GRAZSCHUMPETERCENTRE

GSC Discussion Paper Series

Paper No. 29

Predicting Voter Ideology Using Machine Learning

Patrick Mellacher[1] and Gernot Lechner[2]

[1] University of Graz, Graz Schumpeter Centre,
Universitätsstraße 15/FE, A-8010 Graz

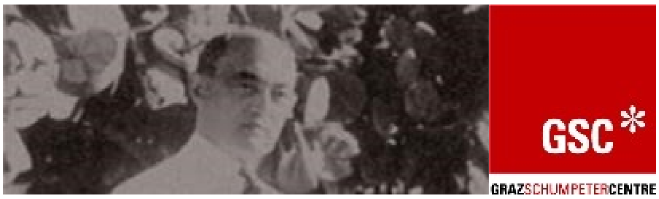
[1] University of Graz, Institute for Operations and Information Systems,
Universitätsstraße 15/E3, A-8010 Graz

Abstract

Many surveys require respondents to place themselves on a left-right ideology scale. However, respondents may not understand the scale or their ‘objective’ position. Furthermore, a uni-dimensional approach may not suffice to describe ideology coherently. We thus propose a novel way to measure voter ideology: Combining expert and voter survey data, we use machine learning to infer how political experts would place voters on three axes: general left-right, economic left-right and social/cultural ‘GAL-TAN’. Our analysis suggests that i) voters are more likely to place themselves at the political center than we would predict experts to do, ii) voters are ideologically most fragmented along the 'GAL-TAN' axis, iii) European countries differ significantly in all ideological dimensions, and iv) ‘objective’ ideology as predicted by our models improves the predictive power of simple spatial voting models even after accounting for the subjective ideological distance between voters and parties as perceived by the voters themselves.

Keywords: machine learning, random forest, voter ideology, political economy, spatial voting

JEL-Codes: C38, D70, D72



Predicting Voter Ideology Using Machine Learning

Patrick Mellacher[1] and Gernot Lechner[2]

[1]Graz Schumpeter Centre, Universitaetsstrasse 15/FE, Graz, 8010, Austria, patrick.mellacher@uni-graz.at
[2]Insitute of Operations and Information Systems, Universitaetsstrasse 15/E3, Graz, 8010, Austria, gernot.lechner@uni-graz.at

January 18, 2023

Abstract

Many surveys require respondents to place themselves on a left-right ideology scale. However, respondents may not understand the scale or their ‘objective’ position. Furthermore, a uni-dimensional approach may not suffice to describe ideology coherently. We thus propose a novel way to measure voter ideology: Combining expert and voter survey data, we use machine learning to infer how political experts would place voters on three axes: general left-right, economic left-right and social/cultural ‘GAL-TAN’. Our analysis suggests that i) voters are more likely to place themselves at the political center than we would predict experts to do, ii) voters are ideologically most fragmented along the ’GAL-TAN’ axis, iii) European countries differ significantly in all ideological dimensions, and iv) ‘objective’ ideology as predicted by our models improves the predictive power of simple spatial voting models even after accounting for the subjective ideological distance between voters and parties as perceived by the voters themselves.

1 Introduction

Studies on voter ideology most often rely on surveys where respondents place themselves on a uni-dimensional scale, in particular left-right (e.g. Knutsen, 1998; De Vries et al., 2013), or, in the case of the US, liberal-conservative (e.g. Gelman et al., 2016).

This approach has three apparent problems: First, respondents may not understand the meaning of ’left’ and ’right’ at all or in the same way as other respondents (Bauer et al., 2017; Palfrey and Poole, 1987; Jessee, 2010). Second, they may have a biased view of their own position, for instance because the structure of their social network makes them believe that their right-wing (or left-wing) views are shared by the majority of the population and thus represent the political center. Such a ’center bias’ is well-documented with regard to individuals’ beliefs about their position in the income distribution (e.g. Fernández-Albertos and Kuo, 2018; Cruces et al., 2013; Bublitz, 2022). Third, a uni-dimensional axis may not suffice to describe a given ideology coherently (Laméris et al., 2018), for instance because a voter may hold views which are economically left-wing and socially conservative.

A second approach is to infer voter ideology based on the ideology of the party they voted for, which itself is assessed either with manifesto data (e.g. Kim and Fording, 2001) or based on the scientific literature (e.g. Funke et al., 2016). While these measures can be very useful, in particular in the face of limited access to voter survey data, this approach cannot separate a vote choice due to ideological proximity from other reasons, such as, e.g., charisma of the favored candidate, which has also shown to play an important role (e.g. Shamir, 1994). Hence, this measure can also not be used for spatial analysis of strategic behavior of political parties (e.g. models in the tradition of Downs 1957), as the ideological stance of a party cannot be separated from the stance of its voters.

