In J. De Bock, A. Cano, E. Miranda, and S. Moral, editors, *Proceedings of the Twelfth International Symposium on Imprecise Probabilities: Theories and Applications*, pp. 201-209, Proceedings of Machine Learning Research (2021) <https://isipta21.sipta.org/papers.html>, last access: July. 21, 2023
4. Kreiss, D., Augustin, T.: Towards a Paradigmatic Shift in Pre-Election Polling Adequately Including Still Undecided Voters – Some Ideas Based on Set-Valued Data for the 2021 German Federal Election. *arXiv preprint* (2021) <https://doi.org/10.48550/arXiv.2109.12069>
5. Rodemann, J.; Kreiss, D.; Hüllermeier, E.; Augustin, T.: Levelwise Data Disambiguation by Cautious Superset Classification. In: Dupin de Saint-Cyr, F., Öztürk-Escoffier, M., Potyka, N., editors, *International Conference on Scalable Uncertainty Management*. Springer Lecture Notes in Artificial Intelligence (2022) [https://doi.org/10.1007/978-3-031-18843-5\\_18](https://doi.org/10.1007/978-3-031-18843-5_18)
6. Kreiss, D., Augustin, T.: Consideration Set Sampling to Analyze Undecided Respondents. Recently submitted. Preprint available on arXiv under: <https://doi.org/10.48550/arXiv.2307.14333>



# Declaration of the author's specific contributions

All contributing papers are the result of a fruitful collaboration with several co-authors. By separately referring to each of the papers, in the following the own contribution of the author of this dissertation is clarified:

**Contribution 1** The paper was written and drafted by Dominik Kreiss, who also implemented the applied methodology. The idea to factorize the likelihood as a framework for forecasting approaches in this setting was developed together, while the three approaches, in particular the one of the homogeneity assumption, were developed by Dominik Kreiss. Both authors did proofreading and revisions.

**Contribution 2** The different machine learning opportunities utilizing the set-valued data were deliberated by Dominik Kreiss, while the background on ontic and epistemic interpretations was developed together with Thomas Augustin. Malte Nalenz contributed to applying random forests for our election data and wrote the chapter in the methods section on random forests. All other parts were drafted and written by Dominik Kreiss. All authors contributed to the proofreading of the paper and revisions.

**Contribution 3** The main idea of compromising between two extremes by combining them in a Bayesian manner was developed by Dominik Kreiss, who also drafted and wrote the paper. Georg Schollmeyer contributed ideas on stochastic dominance and the mathematical properties of the results. All authors contributed to the proofreading and revisions of the paper.

**Contribution 4** The paper was drafted and written by Dominik Kreiss, who applied the approaches to the newly collected data. The data was preprocessed by Dominik Kreiss as well as he implemented the approaches. The ideas on coalitions were developed in joint discussion. Both authors did the proofreading.

**Contribution 5** In most parts, the paper was drafted and written by Julian Rodemann, who also contributed to the idea of twisting-the-tuning and implemented the main approach. Dominik Kreiss contributed the idea of a step-wise narrowing down procedure as

well as the application to the undecided voters and wrote the application section. The idea of how to formally narrow down supersets in section 3 was developed by Julian Rodemann and Dominik Kreiss together. The paper was made possible and improved by the comments of Eyke Hüllermeier. All authors contributed by proofreading the paper.

**Contribution 6** The core ideas of the paper were developed together and initially inspired by Thomas Augustin. Most parts of the paper were drafted and written together, while Dominik Kreiss implemented the methodology and wrote the applied section, and Thomas Augustin contributed the embedding in the theory of random sets. The specific applied approaches were suggested by Dominik Kreiss.

# Chapter 1

## Introduction

Choices and opinions are often surveyed as a crucial part of life. But as pondering between options is characteristic of human beings, not all participants can provide a direct answer as usually demanded in conventional surveys. This dissertation develops a framework we call *consideration set sampling* to collect their valuable information set-valued and provides approaches to utilize it. The structure is visualized in Figure 1.1, beginning with undecided participants in surveys about choice and showing how the set-valued representation can be utilized with two different interpretations.

As undecided survey participants cannot provide the usual single-valued answer, their position must be collected differently. The most common approach is the so-called *don't know option*, drastically reducing the information and often effectively excluding the individual from the analysis (Plass, 2018, p 1 ff.). This approach has the virtues of simplicity yet treats all undecided alike regardless of whether they are entirely indifferent or only pondering between specific options. This leads to a consequential loss of information, as individuals do tend to have a position even when not yet completely determined (Plass et al., 2015b; Oscarsson and Oskarson, 2019).

A further alternative in the opposite direction could ask for personal distributions or rankings over all alternatives. (e.g. (Fürnkranz et al., 2008)) While this is indeed interesting for some applications, for others, it is less practical as; first, it is more time intensive and complicated for participants, and second, not all individuals are capable of providing rankings or distributions.<sup>1</sup>

We thus argue that in some cases, most notably with pre-election polls, a set-valued representation is most suitable due to several advantages. If the undecided participants are provided the option to list all viable options they are still pondering between, they do not have to oversimplify their position. Furthermore, the set-valued response is rather easy and intuitive to provide. This is theoretically substantiated by a prominent theory of choice going back to (Tversky, 1972b) seeing choice as processes in stages stepwise excluding elements which intuitively leads to a set-valued representation of the undetermined individual at a given point in time. Additionally, implementing the set-valued option from a survey

---

<sup>1</sup>More background on this in Chapter 2.1.1

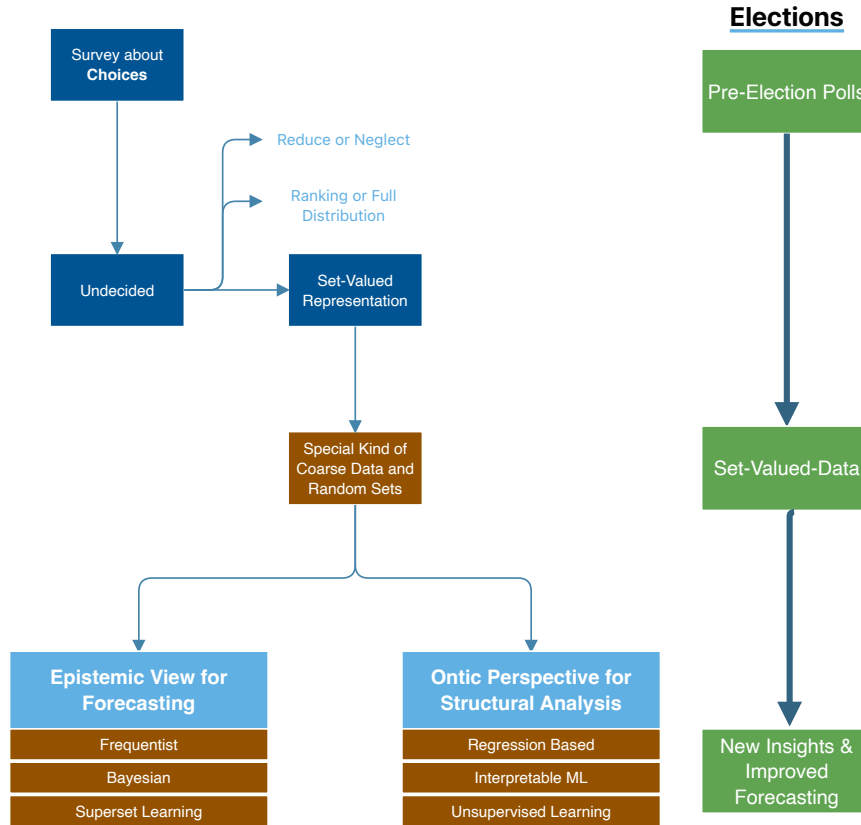


Figure 1.1: Contents of the dissertation with consideration set sampling as the central idea to collect the information of undecided survey participants about choices. The information can be utilized in two ways with approaches for forecasting and structural analysis. On the right side is the paradigmatic example of pre-election polls in the voting context.

perspective is relatively simple and cost-effective, as only one additional question is required to enable a larger portion of the population to be better represented. Throughout this work, we will call this set-valued representation at the given time *Consideration Set* following (Shocker et al., 1991) and (Oscarsson and Oskarsson, 2019).

These theoretical advantages were substantiated by our cooperation with the polling institute *Civey*. We conducted a pre-election poll together for the German federal election according to the current scientific standards and, for the first time, provided the participants with the option of consideration sets.

The newly collected set-valued representation of the undecided's position opens up opportunities for innovative analysis building on the theory of random sets. As a result of this, the state space is extended to the choice combinations, hence to the power set of the original options. So let us say that the original options partitioning the choice space are denoted as  $S = \{1, \dots, s\}$  with a potential choice  $Y$  from  $S$  and eventual realization  $\{Y = l\}$ . Then we observe the consideration set from  $\ell \in P(S)$ , with  $P(S)$  as the power set of the original options. The survey now provides us with a set  $\{\mathcal{Y} = \ell\}$  together with some covariates from  $\mathcal{X}$ . We usually draw a (random) sample of  $\{1, \dots, n\}$  individuals from the underlying population  $\Omega$  and obtain realizations  $(y_1, x_1), \dots, (y_i, x_i), \dots, (y_n, x_n)$  of the random elements.

$$(\mathcal{Y}_i, X_i) : \Omega \longrightarrow P(S) \times \mathcal{X} \quad (1.1)$$

$$i = 1, \dots, n,$$

$$\omega_i \mapsto (y_i, x_i). \quad (1.2)$$

For consideration set sampling, known sampling procedures can be used; only that now we observe from the possible combinations. Depending on the interpretation, two main venues arise in utilizing this new data on the power set level.

First, we can focus on  $\mathcal{Y}_1, \dots, \mathcal{Y}_n$  as a whole called *Ontic* in Figure 1.1. These representations of the true, current position can be used to conduct a structural analysis that adequately includes the undecided. This analysis is subsequently carried out on the state space of the power set or a content-based restriction reducing the set to  $\tilde{P}(S) \subset P(S)/\emptyset$ . Even though there is an implicit ordering of the power set due to the subset structure, different positions are seen as equal, which resolves this structure for our modeling approaches. For instance, if we look at the parties **SPD**, **Green**, and **Left** the consideration sets **SPD/Green** and **SPD/Green/Left** reflect two different positions of their own. This means as the underlying space of options  $S$  is categorical,  $P(S)$  and  $\tilde{P}(S)$  are categorical as well, making the random sets  $\mathcal{Y}_i$ ,  $i = 1, \dots, n$ , categorical random variables. This allows for adjusted transferability of methodology from regression modeling, interpretable machine learning, and unsupervised learning. Hereby, opportunities arise based on the rather simple transfer, now including this pivotal part of the population and exploiting their information for vivid structural analysis. As the undecided usually differ structurally from the decided, new insights about this group of particular interest are now possible, painting a more comprehensive picture of the choice and political landscape.

Second, we can try to derive statements on the single-value  $Y$  from  $S$  written as *Epistemic* in the left orange box of Figure 1.1 building on the epistemic view from (Couso and Dubois, 2014). For this, we focus on the eventual but not observable choice contained in the set-valued representation. The set hereby is interpreted as incomplete information and a coarse version of something initially precise. (e.g. (Couso and Dubois, 2014)) Due to the inherent complex uncertainty, it is necessary to either communicate the uncertainty in



interval-valued results or employ assumptions.<sup>2</sup> In Contribution 1, we develop a factorization of the underlying likelihood, which is the foundation for most of our approaches when predictions are concerned.

$$P(Y = l) = \sum_{(\ell, x) \in (P(S) \times \mathcal{X})} P(Y = l, \mathcal{Y} = \ell, X = x) = \quad (1.3)$$

$$\sum_{(\ell, x) \in (P(S) \times \mathcal{X})} \underbrace{P(Y = l | \mathcal{Y} = \ell, X = x)}_{\text{Transition Probabilities}} \cdot \underbrace{P(\mathcal{Y} = \ell | X = x)}_{\text{Consideration Sets}} \cdot \underbrace{P(X = x)}_{\text{Co-Variables}} \quad (1.4)$$

The second and third parts of the factorization can be directly estimated by the collected data, while the first contains the inherent complex uncertainty that can be estimated single- or interval-valued depending on assumptions made. As shown in Figure 1.1, both Bayesian and Frequentist approaches are possible and will be discussed in Chapter 3. We furthermore introduce one approach building on *Superset Learning* based on (Hüllermeier, 2014), taking a different look at the underlying problem.

In this thesis, we directly build on the ideas of (Plass et al., 2015b), who worked with the same concept and introduced initial approaches. We adopt some ideas like the Dempster Bounds and the distinction between ontic and epistemic views and developed several ourselves. Building on the setting, we initiate a framework for both interpretations of the set-valued data in Contributions 1 (Chapter 3.1.1) and 6 (Chapter 3.3) alongside several approaches and put them to the test with a first state-of-the-art pre-election poll using consideration set sampling for the 2021 German federal election. For structural analysis, we included interpretable and unsupervised machine learning and regularized regression in Contribution 6, while for forecasting, we explicitly addressed the tradeoff, introduced a likelihood factorization as a framework, and extended the ideas to machine learning in Contributions 1, 3, and 5.

Overall, this thesis contributes to improving different analyses by including the undecided in a set-valued manner. Throughout the thesis, the undecided voters in pre-election polls are our application. The early contributions work with artificially generated databased on the idea of Plass et al. (2015b) to contribute to a solution of a “chicken-egg dilemma” (Fink, 2018), resulting from the lack of surveys, including the set-valued question as well as missing methodology, providing practical approaches for such data. The latter ones show their applicability and virtues based on the data from the cooperation with Civey. For the better part, the methodological advances and suggestions could be transferrable to other applications of surveys, however not discussed in this thesis. In the following, we will discuss the methodological background and adjoining fields in chapter 2 and derive the aim of the thesis from this. Then we provide an overview of the contributions in chapter 3 before ending with the concluding remarks in chapter 4.

---

<sup>2</sup>The resulting tradeoff is referred to as Manski’s law of decreasing credibility (Manski, 2003, p. 1)

# Chapter 2

## Background, literature, and aim of this work

### 2.1 Background and General Literature Review

This work builds on several adjoining fields. First, we take a look at the theoretical backgrounds of choice processes and how the emerged uncertainty can be collected. Second, there is a rich theoretical background on interpreting random sets under the *epistemic* and *ontic* view and generally arguing with coarse data. For this, we mostly rely on the ideas by (Couso and Dubois, 2014; Couso et al., 2014) and exhibit some methodological approaches for prediction and structural analysis in a subsequent step. And third, we take a brief look at classical election research and pre-election polls. After addressing these fields in one subsection each, we will derive the aim of this work.

#### 2.1.1 Undecided Survey Participants and Consideration Set Sampling

Choices are linked to uncertainty. Several fields, particularly psychology and marketing research, have tried to understand the underlying processes and developed theories over the years. (Shocker et al., 1991; Stocchi et al., 2016) We invoke two lines of argumentation, why a set-valued representation of undecided voters in pre-election polls reflects the temporary position most suitable before looking at the concrete implementation advantages. First, human choice can be understood as a process that excludes non-eligible options, leading to a compelling subset for undecided voters. (e.g. Tversky (1972a,b); Oscarsson and Oskarsson (2019)) Second, individual choices can be seen as mappings from a latent underlying preference space to one alternative. Hence, undecided individuals can accurately be represented by a subset of the complete set of options.<sup>1</sup> Both argumentations show that a set-valued representation of undecided voters is intelligible and, therefore, preferable to

---

<sup>1</sup>Following the argumentation of random utility models e.g. (McFadden, 1981) or in (Tutz, 2011, ch. 8)

neglecting this valuable information as it is common in conventional surveys. There are several more theoretical frameworks in relation to *random* or *fixed utility model*, while we focus on the two ideas above.

The individual choice process can be seen in stages. This line of thought goes back to two publications by Tversky in 1972 (Tversky, 1972b,a), who introduced something he called *elimination models* next to the in that time usual *random* or *fixed utility model*. (Tversky, 1972a, p. 341) He argues for a probabilistic process of successive elimination by aspect and shows how modeling is possible in the framework of random utility models. (Tversky, 1972a, p. 341) The framework is also extended to ranking models. (Tversky, 1972a, p. 357)) He primarily focuses on the outcome and does not contemplate on a possible final set-valued representation, e.g., a stage of indifference, as we do in our work. Over time, numerous researchers picked up on these ideas with modeling and philosophical considerations. Shocker et al. (1991) argues in the marketing context of a process starting with the *Universal Set*, which is subsequently reduced to the *Awareness Set*, *Consideration Set* and *Choice Set* from which the final choice is made. Excluding not eligible options is a rather easy task while choosing among the compelling options is far more complicated. This process was observed empirically, for instance, by (Edenbrandt et al., 2022) with an eye-tracking study or in (Stocchi et al., 2016) to measure the empirical size of Consideration Sets in specific marketing applications. Recently, in political research, (Oscarsson and Oskarson, 2019) argue for the natural voting process to be in stages within which undecided hold a consideration set before making up their minds. Especially with political parties ranging over the entire political spectrum, most individuals can quite easily and instantly exclude some parties, while the choice amongst the remaining ones is by far more complex. These ideas suggest that characterizing the undecided voters' position by the set of their eligible options is natural and pragmatic. The information concerning the undecided is, at this moment, represented on an adequate level of coarseness, actually reflecting the ambiguity attached to the individual's choice process.

From the second point of view, we can also derive a set-valued representation of undecided voters by regarding voting as a mapping process from a latent preference space to the choice. Political positions are by far more complex than could be captured by a choice between a few alternatives but are rather a high-dimensional, latent, and only partially measurable entity. More concretely, it can be understood as positions in an individually specific and multidimensional preference space reflecting the preferences on any number of political issues. (For more thoughts on this, see the background on random utility models in (Tutz, 2011, ch. 8) or the construction by (Manski, 1977, p 231 ff.)) These preferences might be subconscious as they might be well-defined; we merely assume they exist. In that sense, voting can be seen as compromising on one element of the sparse set of alternatives that resembles the positions as closely as possible. As the offers (or parties) match different individuals differently well, it is only natural that not all individuals can determine one party right away, at a given point in time, before the election. But as the list of alternatives is exhaustive, covering all the potential options, mapping the preferences to a combination of options in the form of a set should be possible. The current position can then be represented by the partition of the complete options on which the political preferences can be

mapped. In other words, it is easier to capture the own complex political position by a set of parties than a single one. Hence, also, from this point of view, providing a set-valued statement about one's political position is easier for an undecided voter.

On a further side note, the political position of a voter might as well be set-valued on election day but is then forced into a single element. If we attempt to research connections between political positions and socioeconomic variables, one could argue that the set-valued representation, even on election day, is better suited than the single-valued one.

Hence, when it comes to concrete implementation, set-valued representation of undecided individuals has various beneficial properties compared to the alternatives. In contrast to the conventional approaches, the undecided can now adequately present their position and no longer have to drop out or convey incorrect information. Furthermore, due to the intuitive set-valued representation of positions, the undecided can provide their position straightforwardly. This further distinguishes the set-valued approach from other attempts, like directly providing a distribution or ranking over the parties, which would be way more complicated or impossible for some participants.

As argued in the introduction, the alternatives would be either a loss of information with the *don't know* option or the attempt for ranking or full distributions. Partial ranking can be a good alternative in some cases. Still, in the case of being undecided about the choice, this would directly lead to set-valued data again, and the ranking of the parties not considered is usually not really of interest. Asking individuals about a probability distribution over all alternatives, e.g., 80% for this party and 20% for that party, would theoretically carry the most information but is, however, impractical. Individuals perceive probabilities very differently, and non-response should increase. (e.g. Gallistel et al. (2014)) Overall, the set-valued representation still carries most of the information while it has the considerable advantage of simplicity.

## Survey Implementation

From a survey point of view, the set-valued data can be collected in different ways. Starting with the established surveys, the least effort would include follow-up questions directed at all participants choosing the *don't know* option. One alternative is to immediately ask all participants to choose all parties they are pondering between and assume that decided individuals choose one. And lastly, one could construct a survey in two stages, first asking about indifference directly and then either asking about one or the set of parties. All of those approaches are possible and yield some advantages and disadvantages. As the third option in two stages enables a strict separation and provides some additional information about this aspect, we opted for this one in cooperation with Civey. We thus conducted a new separate survey that explicitly focusses on the undecided more broadly. For this, a sample size of around 25000 observations is drawn in the first stage resulting in roughly 5000 observations in the stratified and weighted sample, together with 11 ordinal and categorical covariates. We focussed our work primarily on a sample two months before the election.

Overall, the collection of all eligible options for an undecided voter in a pre-election poll

in a set-valued way seems to be the adequate approach to regard their positions. In the case of undecided voters, we could observe the benefits in the dataset provided by Civey.

### 2.1.2 Set-valued Data under Ontic and Epistemic Imprecision

We now take a more thorough look at the set-valued representation  $\mathcal{Y}_i, \dots, \mathcal{Y}_n$  characterizing the undecided voters' current position before focussing on the specific approaches. To recap from the introduction, we can, on the one hand, focus on the indecision itself at the moment of the survey, which is accurately represented by the set as a whole, or on the other focus on the choice outcome, in which case only incomplete information is provided. The **first** interpretation is called *ontic* or *conjunctive* (following Couso et al. (2014)) under which the set is a non-reducible entity, picturing the precise position of the individual at this given point in time. The **second**, called *epistemic* or *disjunctive* (following Couso et al. (2014)), sees the set as a collection of items containing the one the individual ends up choosing. Hence, the ontic view sees the set as a precise representation of something naturally imprecise, while the other sees it as an imprecise representation of something precise. To provide a different example: For the question: "What classes does a student take a given semester" a set is obviously the precise answer. However, if we are interested in the favorite class, the same set only yields incomplete information, even though the true answer lies within that set. Both interpretations are justified and focus on different means of analysis. In both cases, we are in the finite case with our discrete data structures. We will depict either from an applied perspective in two subparts but first briefly connect them to random sets generally. For this, we will broadly follow the notation and embedding in the theory of random sets from (Couso et al., 2014) with the formulation of *random conjunctive sets* for the ontic and *ill-known random variables* for the epistemic (Couso and Dubois, 2014, p. 1504 ff.) and build on remarks by (Plass, 2018) on coarse categorical data. The options completely partitioning the outcome space are denoted as  $S = \{1, \dots, s\}$ . The random conjunctive set is here no more than a generalized random variable consisting of several elements of  $S$  leading to the power set  $P(S)$  or restriction  $\tilde{P}(S)$ . The eventual realization is not of concern, and probability measures can effectively be constructed in a known way.

The ill-known random variable under epistemic interpretation is, however in that sense, hidden within the set-valued observation, for which holds  $l \in \{\mathcal{Y} = \ell\}$  meaning that the true value of interest is within the set-valued observations. Denoting this from the underlying space  $\Omega$  from the sample, the ontic can be written as a mapping  $f : \Omega \rightarrow P(S)$  while for the epistemic, we are interested in the mapping  $f_{precise} : \Omega \rightarrow S$  representing the disjunctive set of mappings. (Plass, 2018, p. 9).  $\{\mathcal{Y} = \ell\}$  is here written over  $\{\omega \in \Omega, \mathcal{Y}(\omega) = \ell\}$ . Probability measures, as well as possibility measures, can be constructed in this setting with examples in (Couso and Dubois, 2014, ch. 6).

In both cases, we rely on an assumed to-be i.i.d. sample (or weighted i.i.d. sample) with observations  $(y_1, x_1), \dots, (y_i, x_i), \dots, (y_n, x_n)$  as written in the introduction. Complex sampling designs (e.g., Skinner and Wakefield (2017)) would theoretically be feasible as well, but we focus on the i.i.d. case in our applications.

### Ontic Applications and Means of Analysis

From an applied perspective, the underlying analysis space becomes the original options' (reduced) power set. If, as in our case, the original set is finite and not ordered, we can interpret the random variable in a way to satisfies the same mathematical properties from a modeling perspective as the original set. This allows for broad transferability of methodology while one has to regard two points.

First, one must determine whether underlying ordering, nesting, or other hierarchical structures exist on the new power set. This depends on the question at hand and the situation as written in the introduction. For instance, in our example of undecided voters, the group undecided between SPD/Green constitutes, politically speaking, a fundamentally different group than those hesitant between SPD/Green/Left. Even though the apparent nesting, there is no further structure here from a modeling point of view, and both groups can, and indeed should, be treated independently from one another. This is different in the example of consumer choice, where there is natural structuring if we compare the groups pondering between two brands or the two brands and one further. Depending on the underlying goal, such ordinal (weak) structures can be included in the modeling process with approaches on ordinal regression. (e.g. Gutiérrez et al. (2016))

Second, the number of possible groups vastly expands with the transformation to the power set. From a practical applied standpoint, this raises issues with small samples in some groups and perfect separation. Especially in the context of regression approaches, regularization seems advisable to get to a feasible number of estimated coefficients as conducted in Contributions 4 and 6. In some cases, groups can also be reduced content dependent.

Established approaches can be transferred to these new questions for interesting insights. To conduct a structural analysis, we have several approaches at our disposal. These can be divided into regression-based ones, the field of interpretable machine learning, and unsupervised or content-related models. How these approaches play out together with advantages and disadvantages is discussed in Contributions 2,4 and 6. Overall, new insights are possible, providing a more complete picture of the political landscape.

### Epistemic Applications and Means of Analysis

Approaches building on the epistemic view of random sets try to contend with the inherent complex uncertainty directly. Hereby, the goal is often to model the underlying coarsening process to some extent to obtain meaningful statements about the true element contained in the set. (Couso and Dubois, 2014, p 1503 ff.) However, as only incomplete information about the eventual choice is provided, one has to either reflect the inherent uncertainty in interval-valued results or make strong assumptions. This is a classic example of *Manski's Law* of decreasing credibility (Manski, 2003, p. 1), as one has to weigh the credibility of the results with the strength of the assumption. This is indeed difficult in applied research, as results do have to be concise enough to be meaningful without neglecting the complex uncertainty attached. Thus, several approaches can address this tradeoff from wide bounds

to single-valued results based on strong, untestable assumptions.

Many approaches are connected to the underlying problem here, stretching from classical statistics to machine learning. As we cannot cover the entire field, we try to highlight some of the important developments from both concise and interval-valued ideas.

Starting with single-valued modeling in classical statistics, we have to rely on strong assumptions and/or further information. Next to not really meaningful assumptions like maximum entropy or best guesses, more sophisticated approaches were developed. Most prominently, the idea of noninformativeness of the coarsening process was developed by (Heitjan and Rubin, 1991), also called *Coarsening at Random*. The coarsening process is hereby assumed to be conditionally random, which can be understood as a more general concept than the missing at random, which means in our notation:

$$\forall \ell : P(\mathcal{Y} = \ell | Y = \ell) = P(\mathcal{Y} = \ell | Y = \ell') \quad \forall \ell, \ell' \in \mathcal{Y} \quad (2.1)$$

Hence, by fixing the coarsening process, modeling can derive a single-valued prediction but only at the cost of the strong underlying assumption. The assumption is similar to our homogeneity assumption from Contribution 1, which is made about the outcome rather than the coarsening process as:

$$\hat{P}(Y = \ell | \mathcal{Y} = \ell, X = x) = \frac{\hat{P}(Y = \ell | X = x, I_d = 1)}{\sum_{a \in \ell} \hat{P}(Y = a | X = x, I_d = 1)} \quad (2.2)$$

with  $I_d = 1$  as the indicator variable for being decided. One major problem with any attempt to completely disambiguate the set is that the underlying assumption is not easy to verify due to the inherent complex uncertainty. This especially holds if, as in our application to undecided voters, the political landscape shifts change the forecasting situation every time.

Interval-valued approaches in the sense of interval-probabilities (Weichselberger, 2000; Walley, 2000), on the other hand, aim to reflect (some) of the inherent uncertainty within the results. This can still be seen as a disambiguation of the set to some extent, depending on the approach. There is a rich theoretical background on how to benefit from an interval-valued representation of the inherent uncertainty (e.g. Augustin et al. (2014)), while again, we only give some examples. The most reliable but least concise approach is based on the work (Dempster, 1967)'s handling of set-valued mappings, constructing best and worst-case scenarios for each group, respectively. These so-called *Dempster Bounds* are constructed in a way reflecting the entire ambiguity without relying on further information like covariates. They provide a range of individuals choosing the parties in  $Y$  in a manner that, without the survey error, the true forecast is certainly contained by shifting the probability mass to the extremes. This can be written for all  $\ell \in P(S)$  as:

$$p_{lower}(Y \in \ell) = \sum_{\ell' \subseteq \ell} p(\mathcal{Y} = \ell'), \quad (2.3)$$

$$p_{upper}(Y \in \ell) = \sum_{\ell' \cap \ell \neq \emptyset} p(\mathcal{Y} = \ell'). \quad (2.4)$$

With this, all elements of the set of all probabilities are considered as potential transition probabilities, which means that  $P(Y = l | \mathcal{Y} = \ell, X = x)$  from equation 1.3 is set to the extreme values independently from the covariates. Application of the Dempster Bounds is also discussed in Contributions 1 and 4.

Ideas on applying partial identification based on considerations of (Manski, 2003) are examined in (Schollmeyer and Augustin, 2015), and related works to categorical data and likelihood inference under epistemic imprecision are discussed in (Plass et al., 2015a) and (Plass et al., 2019). Compromises between point-valued results and wide bounds are possible and discussed in Contributions 3 and 5 and also with (Plass et al., 2015a, 2019) even though from a different angle.

Dealing with incomplete, coarse, or otherwise distorted data is also a big topic in machine learning research. There is a huge variety of different approaches as well as vocabulary differences, and we attempt to highlight some approaches related to our research here. Most of them fall under the umbrella term *Weakly Supervised Learning* used by Zhou (2017). We must distinguish two similar but different angles. In our application in Contributions 1-5, we are interested in explicit data disambiguation, which uses the set in an epistemic manner and outputs some (imprecise) results. However, most of the machine learning literature tries to learn a function that performs well on unseen instances, which in some way benefits from the weakly labeled data. The two things are obviously connected, but often no explicit data disambiguation has to be conducted to learn this function.

Prominent terms and approaches are *Partial Label Learning* (Cour et al., 2011), *Multi-Label Learning* (Zhang and Zhou, 2014), and *Superset Learning* (Hüllermeier, 2014). They are all connected and rely on the same data structure, with the true label being amongst each candidate set and aiming to learn a single-valued model. Their algorithmic groundworks are, however, different. Attempts to communicate the uncertainty in interval-valued results in Machine Learning are sparse but can, for example, be found in extensions of (Zaffalon, 2002) ideas on the naive credal classifier.

There are also several fields related but different from our setting. In semi-supervised learning, both labeled and unlabeled data are used to train one model. (Cour et al., 2011, p.1502) Multi-instance learning potentially assigns a number of labels to one observation and hence closely relates to ontic approaches. (Foulds and Frank, 2010)

Next to the inherent uncertainty, one should not neglect the sampling uncertainty attached to each approach. For reliable forecasting, one should hence communicate the inherent complex uncertainty of the random set next to the potential total survey error (Groves and Lyberg, 2010) consisting for example, of nonresponse error, corrections, and stochastic uncertainty of the corresponding approaches. We are not disregarding other sources of uncertainty and do advocate confidence or credibility intervals, but our main focus is on the uncertainty shown in the response with multiple values.

Overall, several methods developed can be connected to our forecasting setting. A bigger proportion of those is in line with our factorization of the likelihood, namely those dealing with direct data disambiguation. Dealing with inherent complex uncertainty is an ongoing



field; further advances will surely follow in the next years.

### 2.1.3 Classical Election Research and Pre-Election Polls in Germany

In this dissertation, we primarily focus on so-called multiparty voting systems, which in contrast to the majority voting system of the USA, has multiple viable parties as part of parliament. It could be argued in either case to include the option *not voting* and *other parties*, which would extend the state space in the majority voting system as well, but the attached political questions do come more naturally in multiparty systems. In our works, we furthermore primarily address the German voting system, as even though most European countries do have multi-party systems, specific voting modalities differ severely. Transfers of methodology to different systems are straightforward, while for substance matter analysis, one usually has to be aware of each voting system's peculiarities.

In Germany, for example, coalitions play a major role, as they have been necessary to constitute a government in the past decades. Hence, a central question is about a specific coalition reaching the majority rather than a single party. This also suggests conducting an analysis on party combinations, which is effectively identical to our extension of the state space. Furthermore, Germany has a two-vote system, one for a specific representative and one for the party, which could lead to tactical vote splitting (Pappi and Thurner, 2002). More specifically, in Germany, there are currently six relevant parties likely to surpass the 5% hurdle (typically) necessary to win seats in the parliament. Those parties are: The Left, Green Party, SPD, CDU/CSU, FDP, AfD. Overall, Germany has a rather complex, proportional voting, multi-party election system, in which we only focus on the proportion of seats won in the parliament, which almost certainly will be split between those parties.<sup>2</sup>

Depending on the point in time, the country, the specific election, and how the question in the survey is framed, the proportion of still undecided voters differs.<sup>3</sup> Hence, it is difficult to provide wide-ranging and comprehensive numbers on this issue, and even though a trend towards increasing numbers of undecided individuals seems noticeable, it is hard to prove that conclusively. On top of this, there is the issue that many numbers used by newspapers are made available by private institutes that often do not disclose their procedures protecting their methodology, making it hard to validate the results. However, a relevant proportion of voters is still undecided before elections. This was also substantiated by our data from the cooperation with Civey used in Contributions 4, 5, and 6 with for example 22.8% of participants being still undecided two months before the 2021 German

<sup>2</sup>For more information about the German voting system, see: <https://www.bundeswahlleiter.de/bundestagswahlen/2021/informationen-waehler/wahlssystem.html>, last visited July. 21, 2023

<sup>3</sup>See for example differences in German newspapers: <https://www.faz.net/aktuell/politik/bundestagswahl/bundestagswahl-noch-nie-so-viele-unentschlossene-kurz-vor-wahl-17536559.html>; <https://www.sueddeutsche.de/politik/warum-sind-sie-noch-unentschlossen-wen-sie-waehlen-leserdiskussion-1.5413574>; <https://www.spiegel.de/politik/deutschland/superwahljahr-die-lage-wann-machen-sie-ihr-kreuz-a-47f7f84a-428d-47d5-a395-43f5c041320d> last visited: July. 26, 2023

federal election.

Pre-election polls in Germany usually receive a lot of media attention and are used for forecasting and structural analysis. Noticeably, the attached uncertainty is often not or only scarcely communicated. These precise estimates in structural analysis and forecasts might embellish the degree to which the data is conclusive. A good example of this is the popular voter migration analysis<sup>4</sup> based on potentially non-random samples and often assumption-heavy methods utilizing MCMC processes. (e.g. Klima et al. (2017)) In many applied cases of structural analysis, as with the voter migration analysis, the ground truth is not known or shifting, which makes claimed statements hard to falsify.

Amongst others, there are two important issues to accurately forecast elections: The first is the potentially distorted sample by the poll, and the second is undecided voters.

The potential sources for bias in polling data are a known problem. (Bauer, 2014, 2016; Shirani-Mehr et al., 2018)<sup>5</sup> In practice, complicated weighting schemes are usually established to weaken or resolve the biases of the data. (Richter et al., 2022) Civey, as one example, uses post-stratification in the first step drawing a huge number of individuals and subsampling later on, to strive towards a representative sample. Voluntary surveys encounter issues with item and unit non-response (e.g. Spiess (2010); Rubin (2004); Sakshaug et al. (2010)). Especially *Missing Not at Random*, hence a direct dependence of the missing mechanism on the variable of interest, poses problems. An example in the context of pre-election polls is that individuals with extreme political positions or less socially acceptable tend to participate rarer. (Winkler et al., 2006) This issue can not be solved by our consideration set sampling, but it might be attenuated depending on the situation.

The problem with the undecided stems from the conventional *don't know option* used, which has severe implications for the forecasting results. Hereby, the individuals are often disregarded unattended for the forecasting process, which is effectively an implicit *Missing Completely at Random* assumption. The forecast is hereby only based on the decided, which assumes the undecided to not differ in their overall distribution towards the parties. However, the dataset provided by Civey suggested severe differences between those groups concerning their structural properties. Attuning this with weighting is certainly an option, but consideration set modeling should provide more detailed information in this setting. This implicit *Missing Completely at Random* assumption can hence be seen as the benchmark for our forecasting approaches aim to beat, as argued for in Contribution 1.

Including external sources of information is one way to attune to the two problems discussed before (e.g. Graefe (2019); Lewis-Beck and Dassonneville (2015)). We pick up on this idea in Contribution 3. A general problem here is how to communicate uncertainty which stems from different sources of information.

Structural analysis of the political landscape is an important part of electoral research as

---

<sup>4</sup>for one example see: <https://www.spiegel.de/politik/deutschland/bundestagswahl-2021-ergebnis-der-waehlerwanderung-im-detail-a-cebdad34-f727-4f07-b5d1-fe39d1245275>, last visited July. 26, 2023

<sup>5</sup>For more on the *Total Survey Error* and general problems with voluntary survey see Groves and Lyberg (2010)

well. Even with less media coverage compared to forecasting, the findings are interesting for the involved parties and political science. The field can be divided into several parts discussed in Contribution 6. The main goal is usually to find connections between socio-economic variables and specific positions. Examples can be found in (Mauerer et al., 2015; Thurner, 2000; Tutz, 2011).

## 2.2 Aim of this Work

This Ph.D.-thesis aims to provide a sound methodological framework to utilize data collected from undecided responses set-valued. In particular, we promote...

- ...consideration set sampling as a way to accurately represent undecided respondents in pre-election polls
- ...the epistemic and ontic view of the random set as equally possible to utilize the data situation here
- ...methodology for forecasting with a factorized likelihood explicitly addressing the tradeoff between accuracy and conciseness and extensions into the machine learning setting
- ...methodology for structural analysis by formalizing the observations of the consideration set sampling as conjunctive random sets reflecting equally viable positions

# Chapter 3

## About the contributing material: Relations, summaries, and outlooks

This cumulative dissertation explores ways to utilize set-valued data to conduct (imprecise) forecasting or structural analysis. In the following, the author's publications contained in the Ph.D. project are summarized. The main findings are presented together with a critical reflection and remarks on potential further research. As the contributions are naturally linked, there are cross-references throughout. We first discuss the papers on forecasting (e.g., Contribution 1, 3, and 5). Then we cover Contributions 2 and 4, which both take a rather applied perspective arguing how the ontic and the epistemic view are both feasible as discussed in Section 2.1.2. Finally, Contribution 6 provides a rather comprehensive overview of opportunities for structural analysis under the ontic view.

### 3.1 Forecasting with Set-Valued Data

In the following contributions, a framework to utilize set-valued data for predictions (in the election setting called forecasting) is developed, and some new approaches are introduced. We go through the contributions separately and reflect on them each. We want to derive statements about the true element  $\{Y = l\}$  contained in the consideration set  $\{\mathcal{Y} = \ell\}$  either by deploying assumptions or reflecting the results in intervals. We approached the problem from different angles. The approaches can be distinguished by several traits. First, whether they use covariates; second, if they are single or interval-valued; third, if they can be rather assigned to classical statistics or machine learning and fourth if they use external information. Most of them are based on the factorization of the likelihood in our first paper. More modeling ideas are possible while we reflect on those in the corresponding outlook sections of the individual papers.

### 3.1.1 The Framework and Initial Approaches: Contribution 1

Kreiss, D.; Augustin, T.: Undecided Voters as Set-Valued Information – Towards Forecasts Under Epistemic Imprecision. In: J. Davis and K. Tabia, editors, *International Conference on Scalable Uncertainty Management*, pp. 242-250. Springer Lecture Notes in Artificial Intelligence (2020)  
[https://doi.org/10.1007/978-3-030-58449-8\\_18](https://doi.org/10.1007/978-3-030-58449-8_18)

This first paper shows different avenues to utilize set-valued data for election forecasting. The idea of the consideration set sampling building on (Plass et al., 2015b; Oscarsson and Oskarsson, 2019) is developed as a contrast to the traditional handling of undecided participants in pre-election polls. We mostly argue from the loss of information induced by the common neglect, which is used as the main comparison for the developed approaches. From a forecasting view, the position of the undecided at this given point in time can be seen as partial information in the framework of epistemic uncertainty. (Couso and Dubois, 2014).

At the point of this initial paper, no state-of-the-art survey was conducted that enabled undecided participants to provide all viable options, and no sound methodology tailored to this specific situation was developed. This led to the, in the introduction mentioned, "chicken-egg-dilemma" (Fink, 2018), as methodology would be necessary to justify conducting such a survey. We hence worked with artificially constructed data to lay a methodological foundation to show that collecting this data has advantages.

The modeling ideas introduced are all built on a factorized likelihood that can be seen as a foundation of forecasting with set-valued data generally. To recall from the introduction in Equation 1.3, the likelihood is hereby divided in the transition probabilities  $P(Y = l | \mathcal{Y} = \ell, X = x)$  reflecting the epistemic problem and the terms  $P(\mathcal{Y} = \ell | X = x)$  and  $P(X = x)$  which can be estimated from the data without further assumptions.

From there on, the argumentation hence focused on ways to estimate the transition probabilities. At the given point in time, we cannot observe the eventual choice of the undecided and hence have to fall back on either 1) assumptions, 2) further information, or 3) and representation of the uncertainty in an interval-valued manner. Combinations of the three are possible as well. This leads directly to Manski's law of decreasing credibility. (Manski, 2003, p. 1 ff.) We hence have to weigh the strength of the justifiability of the assumption with the conciseness of the results. In practice, this can be difficult, as on the one hand, results have to be at least concise to a certain level to be meaningful, and on the other, we do not want to provide unreliable results.

We provide three different approaches covering the extremes of this tradeoff, to show this dilemma and provide potential starting points.

The **first** approach is the most cautious one possible, only showing what is certain in the data.<sup>1</sup> The idea based on (Dempster, 1967) was already discussed in Section 2.1.2.

<sup>1</sup>Only focusing on the uncertainty induced by the undecided and not on stochastic and survey errors,

This leads to the best and worst-case scenario for each party. Hence, let us say from the perspective of the Green Party, this means, for the best case, that everyone considering voting for them, will actually do so, hence  $\sum_{\ell' \cap \ell \neq \emptyset} p(\mathcal{Y} = \ell')$ . This sets the transition probabilities towards the Green Party of all consideration sets containing the Green Party to one. For the lower bound respectively, nobody will in the end vote for them, constituting the worst case hence:  $\sum_{\ell' \subseteq \ell} p(\mathcal{Y} = \ell')$ . Depending on the data and in our application, the intervals are wide. But they are coherent and reliant intervals.

The **second** approach manifests the other extreme; point-valued identification. For this, we introduce the questionable, but at least somewhat plausible, assumption that, given the covariates, the undecided choose like the decided with the consideration set as the restriction of the possible outcomes. This, from now on, called *homogeneity assumption*, hence estimates the distribution of the decided  $\hat{P}(Y_i = l | X_i = x_i, I_d = 1)$  with  $I_d$  as the indicator function for being decided from the data and uses it for estimation of the transition probabilities:

$$\hat{P}(Y = l | \mathcal{Y} = \ell, X = x) = \frac{\hat{P}(Y = l | X = x, I_d = 1)}{\sum_{a \in \ell} \hat{P}(Y = a | X = x, I_d = 1)} \quad (3.1)$$

This leads to point-valued identification with the cost of the strong assumption. The distribution can be estimated with a multinomial logit model like in this paper or a random forest as used in Contribution 2. For the overall estimation, we predict the outcome of the undecided with the consideration sets as restrictions of the possible values. The paper argues that, despite the potential dubiousness of the assumption, this is still preferable to neglecting the undecided overall with an effectively missing completely at random assumption as argued for in Section 2.1.3.

The least discussed **third** approach can be understood as an outlook to a potential future overcoming the chicken-egg dilemma. We argue that if appropriate data would have been continuously collected, we could estimate transition probabilities from past elections. With  $+$  denoting the previous term, we could then plug in the estimated transition probabilities  $\hat{P}(Y^+ = k | \mathcal{Y}^+ = \mathcal{k}, X^+ = x^+)$ . A point-valued identification would come at the cost of the assumption that the undecided do vote the same way as they did last term. Robustifying the two point-valued approaches could be attempted with neighborhood models. Compromises between the extremes shown in this paper are discussed later in Contributions 3 and 5.

We applied the approaches with data from the *German Longitudinal Election Study* with artificially constructed undecided voters and compared them to the neglecting of the undecided overall. The Dempster Bounds did prove to be rather wide, while the homogeneity assumption delivered seemingly plausible results. But it is important to know that the assumption is not testable here, and also plausible results can be wrong.

---

which should be communicated as well.

### Comments and Outlook

The paper gives a good first impression of the underlying problem and marks out possible approaches by providing one for the extremes each.

The approaches could have been evaluated better by contemplating on potential bias induced, like in Section 2.1.3, especially in contrast to the implicit *missing completely at random* assumption made by neglecting the undecided. Also, quantifying the magnitude of change by neglecting this relevant proportion of the population (32.7 %) in this dataset is interesting to communicate on a higher level. This change hugely relies on whether the undecided differ structurally from the decided. In our survey in cooperation with Civey, we could observe that they do differ concerning nearly every variable. Developing a rule of thumb here could be interesting for further research. (With such a proportion of undecided that differ to a certain extent, one could expect a bias of...)

Reflections on compromises between the extremes are only held short, but we will pick up on those in Contributions 3 and 5. Also, there are many more approaches possible that could have been included in the paper as well. Loss-based approaches to obtain the estimates with the homogeneity assumption are possible and will be conducted in Contribution 2. Including further sources of information is contemplated on in Contribution 3.

This paper solely focuses on the inherent complex uncertainty and disregards the sampling error and stochastic fluctuations. But as discussed in Section 2.1.3, this does play a relevant role as well. In further works, this kind of uncertainty should be done justice as well by providing confidence intervals and including a weighting procedure. Creating a formally sound framework for this takes work but is desirable nevertheless.

As already written in the paper, this opens up a lot of avenues for further research and provides a foundation rather than already resolving all issues.

#### 3.1.2 One Pseudo Bayesian Approach: Contribution 3

Kreiss, D; Schollmeyer, G. and Augustin, T.. Towards Improving Electoral Forecasting by Including Undecided Voters and Interval-Valued Prior Knowledge. In J. De Bock, A. Cano, E. Miranda, and S. Moral, editors, *Proceedings of the Twelfth International Symposium on Imprecise Probabilities: Theories and Applications*, pp. 201-209, Proceedings of Machine Learning Research (2021) <https://isipta21.sipta.org/papers.html>, last access: July. 20th, 2023

This paper contemplates the inclusion of reliable auxiliary information about transition probabilities in the forecasting process. It is common to involve additional sources of information in election forecasting for calibration and stabilization (Graefe, 2019; Lewis-Beck and Dassonneville, 2015) as discussed in Chapter 2.1.3.

One issue with auxiliary information in the voting context is its reliability. As elections are usually held only every few years and the political landscape changes, it is furthermore

challenging to validate such information because of different settings between the years. (Meaning: Only if someone was right before does that lead to lasting credibility?) We hence advocate the idea of asking experts to only provide coarse information that they are sure of. This could, for example, take the form: At least 20% of individuals undecided between the Green and Left Party will vote for the Green and at most 90%. We denote the information with the interval:

$$P(Y = l | \mathcal{Y} = \ell) = [pr_{l,\ell}^{lower} ; pr_{l,\ell}^{upper}]$$

This means that we assume that we have information about every consideration set, with the interval covering  $[0, 1]$  for the least certain or missing prior knowledge.

These intervals can be directly inserted as transition probabilities, already narrowing down the Dempster Bounds of Contribution 1, but results could still be too wide ranges in the forecasts to contain meaningful information. In our primary approach, we hence choose a (pseudo) Bayesian way to find a compromise between a concise and a wide estimate resulting from the experts' intervals. This is motivated by practical applications for which a degree of conciseness is necessary, which is not provided with the experts' intervals in the sense of the tradeoff discussed in Contribution 1. In this case, we suggest narrowing the interval, taking steps in the most plausible direction, and constricting the intervals at the cost of increasingly relying on the homogeneity assumption. The homogeneity assumption from Contribution 3.1.1 in Equation 3.1 is used to calculate estimates on an individual level with a working model that can be understood as a deterministic function from the covariates to the predictions, preserving the assumed i.i.d. structure of the sample.

We thus suggest using these single-valued predictions as data for Bayesian updating of the upper and lower bound with two models, respectively, and calculating a posteriori distribution. Both bounds are hence drawn towards the estimate of the homogeneity assumption, which narrows the interval. It is possible to cover the entire ground of the interval by only using the upper and lower bound due to first-order stochastic dominance. The variance parameter of the prior information can be set manually and regulates the degree of narrowing by the model. As one possible distribution assumption, we suggest a beta distribution over the parameters  $\alpha$  and  $\beta$ , where we rely on the parametrization:  $f_X(x : \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$ . The prior information about the values  $\alpha$  and  $\beta$  can be provided, for example, by (truncated) normal distributions with the convenient variance parameter that directly determines the tradeoff between the extremes. As we do not have a conjugated prior here, we calculate the posteriori over an MCMC process. Other approaches relying on Dirichlet Processes without distributional assumptions would also be possible.

We conducted a simulation study with individuals undecided between three parties, coarse expert knowledge on the transition probabilities, and single-valued predictions from the homogeneity assumption. We insert the expert knowledge in the form of a truncated normal distribution, for which the fixed variance parameter determines the conciseness of the end results and the data from the homogeneity assumption with a beta distribution. The compromise is determined with an MCMC process, narrowing down the initial wide



bounds. Results showed that we indeed can narrow the bounds stepwise by changing the variance parameter.

### Comments and Outlook

The main problem with combining two sources of differently reliable information is how to interpret the compromise. With our approach, we end up in between reliable and concise but unreliable information. We can show this desirable trait due to first-order stochastic dominance in our approach. However, there is no inherent metric to measure how reliable the end results are. Hence, the approach is intuitively feasible but not really possible to validate. The danger of "as concise as absolutely necessary" is that we ultimately might provide unreliable forecasts without proper uncertainty communication. Still, one can argue that this would be better than neglecting the undecided overall, as already mentioned in Chapter 3.1.1.

The tradeoff is a content-driven information fusion idea with one particular application. It would have been interesting to develop a more comprehensive framework for combining different sources of information in the election process. Combining several different interval-valued sources of information is interesting as well. Here we could work with intersections and unions of the intervals provided. Hierarchical models could also be applied when combining multiple sources of information. Ideas to extend the approach to include expert opinion and information of past elections as discussed in the third approach of Contribution 1 is interesting. Next to (pseudo) Bayesian approaches, there are other ideas to combine information as well. Most prominently, Dempster-Shafer's theory of evidence could be used for information fusion in this setting. (e.g. Denoeux (2016))

### 3.1.3 Cautious Superset-Learning: Contribution 5

Rodemann, J.; Kreiss, D.; Hüllermeier, E.; Augustin, T.: Levelwise Data Disambiguation by Cautious Superset Classification. In: Dupin de Saint-Cyr, F., Öztürk-Escoffier, M., Potyka, N., editors, *International Conference on Scalable Uncertainty Management*. Springer Lecture Notes in Artificial Intelligence (2022)  
[https://doi.org/10.1007/978-3-031-18843-5\\_18](https://doi.org/10.1007/978-3-031-18843-5_18)

This paper suggests a specific machine-learning approach to obtain data disambiguation at different levels of coarseness based on both single-valued and set-valued observations. For this, we go through all possible *instantiations* of the set-valued data; hence the values the set can take, which leads to many datasets only containing single-valued observations. We build on the framework of *Optimistic Superset Learning* (OSL) (Hüllermeier, 2014), which originally searches to identify the *most plausible* instantiation contained in the set-valued representation (here called superset). A specific instantiation of all coarse data is

considered more plausible if a given learner makes fewer mistakes or if the structure can be represented by a simpler learner.

We work with the Cartesian product of all possible instantiations, which is a high number in most applications. Furthermore, we utilize a variant of OSL for classification with 0/1 loss and a given model to obtain empirical risks for every possible instantiation. This induces hierarchies on the instantiations and allows a (usually different) set of them with each increase of the admissible loss. We call this  $\mathcal{E}$ -Optimistic Subset with  $\mathbf{Y}$  as the Cartesian product and  $\mathcal{E} \in N$  a pre-defined upper bound for classification errors.

$$\mathbf{Y}_{\mathcal{E}} = \{\mathbf{y} \in \mathbf{Y} \mid n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) \leq \mathcal{E}\} \subseteq \mathbf{Y}$$

based on the notation for an instantiation  $\mathbf{y} \in \mathbf{Y}$ , and the loss function  $L(\cdot)$  with the empirical risk denoted as  $\mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n L(\hat{y}_i^{(\mathbf{h}, \mathbf{y})}(x_i), y_i)$ ,  $x_i \in \mathbf{x}$ ,  $\hat{y}_i \in \hat{\mathbf{y}}^{(\mathbf{h}, \mathbf{y})}(\mathbf{x})$ ,  $y_i \in \mathbf{y} \in \mathbf{Y}$  and hyperparameters  $\mathbf{h}$ .

The most optimistic instantiation is hence possible with the approach, but a context-dependent increase leading to more cautious but imprecise forecasting is possible as well. We provide a visual aid similar to a Scree Plot to assist in the determination of the threshold. We employ a Support Vector Machine and twist the tuning of the hyperparameter  $C$ , indicating the clearness of the split to resolve potential ties. Instead of asking for the optimal  $C$  given the observations, we ask for the instantiation of set-valued observations that leads to the lowest  $C$  with a minimal training error. This can be written based on the notation above with the vector of the model's hyperparameters  $\mathbf{h}$  explicitly containing the hyperparameter  $C$ , i.e.  $\mathbf{h} = (C, \mathbf{h}'_r)'$  as:

$$\mathbf{y}_C^* = \arg \min_{\mathbf{y}^*} \arg \min_C \{\mathcal{R}_{emp}(\mathbf{h}_r, C, \mathbf{x}, \mathbf{y}^*) \mid \mathbf{y}^* \in \mathbf{Y}_{\mathcal{E}}^*\} \quad (3.2)$$

A major issue of this approach is the computational feasibility due to the combinatoric explosion as in the original setting; each instantiation depends on every other one. By grouping the non-singleton observations, this can be attenuated as the number of observations is decreased.

We conducted a simulation study and applied the approach to our data on undecided voters. In the simulation study, we illustrated with a toy example how the approach stepwise narrows down options by a decrease in the empirical risk allowed. In the application of real (but clustered) data from the undecided, we could use the illustration similar to a scree plot to choose how many instantiations we allow context-dependent.

### Comments and Outlook

The idea of obtaining a full hierarchical ordering of all possible instantiations according to plausibility is desirable. However, at least two issues have to be addressed. The first one is the already mentioned computational feasibility which is not given within most realistic applications due to the combinatorial explosion and interdependence of the instantiations. To make this approach feasible, a numerical way has to be developed instead of going

through all instantiations separately. This is not hopeless, as the problem can be written as an optimization in different manners, but further work is necessary here. The second is interpreting and formalizing the term "plausible" used. The results are model dependent, even though the approach itself is model agnostic (e.g., you could use any model). This plausibility is hence evidently not global or objective. This makes it hard to communicate the uncertainty attached to the choice used for forecasting.

Overall, the approach gives a very desirable outcome of a hierarchical family of subsets, which is interesting for some approaches and can be enhanced for further research.

## 3.2 Applying the Ontic and Epistemic Approaches

We take a look at Contributions 2 and 4 and contemplate how the methodology can be applied and which opportunities arise under both ontic and epistemic interpretations of the random set. Hence, we will not discuss new methodology to the same extent as in Contributions 1,3,5, and 6.

### 3.2.1 Initial Ontic and Epistemic Ideas with Machine Learning: Contribution 2

Kreiss, D.; Nalenz, M.; Augustin, T.: Undecided Voters as Set-Valued Information - Machine Learning Approaches under Complex Uncertainty. In: E. Huellermeier and S. Destercke, editors, *ECML/PKDD 2020 Tutorial and Workshop on Uncertainty in Machine Learning* (2020) <https://sites.google.com/view/wuml-2020/program?authuser=0> last access: July. 20th, 2023

This chronically second paper reflects on opportunities to extend the homogeneity assumption to machine learning and, in contrast to the first, suggests methodology under the ontic interpretation as well. It is argued that both the epistemic and the ontic views are justified and merely address two very different questions in the election setting. Crucial is the point in time of interest. The time of the poll reflects the ontic perspective, while the time of the election reflects the epistemic one. From there on, two independent branches arise, yielding opportunities to apply known machine learning procedures as well as develop new methodologies.

For the forecasting, we show how transition probabilities with the homogeneity assumption of Contribution 1 can be estimated with any given supervised learner and conduct the analysis with Random Forests. To obtain the overall forecasts with the estimated multinomial distribution, we repeatedly simulate precise observations in a Monte Carlo manner and average over them.

For the structural analysis, we contemplate on different unsupervised approaches. We suggest spectral clustering to find connections between socioeconomic clusters within the

population to locate trends of indecisiveness. Furthermore, the potential of structural analysis is stressed, arguing that a complete representation of the political landscape is only possible with an adequate representation of this relevant group.

The approaches are applied to the constructed data from the GLES and evaluated. We could find structural dissimilarities between the undecided and decided in the clustering approach and could show that the inclusion of the consideration sets affects forecasting.

### Comments and Outlook

This paper can be seen as a direct extension of Contribution 1, taking a more applied perspective and going towards machine learning methodology. We restricted ourselves to one approach each, but many more could have been discussed and are built on this foundation in the later contributions.

From the ontic view, interpretable machine learning is a natural next step which we carried out in Contribution 6. The clustering can also be extended and embedded in a more holistic view of the political landscape, and other unsupervised learning approaches would be possible as well.

This paper's forecasting (or disambiguation) attempts are restricted to the factorization of the likelihood and the homogeneity assumption. While we reflect on potential extensions in Chapter 2.1.2 and introduced one approach in Contribution 5, this plays only a minor role in this paper. Furthermore, attempts at imprecise disambiguation are an interesting further direction possible in this framework. There are also ideas possible connecting likelihood and loss-based approaches.

The results from the application to data from the GLES showed structural differences between the decided and undecided, particularly with undecided between the Green Party and the SPD, highlighting, once more, that including the undecided voters is necessary for an adequate representation. The results from this paper and Contribution 1 served as the basis for the cooperation with the polling institute Civey so that later Contributions could then work with the carefully constructed survey including undecided voters by their consideration sets.

#### 3.2.2 Application to the 2021 German Federal Election: Contribution 4

Kreiss, D., Augustin, T.: Towards a Paradigmatic Shift in Pre-Election Polling Adequately Including Still Undecided Voters – Some Ideas Based on Set-Valued Data for the 2021 German Federal Election. *arXiv preprint* (2021)  
<https://doi.org/10.48550/arXiv.2109.12069>

This paper was published one day before the 2021 German federal election and is the first one that is based on the data provided by Civey. It hence mostly focuses on the results

and the practical advantages of the consideration set sampling but also argues on the bases of random sets and introduces ideas on working with coalitions. We tried to apply the already-developed methodology from the previous papers in a comprehensible way to build a bridge to practitioners from political science.

Two approaches are new. First, we applied a regularized multinomial logit model to examine the political landscape under the ontic view, and second, we regarded coalitions in the forecasting process.

In political research, Multinomial Logit Models are used to find connections between the parties and structural properties. (Tutz, 2011, Chapter 8) With these models, characteristics of interesting groups can be determined, providing a new opportunity to gain empirically founded insights about undecided voters. Concretely, we used the method described in (Tutz et al., 2015) and implemented it in the R package *MRSP* to perform state-of-the-art regularized choice modeling. Hereby, the groups undecided between given parties are often very different from the respective single ones, showing structural differences between the undecided and decided.

Coalitions play a major role, as they are, as mentioned in section 2.1.3, effectively always necessary to form a government. They consist of two or three parties and hence, as in the consideration set sampling setting, live on the (restricted) power set of the original options. Therefore, some of the originally imprecise set-valued representations are precise for coalition forecasting. The natural bounds for forecasting, including the undecided, are hence narrowed. We further provide an approach narrowing the bounds by assuming that on aggregate, not more than 80% for the upper and not less than 20% for the lower bounds choose the corresponding party. As one example: We assume that at least 20% of the individuals undecided between the SPD and Green party end up voting for the SPD and at most 80%. In this applied case, the original bounds are rather wide and hence carry only little relevant information. But already this relatively weak assumption narrows the bounds substantially.

With forecasting, we could see that including the undecided does influence the prediction with the homogeneity assumption, but rather slightly.

### **Comments and Outlook**

Working with real forecasting data invites a comprehensive evaluation of the approaches. In this paper, we wanted to publish our forecasting and structural analysis results before the election, to actually see our data and approaches performance. A thorough analysis of the bias decreased (or increased) by including the consideration sets is interesting for further work as soon as the results are published. The results of the evaluation could also be used to improve some approaches. Hereby the interconnection between the sampling error and the inherent uncertainty from the undecided is interesting to analyze. As with any poll, the sample quality determines the accuracy of the results as discussed in Chapter 2.1.3.

This paper provided the groundwork for improvements in structural analysis in Contribution 6 and was further built upon in Contribution 5. One example is debating the

adequate modeling approach to deal with rather imbalanced data with few observations in some groups.

Cooperation with political scientists to interpret the results and embed them in content-related analysis would be desirable for the future as well.

### 3.3 Structural Analysis with the Ontic View: Contribution 6

Kreiss, D., Augustin, T.: Consideration Set Sampling to Analyze Undecided Respondents. Recently submitted. Preprint available on arXiv under: <https://doi.org/10.48550/arXiv.2307.14333>

This final paper of the thesis develops a comprehensive framework of possible approaches under the ontic view. Here we argue that from our modeling point of view, there are no further set relations between the consideration sets. We use the example from the introduction that being undecided between, for example, the SPD, Green, and Left Party is something completely different than being undecided between the SPD and Green. They reflect two different, mutually unrelated positions of their own. This means that the properties of the original set (finite, categorical, unordered) are also extended to the (restricted) power set. We further denote a subset of the original power set as  $\tilde{P}(S)$  with  $\tilde{P}(S) \subset P(S) \setminus \emptyset$ . The reduction can be due to technical reasons or context-dependent selections.

Based on this, we suggest three avenues of possible methodology applicable to this situation; regression-based, interpretable machine learning, and unsupervised learning.

Starting with regression approaches we first look at the marginal distribution of  $\mathcal{Y}_1, \dots, \mathcal{Y}_n$  with  $\pi_\ell \stackrel{\text{def}}{=} P(\{\mathcal{Y}_i = \ell\})$ ,  $\ell \in \tilde{P}(S)$ , and  $\sum_{\ell \in \tilde{P}(S)} \pi_\ell = 1$ . The underlying samples can be summarised by an appropriate count statistic. We make use of the counting statistic with  $(N_\ell)_{\ell \in \tilde{P}(S)}$  with  $N_\ell \stackrel{\text{def}}{=} |\{i | \mathcal{Y}_i = \ell\}|$  reflecting how many respondents state category  $\ell \in \tilde{P}(S)$ . For on specific sample  $(n_\ell)_{\ell \in \tilde{P}(S)}$ , this results in a multinomial likelihood

$$\text{lik}((\pi_\ell)_{\ell \in \tilde{P}(S)} || (n_\ell)_{\ell \in \tilde{P}(S)}) \propto \prod_{\ell \in \tilde{P}(S)} (\pi_\ell)^{n_\ell}, \quad (3.3)$$

with the relative frequencies  $\frac{n_\ell}{n}$  of the observed positions  $\ell$  as maximum likelihood estimates of the corresponding probabilities  $\pi_\ell$ . We can then write the multinomial logit model for our categorical random elements following (Tutz, 2011, p. 211 ff.) with linear predictor consisting of the covariate vector  $x_i$  together with the category-specific parameters  $(\beta_\ell)_{\ell \in \tilde{P}(S)}$  in its generic form as:

$$P(\mathcal{Y}_i = \ell | x_i) = \frac{\exp(x_i^T \beta_\ell)}{\sum_{s \in \tilde{P}(S)} \exp(x_i^T \beta_s)} \quad (3.4)$$

Model estimation is then possible with maximum likelihood, Bayesian procedures, or in combination with regularisation. We applied grouped regularized regression on the data and found differences between the groups undecided and decided.

For interpretable machine learning, we take the perspective of the Green Party and follow a strategically important question: What distinguishes my convinced supporters from those who are also considering voting for my biggest rival, the SPD? We train a gradient-boosting model and employed SHAP-values, which recovered some interpretability of the approach, lost due to the flexibility of the learner.

For the unsupervised learning section, we conducted an approach similar to the one in Contribution 2. We employ an unsupervised machine learning model based on tree-based dissimilarity following Shi and Horvath (2006), to establish three different groups and then examined the composition of the political positions in these strata.

With this paper, we gave a framework from different areas of structural analysis with categorical data and interpreted the results, painting a more complete picture of the political landscape due to the adequate inclusion of the undecided voters.

### **Comments and Outlook**

This paper attempted to give a comprehensive but still concise framework for opportunities for structural analysis. There are some natural extensions possible. First, it would be very interesting to analyze a longitudinal data structure with consideration set sampling. With this, we could gain insights into actual choice processes as well as models would be challenging from a technical point of view. Modeling for insights into the political landscape could then include this structural dependency between the waves. From the epistemic view, or even a mixture of both, stepwise prediction of changes in the consideration sets on grouped or single observation levels are possible.

Second, we could deviate from the modeling point of view that each consideration set is a mutually unrelated entity on its own and include the weak structure induced by the subset structure in the modeling process. We could employ ordinal regression here (e.g. Gutiérrez et al. (2016)) and compare this to our modeling approaches.

And third, we could transfer the methodology to other areas, not about elections. The biggest alternative application field is probably consumer choice, where the stepwise decision-making process can equally be observed. For this, we could also use online-tracking data to, for example, understand and model the process of buying a car online. Such information could be used for manufacturers and providers to get a better insight into their consumer base, or maybe even more importantly: those individuals who were pondering buying one of their products but then didn't.

# Chapter 4

## Concluding remarks

This chapter contains an outlook and general reflections on the methodological framework provided by this dissertation. More specific conclusions and outlooks for each contribution can be found in the previous chapter.

Motivated by the common neglect of undecided survey participants, this thesis suggested consideration set sampling for an adequate set-valued representation of undecided participants in pre-election polls and provided a methodological foundation and approaches to work with such data.<sup>1</sup> The data opens up two avenues: a more comprehensive structural analysis including the undecided and improved forecasting with this additional (partial) information. For structural analysis, a variety of methods was carefully transferred to the new state space, while for forecasting the tradeoff between accuracy and concise results following Manski's Law (Manski, 2003, p.1) was addressed with several modeling ideas. This thesis can be seen as a contribution to the solution of the mentioned "chicken-egg" dilemma in Contributions 1 and 2. We both conducted a state-of-the-art survey with the consideration set sampling as well as we established several approaches to work with such data. We hence lay the groundwork for further ideas to utilize this promising set-valued data and give some possible suggestions in the next paragraph.

As mentioned in Chapter 2.1.2, a variety of research is directed at predictions with coarse, set-valued data. And it is possible to directly transfer or modify some approaches to our setting. But this ongoing field offers new potential avenues for further research. Contemplating information fusion in this setting by including several sources to estimate transition probabilities are promising. Furthermore, developing credible, interval-valued approaches in the machine-learning setting is interesting. Quantifying the plausibility of given forecasts is challenging but important nonetheless.

Analyzing this data with a longitudinal structure with consideration set sampling could yield further interesting insights. Here, both choice patterns could be analyzed, as well as step-wise predictions between the consideration sets on an individual or global level.