Finally, Jessee (2010), Wagner and Kritzinger (2012) and Laméris et al. (2018) construct ideological scales based on answers to policy-related questions. This approach has the advantage that voters are arguably more likely to understand their (true) stance on particular policies than on a rather abstract ideological scale. A second advantage provided by this approach is the fact that it is not limited to a single ideological scale. Laméris et al. (2018) formalize this approach with exploratory factor analysis (EFA), which helps to uncover how many different ideological scales may be important and to which policies each of the ideological scales are correlated with in a data-driven way. After uncovering the different ideological axes, they can be interpreted ex post. Measuring ideology with the help of forty 5-point Likert scale items in a Dutch survey, they find that these items are best explained by four distinct latent ideological dimensions. However, this approach does not allow to compare voter ideology with the ideology of political parties – hence, it also cannot be applied to study spatial party competition. In contrast, Jessee (2010) uses ideal point estimation to argue that a single ideological dimension is well-equipped to predict the stances on 10 yes/no policy items related to the US presidential election of 2008. This approach is explicitly designed to put presidential candidates and voters on the same ideological dimension and hence study spatial competition. However, the result that a single dimension suffices to explain ideology may be driven by the data which limits the number of theoretically possible policy combinations to 2^{10}, which is drastically lower than the 5^{40} combinations enabled by the analysis of Laméris et al. (2018). A distinct approach is chosen by Wagner and Kritzinger (2012), who distinguish between 'socio-economic' and 'socio-cultural' policy items from the outset and then conduct a Principal Component Analysis on each separate set of policy questions.

Our paper follows the tradition of constructing ideology based on survey questions, but takes a different route: Instead of constructing the scale based on correlations between policy stances, we model—by applying machine learning—how political experts construct ideological scales on average, and then use this model to predict how experts would evaluate the ideology of each voter. This approach entails two main advantages of the EFA approach compared to voter self placement. First, it does not require a common understanding of the left-right scale between voters. In fact, it does not require that voters understand this scale at all. Instead, they only need to know their positions on specific policies, which is arguably easier. Second, we are not confined to a uni-dimensional analysis, but can also place voters on two other well-established ideological axes, namely the economic left-right axis and the 'GAL-TAN' axis covering social and ecological positions. In contrast to the EFA, we are limited to ideological scales given by the expert survey. However, we also use EFA to confirm that the set of policy stances that we analyze (as well as a wider set of policy stances) are well-described by exactly two factors that mirror the economic left-right and the 'GAL-TAN' axis.

Furthermore, our approach has two main advantages compared to EFA & PCA: First, this scale places voters and parties on the same ideological scale(s). This potentially allows to improve our understanding of ideological congruence between parties and their voters (see e.g. Bakker et al., 2020b) and spatial voting (see e.g. Merrill, 1999). Second, in contrast to the ideology scales produced by EFA & PCA, our results are easier to interpret quantitatively, as they refer to bounded ordinal scales which are commonly known.

Our empirical strategy exploits the fact that some policy-related questions are identical (or nearly identical) in the voter study of the European Election Study (EES) 2019 (Schmitt et al., 2022) and the Chapel Hill Expert Survey (CHES) 2019 (Bakker et al., 2020a). Accordingly, our paper also stands in the tradition of research that combines data from the CHES and the EES, a link which was previously mostly used to study issue and ideology congruence between parties and voters (Rosset and Stecker, 2019; Bakker et al., 2020b), but also, e.g., by Meyer and Wagner (2020) to understand that citizens' (uni-dimensional) left-right placement of parties in the EES depends on the salience of the economic vs. cultural dimension in the CHES. Our approach also has another main advantage, namely that we may be able to construct ideology scales which are more suitable to conduct cross-country comparisons. This is important, because the ideological center in one country may be considered to be left-wing (or right-wing) in another country.

Our approach is inspired by Downs (1957), who argues that political ideology is a way for voters to minimize their transaction costs, as they do not need to know how each party relates to each policy – instead, they only need to know how each ideology is related to each policy stance, and to which ideology each political party subscribes. This view implies that we can create a 'map' that allows us to connect each political ideology with respective policy preferences – even if some voters may only intuitively use it. It is our quest to shed some light on this map.

The rest of this paper is organized as follows: The second section briefly describes our empirical strategy and the data set, as well as the methods that we employ. The third section shows the results. The fourth section discusses some limitations, and section five concludes.