---

<sup>1</sup>This builds on the initial work by (Plass et al., 2015b)



We mostly limited ourselves to methodological considerations without addressing the politological side of undecided voters too much. New applied research in this area would be enabled with the framework of consideration set sampling as well. Customized analysis tailored to content-related questions is possible, as well as the decision process can be examined empirically. One interesting further topic is the communication of the uncertainty beyond the normal sampling error here. Studying how this concept affects nonresponse in different settings is furthermore interesting from a survey point of view.

In our contributions, we only focused on the application case of undecided voters. Some of the approaches suggested could be directly transferred to other applications like consumer choice modeling (as e.g. in Shocker et al. (1991)), while others would have to be adjusted. Under the ontic view, the set has inherent ordering with consumer choice due to the subset structure. This can be included in the modeling process with extensions to ordinal models (e.g. (Gutiérrez et al., 2016)). In this case, modeling with consideration sets could be used to refine target groups and analyze choice patterns.

Overall, we furthered the framework of consideration set sampling together with several modeling approaches. This gives undecided respondents appropriate statistical representation and opens up opportunities for further research.

# Bibliography

- Augustin, T., Walter, G., and Coolen, F. (2014). Statistical inference. In Augustin, T., Coolen, F., de Cooman, G., and Troffaes, M., editors, *Introduction to imprecise probabilities*, pages 135–189. Wiley.
- Bauer, J. (2014). Selection errors of random route samples. *Sociological Methods & Research*, 43(3):519–544.
- Bauer, J. (2016). Biases in random route surveys. *Journal of survey statistics and methodology*, 4(2):263–287.
- Cour, T., Sapp, B., and Taskar, B. (2011). Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536.
- Couso, I. and Dubois, D. (2014). Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning*, 55(7):1502–1518.
- Couso, I., Dubois, D., and Sánchez, L. (2014). *Random sets and random fuzzy sets as ill-perceived random variables*. Springer.
- Dempster, A. (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325 – 339.
- Denoeux, T. (2016). 40 years of Dempster-Shafer theory. *International Journal of Approximate Reasoning*, 79:1–6.
- Edenbrandt, A., Lagerkvist, C.-J., Lüken, M., and Orquin, J. (2022). Seen but not considered? Awareness and consideration in choice analysis. *Journal of choice modelling*, 45:100375.
- Fink, P. (2018). *Contributions to reasoning on imprecise data*. PhD thesis, Departments of Statistics, LMU Munich.
- Foulds, J. and Frank, E. (2010). A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1):1–25.
- Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., and Brinker, K. (2008). Multilabel classification via calibrated label ranking. *Machine learning*, 73:133–153.

- Gallistel, C., Krishan, M., Liu, Y., Miller, R., and Latham, P. (2014). The perception of probability. *Psychological Review*, 121(1):96.
- Graefe, A. (2019). Accuracy of German federal election forecasts, 2013 & 2017. *International Journal of Forecasting*, 35(3):868–877.
- Groves, R. and Lyberg, L. (2010). Total Survey Error: Past, present, and future. *Public Opinion Quarterly*, 74(5):849–879.
- Gutiérrez, P., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernández-Navarro, F., and Hervás-Martínez, C. (2016). Ordinal regression methods: Survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146.
- Heitjan, D. and Rubin, D. (1991). Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244 – 2253.
- Hüllermeier, E. (2014). Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534.
- Klima, A., Küchenhoff, H., Selzer, M., and Thurner, P. (2017). *Exit Polls und Hybrid-Modelle*. Springer.
- Lewis-Beck, M. and Dassonneville, R. (2015). Comparative election forecasting: Further insights from synthetic models. *Electoral Studies*, 39:275–283.
- Manski, C. (1977). The structure of random utility models. *Theory and decision*, 8(3):229.
- Manski, C. (2003). *Partial identification of probability distributions*. Springer.
- Mauerer, I., Pöfnecker, W., Thurner, P., and Tutz, G. (2015). Modeling electoral choices in multiparty systems with high-dimensional data: A regularized selection of parameters using the lasso approach. *Journal of choice modelling*, 16:23–42.
- McFadden, D. (1981). Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications*, 198272.
- Oscarsson, H. and Oskarson, M. (2019). Sequential vote choice: Applying a consideration set model of heterogeneous decision processes. *Electoral Studies*, 57:275–283.
- Pappi, F. and Thurner, P. (2002). Electoral behaviour in a two-vote system: Incentives for ticket splitting in German Bundestag elections. *European Journal of Political Research*, 41(2):207–232.
- Plass, J. (2018). *Statistical modelling of categorical data under ontic and epistemic imprecision: contributions to power set based analyses, cautious likelihood inference and (non-)testability of coarsening mechanism*. PhD thesis, Departments of Statistics, LMU Munich.

- Plass, J., Augustin, T., Cattaneo, and Schollmeyer, G. (2015a). Statistical modelling under epistemic data imprecision: some results on estimating multinomial distributions and logistic regression for coarse categorical data. pages 247–256.
- Plass, J., Cattaneo, M., Augustin, T., Schollmeyer, G., and Heumann, C. (2019). Reliable inference in categorical regression analysis for non-randomly coarsened observations. *International Statistical Review*, 87(3):580–603.
- Plass, J., Fink, P., Schöning, N., and Augustin, T. (2015b). Statistical modelling in surveys without neglecting ‘The undecided’. pages 257–266. SIPTA.
- Richter, G., Wolfram, T., and Weber, C. (2022). Die statistische Methodik von Civey. Online article under <https://civey.com/whitepaper>.
- Rubin, D. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.
- Sakshaug, J., Yan, T., and Tourangeau, R. (2010). Nonresponse error, measurement error, and mode of data collection: Tradeoffs in a multi-mode survey of sensitive and non-sensitive items. *Public Opinion Quarterly*, 74(5):907–933.
- Schollmeyer, G. and Augustin, T. (2015). Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *International Journal of Approximate Reasoning*, 56:224–248.
- Shi, T. and Horvath, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1):118–138.
- Shirani-Mehr, H., Rothschild, D., Goel, S., and Gelman, A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association*, 113(522):607–614.
- Shocker, A., Ben-Akiva, M., Boccara, B., and Nedungadi, P. (1991). Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing Letters*, 2(3):181–197.
- Skinner, C. and Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, 32(2):165–175.
- Spieß, M. (2010). Der umgang mit fehlenden Werten. *Handbuch der sozialwissenschaftlichen Datenanalyse*, pages 117–142. Springer.
- Stocchi, L., Banelis, M., and Wright, M. (2016). A new measure of consideration set size: The average number of salient brands. *International Journal of Market Research*, 58(1):79–94.
- Turner, P. (2000). The empirical application of the spatial theory of voting in multiparty systems with random utility models. *Electoral Studies*, 19(4):493–517.

- Tutz, G. (2011). *Multinomial Response Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Tutz, G., Pößnecker, W., and Uhlmann, L. (2015). Variable selection in general multinomial logit models. *Computational Statistics & Data Analysis*, 82:207–222.
- Tversky, A. (1972a). Choice by elimination. *Journal of mathematical psychology*, 9(4):341–367.
- Tversky, A. (1972b). Elimination by aspects: A theory of choice. *Psychological Review*, 79(4):281.
- Walley, P. (2000). Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2-3):125–148.
- Weichselberger, K. (2000). The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24(2-3):149–170.
- Winkler, N., Kroh, M., and Spiess, M. (2006). Entwicklung einer deutschen Kurzsкала zur zweidimensionalen Messung von sozialer Erwünschtheit. DIW Discussion Papers 579.
- Zaffalon, M. (2002). The naive Credal classifier. *Journal of statistical planning and inference*, 105(1):5–21.
- Zhang, M.-L. and Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.
- Zhou, Z.-H. (2017). A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.

# Attached contributions

*Contribution 1:* p. 34 – 42

*Contribution 2:* p. 43 – 55

*Contribution 3:* p. 56 – 65

*Contribution 4:* p. 66 – 79

*Contribution 5:* p. 80 – 95

*Contribution 6:* p. 96 – 106

## Contribution 1

Kreiss, D.; Augustin, T.: Undecided Voters as Set-Valued Information – Towards Forecasts Under Epistemic Imprecision. In: J. Davis and K. Tabia, editors, *International Conference on Scalable Uncertainty Management*, pp. 242-250. Springer Lecture Notes in Artificial Intelligence (2020)

[https://doi.org/10.1007/978-3-030-58449-8\\_18](https://doi.org/10.1007/978-3-030-58449-8_18)



# Undecided Voters as Set-Valued Information – Towards Forecasts Under Epistemic Imprecision

Dominik Kreiss<sup>(✉)</sup> and Thomas Augustin

Department of Statistics, LMU Munich, Ludwigstr. 33, Munich, Germany  
{dominik.kreiss,thomas.augustin}@stat-uni.muenchen.de

**Abstract.** Increasing numbers of undecided individuals in pre-election polls throughout western democracies impose a severe challenge for election forecasting. While conventionally these voters are neglected relying on presumably unjustified assumptions, we sketch more nuanced approaches incorporating the potential valuable information in a set-valued manner. Hereby, each undecided voter is represented by the set of parties he or she is incapable to choose from. This set, containing one true, but unknown element, enables modelling under so-called *epistemic imprecision*. Depending on further assumptions, (imprecise) transition probabilities between the options can be estimated in order to achieve election forecasting. Starting with Dempster’s upper and lower probabilities as the most cautious approach, two further ideas are introduced, providing initial methodology. Furthermore, extensions including Bayesian modeling are sketched. The theory is applied using data from the *German Longitudinal Election Study* for forecasting concerning the most recent German federal election of 2017. The results are promising, laying the groundwork for further research.

**Keywords:** Epistemic modeling · Election forecasting · Coarse data · Partial identification · Survey methodology

## 1 Introduction

If we think of an election in a multiparty system as a choice of individuals  $i \in \{1, \dots, N\}$  between the options  $\{1, \dots, j\} = S$ , a decided individual in a pre-election poll is capable to single out one element of  $S$  as his or her choice, while an undecided is not. The position of the undecided can therefore be accurately represented by a nonempty subset  $\ell \subset S$  containing all parties the individual is pondering between, hence all options that cannot be excluded.

One advantage of this set-valued information is the rather practical character, as most individuals are capable to state this subset  $\ell$  precisely [8, p. 256 f], providing the opportunity to obtain this information by a pre-election survey. The idea of set-valued response in election choice was recently introduced by [8] in a political science framework, arguing that stepwise exclusion of options is the

natural process of human choice [8, p. 256]. Furthermore, in her work about set-valued data, Plass [9, p. 2–3] argues that providing set-valued response categories might reduce nonresponse substantially. In conventional analysis, the undecided are overall neglected [11, p. 265], not only relying on disputable assumptions about the left out individuals but also missing out on valuable information about their position. Moreover, concerning the question which combination of parties will constitute the government, coalitions can be represented more directly by set-valued information. Despite these reasons, set-valued data is regrettably not yet included in most surveys but first approaches already exist as can be found in [8, 9, 11].

The subset  $\ell$ , further on called *consideration set* following [8], determining the undecided individual’s position, can be seen as a disjunctive random set, containing one ill-known true value. (e.g. [3]) Thus, to predict the undecided’s choice on election day, we can develop models under *epistemic imprecision*, following [3, ch. 2], using the coarse information together with assumptions and further sources of information. A wide range of approaches are possible, reaching from Dempster’s so to say agnostic bounds [4] up to point-valued estimation, relying on strong assumptions. We develop and apply three approaches weighting the justifiability of assumptions with the precision of the results and introduce methodology for overall election outcome forecasting using transition probabilities. We hereby break first ground introducing epistemic methodology to election forecasting.<sup>1</sup>

This paper is structured as follows: First, we briefly recall the underlying epistemic theory in Sect. 2.1 before introducing the general problem in Sect. 2.2 and three modeling approaches in Sect. 2.3. In Sect. 3, we apply the developed approaches to the most recent German federal election. The concluding remarks reflect on the approaches and future possibilities.

## 2 Methods

### 2.1 The Epistemic View of Set-Valued Information

Given the accurate, set-valued representation  $\ell \in \mathcal{P}(S) = 2^S$  of an undecided individual with  $\mathcal{P}(S)$  as the power set of the parties to choose from, there exists one true, yet unknown element  $l \in \ell$  representing the undecided’s choice on election day. The consideration sets  $\ell$  result from individuals excluding their neglectable options, leading to a subset, which by definition contains the true element  $l$ . Hence,  $\ell$  is a set consisting of distinguishable and finite elements containing incomplete information about the true value of interest  $l$ . This is the so-called epistemic view of set-valued information, following [3]. While we are looking for the random variable  $Y(\omega)$  mapping from an underlying space  $\Omega$  to  $S$ , we are only provided with incomplete information in the sense that  $\forall \omega \in \Omega$

<sup>1</sup> In that sense we contribute to a solution of a “chicken-egg dilemma” (Fink), resulting from the lack of surveys including the set-valued question as well as missing methodology, providing applicable approaches for such data.



244 D. Kreiss and T. Augustin

only  $Y(\omega) \in \ell = \mathcal{Y}(\omega)$  is observable, where  $\mathcal{Y}$  is a multi-valued mapping  $\Omega \rightarrow 2^S$  representing the set of mappings  $\{Y : \Omega \rightarrow S, Y(\omega) \in \mathcal{Y}(\omega) \forall \omega\}$ . [1, p. 1504] We thus build an epistemic model of the random variable  $Y(\omega)$ , while for the undecided all that is known is  $Y(\omega) \in \ell$ .

The realization  $l$  can therefore be seen as a realization of an *ill-known random variable* incompletely described by a coarse version in the form of the set  $\ell$ . Due to the lack of information about the true value  $l$ , prediction approaches have to incorporate further information and assumptions in order to obtain more concise or even point-valued results. By [3, p. 1503] this is described as representing both reality as well as knowledge of reality, explicitly accounting for the limited precision. Thus, one has to ponder between imprecise results and the justifiability of assumptions leading to more precise statements.<sup>2</sup>

## 2.2 From Set-Valued Information to Forecasting

Each individual from the sample is determined by both its consideration set  $\ell \in \mathcal{P}(S)$  and its co-variables  $X = x$  in some space  $\mathcal{X}$ , assessing their personal characteristics. The individual's consideration set from the pre-election survey is written as an event  $\{\mathcal{Y} = \ell\}$  with  $\ell \in \mathcal{P}(S)$  and his or her possibly unknown choice on election day  $\{Y = l\}$  with  $l \in S$ . Given the consideration sets of participant  $i \in \{1, \dots, n\}$  in the pre-election poll, we want to obtain the expected frequency of each element of  $S$  within the population, with latent probability distribution  $P(Y = l)$  for all  $l \in S$ , which is a multinomial distribution over the state space with  $|S| - 1$  parameters. The observations  $Y_i$  are assumed to be identically and independently distributed and  $P(Y = l)$  can be written in respect to the consideration sets and co-variables as

$$P(Y = l) = \sum_{(\ell, x) \in (2^S \times \mathcal{X})} P(Y = l, \mathcal{Y} = \ell, X = x) = \quad (1)$$

$$\sum_{(\ell, x) \in (2^S \times \mathcal{X})} \underbrace{P(Y = l | \mathcal{Y} = \ell, X = x)}_{\text{Transition Probabilities}} \cdot \underbrace{P(\mathcal{Y} = \ell | X = x)}_{\text{Consideration Sets}} \cdot \underbrace{P(X = x)}_{\text{Co-Variables}} \quad (2)$$

The probability distribution can therefore be factorized into three parts. First, the from now on so-called *transition probabilities*, determining the probability to vote for a specific party given the consideration set and co-variables. Second, the probability of the consideration sets given the co-variables and third, the one for the co-variables. While the second and third part can be directly estimated from the data of the pre-election survey alone, the first requires further assumptions and/or sources of information, as the eventual choice  $l$  from the options  $\ell$  is not observable amongst the undecided. For the decided individuals, the transition probabilities are naturally one, while for the undecided either point- or interval-valued estimation is necessary. The transition probabilities can be seen as a further (imprecise) multinomial distribution over the individual's consideration set.

<sup>2</sup> See also Manski's Law of Decreasing Credibility [7, p. 1].

There are different directive questions concerning the estimation process of the transition probabilities resulting in several modeling approaches. First, one has to ponder whether results are obliged to be point-valued or not. Second, if the pre-election poll remains the only source of data and third, which assumptions are made in order to determine estimation. In the following section, three approaches relying on different constellations of these issues are discussed. Hereby, basic methodology to brake first ground is introduced and an outlook to improve these ideas is provided.

### 2.3 Approaches to Estimate Transition Probabilities

Starting with the idea of Dempster [4] as the **first approach**, only to use information available in the data alone, not relying on further assumptions nor information, the transition probabilities reflect the entire ambiguity of the individuals. Thus, as no information is available about which element of  $\ell$  constitutes the true one, for every  $\ell$  consisting of more than one element the transition probabilities take the whole range between 0 and 1.<sup>3</sup> Combination with the decided individuals and weighting according to Eq. (2) leads to interval-valued forecasting which tends to be wide. Hence, these so-called Dempster's bounds reach from worst-to best case scenario for each party, while the range of the interval reflects the ambiguity concerning the respective party. Even if the results might not be providing sufficient information depending on the question at hand, all information of the dataset that can be used, not relying on any assumption, is used in the process. The hereby estimated bounds can be seen as the extreme case, resulting from the most cautious way of modeling, leading to rather imprecise results.

As the other extreme, depending on the question at hand and preference, results are required to be point-valued, forcing overall stronger assumptions. Hereby, the parameters of the transition probabilities have to be estimated in a point-valued way to ensure overall point-valued forecasting.

As there is no information about the undecideds' choice provided, for the **second approach** we fall back on the decided individuals. Using the decided, the probability distribution  $P(Y_i = l | X_i = x_i, I_d = 1)$  can be estimated from the data, with  $I_d$  as the indicator function for being decided. To enable point-valued estimation we then assume that, given the co-variables, the undecided choose identical to the decided. The consideration set hereby becomes the restriction of possible outcomes, while the tendency towards a party of the consideration set is predicted using the decided and co-variables as underlying data. Those predictions of affinity towards the parties of the undecided have to be scaled to comply with the multinomial distribution, excluding all options not in  $\ell$ . Therefore, for all  $l \in \ell$  the predicted affinity towards one party is divided by the sum of all the ones in the consideration set resulting in

$$\hat{P}(Y = l | \mathcal{Y} = \ell, X = x) = \frac{\hat{P}(Y = l | X = x, I_d = 1)}{\sum_{a \in \ell} \hat{P}(Y = a | X = x, I_d = 1)} \quad (3)$$

<sup>3</sup> For more details and examples see for instance [11, p. 261] or [4, p. 325 ff].

246 D. Kreiss and T. Augustin

leading to point-valued identification of every parameter.<sup>4</sup> There are several ways to estimate the conditional distributions for each individual necessary for Eq. (3), while we choose the most common approach of linear logit models (e.g. [5, p. 238 ff.]) Even though it is not impossible that the undecided, given the co-variables, behave in average identical to the decided while only excluding options outside the consideration set, some structural differences are likely to be ignored. Nevertheless, it can be argued that the drawbacks from neglecting the undecided overall outweighs this strong assumption.

The **third approach** includes information from the previous election, using data to estimate the transition probabilities of the former election  $P(Y^+ = k | \mathcal{Y}^+ = k, X^+ = x^+)$ , available within the post-election poll, with  $+$  denoting the previous election. Incorporating data from different surveys is controversial, as both the political landscape and the selection of participants might differ severely. In order to obtain point-valued estimates with this information alone, it has to be assumed that the transition probabilities, given the co-variables, are constant between the elections. As this assumption is likely to be violated, there are reasons to rather incorporate the information in another (possibly hierarchical) way together with other sources of information. Nevertheless, point-valued forecasting can be achieved at the cost of these drawbacks.

These three approaches take first steps towards election forecasting including the undecided, while for **further research** each of them can be further developed and improved. Prior information to facilitate the estimation process in the form  $p(Y = l | \mathcal{Y} = \ell)$  could be incorporated in the analysis, as well as set-valued prior information could be used to achieve more plausible interval-valued results. One could assume, building on the third approach, that given specific expert knowledge, the transition probabilities are constant between the elections. Also complex hierarchical Bayesian methodology using the sources of information from the decided individuals, the undecideds' choice of the former period and (set-valued) expert knowledge is possible. A natural way to make such point-valued approaches more robust would be to rely on appropriate neighbourhood models. Another instance where including expert knowledge could be important, is to either weaken assumptions or to deal with the missing not at random structures within the nonresponse of the survey. The three original approaches are computationally rather simple, but even the more complex methods suggested should still be scalable, as typical electoral polls rarely exceed 2000 participants.

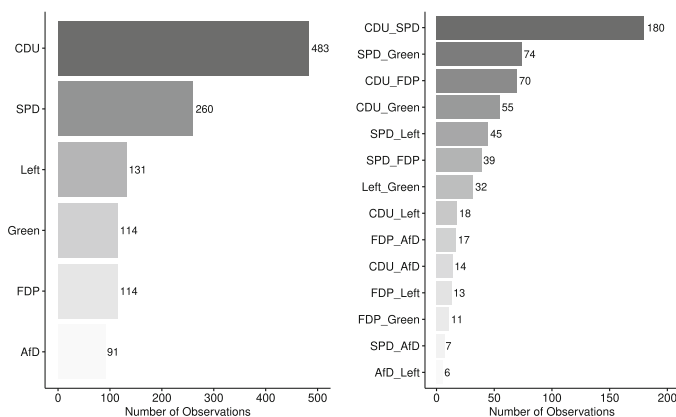
### 3 Application

#### 3.1 The Data from GLES

We applied the ideas developed above for the most recent German federal election of 2017 using the state of the art pre- and post-election surveys provided

<sup>4</sup> Note although intuitively this is a kind of random coarsening assumption, it differs from the usual CAR conditions.

for scientific use by the *GLES*.<sup>5</sup> Set-valued response is regrettably not directly included in this survey, but the assessment of the parties by the individuals as well as their statement about the certainty of their choice are, enabling construction of consideration set as described by [11, p. 261]. To facilitate a proof of concept of our methodology, we only focus on the most common case of indifference between exactly two parties as well as we only use the two binary co-variables *sex* and *residence in east or west Germany*.<sup>6</sup> Moreover, we examine the so-called second vote<sup>7</sup> for the six main parties anticipated to reach at least one seat in the parliament, not including non-voters and small parties. Furthermore, structures of the nonresponse in the dataset are not explicitly adjusted for.



**Fig. 1.** Overview of occurrences of different groups amongst the participants questioned for the 2017 federal election by the GLES.

From the overall 1774 individuals used, 581 are undecided between exactly two parties, constituting about a third of the sample, while from the overall survey 11.87% were undecided between more than two options. Figure 1 illustrates the number of observations within the undecided and decided voters concerning the specific groups.

### 3.2 Results of the Different Approaches

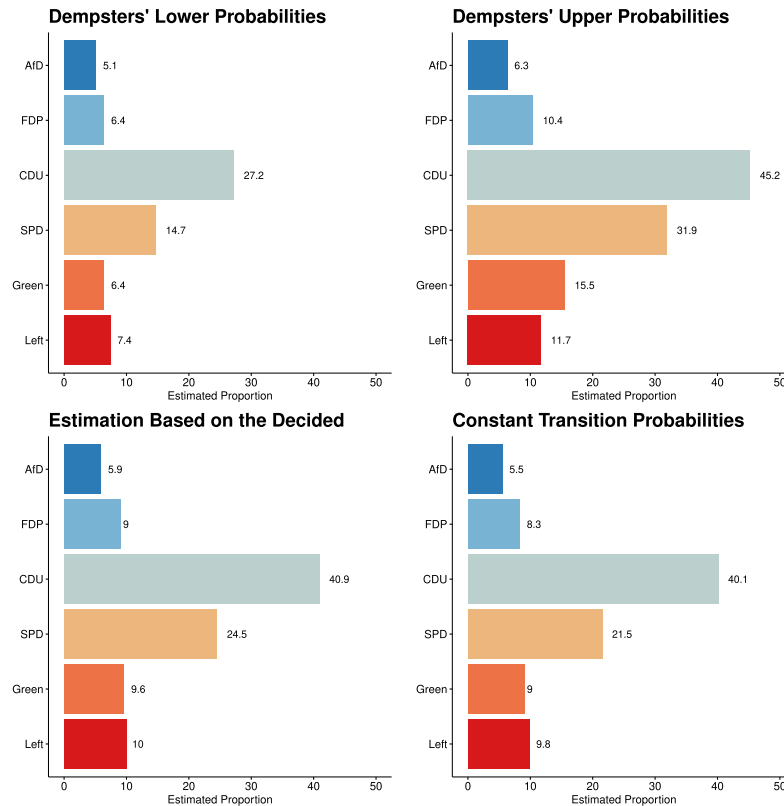
We apply all three approaches discussed in Sect. 2.3 calculating overall forecasts according to Eq. (2), reliant on the same underlying dataset. The results are illustrated in Fig. 2 providing an overview of differences and similarities as well as general tendencies within the approaches.

<sup>5</sup> German Longitudinal Election Study: Pre- and post- election cross-section available under <https://www.gesis.org/wahlen/gles/daten>; last visited: 13.07.20.

<sup>6</sup> We are aware that two variables do not capture the entire structural properties of the individual in our proof-of-concept model, as should be by the co-variables in an ideal scenario to improve estimation for approaches 2 and 3.

<sup>7</sup> Vote for the party, which is usually used for forecasting.

248 D. Kreiss and T. Augustin



**Fig. 2.** Results from epistemic election forecasting of the three approaches based on the same underlying observations.

For the interval-valued Dempster bounds, upper and lower probabilities are illustrated with two separated plots. Hereby, the entire ambiguity is reflected between the upper and lower bounds, thus enclosing the other two approaches and showing the strongest deviation from conventional approaches. The second approach (estimation based on the decided) and the third (assuming constant transition probabilities between this and the last election) seem roughly similar here. The party CDU has by far the highest estimates throughout all approaches, but varies the most between upper and lower bounds. In contrast, the AfD has the lowest turnout with diminishing differences between the approaches. As the non-response structures are not adjusted for, the consideration sets are constructed and the variable selection merely served as a proof of concept, the results should be treated with caution concerning their political implications and validity.

Overall, the methodology has proven to be straightforwardly applicable producing plausible, but not yet sufficient results for final election outcome forecasting. Adjustments, necessary due to the missing not at random structures through weighting or expert knowledge and incorporation of further sources of information should yield substantially improved results.

## 4 Concluding Remarks

In this paper we introduced ideas in order to include the otherwise wasted information of the undecided from pre-election polls, by using their consideration sets. Several approaches are possible, weighting the precision of the results with the justifiability of the assumption, resulting in point- or interval-valued forecasting. We introduced and applied three approaches constituting possible directions, with the most cautious Dempster bounds and two point-valued ones based on different strong assumptions. Reliant on constructed consideration sets and simplifications, our forecasts are not yet perfected, but the potential is considerable. The approaches can be further developed and improved, as already sketched in Sect. 2.3, by making use of supplementary sources of information like expert knowledge or previous elections, for example, in an hierarchical Bayesian manner with imprecise probabilities. One further natural question would be the relationship with other approaches dealing with imprecise data using likelihood or loss minimisation like [2, 6, 10]. In contrast to conventional methodology, the approaches discussed here explicitly address and incorporate the ambiguity of the individuals by making use of their consideration sets, introducing new ideas to election forecasting in times of increasing relevance of undecided voters.

**Acknowledgement.** We are very thankful to the four anonymous reviewers for their helpful remarks.

## References

1. Couso, I., Dubois, D.: Statistical reasoning with set-valued information: ontic vs. epistemic views. *Int. J. Approx. Reason.* **55**, 1502–1518 (2014)
2. Couso, I., Dubois, D.: A general framework for maximizing likelihood under incomplete data. *Int. J. Approx. Reason.* **93**, 238–260 (2018)
3. Couso, I., Dubois, D., Sánchez, L.: *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables*. SAST. Springer, Cham (2014). <https://doi.org/10.1007/978-3-319-08611-8>
4. Dempster, A.: Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38**, 325–339 (1967)
5. Fahrmeir, L., Kneib, T., Lang, S., Marx, B.: *Regression Models*. Springer, Cham (2013). <https://doi.org/10.1007/978-3-642-34333-9>
6. Hüllermeier, E., Destercke, S., Couso, I.: Learning from imprecise data: adjustments of optimistic and pessimistic variants. In: Ben Amor, N., Quost, B., Theobald, M. (eds.) SUM 2019. LNCS (LNAI), vol. 11940, pp. 266–279. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-35514-2\\_20](https://doi.org/10.1007/978-3-030-35514-2_20)
7. Manski, C.: *Partial Identification of Probability Distributions*. Springer, Cham (2003). <https://doi.org/10.1007/b97478>
8. Oscarsson, H., Rosema, M.: Consideration set models of electoral choice: theory, method, and application. *Electoral. Stud.* **57**, 256–262 (2019)
9. Plass, J.: *Statistical modeling of categorical data under ontic and epistemic imprecision*. PhD thesis, LMU Munich (2018). <https://edoc.ub.uni-muenchen.de/22298/> (Accessed 13 July 2020)

250 D. Kreiss and T. Augustin

10. Plass, J., Cattaneo, M., Augustin, T., Schollmeyer, G., Heumann, C.: Reliable inference in categorical regression analysis for non-randomly coarsened observations. *Int. Stat. Rev.* **87**(3), 580–603 (2019)
11. 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 2015*, pp. 257–266. SIPTA (2015)

## Contribution 2

Kreiss, D.; Nalenz, M.; Augustin, T.: Undecided Voters as Set-Valued Information - Machine Learning Approaches under Complex Uncertainty. In: E. Huellermeier and S. Destercke, editors, *ECML/PKDD 2020 Tutorial and Workshop on Uncertainty in Machine Learning* (2020) <https://sites.google.com/view/wuml-2020/program?authuser=0> last access: July. 21, 2023



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

Dominik Kreiss<sup>1</sup>, Malte Nalenz<sup>2</sup>, and Thomas Augustin<sup>3</sup>

<sup>1</sup> LMU Munich, Department of Statistics, Ludwigstr. 33, Munich  
`dominik.kreiss@stat-uni.muenchen.de`

<sup>2</sup> LMU Munich, Department of Statistics, Ludwigstr. 33, Munich  
`malte.nalenz@stat-uni.muenchen.de`

<sup>3</sup> 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<sup>4</sup> urge us to find new ways to deal with these individuals in statistical analysis and empirical election

<sup>4</sup> 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  $\{1, \dots, s\} = S$ , 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  $\ell$  that can be described as the realization of a measurable mapping  $\mathcal{Y} : \Omega \rightarrow P(S)$  from some underlying space  $\Omega$  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  $\Omega \rightarrow P(S)$  now representing the set of mappings  $\{Y : \Omega \rightarrow S, \forall \omega, Y(\omega) \in \mathcal{Y}(\omega)\}$ , 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<sup>5</sup> 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.

<sup>5</sup> 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.<sup>6</sup> 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  $\{Y = \ell\}$  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  $\ell$  are excluded. Therefore, scaling the estimates from the decided to comply with the multinomial distribution results in:

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

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

<sup>6</sup> 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 possess 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 approximated. 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 choose 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 predictive power of random forests with the additional information given by the consideration sets.

<sup>7</sup>

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.

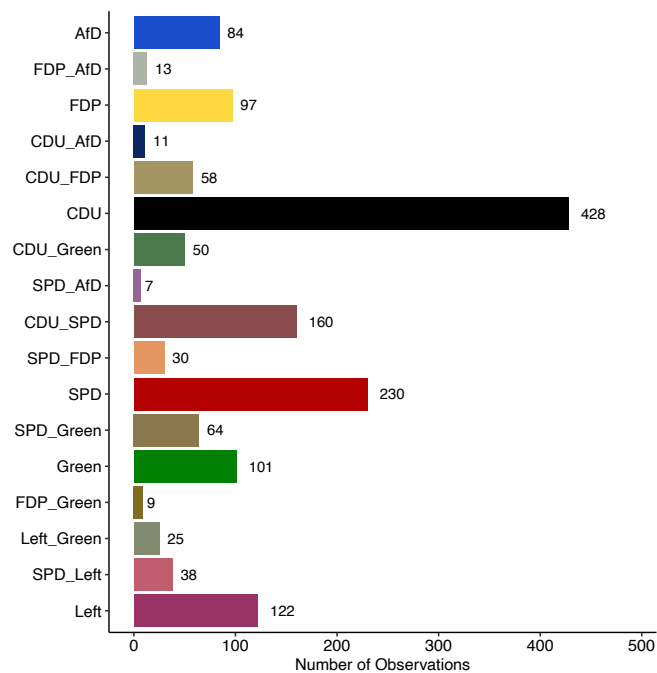
<sup>7</sup> We do not use the undecided in the first level of estimation with some kind of simulation, in order to avoid strong assumptions about the final outcome in the consideration sets.

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*<sup>8</sup>. 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.

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

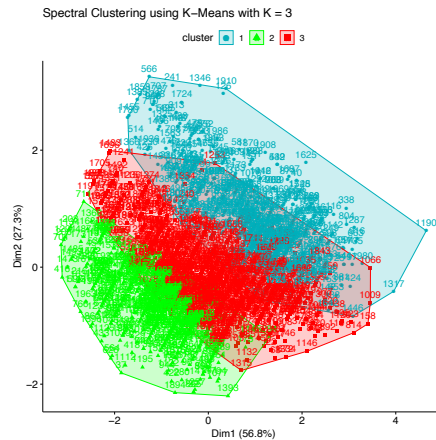
8 Dominik Kreiss, Malte Nalenz and Thomas Augustin

For our analysis, we use the so-called second vote<sup>9</sup> for the six main parties<sup>10</sup> 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.