2 Method & Data

2.1 General Approach

2.2 Method

Even though no standardised process for Data Science has been defined yet due to the dynamics related to this research field (Saltz and Krasteva, 2022), some guiding frameworks are available. We make use of OSEMN—a data science framework which is independent of the domain and discipline: OSEMN stands for *Obtain data*, *Scrub data*, *Explore data*, *Model data*, and *iNterprete results* (Mason and Wiggins, 2010; Esser, 2022). The framework has been referred to in a variety of research projects related to data science in diverse contexts, e.g., nursing (Brennan and Bakken, 2015), photovoltaic systems (David et al., 2020), ethnography (Zafiroglu and Chang, 2018), or traffic classification (Bolanowski et al., 2021). Thus, the chosen procedure follows a typical approach for data science projects covering data acquisition, preprocessing of data, data exploration, feature selection and engineering, training of machine learning models, evaluation of models, and model improvement by tuning hyperparameters. Please note that these steps are not necessarily executed linearly but may require feedback loops: results from a subsequent stage might reveal the necessity to perform further activities in a preceding step. All of the steps are also incorporated in other well-established data science frameworks—like CRISP-DM—and applied analogously in several other studies in various domains (see, for example, Kumar, 2022; Mayo, 2022; Schröer et al., 2021).

2.2.1 Step 1: Obtain data

The main analysis relies on two open access data sets: *The CHES: Chapel Hill Expert Survey 2019* (Bakker et al., 2020a) and the *EES: voter study of the European Election Study 2019* (Schmitt et al., 2022). We refer to the both data sets as CHES 2019 and EES 2019, respectively, in the remaining text.

The aim of the CHES 2019 is to study how experts perceive European political parties in terms of ideology and policy. To this end, 421 political scientists evaluated a total of 277 political parties in the current 27 EU member states plus UK, Iceland, Norway, Switzerland and Turkey between February and May 2020.

The following items are relevant for our purpose, all of which are set up as 11-point Likert scales:

1. an ideological scale called 'general left-right' ranging from 0 (extreme left) to 10 (extreme right), referred to as 'LRGEN';
2. an ideological scale called 'economic left-right', which again ranges from 0 (extreme left) to 10 (extreme right), referred to as 'LRECON';
3. an ideological scale called 'GAL-TAN' (see Hooghe et al., 2002) that aims to capture the social and cultural values of a party ranging from 0 (ranging from green, alternative, libertarian) to 10 (traditional/authoritarian/nationalist), referred to as 'GAL-TAN';
4. questions related to particular policies such as economic redistribution or immigration (these variables serve as independent variables in our models).

A full list of the policy items and their wording, which is copied from the respective codebooks, can be found in the section A.1 of the Appendix. An overview of the data sets as well as related descriptive statistics can be found in section A.2 in the Appendix.

We want to estimate how experts evaluate the ideology of parties in the three dimension using the six independent variables covering policy-related questions. The independent variables and the three dependent ideological scales are hence the basis for all the remaining analyses.

2.2.2 Step 2: Scrub data

Albeit the data sets are prepared in a well-structured format, we need to preprocess the data in order to use it for our analysis. We conducted the following steps in the stated order using Python pandas and scikit-learn (pandas development team, 2020; Pedregosa et al., 2011).

1. *Recategorization*: Since the number of observations available to train our models is rather small compared to the 11 classes given by each ideology scale, we pursue a twofold strategy: We aim to predict both the original 11-point scale, as well as a reduced 3-point scale. While the first exercise aims to give a more detailed picture that is also closer to the data at hand, our second analysis should enhance the accuracy of our predictions, but also allow us to represent ideology in a more comprehensible way in two dimensions.[1] For this purpose, values between 0 and 3 were replaced by 'l' (for left), 4-6 by 'c' (center), and 7-10 by 'r' (right).[2] After this step, we are confronted with two data sets, namely one with an 11-point scale and one with a reduced 3-point scale for the dependent variables. Please refer to section A.3.1 in the Appendix to obtain more information about the resulting data.

2. *Handling missing data of independent variables*: We remove all data rows which contain missing values in any of the six independent variables. This reduces the number of data rows from 3,823 to 2,912. We discuss this step in more detail in section A.3 in the Appendix and show that—even though the loss in the number of observations is substantial—the resulting changes in the structure of the data are marginal.

3. *Partitioning*: Since we want to predict LRGEN, LRECON, and GAL-TAN (i.e. they are the dependent variables in our models), we split the entire data set according to the three ideological scales. This maximizes the number of complete observations per dependent variable. Thus, we are confronted with six different sub-data sets after this step: two types of Likert scales (recategorized 3-point, original 11-point) x three ideological scales (dependent variables).

[1] Together, a 3-point scale for 'economic left-right' and a 3-point scale for 'GAL-TAN' enables us to depict 9 different combinations that arguably include more information than the 11-point scale given by the general left-right axis.

[2] We also experimented with different boundaries, namely that values between 0 and 2 were replaced by 'l', 3-7 by 'c', and 8-10 by 'r', which produced similar results.

4. *Handling missing data of ideological scales*: We remove all data rows where the value for the ideological scale is missing because this data cannot be used to link policy stances to ideology. After this step, the number of observations is different for each data set containing a distinct ideology scale, as we only consider complete observations (i.e. six independent variables + one dependent variable).

5. *Split data sets*: We also take care of a strict validation and testing procedure (cf. Chicco, 2017). Splitting a data set into three independent sub-data sets for training a model, validating its quality, and testing is good practice and state-of-the-art in machine learning: we use a cross-validation approach by separating the data set into independent training and validation parts (see, e.g., James et al., 2021, p. 198ff). The test set is used only after completing the training and parameter optimization phase. It's indispensable that no information is shared between training, validation, and test data set in order to assess the model's performance accurately, avoid overfitting of training models, and thus to create models which are as generalizable as possible (see, e.g., James et al., 2021, p. 32ff). Hence, each of the six different sub-data sets is split into two different parts (90% training and validation at the rate of 4 to 1, 10% test), resulting in twelve sub-data sets.

2.2.3 Step 3: Explore data

In order to increase our understanding of the data at hand and the interrelation between the policy stances, we use data-driven approaches to explore the data by applying basic methods and visualizing relationships in data. To ensure data integrity, as well as to relate our findings to other approaches, we rely on methods known from the literature and expect to obtain similar results. First, we use exploratory factor analysis (EFA) to identify latent constructs in the policy stances included in the CHES 2019 data set. This is important because it may allow us to reveal whether policy stances are indeed driven by some underlying ideologies, hence confirming our approach. Second, we apply a Principal Component Analysis (PCA) to reduce the dimensionality of data while preserving a major part of its information. This makes an initial interpretation of data easier (James et al., 2021, p. 499ff.). PCA reveals potential similarities among different data points by visualizing the originally six-dimensional data on a (two-dimensional) plane.

2.2.4 Step 4: Model data

After data preprocessing and integrity checks, we train different machine learning models based on the CHES 2019. Our working hypothesis, which is supported by EFA and PCA, is that the policy stances of parties as perceived by experts are sufficient to predict how the same experts would place a party on each ideological scale. Thus, we investigate, in a first step, whether the policy-related questions are indeed suitable predictors of the three different ideological axes.

Furthermore, we exploit the fact that the EES 2019 includes six items regarding policy stance that are identical or highly similar to questions included in the CHES 2019. We assume that experts know the relation between policy stances and ideologies, and voters know their policy stance. Hence, we can infer the ideology of voters by estimating how experts categorize parties ideologically given their stance in the multidimensional policy spectrum. In order to do this, we use the data on voter policy stances obtained from the EES to predict how experts would evaluate each voter's ideology and contrast our predictions with their self-reported left-right ideology.

On a technical level, we make use of (multilabel) classification algorithms for supervised learning provided by Python's scikit-learn: Support Vector Machines (SVC), Random Forests

(RFC), Logistic Regression (LRC), and AdaBoost Classifier (ABC).[3] We implemented a grid search including cross validation in order to optimize each algorithm's hyperparameters. Our cross-validation as implemented in scikit-learn is twofold: It optimizes the hyperparameters using the validation data set and then tests each optimized algorithm against a—completely independent—test data set (cf. Chicco, 2017).

2.2.5 Step 5: iNterprete results

We base our interpretation of the models on two pillars: first, we compute several different well-known classification metrics to evaluate the performance of the models. Confusion matrices (see Table 1)—which allow to evaluate the performance of models by opposing actual and predicted values—and the calculation of the mean squared error (Equation 1) for quantifiable values support in understanding the outcomes. As the data sets are imbalanced (i.e. more parties are located in the political center than in the extremes), we add the F1-score (balanced F-score, harmonic mean of precision and recall, see Equation 2) and the balanced accuracy (see Equation 4) to complement the commonly used accuracy (Equation 3). Second, we visualize our predictions in order to compare structural differences of various results.

	Actual Positive	Actual Negative
Predicted Positive	t_p	f_p
Predicted Negative	f_n	t_n

Table 1: Confusion Matrix: true positive t_p, false positive f_p, false negative f_n, true negative t_n

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \text{(with } n \text{ observed } Y_i \text{ and predicted values } \hat{Y}_i\text{)} \tag{1}$$

$$\text{F1S} = \frac{2t_p}{2t_p + f_p + f_n} \tag{2}$$

$$\text{ACC} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{3}$$

$$\text{BAC} = \frac{\frac{t_p}{t_p+f_n} + \frac{t_n}{t_n+f_p}}{2} \tag{4}$$

3 Results

3.1 EFA / PCA

3.1.1 Exploratory Factor Analysis

As a first analysis, we explore the full expert-level data set based on exploratory factor analysis (EFA) using the psych package (Revelle, 2022) for the programming language R (R Core Team, 2022). This is a data-driven approach that allows to identify the latent structure of a data set which was already applied successfully by Laméris et al. (2018) to uncover underlying dimensions of ideology that drive specific policy stances.

We conduct this exercise twice: Once for the six policy items shared by the CHES 2019 (after excluding observations with an NA for any of the variables, n=2912) and the EES 2019 and once for all of the fifteen policy items included in the CHES 2019 (n=1969 complete observations).

[3] For an overview of all available supervised learning algorithms we refer to https://scikit-learn.org/stable/supervised_learning.html.