<sup>9</sup> 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/wahlssystem.html>, last visited 27.07.20

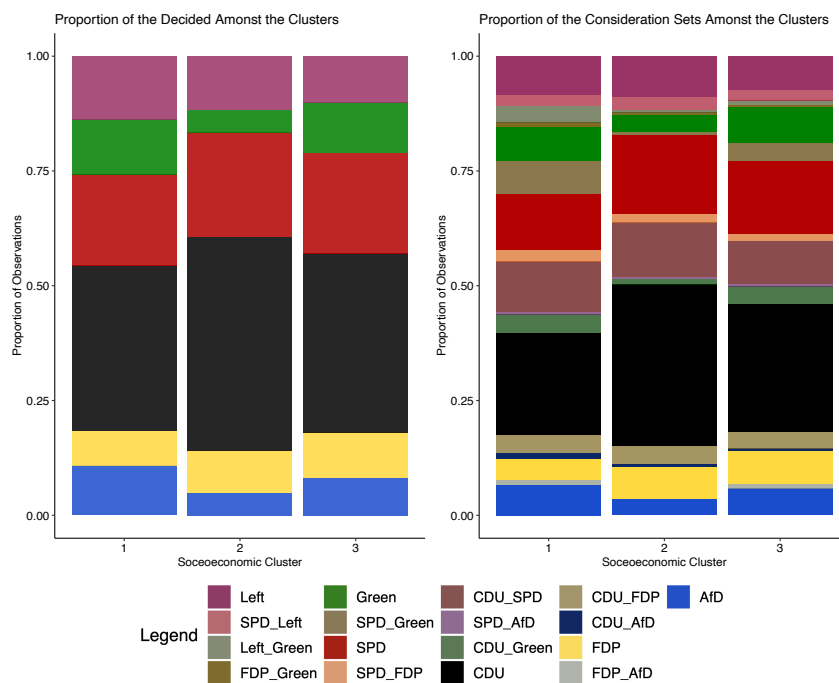
<sup>10</sup> The parties are: AfD, FDP, CDU (including CSU), SPD, Green, Left



Undecided Voters as Complex Uncertainty in Machine Learning 9

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.

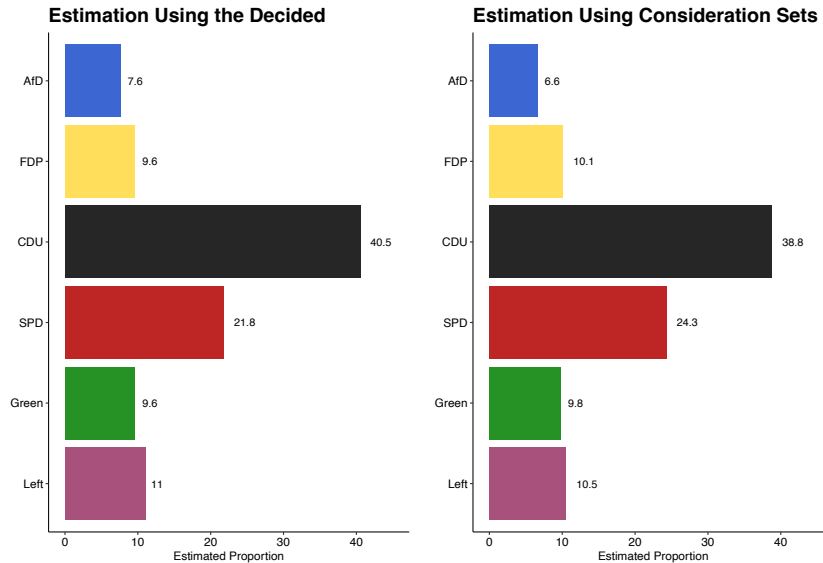
positions are very unevenly distributed amongst the clusters. Notable, for example, is the high proportion of undecided between the Green and other parties within the first cluster, as mentioned above mostly consisting of young voters with comparable high income. The proportion of overall undecided is 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

10 Dominik Kreiss, Malte Nalenz and Thomas Augustin

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 ill-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)

## Contribution 3

Kreiss, D; Schollmeyer, G. and Augustin, T.. Towards Improving Electoral Forecasting by Including Undecided Voters and Interval-Valued Prior Knowledge. In J. De Bock, A. Cano, E. Miranda, and S. Moral, editors, *Proceedings of the Twelfth International Symposium on Imprecise Probabilities: Theories and Applications*, pp. 201-209, Proceedings of Machine Learning Research (2021) <https://isipta21.sipta.org/papers.html>, last access: July. 21, 2023

# Towards Improving Electoral Forecasting by Including Undecided Voters and Interval-valued Prior Knowledge

**Dominik Kreiss**  
**Georg Schollmeyer**  
**Thomas Augustin**

*Institute of Statistics, Ludwig-Maximilians Universität München (LMU), Munich, Germany*

DOMINIK.KREISS@STAT.UNI-MUENCHEN.DE  
 GEORG.SCHOLLMAYER@STAT.UNI-MUENCHEN.DE  
 THOMAS.AUGUSTIN@STAT.UNI-MUENCHEN.DE

## Abstract

Increasing numbers of undecided voters constitute a severe challenge for conventional pre-election polls in multi-party systems. While these polls only provide the still pondering individuals with the options to either state a precise party or to drop out, we suggest to regard their valuable information in a set-valued way. The resulting consideration set, listing all the options the individual is still pondering between, can be interpreted under epistemic imprecision. Within this paper we extend the already existing approaches including this valuable information, by making first steps to utilize interval-valued prior information. Including background information is common in election forecasting while we focus on realistically obtainable and credible interval-valued prior information about transition probabilities from the undecided to the eventual choice. We introduce two approaches utilizing this interval-valued information, weighting the credibility against the precision of the results. For the first approach, we narrow the most cautious and wide so-called Dempster bounds by deploying the prior information on the transition probabilities as new worst and best case scenarios for each party. The second approach applies if these interval-valued results are still too wide for useful application. We hereby narrow them towards a good guess of the eventual choice, estimated by a further model-based source of information making use of the covariates. These single-valued estimates on the individual level are regarded as realizations of an underlying probability distribution, which we combine with the prior knowledge in a Bayesian way. The approach can thus be seen as an attempt to combine two, for the needed outcome by themselves inadequate, sources of information to obtain more concise results. We conduct a simulation study showing the applicability and virtues of the new approaches and compare them to conventional ones.

**Keywords:** epistemic imprecision, election forecasting, undecided respondents, survey methodology, imprecise probabilities

## 1. Introduction

As more and more voters are undecided in pre-election polls, methodology incorporating their valuable information is called for. To this end [Plass et al. \(2015\)](#); [Oscarsson and Rosema \(2019\)](#); [Kreiss and Augustin \(2020\)](#); [Kreiss et al. \(2020\)](#) suggested to regard an undecided voter as the set of options the individual is still pondering between, from now on so-called *consideration sets*. Several arguments are put forward to substantiate this approach, like the reduction of nonresponse, the natural procedure or the more accurate representation of uncertainty. An election in a multi-party-system is hereby a separate choice of  $\{1, \dots, n\}$  individuals between a discrete set of, from the beginning on known, alternatives  $\{1, \dots, s\} = S$ . But at the point in time of the pre-election poll, the undecideds' position can only be characterized by a combination of the original parties representing the options he or she is still pondering between, hence, an element from the state space of the power set of the original options  $\mathcal{P}(S)$ . This set-valued information about a still undecided can be seen as a container of the one true element he or she ends up voting for, thus as a coarse version of that true element. This is called the epistemic interpretation of the consideration set (e.g. [Couso and Dubois \(2014\)](#)), focussing on election forecasting.

Provided with this set-valued information, *transition probabilities* within the consideration sets of the undecided to the final choice can be assessed, in order to obtain overall forecasting together with the decided. As we are faced with inherent, non-stochastic uncertainty which element of their consideration set the individual ends up choosing, the resulting forecasts are naturally interval-valued if no further procedures are deployed. To this end, [Kreiss and Augustin \(2020\)](#); [Kreiss et al. \(2020\)](#) suggested preliminary approaches, reaching from the *Dempster bounds* as the most cautious to point-valued results based on strong additional assumptions.

Within this paper we introduce methodology to incorporate credible interval-valued prior information about transition probabilities as a natural further step to improve forecasting in this setting. Different sorts of background knowledge are commonly used in electoral research. (e.g. [Linzer \(2013\)](#)) Information in the form of probability inter-

vals over the singletons can hereby either stem from experts and/or previous elections. We suggest two approaches utilizing prior information, weighting the credibility against the precision of the results. We hereby use the term prior information in an informal sense, comprising all kinds of background information beyond the data.

The first approach narrows the most cautious upper and lower Dempster bounds deploying the prior information as new best and worst case scenarios for each party. If the interval-valued prior information is credible, we obtain accurate but possibly wide intervals. If the resulting interval-valued forecasts are still too wide for useful application, the prior information alone is inadequate to obtain the necessary outcome. Hence, we have to include further knowledge about the process. This is a rather complicated task as due to the epistemic nature of the problem no entirely reliable information about the eventual choice is available. Therefore, we suggest to include information from another source, using covariates to estimate transition probabilities on an individual level. Hereby, we make use of the information within the covariates about the eventual choice with a working model reliant on presuppositions and trained on supplementary data. This can be seen as our best guess, which on the one hand is presumably biased but on the other hand carries the information within covariates. Towards this guess we can now narrow the interval from the prior knowledge. As the working model is trained on supplementary data, the single-valued estimates on the individual level are obtained by a deterministic function of the covariates. Hence, if the sample is identically independently distributed (i.i.d.), the resulting estimates can be seen as i.i.d. realizations of an underlying distribution characterizing the transition probabilities based on this working model. We combine the interval-valued prior information with the results of the precise working model for each party in each group of undecided separately in a Bayesian way. To achieve this, the interval-valued prior information is not seen infallible but also distributed with a certain variance. This variance determines the impact of the prior information within the overall forecasts.

Depending on distributional assumptions, the lower and upper bounds of the prior information can be respectively deployed as priors for Bayesian models, to obtain two posteriori distributions leading to estimates of the upper and the lower bound of the transition probabilities. As a consequence we obtain narrower bounds for the transition probabilities, combining two, for the needed outcome by themselves potentially inadequate, sources of knowledge. We conduct a simulation study for a simplified but realistic case, showing the applicability and virtues of the new approaches and comparing them to conventional ones.

This paper is structured as follows: First, we discuss the epistemic background of the set-valued information and the basis of overall forecasting with transition probabilities.

Then, we introduce our two approaches based on interval-valued prior information and show the applicability and virtues with a simulation. In the concluding remarks we reflect on the approaches and possible further advancements in this particular field.

## 2. Methods

### 2.1. The Epistemic Interpretation of the Consideration Sets

The consideration sets  $\ell = \mathcal{P}(S)$  characterizing the undecided individuals' position in the pre-election poll, contains all the elements the undecided is still pondering between. Thus, it can be seen as a coarse version of the one true element  $l \in \ell$  contained in the set the individual ends up choosing, which is a particular interesting application of the theory about epistemic imprecision discussed by [Couso et al. \(2014\)](#). Hereby, the set-valued information is an imprecise version of something precise and only incomplete information about the phenomena of interest (the eventual choice) is provided by this consideration set. While we are looking for the random variable  $Y(\omega)$  mapping from an underlying space of the population  $\Omega$  to  $S$ , 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, where  $\mathcal{Y}$  is a multi-valued mapping  $\Omega \rightarrow \mathcal{P}(S)$  representing the set of mappings  $\{Y : \Omega \rightarrow S, Y(\omega) \in \mathcal{Y}(\omega) \forall \omega\}$ . ([Couso and Dubois, 2014](#), p. 1504) We therefore build an epistemic model of the random variable  $Y(\omega)$ , where for the undecided aside from covariates all that is known is  $Y(\omega) \in \ell$ .

Concerning forecasting one can either reflect the uncertainty of  $\ell$  within the final results in an interval-valued manner, or incorporate further information or presuppositions to obtain more concise or even point-valued results. Thus, one has to ponder between imprecise results and the justifiability of assumptions leading to more precise statements.<sup>1</sup> Facing this tradeoff, one sometimes has to comply with an external specification of the maximal degree of imprecision for the results to be usefully applicable. In this case one has to find the most credible approach to comply with the provisions. We can assess (imprecise) *transition probabilities*, as an (imprecise) probability distribution over the elements of  $l \in \ell$ , to obtain forecasts as will be discussed in the following paragraph.

### 2.2. Forecasts Incorporating Consideration Sets

Within the sample of the population in the pre-election poll, the individuals are characterized by one element of the power set  $\ell \in \mathcal{P}(S)$  and the values of the covariates in some space  $\mathcal{X}$ . Starting with the consideration sets and covariates of the  $i \in \{1, \dots, n\}$  participants, we want to estimate by

<sup>1</sup>. See also Manski's Law of Decreasing Credibility ([Manski, 2003](#), p. 1)

an i.i.d. sample the expected frequency of each element of  $S$  within the population, using the generic variables  $Y$  and  $\mathcal{Y}$ . The individual's consideration set from the pre-election survey is written as an event  $\{\mathcal{Y}_i = \ell\}$  with  $\ell_i \in \mathcal{P}(S)$  and the possibly unknown choice on election day  $\{Y = l\}$  with  $l \in S$ . Hence, we estimate the probability distribution  $P(Y = l) \forall l \in S$  over the singletons, which can be seen as a *multinomial distribution* over the state space with  $|S| - 1$  parameters. Hereby, the probability distribution can be factorized according to the chain rule into three parts like discussed in (Kreiss and Augustin, 2020, p. 244):

$$\begin{aligned}
 P(Y = l) &= \sum_{(\ell, x) \in (\mathcal{P}(S) \times \mathcal{X})} P(Y = l, \mathcal{Y} = \ell, X = x) \quad (1) \\
 &= \sum_{(\ell, x) \in (\mathcal{P}(S) \times \mathcal{X})} \underbrace{P(Y = l | \mathcal{Y} = \ell, X = x)}_{\text{Transition Probabilities}} \quad (2) \\
 &\quad \cdot \underbrace{P(\mathcal{Y} = \ell | X = x)}_{\text{Consideration Sets}} \cdot \underbrace{P(X = x)}_{\text{Covariates}} \quad (3)
 \end{aligned}$$

First, the *transition probabilities* determining the probability to vote for a specific party given the consideration set and covariates. Second, the probability of the consideration sets given the covariates and third, the one for the covariates. There are different approaches possible to estimate the second and third part of the factorization in (2) and (3). In this paper we focus on regression methodology. Even though one has to keep in mind sampling and modeling errors, there is sufficient information to estimate these factors right away with established procedures. Therefore, we treat these quantities as fixed and known in the sequel. The first part of the factorization on the other hand, reflects the epistemic problem previously discussed, as the value of  $l \in Y$  among the options of  $\ell$  is not observable for an undecided individual. For every decided individual, the transition probability is naturally one, as there is only one element to choose from, while for an undecided point- or interval-valued assessment is possible, allocating a specific range between 0 and 1 to every party in the consideration set. Hence, we concern ourselves with complex, non-stochastic, inherent uncertainty as there is no clear way to determine the resulting choice. The approaches suggested below distinguish themselves by the presuppositions and information utilized to estimate the transition probabilities from the undecided to the eventual choice, leading to overall different forecasts.

### 2.3. Approaches Incorporating Interval-Valued Prior Information

Due to the periodic nature of elections there is usually prior information about most properties available. This knowledge however is most of the times not precise, as despite overall continuity there are changes and no absolute certainty between the years. Nevertheless, there is usually at least some consistency as well as expertise, which is useful

to improve estimation. In order for this prior information to be credible it has to be provided in an imprecise manner, reflecting the inherent uncertainty interval-valued (e.g. Augustin et al. (2014)). Hereby, the information is the least imprecise version for which we are convinced that it is accurate. Prior information, which can either stem from (a number of) experts or estimated from data of the previous election(s), is commonly used to forecast elections (e.g. Linzer (2013)). Hence, credible imprecise prior information is a natural way to improve (imprecise) election forecasting.

In our case, we employ prior information about the transition probabilities within the groups of undecided, containing information about the choice probabilities. Hereby, for practicability and modeling purposes, we regard the binary case of probabilities about choosing a specific party against choosing a different one. Furthermore, in this paper we only regard explicit probability intervals, stating a possible range over the singletons only. This somewhat restrictive kind of information can be particularly easy provided by experts or known from previous studies. Within each group defined by  $\ell \in \mathcal{Y}$  we assume to be given imprecise knowledge about the probability that an individual chooses a given party, manifested in a probability interval. For example, we are certain that between 30% and 70% of the individuals undecided between the parties  $\{A, B, C\}$  will end up voting for  $A$ . We can now denote the prior information about party  $A$  in the group  $\{A, B, C\}$  as the interval between the extreme points, thus as  $pr_{A, \{A, B, C\}} = [pr_{A, \{A, B, C\}}^{lower}; pr_{A, \{A, B, C\}}^{upper}]$  or more generally as  $pr_{l, \ell} = [pr_{l, \ell}^{lower}; pr_{l, \ell}^{upper}]$ . Therefore, we work with imprecise knowledge manifested as an interval for each party in every group respectively. In some modeling cases it is possible to only regard the upper and lower bound of the interval for all information, while in others it is not. This sort of imprecise and credible prior information will now be deployed to improve forecasting incorporating the undecided.

#### Approach One

The **first approach** builds on the *Dempster bounds* (e.g. Dempster (1967)) like mentioned above, only using the information within the data and not relying on any assumptions nor further information. Therefore, these bounds are completely credible but also reflect the entire uncertainty, hence the best and worst case for every party as interval-valued results. Hereby, the Dempster bounds assign the transition probabilities the entire interval between 0 and 1 to every undecided individual, as this is the only way to reflect the attached uncertainty.

With the first approach we deploy the interval-valued prior information, in order to obtain more concise results. This is at the cost of the assumption that this prior information is accurate, which depending on the source in few



## ELECTORAL FORECASTING WITH INTERVAL-VALUED PRIOR KNOWLEDGE

cases might be disputable. Hereby, we narrow the transition probabilities interval from the original 0 and 1 to the interval-valued prior information. As the prior information in this case directly provides the minimum and maximum of the proportions, we do no longer have to rely on the entire interval between 0 and 1. Under the assumption that the prior information is indeed accurate the transition probabilities can be narrowed, inserting the new values leading to the new transition probabilities:

$$P(Y = l | \mathcal{Y} = \ell) = [pr_{l,\ell}^{lower} ; pr_{l,\ell}^{upper}] \quad (4)$$

Therefore, the new best and worst case are provided with the prior information. In the case of no prior information available, the transition probabilities once again take the whole range between 0 and 1, while for a decided individual they remain 1.

Due to the independence from the covariates, these transition probabilities can directly be inserted in equation (1), leading to overall imprecise forecasts.

Overall, this approach relies on the accuracy of the interval-valued prior information, but if we assume the prior information to be completely reliable we can narrow the bounds without losing any credibility at all. We hereby took a first step narrowing the Dempster bounds towards more concise results.

### Approach Two

Prior information is frequently very imprecise, leading to possibly vague forecasts if we expect it to be entirely true like in approach one. Thus, one may be forced to further narrow the bounds in order for the results to be usefully applicable. Hence, the interval-valued prior informations alone are not enough to obtain a desired level of conciseness and it is necessary to take further measures. As due to the inherent non-stochastic uncertainty more concise results are not evident, we suggest to include another source of information exploiting the information from the covariates. Hereby, we utilize the information contained in the covariates, by training a working model in order to obtain single-valued estimates on the individual level for a first step. We can assume some information about the eventual choice to be within the covariates, even though this information might not be entirely reliable, as is the case in our example in section 3. The resulting estimates can thus be seen as some sort of best guess of single-valued transition probabilities, containing the information about the covariates. Even though this single-valued guess by itself is presumably biased, it carries the valuable additional information of the covariates and provides a direction with which we can achieve narrower results. We thus suggest an approach to combine two sources of knowledge, which by themselves are deficient to obtain an adequate outcome, to find a compromise which meets the external criteria of conciseness. To achieve this, the interval-valued prior information is not seen infallible but also distributed with a

certain variance. With this variance we can determine the influence of the prior information on the overall results. We therefore still assume the prior information to be somewhat accurate, but unlike in the first approach it is not immediately used as the transition probabilities like in equation (4), but combined with the supplementary information.

We now suggest a Bayesian way to combine these sources of information in two steps. In the first step we describe why the predictions on the individual level based on a working model can be seen as i.i.d. data which within the second step can be combined with the prior information. To use the information structure of the covariates, we train a working model on supplementary data to predict the probability of a certain outcome on the individual level. This can either be achieved by regression or machine learning approaches like random forests, showing the working model training data for instance of previous elections or the decided. Proposals how to conduct this in the case of undecided voters reliant on different presuppositions are for example provided by [Kreiss et al. \(2020\)](#); [Kreiss and Augustin \(2020\)](#). We implement the working model below based on regression following equation (5). Hence, each undecided individual is assigned an explicit probability for each party in his or her consideration set by the working model, reflecting the information of the covariates. We therefore obtain the estimates in the form of a multinomial distribution between the elements of the consideration set for each individual. As the working model is trained on supplementary data, it is a deterministic function from the covariates to the predictions, preserving the i.i.d. structure of the sample. Thus, within each group of undecided we have identically and independently distributed observations, characterizing the transition probabilities based on the information of the covariates. We now suggest to regard these predictions as i.i.d. data themselves, following some unknown underlying distribution characterizing the transition probabilities for each group.

In the second step we propose a way to combine these – as data regarded – predictions with the interval-valued prior information. We hereby proceed for each party and group separately. As the working model provides us with estimates which sum up to one from an underlying multinomial distribution we can treat the parties separately, decomposing the multinomial distribution into party specific binomial ones.

Let's say we observe individuals for one specific group undecided between the three parties  $\{A, B, C\}$  and are interested in the proportion of individuals who end up choosing  $A$ . We want to obtain an upper and lower probability how likely an individual in group  $\{A, B, C\}$  ends up choosing  $A$ , combining both sources of information. From the working model we obtain a vector with probabilities how likely these individuals in group  $\{A, B, C\}$  choose party  $A$ , reflecting the information of the covariates. These

probabilities, as argued above, can be seen as i.i.d observations of an unknown underlying probability distribution. Furthermore, we have interval-valued prior information in the form  $[p_{A,\{A,B,C\}}^{lower} = 0.3 ; p_{A,\{A,B,C\}}^{upper} = 0.7]$ . To obtain a posteriori from these two sources of information we can either make distributional assumption about the i.i.d. observations of the supplementary data, understanding the prior knowledge as information about the parameters, or we can work with the empirical probability distribution. (Gelman et al., 2013, ch. 20-23) Either way, we obtain a posteriori combining the information of our two sources, which fulfills two purposes: One, the resulting information now incorporates the information of the covariates and two, the resulting interval gets narrowed as it is combined with single-valued information. The new upper and lower bounds can be seen as a compromise between the sources of information, within the space between the single-valued prediction and the wide bounds.

Not relying on a distributional assumption about the i.i.d. supplementary data, we can work with the empirical probability distribution. Hereby, we can deploy a Dirichlet process Ferguson (1973) with beta distributed priors (as we are in the binary case) characterizing the problem in a nonparametric way. A posteriori distribution can be determined this way, resulting in overall forecasting.

Otherwise we could make a reasonable distribution assumption about the i.i.d. data. One intelligible assumption would be a beta distribution, characterizing the resulting probabilities in a natural way. The distribution is defined over the parameters  $\alpha$  and  $\beta$ , where we rely on the parametrization:  $f_X(x : \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$ . We now want to submit the interval-valued prior information of the experts as knowledge about the parameters. There is a number of distributional assumptions possible to realize this, while we focus on a (truncated) normal distribution, as priories for the parameters  $\alpha$  and  $\beta$ . As the mean of the distribution is defined by  $\frac{\alpha}{\alpha+\beta}$  we choose priori values for  $\alpha$  and  $\beta$  to submit the desired mean. Hence, for a higher prior value we increase  $\alpha$  and decrease  $\beta$  within the prior knowledge about the parameters. This can for example be achieved by demanding the mean values of the parameters to sum up to a given value, ensuring one decreases while the other increases. The parameters can hereby still be simulated independently. The advantage of a (truncated) normal distribution are both its stability concerning sampling as well as the intuitive and explicit variance parameter. Furthermore, we keep the variance parameter constant over the entire interval constituting the prior knowledge. From this the convenient attribute results, that we only have to consider the extreme points of the interval rather than also the entire points in between. As all values within the interval are truncated normal distributed

with identical variance parameter, higher ones hold first order stochastic dominance over lower ones. As the expectation of the posteriori monotonically increases in  $\alpha$  and decreases in  $\beta$  and higher priori values lead to higher  $\alpha$  and lower  $\beta$ , the extreme points of the interval produce all of the interjacent values for the result. This simplifies the process, only demanding two separate models, one for the upper and one for the lower bound. With a high variance parameter we indicate low belief in our prior as with a variance parameter close to zero the priori almost determines the results. A beta distributed prior over the parameters is possible as well, but might not be as intuitive and computationally feasible. As those priors are non-conjugated, we can deploy a MCMC process, drawing samples from the posteriori for estimation. (Gelman et al., 2013, ch. 12). Such a process can for example be easily implemented with *Rstan* (Stan Development Team (2020)).

The resulting upper and lower bounds for the party wise calculated transition probabilities can be directly deployed to calculate overall forecasting with equation (1). We hence obtain overall forecasts using both sources of knowledge to make the most of the information about the undecided. The results are therefore between the extreme points of the initial bounds and the single-valued estimator exploiting the information in the covariates. We can regard this approach, narrowing the initial wide intervals, as a pragmatic attempt. On the other hand, we could also see the process as a regularization of the single-valued estimates towards more credible bounds. In both cases we suggest a tradeoff between, and a combination of, two sources of information. By adjusting the variance parameter, we can determine the influence of the prior information as well as the conciseness of the overall outcome. Hence, the variance parameter effects the accuracy-precision tradeoff, laying more emphasis on the one, or the other information. But as due to the epistemic nature of the problem, we do not know how accurate the estimates based on the covariates are, it is difficult to give general statements about the accuracy-precision tradeoff. Therefore, we described the process as taking steps in hopefully the right direction to make the results as concise as necessary, but do not really generalize how accurate the results are, as this differs from case to case.

In our case we further assume the interval-valued prior information to be overall accurate. But in different applications one would have to be aware of a possible bias from this source as well.

### 3. Simulation Study and Further Details

#### 3.1. Specifying the Simulation

To illustrate the applicability and virtues of the two new approaches we conduct a simulation study, comparing them

## ELECTORAL FORECASTING WITH INTERVAL-VALUED PRIOR KNOWLEDGE

to the Dempster bounds as the most cautious and the point-valued approach neglecting the undecided overall. We consider a scenario in which three parties  $\{A; B; C\}$  can be chosen at the election. Thus, within the pre-election poll, there are three groups of undecided voters  $\{A, B\}; \{B, C\}; \{A, C\}$  resulting in overall six options including the decided voters. We choose a realistic sample size of 1000 individuals and ensure them to be an i.i.d. representation of the underlying truth.

For the first step we draw the individuals from a multinomial distribution describing the proportion of the groups within the population at the pre-election poll. We specified the parameters of this distribution  $P(\mathcal{Y} = \ell)$  as follows:

$$\{p_{\{A\}} = 0.45; p_{\{B\}} = 0.2; p_{\{C\}} = 0.1, \\ p_{\{A,B\}} = 0.15; p_{\{A,C\}} = 0.05; p_{\{B,C\}} = 0.05\}$$

From the resulting data we specify the true transition probabilities with which we simulate the eventual choice of the undecided.

$$\{p(Y = A | \mathcal{Y} = \{A, B\}) = 0.5; \\ p(Y = A | \mathcal{Y} = \{A, C\}) = 0.9; \\ p(Y = B | \mathcal{Y} = \{B, C\}) = 0.5\}$$

This determines the true outcome in our underlying population to be:

$$\{p_A = 0.57; p_B = 0.30; p_C = 0.13\}$$

Furthermore, to mimic the exploitation of the information of the external data, we simulate in addition a covariate which is somehow correlated with the eventual choice. This continuous covariate therefore varies within the respective groups and eventual choices. We hereby use a normal distribution, which contains some, but biased information about the eventual choice. This resembles the realistic scenario in which covariates contain valuable information, but which is only by itself not adequate to produce a reliable prognosis. The variance is fixed within all groups resulting in the parameters in table (3.1). We thus obtain a simplified but

Choice	A	B	C	A	B	B	C	A	C
Set	A	B	C	A/B	A/B	B/C	B/C	A/C	A/C
$\mu$	70	50	30	65	55	55	30	65	30
$\sigma^2$	20	20	20	20	20	20	20	20	20

Table 1: Parameters of the normal distribution of the continuous covariate amongst the different groups and parties

realistic sample with which we can estimate transition probabilities in order to obtain overall forecasts with equation (1).

Furthermore, we have imprecise prior knowledge in the form of probability intervals about the transition

A	Minimum	Maximum	B	Minimum	Maximum
A/B	0.3	0.7	B/A	0.3	0.7
A/C	0.6	0.9	B/C	0.4	0.6

Table 2: Prior knowledge about the probability to choose party A on the left, and to choose party B on the right sight, depending on the underlying groups.

probabilities from an expert illustrated in table (3.1). The prior knowledge concerning party C results from the probabilities of the complements. The interval-valued information is hereby wide enough to be realistically credible, as we expect expert to provide somewhat accurate information with confidence. Hence, the prior information in this case satisfies the realistic criteria to be accurate, but is very imprecise.

Provided with these samples and prior knowledge we can now apply the two approaches discussed above and compare them with the Dempster bounds and the conventional approach.

### 3.2. Applying the Approaches

We simulate and apply the approaches multiple times (50) and average over the results. Hereby, we first calculate the transition probabilities for all three approaches and determine the overall forecasting together with the decided according to equation (1). As we have a representative sample there are means to estimate the second and third part of the factorization in equation (1), we choose a logistic regression approach, estimating the conditional distribution. The prediction resulting from equation (5) is estimated via logistic regression as well. Within our approach we primarily focus on the non-stochastic, complex inherent uncertainty, not elaborating on the sampling and modeling errors induced and treat the estimated quantities as fixed. Nevertheless, an overview of variation between the different samples is provided within the appendix A.

As mentioned above, we regard the parties and groups separately. Thus, for every party in every group we are supplied with a vector of identically independently distributed probabilities from a working model utilizing the simulated covariate here. To achieve this, we follow (Kreiss and Augustin, 2020, p. 245) with a working model identical to the presupposition that the undecided choose identical to the decided given their covariates and consideration sets. The transition probabilities on an individual level are hereby predicted resulting in

$$\hat{P}(Y = l | \mathcal{Y} = \ell, X = x) = \frac{\hat{P}(Y = l | X = x, I_d = 1)}{\sum_{a \in \ell} \hat{P}(Y = a | X = x, I_d = 1)} \quad (5)$$

## ELECTORAL FORECASTING WITH INTERVAL-VALUED PRIOR KNOWLEDGE

with  $I_d$  as the indicator function for being decided.<sup>2</sup> This resulting best guess is now incorporated in a Bayesian way as discussed above. In our application we approximate these realizations in a natural way with a beta distribution. For feasible estimation of the parameters based on the i.i.d. data we need to specify the possible range of  $\alpha$  and  $\beta$  setting it to  $[0, 10]$ . Then, we incorporate the prior knowledge as information about the parameters  $\alpha$  and  $\beta$ . To this end we choose a, strictly speaking truncated, normal distribution, only taking values in the possible range of  $\alpha$  and  $\beta$ . With this (truncated) normal distribution we now specify the prior knowledge about  $\alpha$  and  $\beta$ . The  $\alpha$  and  $\beta$  parameters are hereby simulated independently. To increase one while decreasing the other parameter we demand their expectation to sum up to one. With this we can directly apply the overall expectation of the prior information. Furthermore, it is possible to only regard the upper and lower bounds of the interval, due to the constant variance parameter and the following properties of the (truncated) normal distribution concerning first order stochastic dominance. As an example, following the logic of above, the knowledge  $[p_{A,\{A,B\}}^{upper} = 0.7]$  is transferred into  $\alpha \sim \text{Normal}^*(0.7, \sigma^2); \beta \sim \text{Normal}^*(0.3, \sigma^2)$  with a (truncated) normal distribution controlling the strength of the prior knowledge with the variance parameter. The submitted knowledge therefore constitutes the targeted mean of  $\frac{0.7}{0.7+0.3} = 0.7$ . To give the priori reasonable weight we choose the variance parameter as  $\sigma^2 = 0.05$ . This precise value is admittedly chosen somewhat arbitrary as a subjective consideration concerning the accuracy-precision tradeoff. The approach for the upper bound with this specific prior knowledge can thus be written in a hierarchical way as:

$$\begin{aligned} \alpha, \beta &\in [0, 10] \\ \alpha &\sim N^*(0.7, 0.05) \\ \beta &\sim N^*(0.3, 0.05) \\ \text{Likelihood} &: \text{Beta}(\alpha, \beta) \end{aligned}$$

The posteriori is calculated over a MCMC process implemented with *RStan* [Stan Development Team \(2020\)](#). From the expectation of the posteriori we obtain the overall estimate for the upper bound of the transition probability towards the specific party in the specific group. Hereby, we do not make a parametric assumption about the posteriori distribution but merely take the Monte Carlo expectation of the parameters. This process is repeated for each constellation of upper and lower bounds, groups and parties, which results in transition probabilities and overall forecasts.

2. There is a connection to the work of [Heitjan and Rubin \(1991\)](#), even though the assumption is somewhat different to Coarsening at Random.