The first analysis shows how many dimensions we should consider to meaningfully describe how parties relate ideologies to policy stances according to experts. The second analysis, on the other hand, tests whether our results may be biased by the fact that we are only able to rely on six out of fifteen policy items included in total in the CHES 2019.

As a first step, we have to analyze how many factors are appropriate to represent the given data structure. The literature suggests several criteria to determine the appropriate number of factors. The first criterion is the scree test introduced by Cattell (1966), which plots the number of factors against their eigenvalues. This criterion suggests that we should keep the number of factors before a 'kink' sets in. For both types of data sets (i.e. the one with six and the one with fifteen policy items), the scree test suggests 2 factors (see Figure 1. We can also employ the scree plot to evaluate the appropriate number of factors according to the 'Kaiser' criterion, which suggests that we should keep all factors for which the eigenvalue is above 1 (Kaiser, 1960). According to this criterion, we should choose 1 factor for the data set incorporating 6 variables, but 2 factors for the full data set. Finally, the 'minimum absolute partial correlation' (Velicer, 1976) and the 'very simple structure' (Revelle and Rocklin, 1979) criteria both suggest 2 factors for both data sets (as confirmed using the psych package for R by Revelle 2022).

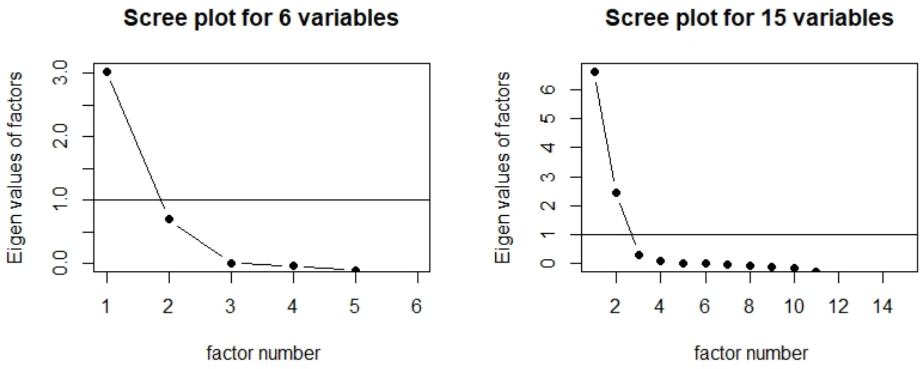


Figure 1: Exploratory Factor Analysis of the six policy stance variables shared by the CHES and EES

These results point to an underlying structure of 2 factors. Figure 2 shows the results of our exploratory factor analyses. For the six policy items shared by the CHES 2019 and EES 2019, the EFA suggests that two (correlated) underlying ideological scales are correlated with 4 and 2 policy stances, respectively. Intuitively, we can see that social, cultural and ecological questions are grouped in one ideological dimension and economic policy stances in a different one. This grouping is mirrored for the data set containing all fifteen policy items. Even for the latter data set, there is only a single variable which is significantly influenced by both factors, namely the question of trade liberalization vs. protectionism. Intuitively, this result suggests that both authoritarian and economically left-wing parties favor protectionism. While the latter result may be driven by the fact that free trade could undermine progressive domestic labor policies, the former may stem from dismissing foreign influence or from politico-economic considerations.

These results suggest i) that the six policy items are sufficiently described by two ideological axes, ii) that this structure generalizes to a much larger set of fifteen policy items, iii) validates the existence and definition of the economic left-right axis and the 'GAL-TAN' axis as largely orthogonal axes, as they seem to also arise from a data-driven approach that does not presuppose any number or definition of ideological axis in particular, and finally iv) also validates our approach, as this analysis suggests a clear relationship between ideology and policy stance, hence supporting us in our endeavor to predict political ideology based on policy stances.

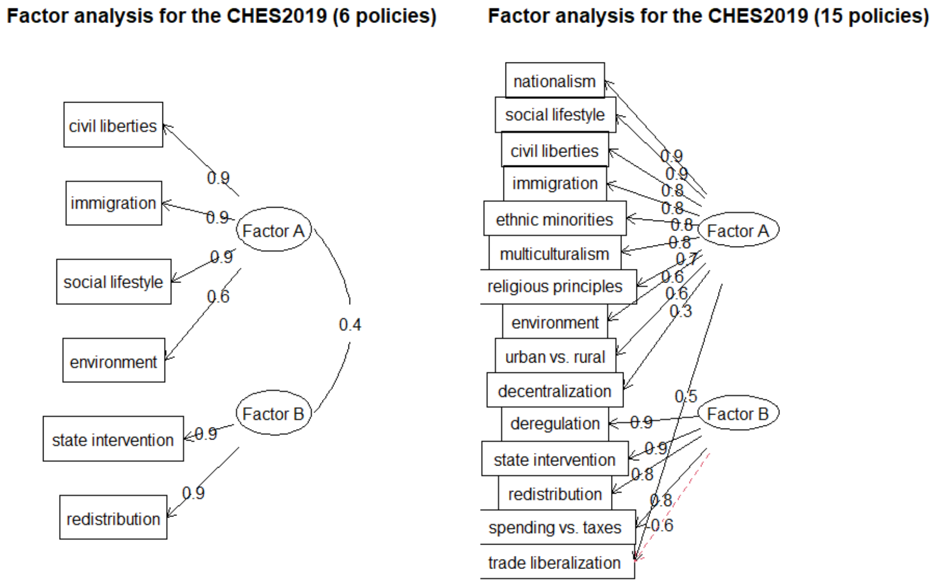


Figure 2: Exploratory Factor Analysis of the six policy stance variables shared by the CHES 2019 and EES 2019 (left) and all fifteen policy stance variables included in the CHES 2019 (right)

3.1.2 Principal Component Analysis

In addition to the EFA using the complete data set, we also conduct a principal component analysis (PCA) on the three training-validation data sets, each containing one of the dependent and all of the six independent variables. A scree plot suggests that the suitable number of principal components is two, and we name them PC1 and PC2. More than 80% of the variance is explained by PC1 and PC2 for each of the data sets. Since our three training-validation data sets are highly similar, the results of the PCA are also highly similar. We thus only visualize the results of the data set containing the general left right ideology (LRGEN) and the six policy stances.

Figure 3 shows not only how our observations are located with regard to the principal components, but also their left-right ideology as assigned by the experts. The data points are not randomly distributed, but they are clustered in some areas on this 2-dimensional planes. Since the large number of data points and categories makes it difficult to see the patterns, in particular for intermediate values of LRGEN, we visualize the centroids as triangles.

Interestingly, the distribution of the left-right scale across the principal components seems to follow an inverted U-curve shape. For instance, the observations containing left-wing radical parties marked by the red dots ('0') are distributed along the bottom-left line of the shape. On the contrary, observations of right-wing radical parties ('10', marked by golden dots) are located on the right boundary of the figure, but their observations also tend to be located at the bottom. This is contrast to 'center' parties, which tend to be in the center of PC1 and the top half of PC2. At the very bottom right of the coordinate system, we can even see that the clusters of left-wing radical and right-wing radical parties slightly overlap.

It is easier to see the underlying clustering according to the general left-right ideology when reducing it to three different levels 'l', 'c', and 'r' (see Figure 4). Although they overlap, the three clusters are clearly visible, and the clustering seems to be mostly driven by PC1.

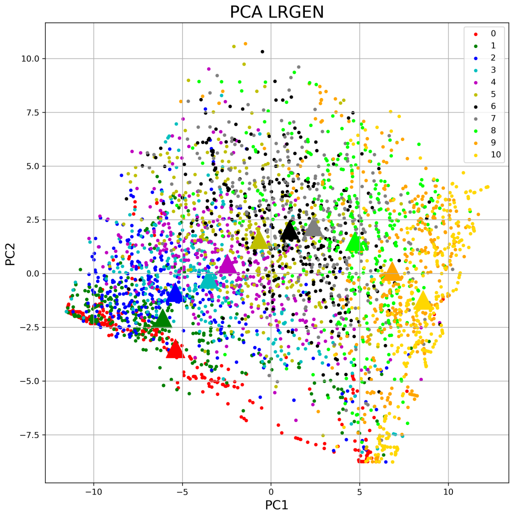


Figure 3: PCA of LRGEN based on an 11-point Likert scale (centroids marked as triangles)

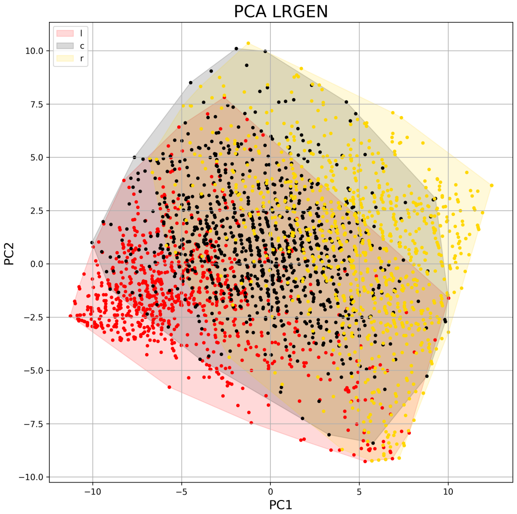


Figure 4: PCA of LRGEN based on a reduced 3-point scale ('l', 'c', 'r')

Based on the results of the exploratory factor analysis and the principal component analysis, two data-driven approaches, we can conclude that the correlation between policy stances indeed indicates that the parties' stances on particular policies are driven by underlying ideologies. This result validates our approach, since it suggests that we can link policies to ideologies and vice versa. Using EFA, we also established that the underlying structure given by six policy stances is mirrored for a much wider set of fifteen policy stances (i.e. all general policy stances included by the CHES), again validating our approach.