3. Normal\* or N\* stands for the truncated normal distribution

The results of all approaches, additionally with the true parameters and the estimate reliant only on the decided are illustrated in figure (3.2). At the left upper

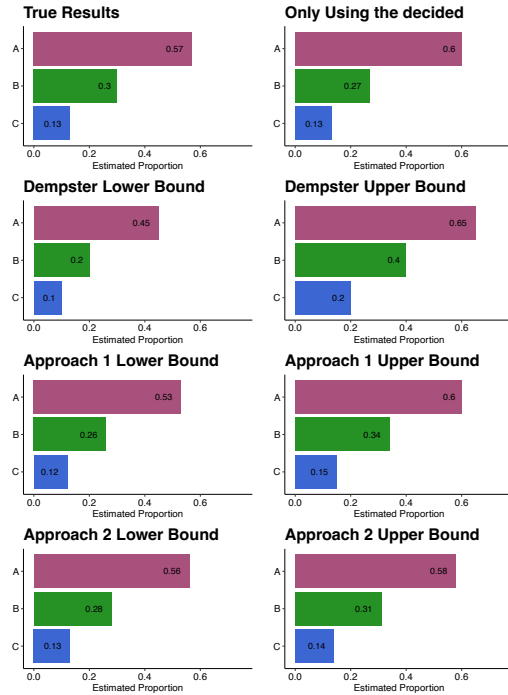


Figure 1: Mean of the result from 50 samples of  $n = 1000$  for three approaches, alongside with the true values and the estimate neglecting the undecided overall in the first row. The variance parameter of the prior knowledge for approach 2 was set to 0.05.

corner we can see the true values the simulation is based on. On the upper right we can examine that the estimate overall neglecting the undecided like in conventional approaches leads to biased results.<sup>4</sup> While the Dempster upper and lower bounds in row two are very wide, they are substantially narrowed with the prior information in approach 1 in the third row. Finally, the second approach in the fourth row ensures the narrowest bounds, incorporating the two sources of information.

The approaches in rows two to four are all possible solutions of the tradeoff between precision of the results and credibility. The Dempster bounds are hereby the most credible but wide results, while approach 2 has the narrowed

4. Whenever in this case the undecided are *missing not at random*, as can usually be expected, such approaches end up biased.

bounds incorporating the information of the covariates. From a practical point of view, we suggest to use the most credible approach still satisfying the necessary criteria of conciseness. Decreasing the variance parameter of the prior within approach 2 would lead to even more concise results. With our realistically chosen parameters all three approaches overlap the true values, emphasizing the credibility of the approaches. The variation of the results reliant on different draws from the simulation are illustrated in the appendix A with a box plot. We can see that the dispersion is not too severe between the results of the different datasets.

Despite the desirable traits of approach 2, it is somewhat complicated to evaluate the accuracy, as it results from a combination of multiple sources of information. One has to choose whether the information provided by the covariates outweighs the potential bias introduced, which in this simplified but realistic scenario is definitively the case. The credibility furthermore depends on how the two sources of information are weighted. Within the simulation the prior information is accurate but quite imprecise. Examining different scenarios lead us to believe that small bias in the prior information does not effect the results severely. Approach two is definitively a strong tool to narrow the initial bounds, which with reasonable weighting of the sources of information should still overlap the true value, as shown in the exemplary simulation.

#### 4. Concluding Remarks

Within this paper we introduced two approaches incorporating interval-valued prior knowledge in order to improve election forecasting including undecided voters. The first one provides narrower bounds in a straight forward manner, only reliant on accurate prior information. Narrowing these bounds further in a credible way is far more complicated, and we address this problem by including further information making use of the informations in the covariates in a Bayesian way. Hereby, we suggest and apply first methodology, regarding the single-valued predictions on an individual level by the covariates as i.i.d. data. The results are in between the initial bounds and the single-valued predictions, incorporating both sources of information. The first results are promising, achieving narrower bounds in a plausible way.

For further research following this train of thought, one could determine the variance parameters for the prior knowledge by demanding a specific precision of the resulting overall forecasts. This can be implemented recursively, increasing or decreasing the variance parameter to obtain more, or less concise overall results. With the extreme points we get the initial bounds or the point-valued estimate of the transition probabilities. This would be one way to explicitly account for the tradeoff between credibility

and precision of the results (e.g. Manski (2003)) in the second approach. Furthermore, highlighting the implications of biased prior information in this context is interesting.

Additionally, there are plenty of directions possible to address undecided voters. For example, regarding the distributions overall, not decomposing it into binary cases and deploying Dirichlet processes for Bayesian modeling is interesting. Furthermore, we could combine the initial bounds with imprecise estimates based on the covariates. The same basic concepts apply, while the procedure is a little more complicated as more combinations arise. Also highlighting connections and differences to other approaches combining evidence, in particular the Dempster-Shafer theory of evidence Denoeux (2016) is interesting for further research.

Within this work we solely focussed on election systems in which the individual casts one vote like common in Europe. Instant-runoff-voting and different ranking voting systems are worthy of exploring further on. Some of the thoughts above can be adopted, but due to different ranking approaches the structure of the underlying state space and with it the methodology changes.

Overall, considering one source of information as i.i.d. data for Bayesian modeling has proven to be a useful measure to combine two, by themselves inadequate, sources of information. This basic concept could be transferred to multiple different applications and is especially useful concerning undecided voters in pre-election polls.

**Acknowledgement.** We are very grateful for the valuable and supportive remarks from the three anonymous reviewers. We further sincerely thank the LMU Mentoring Program, supporting young researchers. Furthermore, we would like to thank Malte Nalenz for the helpful discussions concerning RStan and Bayesian modeling.

#### References

- T. Augustin, G. Walter, and F. Coolen. Statistical inference. In T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 135–189. Wiley, 2014.
- Inés Couso and Didier Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning*, 55(7): 1502–1518, 2014.
- Inés Couso, Didier Dubois, and Luciano Sánchez. *Random sets and random fuzzy sets as ill-perceived random variables*. Springer, 2014.
- Arthur Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

Thierry Denoeux. 40 years of Dempster–Shafer theory. *International Journal of Approximate Reasoning*, 79: 1–6, 2016.

Thomas Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Daniel Heitjan and Donald Rubin. Ignorability and coarse data. *The Annals of Statistics*, pages 2244–2253, 1991.

Dominik Kreiss and Thomas Augustin. Undecided voters as set-valued information, towards forecasts under epistemic imprecision. In Jesse Davis and Karim Tabia, editors, *Scalable Uncertainty Management 2020*, pages 242–250. Springer, 2020.

Dominik Kreiss, Malte Nalenz, and Thomas Augustin. Undecided voters as set-valued information, machine learning approaches under complex uncertainty. In Eyke Huellermeier and Sebastian Destercke, editors, *ECML/PKDD 2020 Tutorial and Workshop on Uncertainty in Machine Learning*. 2020. URL <https://drive.google.com/file/d/1abrLGZ154htGuYz8HzylQzJ8vyc3kr2K/view>.

Drew Linzer. Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, 108(501):124–134, 2013.

Charles Manski. *Partial identification of probability distributions*. Springer, 2003.

Henrik Oscarsson and Martin Rosema. Consideration set models of electoral choice: Theory, method, and application. *Electoral Studies*, 57:256–262, 2019.

Julia Plass, Paul Fink, Norbert Schoening, and Thomas Augustin. Statistical modelling in surveys without neglecting ‘The undecided’. In Thomas Augustin, Serena Doria, Enrique Miranda, and Erik Quaeghebeur, editors, *ISIPTA 15*, pages 257–266. SIPTA, 2015.

Stan Development Team. RStan: the R interface to Stan, 2020. URL <http://mc-stan.org/>. R package version 2.21.2.

## Appendix A. Boxplot of the simulations

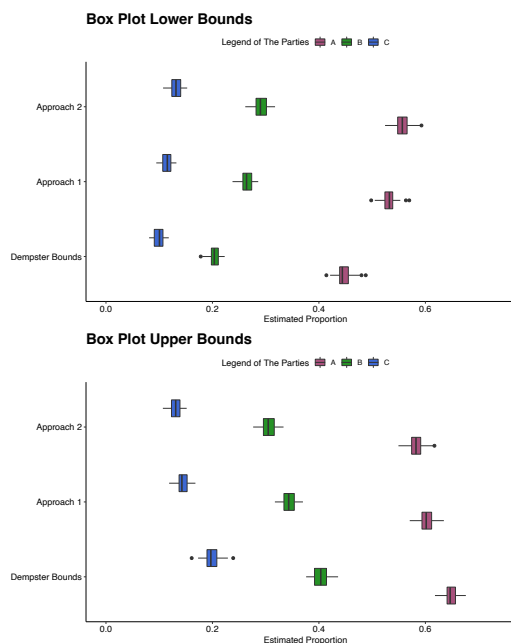


Figure 2: Box Plots illustrating the results from the different simulation iterations.

## Contribution 4

Kreiss, D., Augustin, T.: Towards a Paradigmatic Shift in Pre-Election Polling Adequately Including Still Undecided Voters – Some Ideas Based on Set-Valued Data for the 2021 German Federal Election. *arXiv preprint* (2021)  
<https://doi.org/10.48550/arXiv.2109.12069>

---


# TOWARDS A PARADIGMATIC SHIFT IN PRE-ELECTION POLLING ADEQUATELY INCLUDING STILL UNDECIDED VOTERS – SOME IDEAS BASED ON SET-VALUED DATA FOR THE 2021 GERMAN FEDERAL ELECTION

---

TECHNICAL REPORT

 **Dominik Kreiss**  
Department of Statistics  
LMU Munich

dominik.kreiss@stat.uni-muenchen.de

 **Thomas Augustin**  
Department of Statistics  
LMU Munich

thomas.augustin@stat.uni-muenchen.de

September 27, 2021

## ABSTRACT

Within this paper we develop and apply new methodology adequately including undecided voters for the 2021 German federal election. Due to a cooperation with the polling institute Civey, we are in the fortunate position to obtain data in which undecided voters can state all the options they are still pondering between. In contrast to conventional polls, forcing the undecided to either state a single party or to drop out, this design allows the undecided to provide their current position in an accurate and precise way. The resulting set-valued information can be used to examine structural properties of groups undecided between specific parties as well as to improve election forecasting. For forecasting, this partial information provides valuable additional knowledge, and the uncertainty induced by the participants' ambiguity can be conveyed within interval-valued results. Turning to coalitions of parties, which is in the core of the current public discussion in Germany, some of this uncertainty can be dissolved as the undecided provide precise information on corresponding coalitions. We show structural differences between the decided and undecided with discrete choice models as well as elaborate the discrepancy between the conventional approach and our new ones including the undecided. Our cautious analysis further demonstrates that in most cases the undecideds' eventual decisions are pivotal which coalitions could hold a majority of seats. Overall, accounting for the populations' ambiguity leads to more credible results and paints a more holistic picture of the political landscape, pathing the way for a possible paradigmatic shift concerning the adequate inclusion of undecided voters in pre-election polls.

**Keywords** Undecided Voters · Set-Valued Data · Election Forecasting · Epistemic Imprecision · Ontic Imprecision · Random Sets · Voting Research · Partial Identification · Dempster Bounds · Consideration Sets · Ambiguity of Choice · Questionnaire Design

## 1 Introduction

As tough choices usually demand a consideration stage, several individuals can not state a precise intent which party to vote for in pre-election polls. These undecided voters, still pondering between options, induce a new source of uncertainty going beyond the common survey error. This is especially visible before this years German federal election, as the amount of indecisiveness seems to have reached a peak and conventional forecasts for specific parties skyrocketed and plunged in short periods of time. As conventional polls force undecided individuals to either state a single party choice or to drop out, this ambiguity within the population is not represented in resulting forecasts and other analysis. To face this issue, we suggest to provide undecided voters with the option to state all the parties he or she is still pondering between, hence accurately providing their current position set-valued. This way of regarding undecided voters yields



several advantages: Stepwise exclusion of options until arriving at the final element is a natural human decision process (see f.e. [Oscarsson and Rosema, 2019, p. 256]). Thus, participants can intuitively provide the set-valued information. Furthermore, concerning forecasting, this valuable partial knowledge from the undecided is preferable to wasting it overall. The exclusion further makes the implicit assumption that undecided voters do not structurally differ, which is highly questionable. Additionally, new insight into properties of groups undecided between specific parties can be analyzed using the set-valued data. And last, there is a rich theoretical groundwork laid how to utilize this set-valued data as well as adequately regarding the uncertainty attached in interval-valued results.

We introduced some ideas and methodology how to utilize this information in our foregoing works [Kreiss and Augustin, 2020], [Kreiss et al., 2020] and [Kreiss et al., 2021] as well as we build on previous provisional ideas in the direction of characterizing the undecided set-valued (f.e. [Oscarsson and Oskarson, 2019] and [Plass et al., 2015]). The resulting set-valued information can be interpreted in two ways, dependent on the question at hand. First, focusing on forecasting, a set of choices can be seen as a coarse version of one true but at the time unknown element contained in the set, providing incomplete information on the later choice. Following [Couso and Dubois, 2014], this is the so-called *epistemic* (or disjunctive) view. Second, focusing on the analysis of structural properties, the set is understood as representing 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 put to use, dealing with complementary issues.

With the ontic approach, regarding the undecided between specific parties as positions of their own, we examine new structural properties concerning the political landscape, using regularized Discrete Choice Models. For the epistemic view, we apply self-developed forecasting approaches weighting the justifiability of assumptions with the precision of the results.<sup>1</sup> We both provide point-valued forecasts, as well as interval-valued ones, reflecting the ambiguity of the undecided within the final results. Forecasting the proportion of votes for specific coalitions plays hereby an interesting role, as this ambiguity is reduced automatically in the process: indecisiveness between certain parties induces a precise vote for coalitions containing those parties.

The polling institute Civey generously provided us with a first custom made advanced pre-election poll regarding the undecided voters set-valued. This gives us the opportunity of direct implementation of our methodology developed for the 2021 German federal election. From the Civey survey we obtain data in three different waves, each providing a stand alone sample for a given point in time. With this novel type of data, we first take a good look at the undecided voters, analyzing structural properties and connections to socioeconomic variables with discrete choice models. Subsequently, we give our election forecasts, utilizing the newly obtained valuable information of the undecided, also reflecting the ambiguity resulting from the inherent complex uncertainty within interval-valued results. Furthermore, we analyze coalitions in which the uncertainty is *eo ipso* reduced in a natural way.

In more detail, this paper is structured as follows. After discussing the implementation and sketching the theoretical background of the survey and the emerging data in chapter 2, we take a detailed look at the set-valued data and connections to socioeconomic variables in chapter 3. In chapter 4 we then focus on the election forecasting utilizing the information of the undecided. Further possibilities and challenges of the approaches are discussed in chapter 5. The main text presents the empirical results and gets along with an informal description of our methodology; all technical notation and mathematical background is put into boxes and can thus be easily skipped or enjoyed according to the readers' preference.

## 2 Set-Valued Data Characterizing Undecided Voters

We are provided with three different stand alone waves of data with a sample size around 5000 observations each. The first wave is conducted two months, the second one month and the third one week before the election. Within each poll, the participants are first asked whether or not they are certain about their election choice. Those not certain were then asked for all the parties they are still considering for their choice, while for the others, the poll with the selected single party is used. Hence, for all participants we are provided with the set of parties he or she is still pondering between – in the case of a decided consisting of one, in the case of an undecided of several parties. Thus, every participant can provide their current position both accurately and as precisely as he or she is capable of.

Civey strives for a representative sample in each wave with the use of a quota sample from an initial selection as well as weighting the individuals bases on covariates.<sup>2</sup> Hence, within all our analysis we rely on this sample provided by the polling institute, containing the set-valued response option. We are aware that no voluntary poll is beyond some survey

<sup>1</sup>See also see Manski's Law of Decreasing Probability [Manski, 2003, p. 1].

<sup>2</sup>More on the methodology of Civey to obtain a representative sample can be found in [Richter et al., n.d.]

error, induced by randomness of choice and structural nonresponse patterns. To at least ease some of the errors induced by the structural response patterns, we employ weighting provided by Civey.

This paper predominantly concerns itself with the last wave, closest to the election, while thoughts on individual changes of opinion and the exclusion process by the individuals will be discussed in further works succeeding this paper. The corresponding results for the second wave are shown in the appendix, while the first wave is not covered as there are no weights available.

Furthermore, we only focus on the six main parties in Germany likely to surpass the 5% hurdle, (typically) necessary to win seats in the parliament. Germany has a rather complex, proportional voting, multi-party election system, in which we only focus on the proportion of seats won in the parliament, which almost certainly will be split between those parties.<sup>3</sup> We are aware that this does not completely do justice to the rather complex German voting system, but this simplification is commonly used in election polling and still conveys the important messages.

Overall, we primarily see ourselves as providers of new methodology, introducing ideas including undecided voters with set-valued data in pre-election polls, and hence will not contemplate lengthy about politological interpretations.

In the third wave 533 of overall 4730 individuals are still undecided and provided set-valued answers, while for the second wave 837 of 5001 and in the first 1311 of 5076 individuals were still pondering between options. This decrease of undecided individuals is logical, as closer to the election more and more people make up their mind. This trend is further supported by the current situation in which a high proportion of individuals votes by post and thus might have already voted at the point in time the poll is conducted. With this data we can see that including the undecided is more important the farther away the election is, but even immediately before still more than 10% did not make up their minds. The 15 biggest groups on these individuals from the third wave are shown in figure 1.

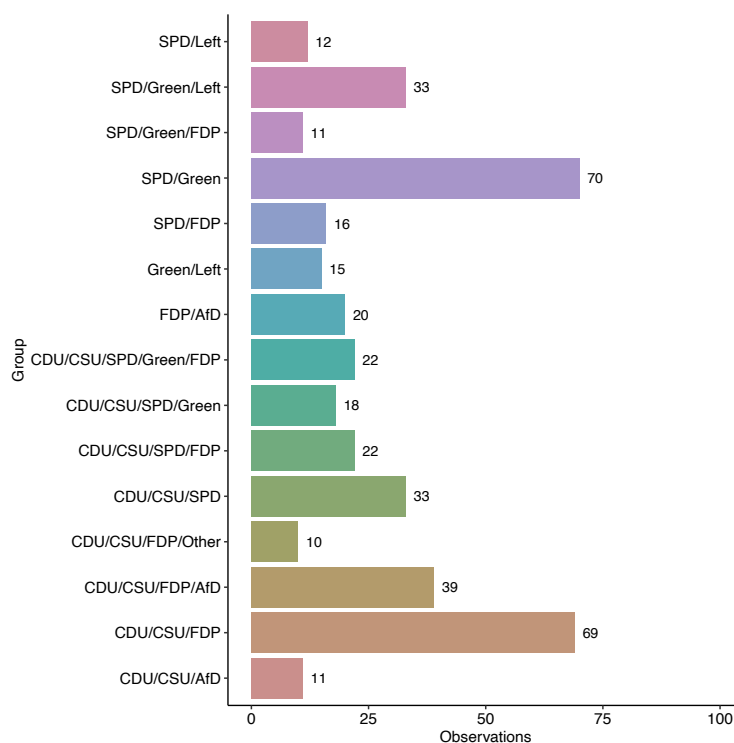


Figure 1: Numbers of observations for the 15 biggest groups of individuals undecided between specific parties

<sup>3</sup>For more information about the German voting system see: <https://www.bundeswahlleiter.de/bundestagswahlen/2021/informationen-waehler/wahlssystem.html>, last visited 22.09.21

As we can see, most individuals are pondering between two parties and only very few between more. The biggest group in this wave is undecided between the two, closely associated Parties SPD and Green party, immediately followed by the two parties on the other side of the spectrum CDU/CSU<sup>4</sup> and FDP.

For the notation, the state space of the consideration sets consists of all possible combinations of the original options, which can naturally be represented by the power set  $P(S)$  of the set  $S$  of the original options. Hence, in the case of an undecided, we observe a set  $\ell$  that can be described as the realization of a measurable mapping  $\mathcal{U} : \Omega \rightarrow P(S)$  from some underlying space  $\Omega$  into the set of all combinations. The ontic view sees the set-valued data as a non-reducible entity of its own, characterizing a specific political position, while the epistemic view interprets it as a collection of elements within which the true value lays.

### 3 Analyzing Groups of Undecided Voters – Ontic Approaches

Within this chapter, we focus on the individuals' position at the point in time of the poll one week before the election. As argued above, at this given point in time, an undecided individual's position is best characterized by the set of parties he or she is still pondering between. This set cannot be reduced or improved in any way and hence is the most accurate information an undecided is capable to provide. Hereby, each set is one viable position of its own, equal to the decided individuals with only one party in their consideration sets. In other words, the set is a precise representation of something naturally imprecise, and this is called the ontic view.

Following the notation of above, the elements of the set  $\mathcal{U}$  can be understood as the most suitable operationalization of the individuals' political position. As  $S$  is a finite, not ordered, discrete space,  $P(S)$  satisfies the same basic mathematical principles as the original choice set, and  $\mathcal{U}$  can be treated as any other discrete random element.

Provided with the individuals' positions we want to examine these groups, in order to find interesting and new insights into the political landscape, gaining information about the undecided. To this end, we examine relationships between socioeconomic variables and the different groups undecided between specific parties with *Discrete Choice Models*. With these models, characteristics of interesting groups can be determined, providing a new opportunity to gain empirically founded insights about undecided voters. Such information is compelling not only to the involved parties but also from a sociological and political science point of view.

Concretely, we use the method described in [Tutz et al., 2015] and implemented in the R package *MRSP* based on it, in order to perform state of the art regularized choice modeling. Further reading on Discrete Choice Models and regularization can be found in [Tutz, 2011, ch. 8] and first application with set-valued data in the election context in [Kreiss, 2019]. Fortunately, this established methodology can hereby be directly transferred to the set-valued data, as the new state space satisfies the same mathematical properties as the original one. The modeling is conducted including the five groups of undecided voters with the most observations illustrated in figure 1, together with the already decided individuals.

Further, we use the five independent variables *sex*, *age*, *resident of former east or west Germany*, *purchasing power* and *population density* from the data. All variables are regarded binary in order to avoid trouble with perfect separation and limited degrees of freedom. For the model, we chose a symmetric constraint, losing one degree of freedom for better interpretability. The results do hereby not rely on a reference category, but are interpreted in contrast to the data itself. Furthermore, we use a Categorically Structured Lasso with group penalties and cross-validation to determine  $\lambda$ . More on this topic can be found in [Tutz et al., 2015, p. 209 ff.].

The results of the regularization are shown in figure 2, and the estimates are illustrated in table 1. Due to their direct connection to the target variable, none of the covariates is regularized exactly to zero.

With our ontic model we are able to determine new insights, analyzing structural connections with undecided voters and socioeconomic variables. Hereby, the groups undecided between given parties are often very different from the respective single parties, showing structural differences between the undecided and decided. As an example we see within our model that with an age over 65 the chance to choose the category SPD/Green decreases rapidly in contrast to the categories containing the single parties SPD or Green. Such findings are only possible including the undecided and stress the importance of properly doing so. As differences to the conventional model are apparent, the necessity of the new approach for forecasting is confirmed. This more differentiated and accurate approach furthermore provides more detailed information by including the groups undecided between specific parties.

<sup>4</sup>As common in German election polling, the CDU/CSU is treated as a single party, because, depending on the place of residence, one can vote either only for the CSU or only for the CDU.

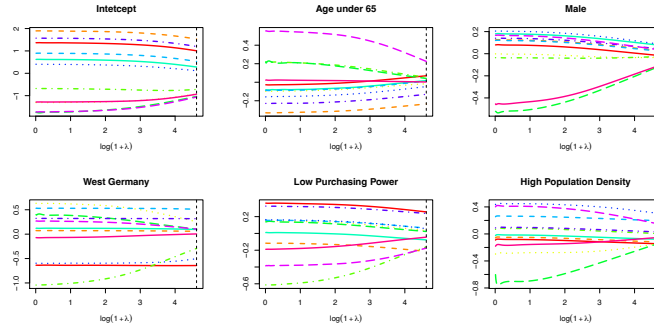


Figure 2: Illustration of the regularization results conducted to the model. None of the variables are exactly reduced to zero

Ontic Groups	Age under 65	Male	West Germany	Low Purchasing Power	High Population Density
AfD	0,069	-0,013	-0,641	0,255	-0,141
CDU/CSU	-0,233	0,033	0,057	-0,213	-0,118
CDU/CSU/FDP	-0,017	-0,032	0,25	0,035	-0,143
CDU/CSU/FDP/AfD	0,054	-0,026	-0,294	-0,162	0,008
CDU/CSU/SPD	0,045	-0,129	0,105	0,024	-0,157
FDP	0,018	0,08	0,095	-0,076	-0,075
Green	0,015	0,027	0,509	0,064	0,198
Left	-0,046	0,089	-0,502	0,056	0,299
SPD	-0,13	0,048	0,313	0,237	0,032
SPD/Green	0,222	0,04	0,103	-0,174	0,15
SPD/Green/Left	0,002	-0,117	0,004	-0,045	-0,053

Table 1: Estimates for the regularized discrete choice model conducted with the six main parties as well as the five biggest groups of individuals undecided.

## 4 Forecasting Utilizing the Undecided – Epistemic Approaches

In this chapter we utilize the set-valued information of the undecided voters within forecasting for two purposes: First, not losing out on this valuable partial knowledge about party preferences, and second, communicating the uncertainty which results from the individuals' ambiguity extending beyond the usual survey error. To achieve this, after briefly describing the methodological framework and calculating the conventional approach, we start off with an intuitive one, providing point-valued estimates exploiting the partial information together with the covariates reliant on a rather strong assumption. Afterwards, we show how the ambiguity affects the precision of the results if no, or weak assumptions on the undecideds' eventual choice are made. The interval-valued ideas are also deployed for coalitions, resolving some of the ambiguity due to the fact that the members of a coalition are considered together. Further thoughts on how to narrow the intervals with some quite plausible assumptions are realized later on.

### 4.1 Methodological Framework

The epistemic approach, as discussed in the introduction, concerns itself with the yet unknown element in the consideration set the individual ends up voting for. In contrast to the ontic view, we hereby have imprecise information about something precise (the eventual choice) in the form of a set. To obtain statements about the precise values of interest, one would need perfect external information about the (outcome of the) eventual individual decision processes. Assumptions about these processes have to be made with greatest care and must be founded well on external knowledge: Such assumptions can shown to be eo ipso not testable by any statistical test and thus, even if they are misleading, as a matter of principle, are not refutable by the data. Thus, making assumptions motivated solely by mathematical convenience or for the sake of ease of interpretation may substantially jeopardize the relevance of the

results achieved. Avoiding spurious precision by a careful reflection of all the uncertainty involved and communicating it by interval-valued results shall become good scientific practice. (e.g. [Manski, 2015]) Implicitly, our development here is grounded on the general methodological frameworks of partial identification (e.g. [Manski, 2003]) and imprecise probabilities (e.g. [Augustin et al., 2014]), handling complex uncertainty by considering the set of all traditional models compatible with the data and additional information as the basic entity.

In our case, 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  $\Omega \rightarrow P(S)$  now representing the set of mappings  $\{Y : \Omega \rightarrow S, \forall \omega, Y(\omega) \in \mathcal{Y}(\omega)\}$ , where we assume one of each is the true underlying mapping (e.g. [Couso and Dubois, 2014, p. 1504]).

To obtain overall forecasting, the distribution can conveniently be factorized into three parts: First, the from now on so-called *transition probabilities*, determining the probability to vote for a specific party given the consideration set and co-variables. Second, the probability of the consideration sets given the co-variables and third, the one for the co-variables. For more information see [Kreiss and Augustin, 2020].

Each individual from the sample is determined by both its consideration set  $\ell \in \mathcal{P}(S)$  and its co-variables  $X = x$  in some space  $\mathcal{X}$ , assessing their personal characteristics. The individual's consideration set from the pre-election survey is written as an event  $\{\mathcal{Y} = \ell\}$  with  $\ell \in \mathcal{P}(S)$  and his or her possibly unknown choice on election day  $\{Y = l\}$  with  $l \in S$ . Given the consideration sets of participant  $i \in \{1, \dots, n\}$  in the pre-election poll, we want to obtain the expected frequency of each element of  $S$  within the population, with latent probability distribution  $P(Y = l)$  for all  $l \in S$ , which is a multinomial distribution over the state space with  $|S| - 1$  parameters. The observations  $Y_i$  are assumed to be identically and independently distributed copies of the generic variable  $Y$ , and  $P(Y = l)$  can be written in respect to the consideration sets and co-variables as

$$P(Y = l) = \sum_{(\ell, x) \in (2^S \times \mathcal{X})} P(Y = l, \mathcal{Y} = \ell, X = x) = \quad (1)$$

$$\sum_{(\ell, x) \in (2^S \times \mathcal{X})} \underbrace{P(Y = l | \mathcal{Y} = \ell, X = x)}_{\text{Transition Probabilities}} \cdot \underbrace{P(\mathcal{Y} = \ell | X = x)}_{\text{Consideration Sets}} \cdot \underbrace{P(X = x)}_{\text{Co-Variables}} \quad (2)$$

As argued above we only focus on the third survey provided by Civey, one week and thus closest to the election. We do this to come closest to what can be called election forecasting, even though we strictly speaking pursue nowcasting. Like most forecasts we implicitly make the assumption that within the final week the situation on aggregate stays the same and hence can be generalized to the future. [Bauer et al., 2021, ch. 3]

Furthermore, we focus on the complex non-stochastic uncertainty induced by the individuals' ambiguity and not on survey errors and confidence intervals. We apply the weights provided by Civey as a state of the art approach to minimize survey error effects without going into further detail on the usual issues related to voluntarily surveys. As mentioned above, about 11% of the individuals are still pondering between parties this close to the election and induce this further source of uncertainty.

In the following chapters we conduct and compare different approaches, starting with the conventional one neglecting the undecided as reference for the others illustrating the benefits of including the undecided in different manners.

## 4.2 Neglecting the Undecided

In order to have a comparison to the approaches including the undecided in a set-valued manner, we start off with the approach based on conventional data, which excludes the undecided voters overall. By this, the partial information from the undecided is not only wasted, reducing the sample from 4730 to 4197 observations, but there is also an implicit assumption made that the undecided do not structurally differ from the decided in their voting behavior. But as we could show that the undecided systematically differ from the decided with our analysis in chapter 3, this does not hold in our case. Hence, the undecided provide not only additional, but also different information for forecasting.

Nevertheless, the point-valued results neglecting the undecided are illustrated in figure 3.

These forecasts are somewhat similar to the ones provided by other polling institutes<sup>5</sup>, showing new strength of the SPD and diminishing numbers concerning the CDU/CSU. Without going into detailed description of the results, the forecasts neglecting the undecided serve as comparison for our other approaches.

<sup>5</sup>For frequently updated election forecasting see: <https://de.statista.com/statistik/daten/studie/30321/umfrage/sonntagsfrage-zur-bundestagswahl-nach-einzelnen-instituten/>, last visited 21.09.21

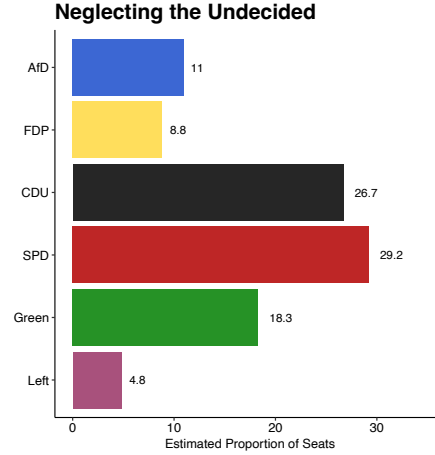


Figure 3: Results for the estimation proportion of seats in the parliament for the six main parties, neglecting the undecided

### 4.3 Point-Valued Forecasting with a Homogeneity Assumption

To establish a point-valued alternative to wasting the information of the undecided, we have to make additional assumptions on the hidden process towards the eventual choice. Such assumptions have to be on the one hand plausible but on the other rather restrictive when they should ensure point-valued results. Several ones are possible, but plausible ones are rare, indeed going beyond what a researcher can deliver with certainty, making such assumptions a kind of best guess driven by the overburdening goal to achieve a precise statement in a situation of complex uncertainty. Overall, the assumption has to be preferable to the one that the undecided do not structurally differ from the undecided, which is not too high of a hurdle.

For our approach we suggest a homogeneity assumption exploiting the covariates together with the information of the decided.<sup>6</sup> The undecided are assumed to behave on average like the decided conditional on the covariates, with their consideration set as restriction of the possible outcomes. This assumption is both disputable and intuitive. On the one hand, it is plausible that, given covariates, the undecided choose amongst their consideration set similar to the decided. But on the other hand complete homogeneity will probably not hold up in practice. Nevertheless, this approach appealingly regards the entire information of the consideration set as well as the one of the covariates, which can easily argued to be better than neglecting the information of the undecided overall.

Using the decided, the probability distribution  $P(Y_i = l | X_i = x_i, I_d = 1)$  can be estimated from the data, with  $I_d$  as the indicator function for being decided. The consideration set in this approach becomes the restriction of possible outcomes, while the tendency towards a party of the consideration set is predicted using the decided and co-variables as underlying data. Those predictions of affinity towards the parties of the undecided have to be scaled to comply with the multinomial distribution, excluding all options not in  $\ell$ . Therefore, for all  $l \in \ell$  the predicted affinity towards one party is divided by the sum of all the ones in the consideration set resulting in

$$\hat{P}(Y = l | Y = \ell, X = x) = \frac{\hat{P}(Y = l | X = x, I_d = 1)}{\sum_{a \in \ell} \hat{P}(Y = a | X = x, I_d = 1)} \quad (3)$$

leading to point-valued identification of every parameter. The prediction can be obtained using a variety of methods, while we choose a regression approach.

This results in point-valued estimates illustrated in figure 4, which come at the cost of having made a strong, untestable assumption about the individual decision process. As covariates we used the same ones as in chapter 3.

Looking at the results, only slight differences to the conventional approach can be found. While the proportion of the FDP increases, the one for the AfD decreases. As the number of undecided individuals decreases closer to the election the differences between the conventional and the homogeneity assumption approach declines as well. Hence, for the first and second wave the differences are higher.

<sup>6</sup>This assumption was developed in [Kreiss and Augustin, 2020] and thoroughly discussed and compared to different ones.

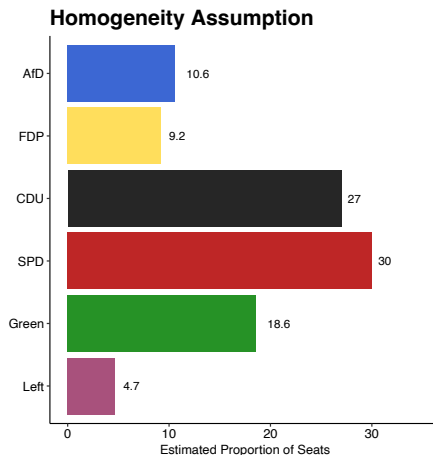


Figure 4: Results for the estimation proportion of seats in the parliament for the six main parties, utilizing the information of the undecided together with covariates and a homogeneity assumption

This approach does not communicate the uncertainty induced by the undecideds' ambiguity, but provides clearly at least a serious alternative to the conventional approach.

#### 4.4 Interval-Valued Credible Forecasting with the Dempster Bounds

To achieve reliable results, not having to rely on a strong assumption like in the point-valued approach of above, we can reflect the ambiguity of the undecided within interval-valued results. The so-called Dempster Bounds, in the spirit of [Dempster, 1967]'s handling of set-valued mappings, constitute hereby the most cautious approach. This results in the most accurate but also coarse forecasts, reflecting the entire ambiguity induced by the undecided within interval-valued results. Thus, as no information is available about which party from the consideration set is the eventual choice, these bounds reach from the worst case (everyone pondering between parties chooses the other one) to the best case (no one does) for every single party, describing so-to-say the continuum between the guaranteed seats and the still potentially achievable seats. Hence, the bounds tend to be wide, showing the entire ambiguity within the population.

With the Dempster Bounds a range for the proportion of individuals choosing the parties in  $Y$  is conveyed, in which (leaving out the survey error) the true one is contained in. The range emerges from shifting the probability mass to the extremes. This can be written for all  $\ell \in P(Y)$  as:

$$p_{lower}(Y \in \ell) = \sum_{\ell' \subseteq \ell} p(Y = \ell'), \quad (4)$$

$$p_{upper}(Y \in \ell) = \sum_{\ell' \cap \ell \neq \emptyset} p(Y = \ell'). \quad (5)$$

In this approach all elements of the set of all probabilities are considered as potential transition probabilities, which means that  $P(Y = \ell | Y = \ell, X = x)$  from equation 1 is set to the extreme values independently from the covariates.

The distances between worst and best case shown in the Dempster Bounds are interesting identification numbers as well, as this quantifies the amount of ambiguity from the undecided voters. If we let aside the survey error and possible structural changes until the election the results are completely credible, guaranteeing the eventual choice to lay in within those bounds.

The resulting Dempster Bounds from our data are illustrated in figure 5.

As we can see, the bounds are wide, especially in the case of the FDP relatively to its size, indicating that a lot of individuals are still pondering between this and other parties. As argued above, setting aside the survey error, for a respective party the lower bound can be seen as the guaranteed minimum of votes, while the upper bound shows its potential if all undecided not excluding this party indeed can be convinced to vote for that party.

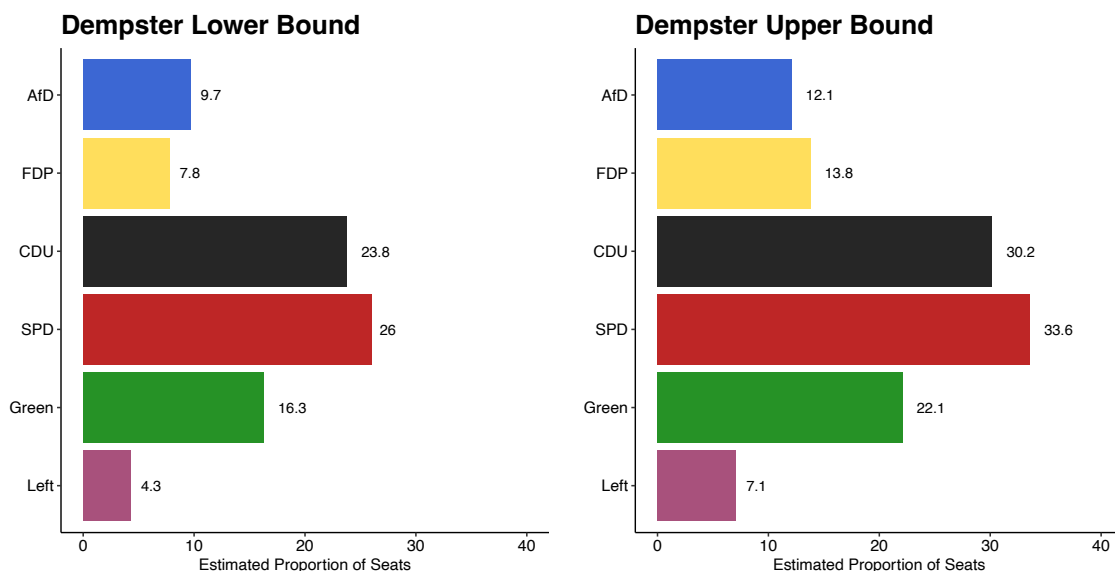


Figure 5: The Dempster Bounds reflection the entire ambiguity of the undecided within broad interval-valued results for each party

The magnitude of the uncertainty induced by the undecided voters' ambiguity even shortly before the election is shown in the width of the bounds. Within the second wave the bounds are naturally wider, as still more individuals are undecided. The plot concerning this matter can be found in the appendix.

On the other hand, one has to keep in mind that the Dempster Bounds reflect the entire uncertainty, always reaching from best to worst case. One of these extreme scenarios for one party is very unlikely to happen, as in aggregate not all individuals pondering between specific parties will end up voting for the same. Thus, we further provide an approach narrowing the bounds by assuming that on aggregate not more than 80% for the upper, and not less than 20% for the lower bounds choose the corresponding party. As one example: We assume that at least 20% of the individuals undecided between the SPD and Green party end up voting for the SPD and at most 80%. Already this rather weak assumption narrows the bounds substantially resulting in the forecasts illustrated in figure 6.<sup>7</sup>

These narrowed bounds frame a realistic range of outcomes, and delivers useful results. It can be argued that the bounds are close enough to provide meaningful statements, without too strict assumptions and can hence be seen as a compromise between the point-valued results and the Dempster Bounds.

#### 4.5 Forecasting the Strength of Coalitions – A New State Space for Epistemic Approaches

One very important subject concerning the German federal election are potential coalitions and if they could collect more than 50% of the votes in order to be capable to form a new government. Hence, forecasting the strength of specific party combinations is of interest. As the coalitions result from party combinations, the state space is extended in a way similar to the structure of our set-valued data. This extension of the state space towards our set-valued data has the fortunate property of dissolving some of the ambiguity within our data. To make one example, if a person is indifferent between the Green Party and the SPD, he or she will definitely provide a vote for the coalition of Green/SPD. Hence, there is no more uncertainty induced by the ambiguity of this person, and the originally partial information becomes precise. This only holds if individuals are undecided between parties of one coalition, but this is frequently the case, due to content-related similarities between parties that intend to form a coalition. There are many coalitions possible, while we focus on the ones frequently discussed and at least somewhat plausible. The results for the Dempster Bounds for these coalitions are illustrated in figure 7.

<sup>7</sup>Methodologically, such strengthening of the bounds by adding additional knowledge is in the core of the framework of partial identification, where, dependent on the context and the problem setting, a specific balance has to be found between a practically relevant precision of the result and its credibility by using only well-supported assumptions; see, e.g., [Manski, 2003]



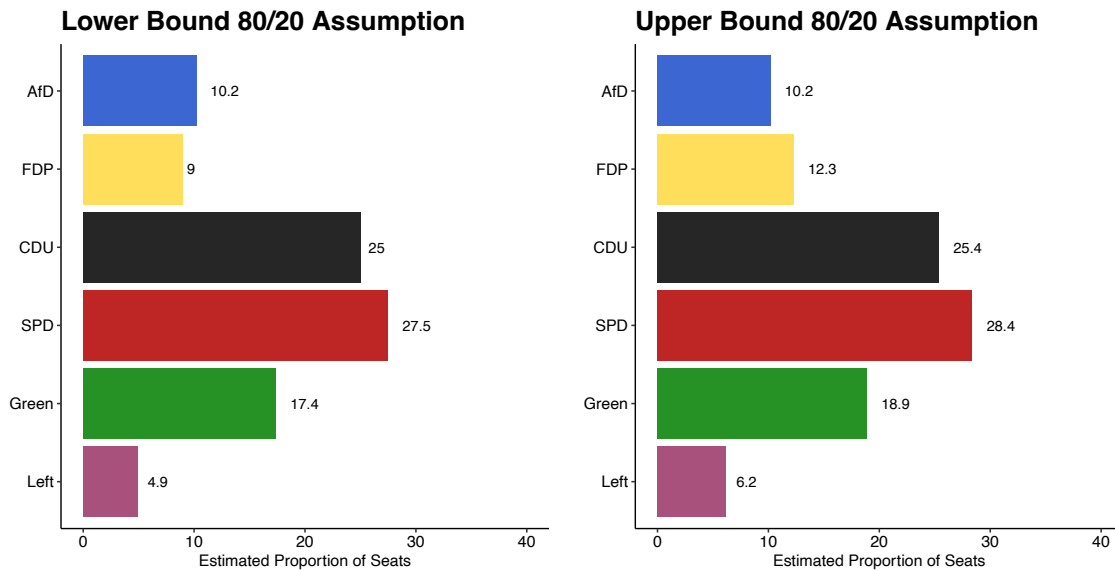


Figure 6: Modified Dempster Bounds reliant on the assumption that at least 20% and at most 80% choose one party from the consideration set

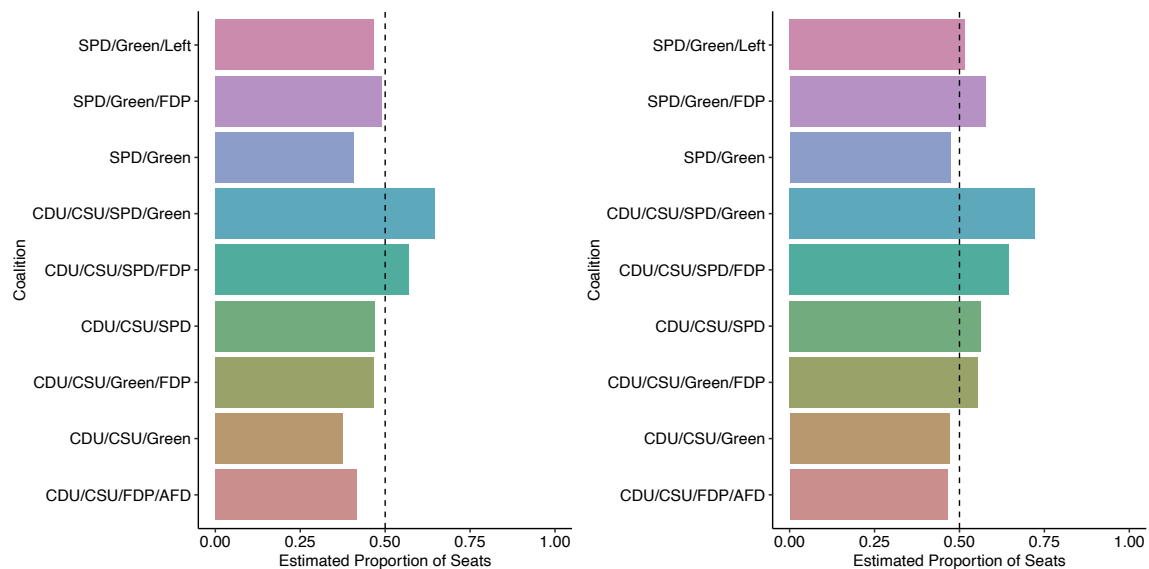


Figure 7: The Dempster Bounds for several possible party coalitions, with the lower bound on the left.

The bounds can, like above, be seen as the space between the the guaranteed minimum and the full potential of the coalition's strength. But in this case the bounds lay closer to each other as they did in figure 5, due to the reduction of ambiguity with the new state space. Concerning coalitions the attention is predominantly payed to whether or not coalitions collect at least 50% of the seats. The results show, that somewhere in between six and two coalitions will be capable to form a new government.

Equally to above, these bounds can be further narrowed by assuming on average at least 20% and at most 80% choose one specific party from the consideration set. These narrowed bounds are illustrated in figure 8.

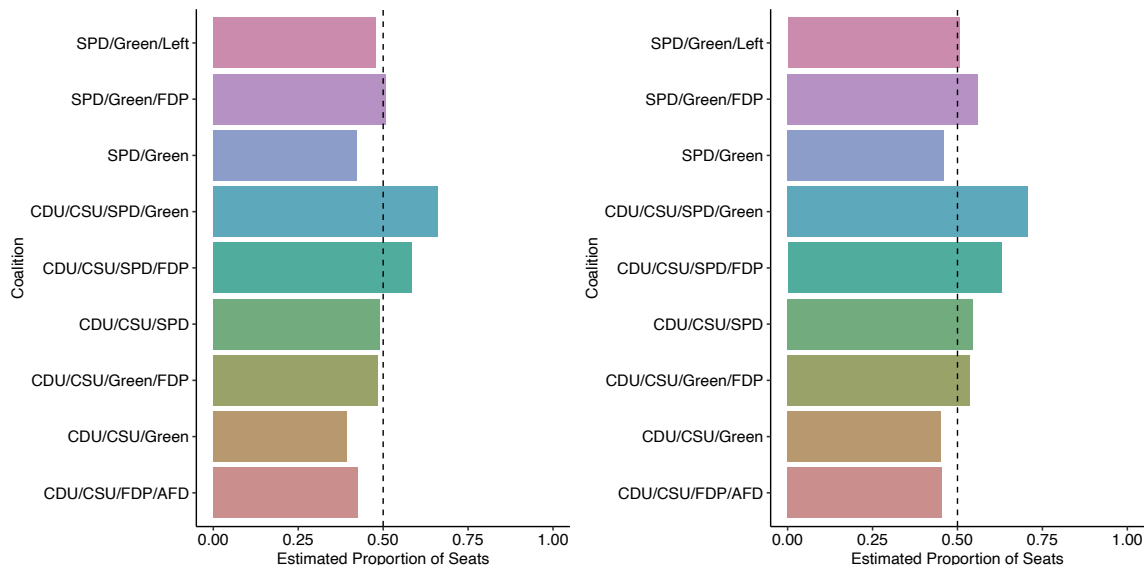


Figure 8: Modified Dempster Bounds for the possible coalitions reliant on the 80-20 assumption.

The bounds are indeed narrowed, but not as much as for the approach for the single parties, as some of the ambiguity is already dissolved by the new state space. Within the results (letting aside the survey error), for the Dempster Bounds two and for the modified three coalitions should reach at least 50% according to the lower bounds. In both cases, only combinations of three parties reach the necessary 50% with the lower bound. For the upper bound, with both approaches at least six parties can collect more than the necessary 50%.

These findings stress the potential of the new approach including the undecided voters, as meaningful statements concerning coalitions are possible even with none or very weak assumptions.

## 5 Outlook

Within this paper we could show with data about the 2021 German federal election that undecided voters can make a valuable contribution to election research if regarded set-valued. This information can on the one hand be used, not only to improve point-valued election forecasting, but also to communicate the uncertainty arising from ambiguity within the population in interval-valued forecasting. On the other hand, new insights into the political landscape and properties of individuals undecided between specific parties can be obtained.

This paper can be seen as a contribution to the solution of a Chicken-Egg dilemma of the past, as up until recently neither methodology nor data was available on this new way to include undecided voters. As this paper brought both together, on the one hand data from a first German pre-election poll regarding undecided voters set-valued and on the other our methodology developed, nothing stands in the way of further research in this direction. Building on the foundation laid with this and our previous works, methodology has to be further developed and improved. There are numerous possibilities, weighting the preciseness of the results and the credibility of the underlying assumption. Incorporating partial expert knowledge and other sources of information is for example one promising possible direction. Further approaches utilizing the longitudinal structure of our data and examining the undecided votes more thoroughly are very interesting as well.

This paper can be seen as a potential first step towards a paradigmatic shift concerning election research, in which the growing group of undecided is no longer neglected, but seen as the valuable part of the political landscape they are.

**Acknowledgement.** This project relies heavily on the cooperation with Civey who integrated our new survey design directly addressing the undecided. We are most grateful for the cooperation and especially thank Anna-Lena Disterheft and Gerrit Richter for their generous support. Dominik Kreiss is further very thankful to the LMU Mentoring Program supporting young researchers.

## References

- T. Augustin, F. P. A. Coolen, G. de Cooman, and M. Troffaes, editors. *Introduction to Imprecise Probabilities*. Wiley, Chichester, 2014.
- A. Bauer, A. Klima, J. Gauß, H. Kümpel, A. Bender, and H. Küchenhoff. Mundus vult decipi, ergo decipiatur: Visual communication of uncertainty in election polls. *Arxiv*, 2021.
- I. Couso and D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning*, 55(7):1502–1518, 2014.
- A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.
- D. Kreiss. Examining undecided voters in multiparty systems. Master’s thesis, LMU Munich, Department of Statistics, 2019. URL <https://epub.ub.uni-muenchen.de/70668/>.
- D. Kreiss and T. Augustin. Undecided voters as set-valued information, towards forecasts under epistemic imprecision. In J. Davis and K. Tabia, editors, *Scalable Uncertainty Management 2020*, pages 242–250. Springer, 2020.
- D. Kreiss, M. Nalenz, and T. Augustin. Undecided voters as set-valued information, machine learning approaches under complex uncertainty. In E. Huellermeier and S. Destercke, editors, *ECML/PKDD 2020 Tutorial and Workshop on Uncertainty in Machine Learning*. 2020. URL <https://drive.google.com/file/d/1abrLGZ154htGuYz8HzYLQzJ8vyc3kr2K/view>.
- D. Kreiss, G. Schollmeyer, and T. Augustin. Towards improving electoral forecasting by including undecided voters and interval-valued prior knowledge. In J. De Bock, A. Cano, E. Mirande, and S. Moral, editors, *Proceedings of the Twelfth International Symposium on Imprecise Probabilities: Theories and Applications*, Proceedings of Machine Learning Research, Granada, Spain, 06–09 Jul 2021. PMLR.
- C. Manski. *Partial identification of probability distributions*. Springer, 2003.
- C. F. Manski. Communicating uncertainty in official economic statistics: An appraisal fifty years after morgenstern. *Journal of Economic Literature*, 53(3):631–53, 2015.
- H. Oscarsson and M. Oskarson. Sequential vote choice: Applying a consideration set model of heterogeneous decision processes. *Electoral Studies*, 57:275–283, 2019.
- H. Oscarsson and M. Rosema. Consideration set models of electoral choice: Theory, method, and application. *Electoral Studies*, 57:256–262, 2019.
- J. Plass, P. Fink, N. Schöning, and T. Augustin. Statistical modelling in surveys without neglecting ‘The undecided’. In T. Augustin, S. Doria, E. Miranda, and E. Quaeghebeur, editors, *ISIPTA 15*, pages 257–266. SIPTA, 2015.
- G. Richter, T. Wolfram, and C. Weber. Die statistische methodik von civey. n.d. URL <https://civey.com/whitepaper>.
- G. Tutz. *Regression for categorical data*, volume 34. Cambridge University Press, 2011.
- G. Tutz, W. Pöbnecker, and L. Uhlmann. Variable selection in general multinomial logit models. *Computational Statistics & Data Analysis*, 82:207–222, 2015.

# Appendices

## A Results for the Second Wave of Data

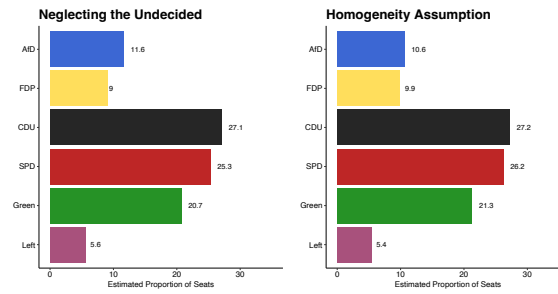


Figure 9: Second Wave Conventional and Homogeneity Results Plot

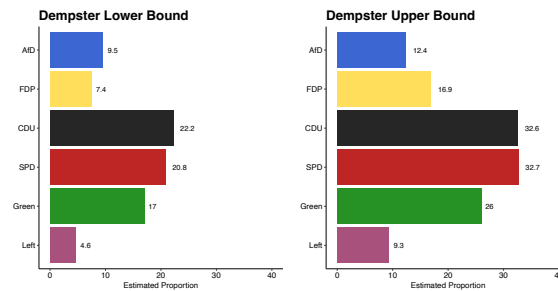


Figure 10: Second Wave Dempster Bounds

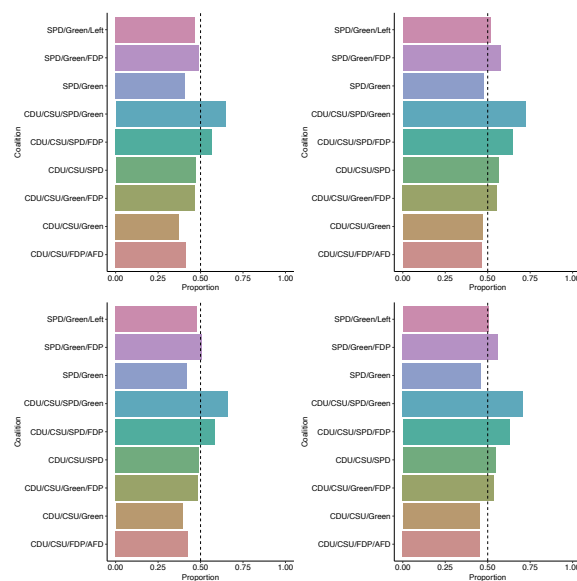


Figure 11: Results concerning coalitions from the second wave. On the top Row the Dempster and on the bottom row the modified Dempster Bounds are illustrated

## Contribution 5

Rodemann, J.; Kreiss, D.; Hüllermeier, E.; Augustin, T.: Levelwise Data Disambiguation by Cautious Superset Classification. In: Dupin de Saint-Cyr, F., Öztürk-Escoffier, M., Potyka, N., editors, *International Conference on Scalable Uncertainty Management*. Springer Lecture Notes in Artificial Intelligence (2022)  
[https://doi.org/10.1007/978-3-031-18843-5\\_18](https://doi.org/10.1007/978-3-031-18843-5_18)

---

## Levelwise Data Disambiguation by Cautious Superset Classification\*

Julian Rodemann<sup>1</sup>, Dominik Kreiss<sup>1</sup>, Eyke Hüllermeier<sup>2</sup>, and Thomas Augustin<sup>1</sup>

<sup>1</sup> Department of Statistics, LMU Munich, Germany

{julian.rodemann,dominik.kreiss,thomas.augustin}@stat.uni-muenchen.de

<sup>2</sup> Department of Computer Science, LMU Munich, Germany  
eyke@lmu.de

**Abstract.** Drawing conclusions from set-valued data calls for a trade-off between caution and precision. In this paper, we propose a way to construct a hierarchical family of subsets within set-valued categorical observations. Each subset corresponds to a level of cautiousness, the smallest one as a singleton representing the most optimistic choice. To achieve this, we extend the framework of Optimistic Superset Learning (OSL), which disambiguates set-valued data by determining the singleton corresponding to the most predictive model. We utilize a variant of OSL for classification with 0/1 loss to find the instantiations whose corresponding empirical risks are below context-dependent thresholds. Varying this threshold induces a hierarchy among those instantiations. In order to rule out ties corresponding to the same classification error, we utilize a hyperparameter of Support Vector Machines (SVM) that controls the model's complexity. We twist the tuning of this hyperparameter to find instantiations whose optimal separations have the greatest generality. Finally, we apply our method on the prototypical example of yet undecided political voters as set-valued observations. To this end, we use both simulated data and pre-election polls by Civey including undecided voters for the 2021 German federal election.

**Keywords:** Optimistic Superset Learning · Set-Valued Data · Support Vector Machines · Data Disambiguation · Epistemic Imprecision · Undecided Voters.

---

\*We sincerely thank the polling institute Civey for providing the data as well as the anonymous reviewers for their valuable feedback and stimulating remarks. DK further thanks the LMU mentoring program for its support.

2 J. Rodemann et al.

## 1 Introduction

Within many applied learning settings, data is not available with the level of precision required for conventional methodology. This coarseness can arise from insufficient information about an existing truth as within sensor imprecision or can be due to inherently unacquaintable table structures like temporary indecisiveness between viable choice options. Either way, we are often provided with a set of viable candidates as a coarse version of one true value. Predicting the true value out of the set of candidates or training an overall model is difficult, as one has to account for the uncertainty either cautiously or has to rely on possibly untenable strong assumptions.

Technically, such data are described by so-called disjunctively or epistemically interpreted random sets (see, e.g., [3]). Without any further assumptions or underlaid structure, the empirical distribution of the underlying true values is only partially identified [16,17]. The field of *superset learning* (also known under different names, such as partial label learning) provides a methodological framework to incorporate set-valued data in the learning process, (re)interpreting and utilizing its information in different manners. The goal is predominantly to obtain one overall best model (e.g. [18]) or an optimal set of models (e.g. [4]) by incorporating the imprecise information. Different ideas building on maximum likelihood from fuzzy data were suggested by [5]. [20] show that the direct profile likelihood of set-valued categorical data naturally has a set-valued maximum, while underlying further parametric modelling structures (for instance, a non-saturated multinomial logit model) may substantially reduce imprecision in the result, even possibly leading to single-valued parameter estimates, see also [22] or the marrow region of [24].

[8] introduced Optimistic Superset Learning (OSL). Combining model identification and data disambiguation, OSL searches for and relies on the most plausible instantiation, i.e. a singleton (precise) representation of set-valued (imprecise) observations. The idea is to quantify the plausibility of possible data instantiations by the discriminative power of a given model when trained on it.

In this paper we build on OSL, constructing hierarchical set-valued variants of it: Instead of possibly over-optimistically determining only one single instantiation, we consider the set(s) of all instantiations whose empirical risk lies below a (varying) context-dependent threshold and focus on data disambiguation. To this end, we utilize a variant of OSL with the 0/1 loss, resulting in the full set of alternatives first, and narrowing down those alternatives in a hierarchical manner by decreasing the threshold in a step-wise manner. In order to rule out ties, we use a hyperparameter of Support Vector Machines that controls the model's complexity to obtain the instantiation whose separation has the most clarity. Provided with this hierarchical family of subsets, the practitioner can now choose the threshold to induce the level of conciseness desired for their application. We further provide a visual aid similar to a Scree Plot to assist the choice of the context-dependent threshold. We illustrate our method in a simulation study and later apply the new approach to undecided voters in a pre-election poll for the 2021 German federal election. Within this prototypical situation of

complex inherent uncertainty, we characterize still undecided voters with their set of viable options, as suggested by [13, 14, 19, 21], instead of neglecting them like in conventional polls.

This paper is structured as follows. After formalizing data disambiguation and discussing OSL in Section 2, we introduce our extensions narrowing down the supersets in Section 3 and resolving potential ties in Section 4. The proposed methodology is then applied on simulated as well as on real-world survey data in Section 5. Finally and in light of the presented results, we conclude by discussing some potential venues for future work in Section 6.

## 2 Data Disambiguation by Optimistic Superset Learning

Consider a set of observations  $\Theta = \{(x_i, Y_i)\}_{i=1}^n \in (\mathcal{X} \times 2^{\mathcal{Y}})^n$ , where  $\mathbf{x} = (x_1, \dots, x_n)$  are singleton observations of covariates and  $Y_i$  set-valued observations of target variables.<sup>3</sup>  $\mathcal{X}$  is the covariate space and  $\mathcal{Y}$  is the target space. Leaning on the idea of Optimistic Superset Learning (OSL) as proposed by [8],  $Y_i$  is regarded a coarse representation (a superset) of a true underlying singleton  $y_i \in \mathcal{Y}$ . In what follows,  $\mathcal{Y}$  is assumed to be categorical. Let  $\mathbf{Y} = Y_1 \times Y_2 \times \dots \times Y_n$  be the Cartesian product of the observed supersets, and denote the number of different observed categories by  $q$ .<sup>4</sup> Then any singleton vector  $\mathbf{y} = (y_1, \dots, y_i, \dots, y_n)' \in \mathbf{Y}$  is called an *instantiation* of the observed set-valued data.

In practice, the set of candidate instantiations might be restricted to a subset of  $\mathbf{Y}$ , thereby allowing for the incorporation of domain knowledge in the form of constraints, for example, that observations with similar covariates and supersets ought to be instantiated with the same value for the target variable,<sup>5</sup> see Section 5.1. We regard further research concerning the restriction of  $\mathbf{Y}$  as powerful and briefly touch upon it in Section 6. In the following, for ease of exposition, we simply assume  $\mathbf{Y}$  to be the full Cartesian product of the individual set-valued observations.

Consider an instantiation  $\mathbf{y} \in \mathbf{Y}$ , a loss function  $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$  and a model's predictive function  $\hat{\mathbf{y}}^{(\mathbf{h}, \mathbf{y})}(\mathbf{x})$  when trained on this instantiation  $\mathbf{y}$ . The latter is found by minimizing the empirical risk with a suitable loss function. Vector  $\mathbf{h}$  shall denote the predictive model's hyperparameters, which are assumed to be fixed for now but will turn out to be of some relevance in Section 4. Now denote by  $\mathbb{P}(\mathbf{x}, \mathbf{y})$  the underlying joint probability measure of  $\mathbf{x}, \mathbf{y}$  and  $\mathcal{R}(\mathbf{h}, \mathbf{x}, \mathbf{y}) = \int L(\hat{\mathbf{y}}^{(\mathbf{h}, \mathbf{y})}(\mathbf{x}), \mathbf{y}) \, d\mathbb{P}(\mathbf{x}, \mathbf{y})$  the (theoretical) risk, which is estimated by the empirical risk  $\mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n L(\hat{y}_i^{(\mathbf{h}, \mathbf{y})}(x_i), y_i)$ ,  $\hat{y}_i \in \hat{\mathbf{y}}^{(\mathbf{h}, \mathbf{y})}(\mathbf{x})$ ,  $y_i \in \mathbf{y}$ ,  $x_i \in \mathbf{x}$ . Based on OSL we then consider

$$\mathbf{y}_{\mathcal{R}_{emp}}^* = \arg \min_{\mathbf{y} \in \mathbf{Y}} \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) \quad (1)$$

<sup>3</sup>Note that this formalization allows  $Y_i$  to also (partially) consist of singletons.

<sup>4</sup>Notably,  $q = |\mathcal{Y}| - k$ , where  $k$  is the number of categories in  $\mathcal{Y}$  that are not present in the data.

<sup>5</sup>This subsetting of  $\mathbf{Y}$  can be seen as a form of “data choice” similar to model choice.



4 J. Rodemann et al.

for a pre-defined  $\mathbf{h}$  the most plausible instantiation(s).<sup>6</sup> That is, we opt for those instantiation(s)  $\mathbf{y} \in \mathbf{Y}$  that make a given model the most predictive one when trained and evaluated on those instantiation(s). Its predictive function  $\hat{\mathbf{y}}^{(\mathbf{h}, \mathbf{y})}(\mathbf{x})$  might also output set-valued predictions as long as they can be evaluated by a real-valued loss function. In Section 3, we will explicitly estimate  $\hat{\mathbf{y}}^{(\mathbf{h}, \mathbf{y})}(\mathbf{x})$  for each instantiation  $\mathbf{y} \in \mathbf{Y}$  to find  $\mathbf{y}_{\mathcal{R}_{emp}}^*$  from (1), i.e. minimize the empirical risk for all  $\mathbf{y} \in \mathbf{Y}$ . This is in contrast to minimizing a generalized empirical risk function, the “optimistic superset loss”<sup>7</sup>

$$\text{OSL}(\hat{\mathbf{y}}, \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n L^*(\hat{y}_i, Y_i) = \frac{1}{n} \sum_{i=1}^n \min_{y \in Y_i} L(\hat{y}_i, y), \quad (2)$$

as done in the original OSL method [8, Section 4.1]. Our approach will allow us to hierarchically distinguish between instantiations in  $\mathbf{Y}$  with regard to their  $\mathcal{R}_{emp}$ . Computationally, this comes at the cost of estimating up to  $q^n$  models. Solving the optimization problem in equation (1) thus has exponential computational complexity. However, we will suggest a variant of (1) in Section 5 for socio-economic applications that reduces this number by clustering observations.

Further note that OSL, in addition to the optimal instantiation, returns a predictive model with minimal risk, i.e. a model producing predictions  $\hat{y}_i$  minimizing (2). Thus, OSL performs model identification and data disambiguation simultaneously [10, Section 2.2]: The model provides information about the data and, vice versa, the data about the model.<sup>8</sup> Nevertheless, model identification could be regarded as less general than data disambiguation in practical applications. This is due to the fact that the found risk-minimal instantiation(s)  $\mathbf{y}_{\mathcal{R}_{emp}}^*$  can be used to train other models regardless of the one used in OSL. Theoretically, the same holds vice versa. Yet, it might be hard to access other (that is, new) data in practice. Furthermore, our approach of disambiguating data by providing subsets rather than singletons can be regarded as a way of loosening the degree to which we rely on the model. This is why we will focus on data disambiguation rather than model identification in the following.

<sup>6</sup>Criterion (1) aims at a unique minimum. In general, in the light of the next section, we understand  $\arg \min$  potentially in a set-valued manner, i.e. giving the set of all elements where the minimum is attained.

<sup>7</sup>The loss is called optimistic due to the minimum in (2): each prediction  $\hat{y}_i$  is assessed optimistically by assuming the most favorable ground-truth  $y \in Y_i$ .

<sup>8</sup>Notably, some models can be more informative on certain aspects of the data generating process than others. For instance, naive Bayes classifiers model the joint distribution  $\mathbb{P}(x, y)$  as opposed to standard regression models that are typically concerned with the conditional distribution  $\mathbb{P}(y|x)$ .