Our analysis suggests that a two-dimensional approach to ideology is more suitable than a uni-dimensional one, as cultural, social and ecological values seem to be largely separated from economic ones. However, the uni-dimensional approach also has some merits, as suggested by the 'Kaiser' criterion for EFA, as well as by the fact that the two-dimensional PCA shows that parties are clustered according to their general left-right ideology. We can thus proceed to train and evaluate respective machine learning-models to make use of the full information provided by the six independent variables.

3.2 Predicting voter ideology: Evaluation of methods

Since most of the algorithms that we employ use hyperparameters, we perform a grid search-approach to search for the best parameter calibration over a given search space. Even though this approach does not guarantee to find an optimal result, it allows us to identify well-performing model candidates under the constraint of limited computational power.

We use balanced accuracy (BAC) as the (single) scoring method for evaluating the various models resulting from the grid search during the training-validation phase. We choose this metric as we want to avoid a bias resulting from 'overfitting' our models to the most prevalent values of the dependent variables (which are usually located in the center). The trained and validated models are subsequently tested using test data sets which are independent of the training-validation data set. This approach ensures that no information is transferred from the training and validation stages to the final test. Again, we use this approach of cross-validation in order to avoid overfitting. Apart from balanced accuracy, we also show other common metrics to assess the models' accuracy: accuracy (ACC), F1-score (F1S), and—if reasonable—mean squared error (MSE). Please note that a higher predictive power is indicated by higher values in all metrics except for the MSE.

3.2.1 Results of model evaluation: 11-point Likert scale

As shown in Table 2—best scores are marked by a bold font—, Random Forest Classification (RFC) outperforms all other methods for all investigated metrics. Albeit the accuracy of close to 40% does not seem to be very impressive at first glance, we argue that it in fact is: there are eleven classes in total to predict, and the maximum share of one class related to the total respective ideological scale is 21.18% (class 5, GAL-TAN). Thus, our approach provides much better results than simple strategies, as, e.g., a random draw or simply choosing the most prevalent value.

	LRGEN				LRECON				GALTAN			
	SVC	RFC	LRC	ABC	SVC	RFC	LRC	ABC	SVC	RFC	LRC	ABC
ACC	0.3149	**0.3772**	0.3495	0.3114	0.3160	**0.3924**	0.3194	0.2535	0.3045	**0.3841**	0.3322	0.3045
BAC	0.3374	**0.4425**	0.4245	0.3481	0.3414	**0.4116**	0.3575	0.3140	0.3078	**0.3820**	0.3184	0.2964
F1S	0.3086	**0.3695**	0.3459	0.2739	0.3143	**0.3855**	0.3241	0.1564	0.3031	**0.3840**	0.3272	0.2492
MSE	2.8270	**1.8339**	2.3599	2.2976	2.6979	**1.7674**	2.4167	4.3611	3.3287	**2.1384**	2.2976	3.6125

Table 2: Overview of classification results: 11-point Likert scale

In order to better understand where the missed predictions are located, we present the confusion matrix related to the 11-point Likert LRGEN-scale in Table 3. The confusion matrix compares the actual values with the predicted values and counts each comparison. Correct predictions are located on the principal diagonal marked in bold font (left-upper corner to right-bottom corner). We can see that almost all values are on or very close to the principal diagonal. Thus, the wrong predictions mostly concern 'near misses' and are actually located closely to the true values.

		Predicted values										
		0	1	2	3	4	5	6	7	8	9	10
Actual values	0	**5**	0	0	0	0	0	0	0	0	0	0
	1	3	**9**	4	2	1	0	0	0	0	0	0
	2	0	4	**17**	7	1	0	0	0	1	0	0
	3	0	1	9	**5**	3	1	3	0	0	0	0
	4	1	0	3	8	**15**	4	1	3	0	0	0
	5	0	0	0	5	11	**9**	9	5	1	0	0
	6	0	0	0	0	6	10	**11**	7	3	1	0
	7	0	0	0	0	1	2	7	**15**	12	0	0
	8	0	0	0	1	0	1	4	8	**11**	6	1
	9	0	0	0	0	0	0	0	1	7	**3**	8
	10	0	1	0	0	0	0	0	0	0	2	**9**

Table 3: Confusion Matrix: LRGEN ideological scale / 11-point Likert scale

In order to better understand this effect quantitatively, we propose another metric that we call 'relaxed accuracy' (RAC). While standard accuracy (ACC, see Equation 3) measures the share of correct predictions, RAC measures the share of predictions that are at maximum 1 below or above the target. Hence, a prediction of '3' for an actual value of '2' counts towards RAC (we call this a 'near miss'). Since our machine learning models aim to predict evaluations by single experts which are ultimately subjective opinions that also may be prone to 'error'—even if just in the sense that there may exist more than one plausible interpretation of each ideological scale—we argue that RAC is a valuable additional metric for our purpose, as getting it 'close enough' on an 11-point scale should be sufficient for most research purposes.