### 3 Narrowing Down Supersets

As in some situations it is preferable to be cautious rather than optimistic, we attempt to narrow down the supersets in a hierarchical manner ranging from least to most concise. Recall that for an instantiation  $\mathbf{y} \in \mathbf{Y}$ , criterion (1) depends on the loss function  $L(\cdot)$  through the empirical risk  $\mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n L(\hat{y}_i^{(\mathbf{h}, \mathbf{y})}(x_i), y_i)$ ,  $x_i \in \mathbf{x}$ ,  $\hat{y}_i \in \hat{\mathbf{y}}^{(\mathbf{h}, \mathbf{y})}(\mathbf{x})$ ,  $y_i \in \mathbf{y} \in \mathbf{Y}$ . Let  $\hat{\mathbf{y}}^{(\mathbf{h}, \mathbf{y})}(\mathbf{x})$  be the predictive function of a specific linearly representable model with fixed hyperparameters  $\mathbf{h}$  trained on  $\mathbf{y} \in \mathbf{Y}$  with any suitable loss function.

For the 0/1-loss  $L(\hat{y}_i^{(\mathbf{h}, \mathbf{y})}(x_i), y_i) = I(\hat{y}_i^{(\mathbf{h}, \mathbf{y})}(x_i) \neq y_i)$ ,  $I$  the indicator function, we can evaluate the model of an instantiation  $\mathbf{y} \in \mathbf{Y}$  by  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y})$  the number of misclassifications. Hence, we are able to compare all instantiations with regard to their induced number of misclassifications  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y})$  or misclassification rate  $\mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y})$ . Formally, we end up with a total order. This is due to  $(\mathbb{N}, \leq)$  being a total order and the fact that any subset of a totally ordered set is a total order with the restriction of the order on the subset.<sup>9</sup>

We use this very order to provide the decision-maker with a hierarchy of sets of instantiations ranging from complete ambiguity to a concise optimistic interpretation of set-valued observation. To prepare this, two definitions are given. The first one looks at the respective number of misclassifications  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y})$  to introduce the notion of an optimistic subset; the second one describes the resulting set on the level of individual observations.

**Definition 1 ( $\mathcal{E}$ -Optimistic Subset).** *Let  $\mathbf{Y}$  be the Cartesian product of the observed supersets as above and  $\mathcal{E} \in \mathbb{N}$  a pre-defined upper bound for classification errors. Then*

$$\mathbf{Y}_{\mathcal{E}} = \{\mathbf{y} \in \mathbf{Y} \mid n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) \leq \mathcal{E}\} \subseteq \mathbf{Y}$$

*shall be called  $\mathcal{E}$ -optimistic subset of  $\mathbf{Y}$ .*

**Definition 2 ( $i$ -th Consideration Function).** *Let  $y_i \in \mathbf{y} \in \mathbf{Y}_{\mathcal{E}}$  be the class of a fixed observation  $i \in \{1, \dots, n\}$  in an instantiation  $\mathbf{y} \in \mathbf{Y}_{\mathcal{E}}$ . For varying  $\mathcal{E}$ , the function*

$$\begin{aligned} f_i: \mathbb{N} &\rightarrow 2^{\mathcal{Y}} \\ \mathcal{E} &\mapsto \{y \in \mathcal{Y} \mid \exists \mathbf{y} \in \mathbf{Y}_{\mathcal{E}} : y = y_i, y_i \in \mathbf{y}\} \end{aligned}$$

*shall be called consideration function of observation  $i$ .*

Verbally,  $f_i(\mathcal{E})$  gives the set of possible classes of an observation  $i$  in all instantiations in  $\mathbf{Y}_{\mathcal{E}}$ , i.e. so-to-say the set still under consideration given an overall error  $\mathcal{E}$ . Note that the above described total order of  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y})$ -values induces a partial order  $(\mathbf{Y}_{\mathcal{E}}, \subseteq)$ , which is part of the following proposition's proof.

<sup>9</sup>Note that  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) \in \mathbb{N}$ .

6 J. Rodemann et al.

**Proposition 1.** *Function  $g_i(\varepsilon) = |f_i(\varepsilon)|$  is monotonically non-decreasing.*

*Proof.* Let  $\tilde{\mathbf{y}} \in \mathbf{Y}_{\varepsilon_1}$ . Definition 1 directly delivers that  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \tilde{\mathbf{y}}) \leq \varepsilon_1$ . With  $\varepsilon_1 < \varepsilon_2$  by assumption, we trivially have  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \tilde{\mathbf{y}}) \leq \varepsilon_2 \implies \tilde{\mathbf{y}} \in \mathbf{Y}_{\varepsilon_2}$ . Thus, for any two  $\varepsilon_1, \varepsilon_2 \in \mathbb{R}$  with  $\varepsilon_1 < \varepsilon_2$  it holds  $\mathbf{Y}_{\varepsilon_1} \subseteq \mathbf{Y}_{\varepsilon_2}$ . Since  $f_i(\varepsilon)$  only contains classes of instantiations in  $\mathbf{Y}_{\varepsilon}$ , the assertion follows.

The  $\varepsilon$ -optimistic subset  $\mathbf{Y}_{\varepsilon} \subseteq \mathbf{Y}$  can be interpreted as those instantiations that are (optimistically) plausible given models that make less than  $\varepsilon$  classification errors. Practitioners might either *a priori* select an application-dependent level of tolerable errors  $\varepsilon \in \mathbb{N}$  and proceed with the corresponding instantiations  $\mathbf{Y}_{\varepsilon} \subseteq \mathbf{Y}$ . They might as well decide *a posteriori* by visual support of plotting  $|\mathbf{Y}_{\varepsilon}|$  (the number of instantiations in the subset) against  $\varepsilon$ , see Section 5. Generally, this order will include ties for instantiations that are separable by the classifier with the same misclassification error:  $\mathbf{Y}_{\varepsilon}^* \stackrel{def}{=} \{\mathbf{y}^* \mid n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}^*) = \varepsilon\}$ ,  $\mathbf{Y}_{\varepsilon}^* \subseteq \mathbf{Y}$ , see the weak monotonicity of  $g_i(\varepsilon)$  in Proposition 1. For a given  $\varepsilon$ , they can be thought of equally optimistic instantiations. Instead of forcing to identify a singleton instantiation in the set-valued observations, the practitioner can make his choice how to work with this set of instantiations.

However, in some applications, it might also be beneficial to at least have the opportunity to decide for a “most optimistic” singleton from all instantiations in  $\mathbf{Y}_{\varepsilon}^*$  that are *prima facie* equally optimistic with regard to their corresponding values of  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y})$ . This option seems especially relevant for the smallest non-trivial (that is, non-empty) set in the hierarchy induced by Proposition 1, e.g. for all instantiations separable by the classifier, that is for the set  $\mathbf{Y}_0^*$  ( $\varepsilon = 0$ ).

In the following, we will introduce methodology to decide for such a “most optimistic” instantiation from  $\mathbf{Y}_{\varepsilon}^*$  while maintaining the interpretable and intuitive number of misclassifications  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) \in \mathbb{N}$  as order criterion. Whilst the method of narrowing down supersets is generally applicable to any classifiers, we will restrict ourselves to Soft-margin Support Vector Machines (SVMs) in the following. This is due to their hyperparameter  $\mathcal{C}$  that has an exciting interpretation, which we will utilize for a second-level-criterion. In doing so, we will seek inspiration from [9, Section 3.2], where the model architecture (that is, hyperparameters) is (visually) taken into account and instantiations are compared based on models of varying complexity.<sup>10</sup>

<sup>10</sup>However, in [9, section 3.1] the class of models, thus the model’s hyperparameters, is fixed.

## 4 Resolving Ties by Twisted Tuning of SVMs

Support Vector Machines (SVMs) [1] transform input vectors to a high-dimensional covariate space, where a linear classification hyperplane is constructed.<sup>11</sup> Soft-margin SVMs [2] allow violations of this hyperplane. In order to penalize such misclassifications, a hyperparameter  $C$  is used to control the trade-off between maximizing the margin  $M$  (the minimal distance from the separating hyperplane to the data) and minimizing the number of violations – or, in the original words of [2], between “complexity of decision rule and frequency of error.” To be a bit more precise, the classification hyperplane is found by minimizing a weighted sum of the margin  $M$  and the loss function that penalizes misclassifications. The hyperparameter  $C$  is the weight of that loss function.

For a given classification problem, the  $C$  that minimizes the training error can be thought of as a proxy for the clarity of optimal separation. In other words, the larger the hyperparameter  $C$ , the more sensitive the SVM is towards violations of the hyperplane; the lower  $C$ , the more the SVM focuses on finding maximal margins, respectively.<sup>12</sup>

The latter is the starting point for our deliberations regarding hyperparameter tuning of  $C$ , which is usually not learned by the data, but set *a priori* by human choice. The emerging field of automated machine learning, however, aims at an unmanned optimal selection of hyperparameters. This is typically achieved by optimizing the generalization error through cross-validation.

We twist this tuning of  $C$ : Instead of asking for the optimal  $C$  given the observations, we ask for the instantiation of set-valued observations that leads to the lowest  $C$  when chosen in order to minimize the training error. In other words, among the set  $\mathbf{Y}_\varepsilon^* = \{\mathbf{y}^* \mid n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}^*) = \varepsilon\}$ , see Section 3, of instantiations that correspond to the same  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y})$  with 0/1-loss, we search for that version of the data whose optimal separation has the greatest clarity. To make things more tangible, recall the set of candidate instantiations  $\mathbf{Y}$  from Section 2 and 3. Note that the vector of the model’s hyperparameters  $\mathbf{h}$  now contains, possibly among others,  $C$ , i.e.  $\mathbf{h} = (C, \mathbf{h}_r)'$ . We abstract from the remaining hyperparameters  $\mathbf{h}_r$  and assume them to be manually set. Among those instantiations in  $\mathbf{Y}_\varepsilon^*$  we propose to select the instantiation whose most predictive model on the training data has the least complex decision rule, see equation (3).

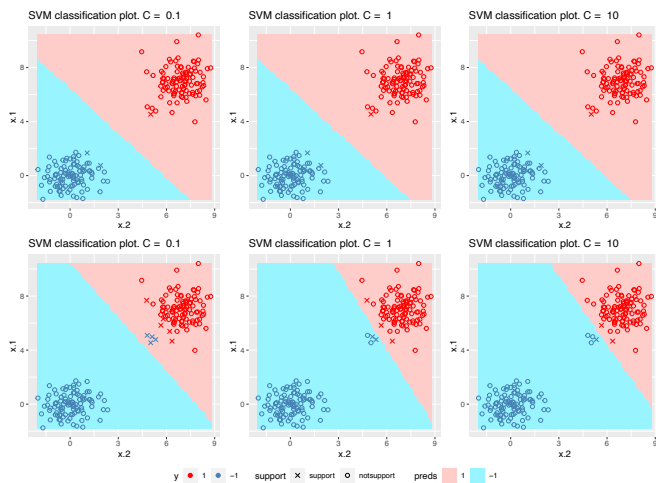
$$\mathbf{y}_C^* = \arg \min_{\mathbf{y}^*} \arg \min_C \{\mathcal{R}_{emp}(\mathbf{h}_r, C, \mathbf{x}, \mathbf{y}^*) \mid \mathbf{y}^* \in \mathbf{Y}_\varepsilon^*\} \quad (3)$$

<sup>11</sup>For multi-class classification (as in Section 5), hyperplanes from one-versus-all classifications are combined by a voting scheme and Platt scaling, for details see [11, pages 8-9]. When tuning with regard to  $C$ , one common  $C$ -value is used for all one-versus-all classifications.

<sup>12</sup>For kernelized versions of SVMs this hyperplane is generally only linear in the transformed feature space. However, we can still think of  $C$  as a proxy for the generality of optimal separation in that transformed space.

8 J. Rodemann et al.

In other words, we perform hyperparameter-tuning<sup>13</sup> with 0/1-loss with regard to  $C$  of all those models that were trained on instantiations from  $\mathbf{Y}_\xi^* = \{\mathbf{y}^* \mid n \cdot \mathcal{R}_{emp}(\mathbf{h}_r, C, \mathbf{x}, \mathbf{y}^*) = \xi\}$ . We then choose the instantiation(s) corresponding to the model(s) with the lowest  $C$ . Notably, we fix the hyperparameters  $\mathbf{h}$  including  $C$  in order to find instantiations in  $\mathbf{Y}_\xi^*$ , see Section 3, only to optimize with regard to it later. We select the minimal  $C$  in the set of all  $C$  values that minimize the training or generalization error:  $\arg \min_C$ . This minimal  $C$  has a sound interpretation: It tells us, for a given instantiation, how general we can make the decision rule while maintaining optimal classification. Figure 1 illustrates this very idea in a specific context: Depicted are  $n = 200$  singleton observations of a binary target variable in a two-dimensional covariate space, of which 100 belong to class  $-1$  (blue) and 100 to class  $1$  (red). Four observations (situated around  $(5, 5)$ ) are set-valued and might be interpreted as indecisive between blue and red. In the upper row, all of these four observations are instantiated as  $1$  (red) and in the lower row as  $-1$  (blue), respectively. Note that for both instantiations we have  $\mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) = 0$ , since the corresponding data set is linearly separable in the covariate space. Each column in Figure 1 shows the predictions of an SVM with varying  $C$ . It becomes evident that the red instantiation can be separated even for  $C = 0.1$  (left), while the blue instantiation requires higher  $C$  values, i.e. more complex decision rules, in order to be classified correctly.



**Fig. 1.** Different instantiations of set-valued observations require different levels of  $C$  in order to be classified correctly.

<sup>13</sup>We use Grid Search for solving this minimization problem. When evaluations are rather expensive, Bayesian Optimization, Simulated Annealing or Evolutionary Algorithms might be preferred. For an overview of these heuristic optimizers and their limitations, see [23, chapter 10].

Distinguishing between instantiations in  $\mathbf{Y}_\varepsilon^*$  by (3) gives rise to the following preference function.

**Definition 3 (*i*-th Preference Function for level  $\varepsilon$ ).** Let  $y_i \in \mathbf{y}^* \in \mathbf{Y}_\varepsilon^*$  be the class of a fixed observation  $i \in \{1, \dots, n\}$  in an instantiation  $\mathbf{y}^* \in \mathbf{Y}_\varepsilon^*$ . For a given  $\varepsilon$ , the function

$$p_i^{(\varepsilon)}: \mathcal{Y} \rightarrow \mathbb{R}$$

$$y \mapsto \min\{C \mid C = \arg \min_{\underline{C}} \{\mathcal{R}_{emp}(\mathbf{h}_r, C, \mathbf{x}, \mathbf{y}^*) \mid \mathbf{y}^* \in \mathbf{Y}_\varepsilon^* \wedge y = y_i \in \mathbf{y}^*\}\}$$

shall be called *preference function of observation  $i$  for subset  $\mathbf{Y}_\varepsilon^*$* .

Verbally, the *i*-th Preference Function outputs for class  $y$  the minimal  $C$  from all those minimal  $C$ -values that correspond to such instantiations in  $\mathbf{Y}_\varepsilon^*$  that assign class  $y$  to observation  $i$ . The following proposition then entitles us to provide the user with a ranking of classes according to their plausibility in  $\mathbf{Y}_\varepsilon^*$  for the *i*-th individual. The induced total order can be used to rank all classes present in  $\mathbf{Y}_\varepsilon^*$  for observation  $i$ .

**Proposition 2.** For any fixed  $i$ , the element-wise composition  $p_i^{(\varepsilon)} \odot f_i$  induces a total order.

*Proof.* Since  $p_i^{(\varepsilon)}$  maps to  $\mathbb{R}$ , we have  $p_i^{(\varepsilon)} \odot f_i(\varepsilon) \in \mathbb{R}^d$ , where  $d \leq |\mathcal{Y}|$  is the dimension of the output of  $p_i^{(\varepsilon)}$ . Since any subset of the total order  $(\mathbb{R}, \leq)$  is a total order with the restriction of the total order on the subset, one single output vector  $p_i^{(\varepsilon)} \odot f_i(\varepsilon) \in \mathbb{R}^d$  has elements that are totally ordered.

Notably, using the Hinge loss function [7] in OSL would also allow for disambiguation of instantiations in  $\mathbf{Y}_\varepsilon^*$ , since it accounts for margin maximization. Deploying OSL with hinge loss in the first place, however, typically does not induce ties, since then  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y}) \in \mathbb{R}$ . What is more, the real-valued  $n \cdot \mathcal{R}_{emp}(\mathbf{h}, \mathbf{x}, \mathbf{y})$  is not as interpretable as in the countable case of the 0/1-loss and thus a pre-defined and context-dependent level of acceptable errors  $\varepsilon$  might be hard to specify for the decision maker. Still, OSL with hinge loss could be used to eventually rule out ties after having sequentially narrowed down supersets by means of 0/1-loss. However, the simultaneous model identification would not take into account  $C$  and could thus be regarded less general. In light of this, we recommend further research on the interaction of margin maximization induced by the hinge loss and the optimal level of generality represented by  $C$ .

10 J. Rodemann et al.

## 5 Applications to Undecided Voters

### 5.1 Clustering

In what follows, we will abstain from considering all  $q^n$  possible instantiations (with  $q$  again as the number of different observed classes) by only considering the candidate instantiations in  $\mathbf{Y}$ . Instead, for the sake of both interpretability and computational convenience, we cluster all (non-singleton) set-valued observations to  $k$  groups of observations  $G_1, \dots, G_k$  according to their covariates.<sup>14</sup> Generally, for each common set of classes one would need to perform a cluster analysis separately. With  $q$  observed classes, we would have  $q_c = 2^q - q - 1$  clusterings to be done, since we exclude  $q$  singletons and the empty set from the power set of observed classes. In the following application, however, we will only deal with individuals that are fully ambiguous among all options, i.e.  $q_c = q$ . We then disambiguate all observations in a cluster in the same way, i.e. assign all of them to the same class. This reduces  $Y_1 \times Y_2 \times \dots \times Y_n$  to  $G_1 \times G_2 \times \dots \times G_k$  with  $q_c^k$  instead of  $q^n$  possible instantiations.

### 5.2 Simulations

We simulate 120 observations with two metric socio-economic covariates in the set-up of a pre-election polling survey with individuals undecided between three parties. Figure 2 illustrates the distribution of the observations in the covariate space. The 60 still undecided (among all three parties) voters hereby disaggregate in three clusters, with from now on called *Cluster I* (triangles) in the left lower corner, *Cluster II* (squares) in the upper middle and *Cluster III* (crosses) in the right upper corner. One might think of these groups as sociodemographic clusters or social milieus. With the three parties we obtain  $3^3 = 27$  possible instantiations for the clusters.

<sup>14</sup>Any clustering algorithm can be used. In our applications in Section 5, we opt for k-means clustering as proposed by [15].

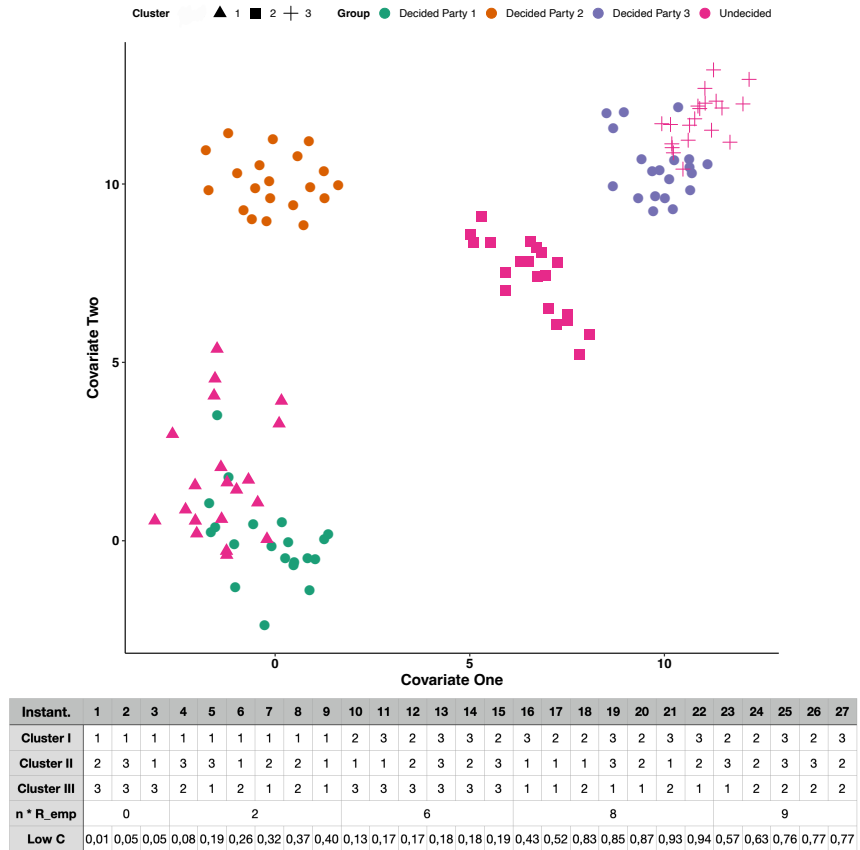


Fig. 2. Simulation setting: 120 observations in a two-dimensional covariate space with three parties, among which 60 are undecided. Simulation results: The 27 possible instantiations are ordered by their  $R_{emp}$  and the lower bound of their  $C$  value.

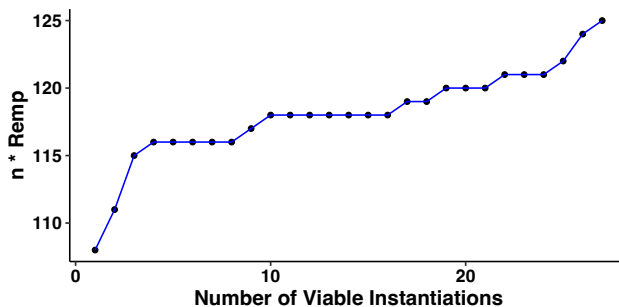
With our approach we obtain for each instantiation its  $R_{emp}$  as well as a lower bound of its  $C$  value. We can thus order them as illustrated in Figure 2. Depending on the application, the practitioner can now decide which level of imprecision is adequate and choose the viable instantiations correspondingly. A tolerable number of misclassifications of  $\varepsilon = 2$  would for example induce Cluster I to be always assigned to Party 1 while the other clusters might be assigned to either of the three. The resulting ties in  $n \cdot R_{emp}$  can furthermore be resolved by taking into account the lower bound of the  $C$  value.



12 J. Rodemann et al.

### 5.3 German Pre-Election Polls

We are in the fortunate position of cooperating with the polling institute Civey to explicitly account for undecided voters in a set-valued manner. Hence, we have first-hand access to polling data for the 2021 German federal election two months before the election, in which all still pondering individuals are represented by the set containing their viable options. We employ our methodology to the three center-left-/ left-leaning parties *SPD*, *Greens* and *The Left*, while it could straight forwardly be generalized to overall forecasting by addressing all groups of undecided sequentially following [12, p. 245]. In the provided polling data we have 935 participants determined to vote for the Greens, 592 determined to vote for the SPD, 168 for The Left and 66 still pondering between the three parties, thus  $n = 1761$ . Furthermore, we are provided with 10 socio-economic covariates capturing the socioeconomic status (education, population density of place of residence, purchasing power, employment status and the like) in an ordinal and nominal manner.<sup>15</sup>



**Fig. 3.** Results from the application on polling data. Party Legend: Lef = Left, Gre = Green. The 27 possible instantiations are again ordered by their  $n \cdot R_{emp}$  and the lower bound of their  $C$  value.

Clustering finds three socio-economic groups. Based on the covariates they can be roughly subsumed as older population with low and medium income (1), top-earning academics on the countryside (2) and a small group of urbanites without paid employment (3). Figure 3 shows the results for all  $3^3 = 27$  instantiations in the same manner as for the simulated data in Figure 2. It also entails

<sup>15</sup>The covariates appear to be generally of rather low predictive power: Training and generalization error, even exclusively for the decided, are high.

a plot of the number of instantiations in the  $\varepsilon$ -optimistic subsets  $|\mathbf{Y}_\varepsilon|$  and their respective  $\varepsilon = n \cdot \mathcal{R}_{emp}$  that can be used as decision support when opting for a level of tolerable misclassifications  $\varepsilon$ .

It becomes evident that even for realistically large datasets ( $n = 1761$ ) we can obtain ties with OSL and the 0/1-loss. In other words, we end up with non-singleton (and non-empty) sets  $\mathbf{Y}_\varepsilon^*$  for  $\varepsilon \in \{116, 118, 119, 120, 121\}$ . Here, twisted tuning can offer decision support. Applying a modified version of the  $i$ -th Consideration Function (Definition 2) to clusters rather than to individuals offers additional insight: The first (older population with low and middle income) and the third (unemployed townspeople) socio-economic clusters are not instantiated as Greens voters for  $\varepsilon$  lower than 116. Rural top-earners (2), however, are. For the latter group it is more plausible to vote Green, given the model and the available covariates, than for the other two groups, which appears to be an empirical insight that is in line with socio-politological literature on previous German elections, see [6] for instance.

## 6 Discussion

As underpinned by the application on polling data in Section 5, considering several instantiations can be an attractive extension to classical OSL, as it offers additional insights into (groups of) undecided voters. Moreover, it might prevent forecasters from over-optimistic predictions. Generally, we consider our level-wise approach to data disambiguation a practically powerful alternative to exclusively relying on a singleton instantiation of set-valued data.

However, with increasing  $n$ , stronger efforts are indispensable to ensure computational feasibility. This opens up venues for further work, extending our approach of homogeneous treatment of found clusters to general approaches of “data selection”. For instance, one could integrate the restrictions describing the reduced sets of instantiations  $Y_1 \times Y_2 \times \dots \times Y_n$  (see Section 2) as side-constraints in the minimization of the generalized empirical risk [8, Section 4.1] for classical Optimistic Superset Learning.

Furthermore, decision criteria beyond the total (lexicographic) order on the  $\varepsilon$ -optimistic subsets to moderate the trade-off between accuracy and generality should be investigated in detail. One could argue, for instance, that considering instantiations corresponding to higher  $\varepsilon$  is justified if they can be separated with sufficiently lower  $C$ , i.e. more general hyperplanes. Clear recommendations for different decision rules tailored to specific applications would be of high practical value. In addition, we see potential in a more versatile approach by not forcing a precise disambiguation of inconclusive cases in  $\mathbf{Y}_\varepsilon^*$ . This could either be achieved by two-stage criteria that account for further hyperparameters of the model or by considering some interval  $[C - \epsilon, C + \epsilon]$  instead of  $C$ .

14 J. Rodemann et al.

## References

1. Boser, B., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. pp. 144–152 (1992)
2. Cortes, C., Vapnik, V.: Support-vector networks. *Machine Learning* **20**(3), 273–297 (1995)
3. Couso, I., Dubois, D.: Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning* **55**, 1502–1518 (2014)
4. Couso, I., Sánchez, L.: Machine learning models, epistemic set-valued data and generalized loss functions: an encompassing approach. *Information Sciences* **358**, 129–150 (2016)
5. Dencœux, T.: Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on Knowledge and Data Engineering* **25**(1), 119–130 (2011)
6. Faas, T., Klingelhöfer, T.: The more things change, the more they stay the same? The German federal election of 2017 and its consequences. *West European Politics* **42**(4), 914–926 (2019)
7. Gentile, C., Warmuth, M.: Linear Hinge loss and average margin. *Advances in Neural Information Processing Systems* **11** (1998)
8. Hüllermeier, E.: Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning* **55**, 1519–1534 (2014)
9. Hüllermeier, E., Cheng, W.: Superset learning based on generalized loss minimization. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 260–275. Springer (2015)
10. Hüllermeier, E., Destercke, S., Couso, I.: Learning from imprecise data: adjustments of optimistic and pessimistic variants. In: International Conference on Scalable Uncertainty Management. pp. 266–279. Springer (2019)
11. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: kernlab—an S4 package for kernel methods in R. *Journal of Statistical Software* **11**(9), 1–20 (2004)
12. Kreiss, D., Augustin, T.: Undecided voters as set-valued information, towards forecasts under epistemic imprecision. In: International Conference on Scalable Uncertainty Management, pp. 242–250. Springer (2020)
13. Kreiss, D., Augustin, T.: Towards a paradigmatic shift in pre-election polling adequately including still undecided voters—some ideas based on set-valued data for the 2021 German federal election. arXiv preprint arXiv:2109.12069 (2021)
14. Kreiss, D., Nalenz, M., Augustin, T.: Undecided voters as set-valued information, machine learning approaches under complex uncertainty. In: ECML/PKDD 2020 Tutorial and Workshop on Uncertainty in Machine Learning (2020)
15. Lloyd, S.: Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**, 129–137 (1982)
16. Manski, C.: *Partial Identification of Probability Distributions*. Springer (2003)
17. Molchanov, I., Molinari, F.: *Random Sets in Econometrics*. Cambridge University Press (2018)
18. Nguyen, N., Caruana, R.: Classification with partial labels. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 551–559 (2008)

19. Oscarsson, H., Oskarson, M.: Sequential vote choice: Applying a consideration set model of heterogeneous decision processes. *Electoral Studies* **57**, 275–283 (feb 2019)
20. Plass, J., Cattaneo, M., Augustin, T., Schollmeyer, G., Heumann, C.: Reliable inference in categorical regression analysis for non-randomly coarsened observations. *International Statistical Review* **87**, 580–603 (2019)
21. Plass, J., Fink, P., Schöning, N., Augustin, T.: Statistical modelling in surveys without neglecting ‘The undecided’. In: *ISIPTA 15*, pp. 257–266. *SIPTA* (2015)
22. Ponomareva, M., Tamer, E.: Misspecification in moment inequality models: back to moment equalities? *The Econometrics Journal* **14**, 186–203 (2011)
23. Rodemann, J.: *Robust Generalizations of Stochastic Derivative-Free Optimization*. Master’s thesis, LMU Munich (2021)
24. Schollmeyer, G., Augustin, T.: Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *International Journal of Approximate Reasoning* **56**, 224–248 (2015)

## Contribution 6


Kreiss, D., Augustin, T.: Consideration Set Sampling to Analyze Undecided Respondents. Recently submitted. Preprint available on arXiv under:  
<https://doi.org/10.48550/arXiv.2307.14333>

---


# CONSIDERATION SET SAMPLING TO ANALYZE UNDECIDED RESPONDENTS

---

A PREPRINT

 **Dominik Kreiss**  
Department of Statistics  
LMU Munich

dominik.kreiss@stat.uni-muenchen.de

 **Thomas Augustin**  
Department of Statistics  
LMU Munich

thomas.augustin@stat.uni-muenchen.de

July 27, 2023

## ABSTRACT

Researchers in psychology characterize decision-making as a process of eliminating options. While statistical modelling typically focuses on the eventual choice, we analyze consideration sets describing, for each survey participant, all options between which the respondent is pondering. Using a German pre-election poll as a prototypical example, we give a proof of concept that consideration set sampling is easy to implement and provides the basis for an insightful structural analysis of the respondents' positions. The set-valued observations forming the consideration sets are naturally modelled as random sets, allowing to transfer regression modelling as well as appropriate machine learning procedures.

**Keywords** choice models · consideration set · pre-election poll · random set · undecided voters

## 1 Introduction

It is characteristic for human beings to ponder between options before finally making up their minds. At a given point in time prior to the eventual choice, individuals are often not yet determined but have already narrowed down their viable options to a compelling subset. Rationals on how this set, often called *consideration set*, emerges and which properties it satisfies have been developed in psychology and marketing research. (e.g. Stocchi et al. [2016], Shocker et al. [1991]) In this vein, one understands choice as a process of successively eliminating options by aspects, firstly excluding alternatives from the complete choice set and later on choosing from the emerged consideration set. [Tversky, 1972, Oscarsson and Rosema, 2019] Surprisingly, although substantial information concerning individual preferences and positions may be expected to lie in such consideration stages, most research has confined itself to the eventual choice.

A prototypical example of a choice process with a substantial amount of indecisiveness is voting in a multiparty system. Even though it is generally known that a relevant proportion of voters in multiparty systems is still undecided in pre-election polls (e.g. Arcuri et al. [2008]), current polling designs do not allow those still undecided voters to provide their current position accurately. Indeed, these respondents are confronted with, respectively unsatisfactory, possibilities: they can give a spuriously precise answer by more or less arbitrarily selecting one of the parties from their consideration set; they can drop out; or they can tick the basically meaningless 'don't know' category, comprising a mélange of the most diverse opinions. All these types of responses are detrimental, as either error is induced or valuable information is lost for a proper picture of the political landscape. As undecided voters appear to relevantly differ from the undecided as worked out by Oscarsson and Oskarsson [2019], simply neglecting them leads to a restricted and biased analysis.

In this paper, we argue that it is possible to exploit the valuable information in the consideration stage(s) in a comprehensive way, providing a vivid basis for insightful structural modelling of the respondents' preferences, positions, and attitudes. We advocate and introduce a direct *consideration set sampling*, where – without the need to change the

sample design itself – the respondents are enabled to provide their current position by listing all alternatives they are still pondering between.<sup>1</sup> This extension has several appealing properties.

- First of all, the information is collected in an undistorted way on an adequate level of coarseness, taking properly into account the fact that the respondents typically ponder between a few of the parties only and thus are neither completely indifferent nor already capable of stating one single party.
- A second main advantage is the intuitive character of the answer for the undecided, who are, as argued above, naturally thinking in set-valued structures building on the choice process in stages (e.g. Oscarsson and Oskarson [2019], Plass et al. [2015]) and thus directly find themselves with their position in the questionnaire.
- Thirdly, as concretely practised in this work in cooperation with the German polling institute Civey, implementing consideration set sampling in a survey proves to be relatively simple and cost-effective, as only one additional question is required to properly record the individual set-valued information.
- And, last but not least, we show that it is immediately possible to embed methodologically the set-valued observations arising from consideration set sampling into the framework of so-called (conjunctive, ontically interpreted) random sets (e.g., Couso et al. [2014]). Since in this formalization process, the underlying random variable proves to be again of a categorical scale of measurement, popular methods for categorical data analysis like discrete choice modelling as well as standard machine learning procedures can successfully be transferred.

Indeed, consideration set sampling provides a powerful basis for an insightful structural analysis of the respondents' genuine positions and attitudes, as we will demonstrate with the example of a German pre-election poll. Closely cooperating with one of the biggest German polling institutes Civey<sup>2</sup>, we are able to provide a carefully implemented first proof of concept of our methodological concepts. Eight weeks prior to the German 2021 Bundestag Election, 5076 voters participating in the survey have been asked for their consideration set, whereby 22.8% indeed were still undecided, providing a non-singleton consideration set. Our structural modelling of the political landscape based on the consideration sets, in particular, contains an adoption to our context of electoral choice models, several exemplary interpretable machine learning methods, and a cluster analysis as a prototypical unsupervised technique. Thus, this work operationalizes, puts into practice, and statistically fertilizes fundamental theoretical insights into the choice process in general (e.g. Tversky [1972]) as well in voting research (e.g. Oscarsson and Rosema [2019]), hereby systematizing and substantially extending some first elementary corresponding deliberations presented at a symposium, a workshop and in a preliminary exploratory analysis prior to the election. [Plass et al., 2015, Kreiss and Augustin, 2020, 2021]

Concretely, this paper proceeds as follows: In Section 2.1, we formalize consideration set sampling by random sets and show under which circumstances methodology can generally be adapted for structural analysis. Then, in Section 2.2, we suggest in our framework some explicit approaches. This includes regression-based modelling (Section 2.2.1) as well as techniques from interpretable machine learning (Section 2.2.2) and unsupervised learning (Section 2.2.3). Afterwards, in Section 3, we apply the discussed methods to real data collected in cooperation with Civey and show how new insights can be obtained. In the concluding remarks in Section 4, we reflect on further avenues and opportunities enabled by consideration set sampling.

## 2 Methodical Framework for a Structural Analyses of Consideration Sets

### 2.1 Consideration Set Sampling and Random Sets

To describe and implement our consideration set sampling, consider a finite set  $\mathcal{S} = \{a, b, c, \dots\}$  of unordered elements, for instance, the major parties in a pre-election poll in a multi-party system,  $P(\mathcal{S})$  its power set, and a space  $\mathcal{X}$  collecting potential values of the covariates, in particular, socio-demographic characteristics in our example. We sample  $n$  units from the underlying population  $\Omega$ , whereby we suppose the population to be large enough to allow for treating it as infinite, neglecting specific finite population effects. The deliberations in the sequel rely on simple random sampling (with replacement) producing i.i.d. random variables/random elements.<sup>3</sup> For technical reasons or context-dependent choices, the set of all options of the power set usually has to be reduced to a subset, denoted as  $\tilde{P}(\mathcal{S})$ , such that  $\tilde{P}(\mathcal{S}) \subset P(\mathcal{S}) \setminus \emptyset$  and  $\{q\} \in \tilde{P}(\mathcal{S})$ , for all  $q \in \mathcal{S}$ .

<sup>1</sup>Asking participants for rankings of the options as suggested, for example, by Fürnkranz and Hüllermeier [2011] does not address the underlying issue here, as the order would either induce a first choice or partial ranking leading again to set-valued data for the decision of interest.

<sup>2</sup>More information is available under <https://civey.com/ueber-civey>, last visited 2023.07.25

<sup>3</sup>Techniques to adopt to complex sampling designs (e.g., Skinner and Wakefield [2017]) should be transferable in principle; a detailed discussion of this issue, however, is left to further research.

We hence obtain realizations  $(\varphi_1, x_1), \dots, (\varphi_i, x_i), \dots, (\varphi_n, x_n)$  of the random elements<sup>4</sup>

$$(\mathcal{Y}_i, X_i) : \Omega \longrightarrow \tilde{P}(\mathcal{S}) \times \mathcal{X} \quad (1)$$

$$\omega_i \mapsto (\varphi_i, x_i). \quad i = 1, \dots, n, \quad (2)$$

The random elements  $\mathcal{Y}_i$  are random sets (e.g., Molchanov [2005]), formalizing the crucial aspect that the observed considerations sets typically are set-valued, i.e. consisting of several elements of  $\mathcal{S}$  and thus being an element of the reduced power set  $\tilde{P}(\mathcal{S})$ . For the sake of a unified representation, the response of an already decided respondent is described as a singleton observation, i.e. the already taken decision for party  $q \in \mathcal{S}$  is identified with the consideration set  $\{q\}$ .

Random sets have two fundamentally different interpretations. Following the terminology in Couso et al. [2014], an *ontic* and an *epistemic* view have to be distinguished. The ontic view understands a set-valued observation as a holistic entity and thus as an irreducible, precise observation per se. The epistemic view, in contrast, sees the set as an imprecise, coarsened description of a genuinely precise outcome, which would correspond to the eventual choice in our context. Apart from some brief remarks in the Concluding Remarks in Section 4, we focus throughout our paper on the ontic point of view, being interested in a structural analysis of indecision between certain parties as a specific political position of its own. Consequently, set relations between the consideration sets are (taken as) meaningless: For instance, indecision between two parties  $a_1$  and  $a_2$  is in no way set-theoretically related to indecision between  $a_1, a_2$  and  $a_3$ ; the consideration sets  $\{a_1, a_2\}$  and  $\{a_1, a_2, a_3\}$  reflect two different, mutually unrelated positions of their own. Technically, this means that – with the underlying space of options  $\mathcal{S}$  being categorical –  $\tilde{P}(\mathcal{S})$  is categorical as well, making the random sets  $\mathcal{Y}_i, i = 1, \dots, n$ , categorical random elements, to which statistical methods for analyzing categorical variables, therefore, can be transferred. We make direct use of this crucial methodological fact. Thus, in the sequel we build our structural analysis on categorical regression with outcome space  $\tilde{P}(\mathcal{S})$  as well as on interpretable learning methods with  $\tilde{P}(\mathcal{S})$  as their target space and unsupervised learning techniques with underlying space  $\tilde{P}(\mathcal{S}) \times \mathcal{X}$ . For the sake of clarity, we will rely on the following notational convention: small standard Latin letters denote elements of  $\mathcal{S}$  (cf. above), while small calligraphic letters like  $\varphi$  stand for elements of  $\tilde{P}(\mathcal{S})$ .

## 2.2 Structural Analysis based on Consideration Set Sampling

Using the set-valued information about the current position, we want to gain a more comprehensive picture in our example of the political landscape and political positions. Building on the general concepts developed in the previous section, we now provide explicit modelling approaches demonstrating how applied research can benefit from consideration set sampling. We first discuss the approaches in a general setting and then apply them to our data set. Hereby we suggest methods from regression-based choice modelling as well as interpretable and unsupervised learning, all based on the i.i.d. sample  $(\varphi_1, x_1), \dots, (\varphi_i, x_i), \dots, (\varphi_n, x_n)$ .

### 2.2.1 Multinomial Regression Approaches

As argued in Section 2.1, the random sets  $\mathcal{Y}_i, i = 1, \dots, n$ , are categorical random elements, and thus we can directly transfer the common multinomial settings to our situation. We will briefly investigate the marginal distribution of the considerations sets, i.e. the positions  $\mathcal{Y}_i$ , and then turn to regression-based approaches, where the positions are seen as dependent and the covariates  $X_i$  as independent variables.

We firstly look at the marginal distribution of  $\mathcal{Y}_1, \dots, \mathcal{Y}_n$  with  $\pi_\varphi \stackrel{\text{def}}{=} P(\{\mathcal{Y}_i = \varphi\}), \varphi \in \tilde{P}(\mathcal{S})$ , and  $\sum_{\varphi \in \tilde{P}(\mathcal{S})} \pi_\varphi = 1$ . The underlying samples can be summarised by an appropriate count statistic  $(N_\varphi)_{\varphi \in \tilde{P}(\mathcal{S})}$  with  $N_\varphi \stackrel{\text{def}}{=} |\{i \mid \mathcal{Y}_i = \varphi\}|$  reflecting how many respondents state category  $\varphi \in \tilde{P}(\mathcal{S})$ . With the corresponding sample denoted by  $(n_\varphi)_{\varphi \in \tilde{P}(\mathcal{S})}$ , this results in a multinomial likelihood

$$\text{lik}((\pi_\varphi)_{\varphi \in \tilde{P}(\mathcal{S})} \mid (n_\varphi)_{\varphi \in \tilde{P}(\mathcal{S})}) \propto \prod_{\varphi \in \tilde{P}(\mathcal{S})} (\pi_\varphi)^{n_\varphi},$$

yielding naturally the relative frequencies  $\frac{n_\varphi}{n}$  of the observed positions  $\varphi$  as maximum likelihood estimates of the corresponding probabilities  $\pi_\varphi$ .

<sup>4</sup>Measureability is not explicitly problematized here. Working with an underlying measurable space  $(\Omega, \mathcal{A})$  and an appropriate  $\sigma$ -field  $\mathcal{F}$  over  $\mathcal{X}$ , we rely canonically on the smallest  $\sigma$ -field generated by  $P(\tilde{P}(\mathcal{S})) \times \mathcal{F}$ .



We base our regression analysis on the adaption of multinomial regression models, which are common in marketing and voting research (e.g. Tutz [2011], Hensher and Johnson [2018]) as they enable a natural linear interpretation of the coefficients. For our categorical random elements, we can then write the multinomial logit model following [Tutz, 2011, p. 211 ff.] with linear predictor consisting of the covariate vector  $x_i$  together with the category-specific parameters  $(\beta_{\mathcal{J}})_{\mathcal{J} \in \tilde{P}(\mathcal{S})}$  in its generic form as:

$$P(\mathcal{Z}_i = \mathcal{J} | x_i) = \frac{\exp(x_i^T \beta_{\mathcal{J}})}{\sum_{\mathcal{J} \in \tilde{P}(\mathcal{S})} \exp(x_i^T \beta_{\mathcal{J}})}, \quad (3)$$

whereby fixing a reference category or symmetric side constraint is necessary to ensure well-definiteness. Estimation can be obtained via maximum likelihood or along the Bayesian paradigm.

In most settings with consideration set sampling, the number of potential groups is rather large, and even after the original reduction from  $P(\mathcal{S})$  to  $\tilde{P}(\mathcal{S})$  observations might get stretched thin. From an applied standpoint, it is recommended to limit the degrees of freedom needed for example by additional regularization.

Such regularization can be conducted with different concepts. With the penalization parameter  $\lambda$ , the penalized log-likelihood  $pl(\beta)$  can be written in the general form with regards to the usual log-likelihood contribution of the  $i$ -th observation  $l_i(\beta)$  as:  $pl(\beta) = \sum_{i=1}^n l_i(\beta) - \frac{\lambda}{2} J(\beta)$  with  $J(\beta)$  as a function penalizing the parameters. [Tutz, 2011, p. 233 ff.] Both  $L1$  and  $L2$  penalization are possible as well as it can be beneficial to group coefficients to remove some covariates entirely, following Tutz et al. [2015] or Vincent and Hansen [2014]. The penalization parameter  $\lambda$  is usually determined by cross-validation, but relying on the Akaike or Bayesian information criterion is also possible. All covariates outlasting the regularization do have some contribution but are not necessarily significantly different from zero. Confidence intervals and hypothesis testing are impeded by regularization but not impossible. (see e.g. Minnier et al. [2011])

An adequate model has to be fitted context-dependent. Further extensions to additive models (e.g. Ravikumar et al. [2009]), ordinal covariates or alternative specific coefficients (e.g. Hensher and Johnson [2018]) can be incorporated as well. With such a regression model, characteristics of the interesting positions can be analyzed, providing a new opportunity to gain empirically founded insights about the undecided.

### 2.2.2 Interpretable Machine Learning Methods

The field of interpretable machine learning provides a range of approaches to examine structural properties in this data situation parallel to the regression approaches. For this, we learn a classifier amongst the options of  $\mathcal{Z}_i$  with observations  $(\mathcal{J}_1, x_1), \dots, (\mathcal{J}_n, x_n)$ , in the classical supervised learning setting and then examine according to which structures the model classifies. The advantage of supervised machine learning lies in the relaxation of model classes, which allows for non-linear relations and (higher-order) interactions between the covariates. Additionally, ordinal covariates and restrictions with degrees of freedom can be easier included. A carefully chosen learner should hence represent the connection between the positions and covariates more flexibly than the regression approaches suggested above, however, at the cost of the straightforward interpretation. In order to regain (some of) the interpretability, several approaches have been developed. (e.g. Molnar [2022]) They cover global methods like partial importance (PI) or the permutation feature importance (PFI) and local methods like *Local Interpretable Model-Agnostic Explanations* (LIME) or *SHapley Additive exPlanations* (SHAP). [Molnar et al., 2020, p. 2] Depending on the underlying question, different approaches are worth considering.

For the corresponding model, it is then possible to look at different approaches and model classes. While the Feature Importance gives an immediate intuition about the influence of covariates and the Partial Dependence Plots provide a nice but simplified picture, the local and more complex approaches are more thorough. For example, *SHAP-Values* introduced by [Lundberg and Lee, 2017] from coalitional game theory, try to determine the contribution of one feature to the overall prediction. One has to keep in mind that the results are reliant on the applied model and, with it, inevitably on the underlying data. Fitting problems, bad generalizations, or issues with imbalanced data must be addressed, due to their influence on the interpretation. [Molnar et al., 2020, p. 5] Especially, as with increasing groups, the data becomes more imbalanced one has to be cautious with statements about the smaller groups, and oversampling methods like *SMOTE* [Chawla et al., 2002] can be considered in the multi-group case. If the learner is chosen and tuned correctly, these approaches provide an interesting and additional insight into patterns and connections between the political positions and the covariates of interest.

### 2.2.3 Unsupervised Learning and Partitions of the Population

Unsupervised machine learning approaches take a further and different look at the data situation of the underlying space  $\tilde{P}(\mathcal{S}) \times \mathcal{X}$ . With this, we are interested in finding patterns and structural peculiarities in the data with a connection

to the different groups of undecided participants. One interesting way to illustrate the partitioning of the underlying population is first to determine socio-demographic clusters and then examine the choice positions in these clusters. This gives an impression and intuition about the interconnection of the positions and the covariates. There are several ways to determine clusters in the population. Either they are already given due to content-related restrictions, or machine learning approaches can be applied for the determination. (e.g. Ghahramani [2003], Hastie et al. [2009]) This is one way to discriminate the participants according to their features and illustrates to what extent distinctive groups can be associated with particular political positions. Another idea is to regard particular natural population partitions from sociological or political research and then illustrate the groups amongst these strata.

In the following chapter, we implement approaches from all three methodological categories to a pre-election poll for the 2021 German federal election, demonstrating how new insights can be obtained due to the accurate representation of undecided voters.

### 3 Application to the 2021 German Federal Election

#### 3.1 The Data

We apply exemplary approaches from all three methodological categories discussed above to the sample two months before the 2021 German federal election. A two-stage survey format with consideration set sampling was developed in cooperation with the polling institute Civey<sup>5</sup>, collecting the set-valued information for the undecided next to 11 categorical or ordinal features about socioeconomic status and demographics. First, participants are distinguished whether or not they are uncertain about their vote and, in a subsequent step, asked about either that particular party or all of their viable options. Unfortunately, no questions concerning positions on political issues were included in the survey, despite their perceived increasing importance for voters' choice.

In Germany, there are currently six relevant parties likely to surpass the 5% hurdle (typically) necessary for being eligible for a seat in parliament.<sup>6</sup> Those are: The Left, Green Party, SPD, CDU/CSU, FDP, AfD.<sup>7</sup> There is no natural ordering of the parties as classification like left-to-right scale or liberal-conservative somewhat lost meaning in the shifting political spectrum. (e.g. Dippel et al. [2022]) We focus only on the main parties here, resulting in a six-dimensional initial space  $\mathcal{S}$ .

Civey constructs online surveys, with the benefits of large sample sizes but the downside of an initially non-random sample of the voting population. To address this, they implement a post-stratification process by sub-sampling from an initial five times bigger sample to achieve approximate representativity.<sup>8</sup> Regardless of the potential error induced hereby and possibly missing not at random structures, Civey established itself with this procedure amongst other polling institutes in Germany. They hence provide us with a state-of-the-art sample with 5076 observations two months before the German federal election, which we treat as an i.i.d. sample in coherence with Civey's post-stratification in the following applications.

#### 3.2 Application

In this section, we illustrate how the modelling approaches in chapter 2.2 provide interesting insights into both decided and undecided voters two months prior to the German federal election. While the basic methodological considerations are our main interest, we also provide brief interpretations of the results for each corresponding approach without pushing us in the field of political research.

##### 3.2.1 Applied Grouped Regularized Regression

Starting with the regression approach, we require a feasible minimum of observations in each group as well as sparse covariates to avoid perfect separation. We examine the six relevant parties together with the three biggest consideration sets of the individuals undecided between SPD/Green/Left, SPD/Green and CDU/CSU/FDP for an overview. For this, we binarize covariates and employ grouped L1 regularization from the *glmnet* package [Friedman et al., 2010] and determine the regularization parameter  $\lambda$  with cross-validation. This reduces the number of the coefficients illustrated

<sup>5</sup>See also the news article about our cooperation under <https://civey.com/ueber-civey/unsere-methode/artikel/civey-unterstuetzt-forschung-zur-mitberuecksichtigung-unentschlossener>, last visited 23.07.23

<sup>6</sup>For more information about the German voting system see: <https://www.bundeswahlleiterin.de/bundestagswahlen/2021/informationen-waehler/wahlssystem.html>, last visited 2023.07.25

<sup>7</sup>Note that CDU/CSU is seen as one party here, even though it is assembled from different regions of Germany

<sup>8</sup>Further information about the Civey procedure can be found in Richter et al. [2022]

in Table 1, which can be interpreted as log odds with a symmetric constraint to enable estimation.<sup>9</sup> Due to the grouping, four covariates were regularized precisely to zero.

	Left	SPD/ Green/ Left	Green	SPD/ Green	SPD	CDU/ CSU	CDU/ CSU/ FDP	FDP	AfD
(Intercept)	-0.30	-1.47	0.81	-0.93	0.63	1.58	-1.11	0.05	0.73
City Habitant	0.01	0.03	0.09	-0.02	0.08	-0.07	-0.03	-0.00	-0.08
Low Purchasing Power	0.04	0.01	-0.04	-0.01	0.03	-0.06	-0.02	0.00	0.04
University Degree	0.02	0.01	0.18	-0.00	-0.09	-0.06	0.04	0.03	-0.13
Married	-0.09	-0.11	-0.12	0.01	-0.01	0.20	0.03	0.04	0.05
Non-Religious	0.19	0.01	0.07	-0.02	-0.10	-0.32	-0.03	0.02	0.19
Former West Germany	-0.44	0.09	0.37	0.16	0.30	-0.06	0.10	0.05	-0.56

Table 1: Results from a (group)-regularized multinomial logit model with symmetric constraint and binarized covariates. Four covariates are regularized precisely to zero and are not included in the table.

Two major insights occur, stressing the benefits of including the undecided voters. First, we are now capable of analyzing the groups undecided between specific parties. Second, we can determine differences between the positions of the single parties and consideration sets containing those parties. This shows how individuals undecided between specific parties really constitute new groups. If we, for example, take the coefficient of the covariate `Married`, we see a slightly positive coefficient for being undecided between the SPD and the Green Party, while the one for those determined to vote for Green and those undecided between the SPD, Green, and Left is negative. The `University Degree` seems to distinguish the Green Party from those undecided between the SPD and Green Party. By examining various variables, both differences and similarities can be identified. This shows that including groups undecided between specific parties contributes something new, highlighting interesting aspects hidden before.

### 3.2.2 Applied Interpretable Machine Learning

We now focus on two groups of particular interest and exemplary employ SHAP-values for structural insights here. We take the perspective of the Green Party and follow a strategically important question: What distinguishes its convinced supporters from those who are also considering voting for my biggest rival, the SPD?

As the underlying model, we choose a gradient boosting tree for binary classification between Green Party and SPD/Green, implemented with the package `h2o Landry` [2022], leading to an error rate of 4.4% on the training and 9.6% on the test data in our exemplary setting. We subsequently investigate how the model made the classification choice by looking at the SHAP-values.

A summary plot of the SHAP-values is given in Figure 1. Here we see the SHAP-based connection of the corresponding variable to the binary classification. The extent of the relation is visualized but without communicating the attached uncertainty. According to the model, the covariate `purchasing power` has the strongest connection to the model’s decision. Especially those with a medium purchasing power seem to be more often classified as undecided. Furthermore, a good distinction can be observed with the covariate `male`. Results like this, potentially accompanied by thorough descriptive analysis, provide essential information for election campaign strategy and give potential starting points to orientate a campaign at. They are furthermore of interest to the neutral observer as well.

### 3.2.3 Applied Unsupervised Learning

We now take a broader look at different socioeconomic classes in our dataset. We strive for a simple overview of the political positions’ connection to socioeconomic classes. To this end, we employ an unsupervised machine learning model relying on tree-based dissimilarity following Shi and Horvath [2006], to establish three different groups with the 11 covariates available as discussed in Section 2.2.3. These three groups are now examined concerning the proportions of political positions. We illustrate the results in Figure 2, with the proportions of political groups in each cluster.

We get intuition and first impressions about the covariates and their connections with the political positions. There are some differences between the groups like a slight overrepresentation of individuals favoring the Left Party in the first cluster, mostly consisting of younger and well-educated participants, and a diminishing proportion of individuals indifferent between the SPD, Green and Left in the third cluster. Overall some differences occur but the clusters seem rather similar, which hints towards a relatively weak discriminative power of the features. This supports the

<sup>9</sup>As mentioned above, hypothesis testing is somewhat controversial in combination with regularization. We did not provide confidence intervals.

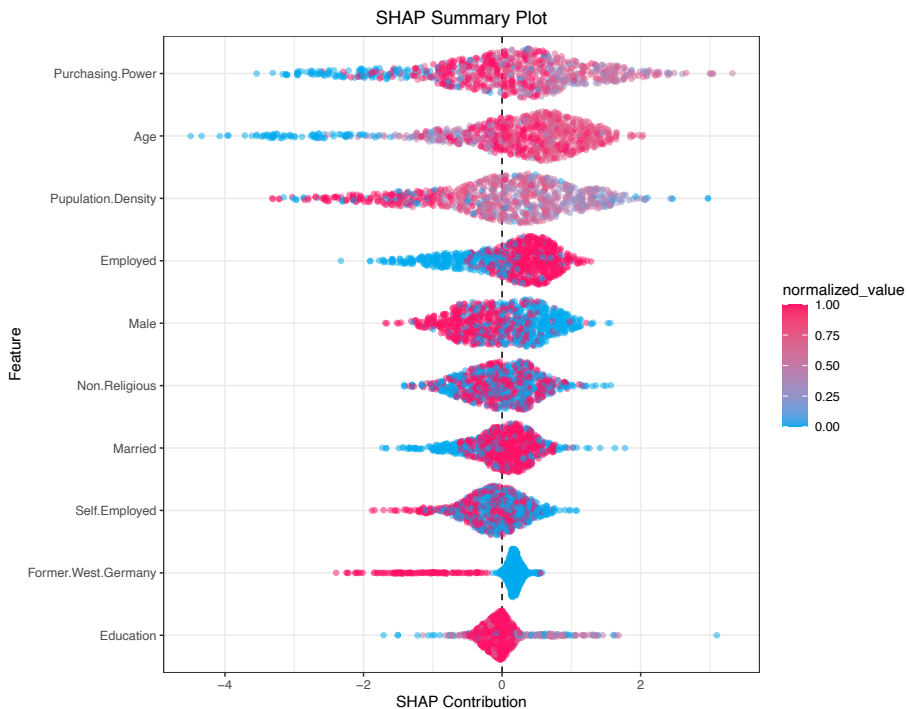


Figure 1: SHAP values for all covariates for every observation in the data. For each covariate, the associated SHAP-value is plotted on the x-axis. Each point represents a single model decision. The more influential a feature is, the more negative or positive its associated SHAP-value. A red point goes in the direction of being undecided between the SPD/Green

idea currently discussed in public that socioeconomic status and demographics do not determine political positions as strongly as it has some decades ago in Germany. The clusters are quite intermixed, hinting that other properties besides socioeconomic ones determine the election choice. One could guess that the polarization within society induced by the pandemic as well as the migration crises of 2015 in Germany, plays a key role here and should be examined in further research. Overall the data gives the impression that socioeconomic status does indeed play a role but only explains the choice process of voting to a rather limited extent.

#### 4 Concluding Remarks

Motivated by the common neglect of undecided survey participants, we developed the framework of consideration set sampling for comprehensive structural analysis. By allowing a set-valued characterization of the relevant group of not yet fully decided respondents, we obtained an accurate representation of their current position that can be interpreted as conjunctive random sets. With this, the established methodology can be successfully transferred. We developed a collection of methodological suggestions tailored to the new situation containing regularized regression and interpretable and unsupervised machine learning. In our application to pre-election polls with a self-constructed survey, a more complete picture of the political landscape arises.

This work can be extended methodologically in multiple directions. Most evidently, approaches under the epistemic view, utilizing consideration sets as partial information to improve forecasting and predictions about the eventual choice, can be developed. First ideas were introduced by Plass et al. [2015], Kreiss and Augustin [2020], Kreiss et al. [2020], but, taking a broader perspective, the framework of consideration set sampling can be embedded both in the theory of learning from imperfect data in statistics and machine learning. In the machine learning context, one could follow up ideas on *Partial Label Learning* [Cour et al., 2011], *Multi-Label Learning* [Zhang and Zhou, 2014], or *Superset Learning* [Hüllermeier, 2014] as explored in Rodemann et al. [2022]. In statistics, the limitations of the strong assumption of *Coarsening at Random* [Heitjan and Rubin, 1991] are interesting. In this setting, a tradeoff between the plausibility of the underlying assumptions and the conciseness of the results has to be addressed, following Manski’s law of decreasing

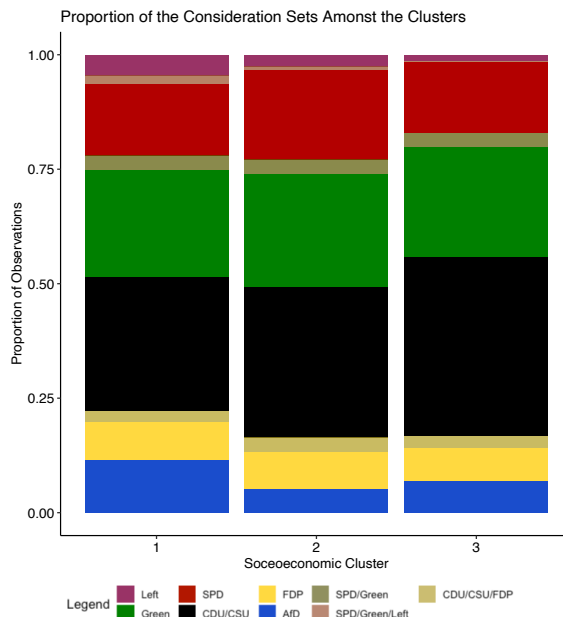


Figure 2: Three socio-demographic clusters determined by tree-based dissimilarity. In each cluster, the proportion of decided and undecided survey participants is shown.

credibility [Manski, 2003, p. 1], opening up avenues for approaches with partial identification [e.g. Molinari, 2020, Arpino et al., 2014, Li et al., 2023].

As a connection between the ontic and the epistemic view, analyzing the shifts of consideration sets over time in a longitudinal manner could provide further new insights into the choice process.

We believe that interpreting consideration set as conjunctive random sets gives undecided respondents appropriate statistical representation. The advantages of structural analysis are evident: simple implementation, undistorted information collection as well as new insights on socio-demographic determinants of both undecided and decided voters.

## 5 Competing interests

No competing interest is declared.

## 6 Author contributions statement

The core ideas of the paper were developed together and initially inspired by TA. Most parts of the paper were drafted and written together, while DK implemented the methodology and wrote the applied section, and TA contributed the embedding in the theory of random sets. The specific applied approaches were suggested by DK.

## 7 Acknowledgments

We are most grateful to Civey for the intensive cooperation, the concrete implementation of our ideas, sharing of the data and continuous support. Financial and general support from the LMU Mentoring Program (DK) and the Federal Statistical Office of Germany within the cooperation “Machine Learning in Official Statistics” (TA and DK) is gratefully acknowledged. Furthermore, we thank Julian Rodemann and Gunnar König for the fruitful discussions.

## References

- L. Arcuri, L. Castelli, S. Galdi, C. Zogmaister, and A. Amadori. Predicting the vote: Implicit attitudes as predictors of the future behavior of decided and undecided voters. *Political Psychology*, 29:369–387, 2008.
- B. Arpino, E. De Cao, and F. Peracchi. Using panel data for partial identification of human immunodeficiency virus prevalence when infection status is missing not at random. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 177(3):587–606, 2014.
- N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
- T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12: 1501–1536, 2011.
- I. Couso, D. Dubois, and L. Sánchez. *Random sets and random fuzzy sets as ill-perceived random variables*. Springer, 2014.
- A. Dippel, L. Hetzer, and A. Burger. Links oder rechts? Die ideologische Selbstverortung von Wähler:innen und ihre Wahrnehmung von Parteien in Deutschland. *easy\_social\_sciences*, 67:19–29, 2022.
- J. Friedman, R. Tibshirani, and T. Hastie. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.
- J. Fürnkranz and E. Hüllermeier. Preference learning and ranking by pairwise comparison. In J. Fürnkranz and E. Hüllermeier, editors, *Preference Learning*, pages 65–82. Springer, 2011.
- Z. Ghahramani. Unsupervised learning. In *Summer school on machine learning*, pages 72–112. Springer, 2003.
- T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, and J. Friedman. Unsupervised learning. *The elements of statistical learning: Data mining, inference, and prediction*, pages 485–585, 2009.
- D. Heitjan and D. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244–2253, 1991.
- D. A. Hensher and L. W. Johnson. *Applied discrete-choice modelling*. Routledge, 2018.
- E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014. Special issue: Harnessing the information contained in low-quality data sources.
- D. Kreiss and T. Augustin. Undecided voters as set-valued information, towards forecasts under epistemic imprecision. In J. Davis and K. Tabia, editors, *International Conference on Scalable Uncertainty Management*. Springer, 2020.
- D. Kreiss and T. Augustin. Towards a paradigmatic shift in pre-election polling adequately including still undecided voters: Some ideas based on set-valued data for the 2021 German federal election. *arXiv preprint arXiv:2109.12069*, 2021.
- D. Kreiss, M. Nalenz, and T. Augustin. Undecided voters as set-valued information, machine learning approaches under complex uncertainty. In E. Huellermeier and S. Destercke, editors, *ECML/PKDD 2020 Tutorial and Workshop on Uncertainty in Machine Learning*. 2020. URL <https://drive.google.com/file/d/1abrLGZ154htGuYz8HzyLQzJ8vyc3kr2K/view?pli=1>. last visited: 2023.07.26.
- M. Landry. *Machine Learning with R and H2O*, October 2022. URL <https://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/RBooklet.pdf>. last visited 2023.07.25.
- H. Li, D. L. Millimet, and P. Roychowdhury. Measuring economic mobility in India using noisy data: a partial identification approach. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(1):84–109, 2023.
- S. Lundberg and S. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4768–4777, 2017.
- C. Manski. *Partial identification of probability distributions*. Springer, 2003.
- J. Minnier, L. Tian, and T. Cai. A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106(496):1371–1382, 2011.
- I. Molchanov. *Theory of random sets*. Springer, 2005.
- F. Molinari. Microeconometrics with partial identification. In S. Durlauf, L. P. Hansen, J. Heckman, and R. Matzkin, editors, *Handbook of Econometrics, Volume 7A*, volume 7 of *Handbook of Econometrics*, pages 355–486. Elsevier, 2020.
- C. Molnar. *Interpretable Machine Learning*. 2022. 2023 2nd edition (Version: 2023-07-22) <https://christophm.github.io/interpretable-ml-book>.

- C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, editors, *International workshop on extending explainable AI beyond deep models and classifiers*, pages 39–68. Springer, 2020.
- H. Oscarsson and M. Oskarson. Sequential vote choice: Applying a consideration set model of heterogeneous decision processes. *Electoral Studies*, 57:275–283, 2019.
- H. Oscarsson and M. Rosema. Consideration set models of electoral choice: Theory, method, and application. *Electoral Studies*, 57:256–262, 2019.
- J. Plass, P. Fink, N. Schöning, and T. Augustin. Statistical modelling in surveys without neglecting ‘The undecided’. In T. Augustin, S. Doria, E. Miranda, and E. Quaghebeur, editors, *ISIPTA 15*, pages 257–266. SIPTA, 2015.
- P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030, 2009.
- G. Richter, T. Wolfram, and C. Weber. Die statistische Methodik von Civey. 2022. Online article under <https://civey.com/whitepaper>, last visited: 2023.07.22.
- J. Rodemann, D. Kreiss, E. Hüllermeier, and T. Augustin. Levelwise data disambiguation by cautious superset classification. In F. Dupin de Saint-Cyr, M. Öztürk Escoffier, and N. Potyka, editors, *International Conference on Scalable Uncertainty Management*, pages 263–276. Springer, 2022.
- T. Shi and S. Horvath. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1):118–138, 2006.
- A. Shocker, M. Ben-Akiva, B. Boccara, and P. Nedungadi. Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing Letters*, 2(3):181–197, 1991.
- C. Skinner and J. Wakefield. Introduction to the design and analysis of complex survey data. *Statistical Science*, 32(2):165–175, 2017.
- L. Stocchi, M. Banelis, and M. Wright. A new measure of consideration set size: The average number of salient brands. *International Journal of Market Research*, 58(1):79–94, 2016.
- G. Tutz. *Regression for Categorical Data*. Cambridge University Press, 2011.
- G. Tutz, W. Pöbnecker, and L. Uhlmann. Variable selection in general multinomial logit models. *Computational Statistics & Data Analysis*, 82:207–222, 2015.
- A. Tversky. Elimination by aspects: A theory of choice. *Psychological Review*, 79(4):281, 1972.
- M. Vincent and N. R. Hansen. Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis*, 71:771–786, 2014.
- M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

**Eidesstattliche Versicherung**

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

**Kreiß, Dominik**

-----  
Name, Vorname

**München, 19.10.23**

Ort, Datum

Dominik Kreiß

Unterschrift Doktorand/in