Table 4 provides the respective figures. It turns out that the ratio of 'near misses' is more than 40% for all dependent variables, boosting RAC to close to 80% (GAL-TAN) or above 80% (LRGEN and LRECON). Hence, based on the policy stances that an expert ascribes to a party, our models can predict in about 80% of the cases how she would place a party on any of the three ideology scales with an accuracy of +/-1.

	LRGEN	LRECON	GAL-TAN
Accuracy	37.72%	39.24%	38.41%
'Near miss'	44.64%	42.01%	41.52%
Relaxed Accuracy	82.35%	81.25%	79.93%

Table 4: Accuracy, 'near misses' and Relaxed Accuracy for LRGEN, LRECON, and GAL-TAN

3.2.2 Results of model evaluation: reduced 3-point scale

As a complementary analysis, we also condense the 11-point scales to a 3-point scale 'l', 'c', and 'r'.

As expected and shown in Table 5, reducing the scale results in an increase of the relevant metrics that describe the models' predictive power. All ACC, BAC, and F1S are above 77% for the best models, even exceeding 80% for the LRGEN-scale. The comparison of predicted values and actual values in the confusion matrix (Table 6) related to this scale demonstrates the accuracy of the prediction model. Only a single left-wing party is wrongly predicted to be a right-wing party, and no right-wing parties are predicted to be left-wing parties.

	LRGEN				LRECON				GALTAN			
	SVC	RFC	LRC	ABC	SVC	RFC	LRC	ABC	SVC	RFC	LRC	ABC
ACC	**0.8028**	**0.8028**	0.7612	0.7924	**0.7778**	0.7674	0.7569	0.7431	0.7855	**0.7958**	0.7889	0.7612
BAC	**0.8153**	0.8142	0.7781	0.8051	**0.7851**	0.7736	0.7650	0.7553	0.7733	**0.7792**	0.7697	0.7417
F1S	**0.8010**	**0.8010**	0.7569	0.7890	**0.7771**	0.7661	0.7558	0.7410	0.7909	**0.7972**	0.7910	0.7653

Table 5: Overview of classification results: reduced 3-point scale

		Pred. val.		
		l	c	r
Act. val.	l	**68**	7	1
	c	17	**80**	16
	r	0	16	**84**

Table 6: Confusion Matrix: LRGEN ideological scale / reduced 3-point scale

Although SVC scores better than RFC in the case of the reduced 3-point scale with regard to the economic left-right and general left-right variables (but not w.r.t. GAL-TAN), we proceed with RFC for the remaining model applications for two reasons: First, the performance gap is rather small and thus seems to be negligible, and second, we want to exclude any uncertainty that could possibly result from using a different method in the subsequent experiments.

3.3 Predictions and experiments using EES 2019

After having established that the random forest classification algorithm performs best with regard to its predictive power in terms of the 11-point scales and best (or second-best) in terms of the 3-point scales, we can now use the best models to predict how experts would have evaluated the ideological position of survey respondents of the voter study of the European Election Studies (EES) 2019.

3.3.1 Self-assessment with EES 2019

As a start, we predict the whole European data set and compare it to the self-described ideology of the survey respondents. Figure 5 shows that the voter self-placement seems to exhibit a significant and sizeable 'center bias', i.e. respondents are more likely to place themselves at the ideological center than experts are predicted to do so. Already Knutsen (1998) showed that voters had an increasing tendency to place themselves at the center. Our results suggest that this may be (at least partly) unwarranted.

This suspicion is fueled by the fact that there has been a comprehensive literature on the so-called 'central tendency bias' (also known as 'regression effect', see Stevens and Greenbaum 1966), a cognitive bias that induces respondents to choose the mean of the distribution they are presented with (see Crosetto et al., 2020). This bias is particularly well-known for Likert scale items (Douven, 2018), but has long been known to affect many different kinds of responses (see, e.g., Hollingworth, 1910).

Interestingly, however, this cognitive bias should also affect our dependent variables, as they, too, are 11-point Likert scale items. If we look at the distributions of the policy stances shown in Figure 13 in the Appendix A.2.2, we can indeed see that the distribution of any policy variable has a local maximum at '5' (but also at the extreme values '0' and '10'). Interestingly, however, for half of the policy stances, a relative majority of respondents place themselves at one of the extreme ends of the spectrum. We hypothesize that this effect is driven by an information deficit. Respondents who are unsure about where they are located exactly at a given scale will tend to choose the mean. People who are well-informed may hold beliefs which are stronger, therefore emphasizing the extreme ends.

Another feature of our predicted voter distribution is that it seems to be skewed to the right. A country-level analysis presented in the Appendix (Figure 16-18), however, shows that both the center bias and the tendency towards the right-wing do not necessarily hold in every country.

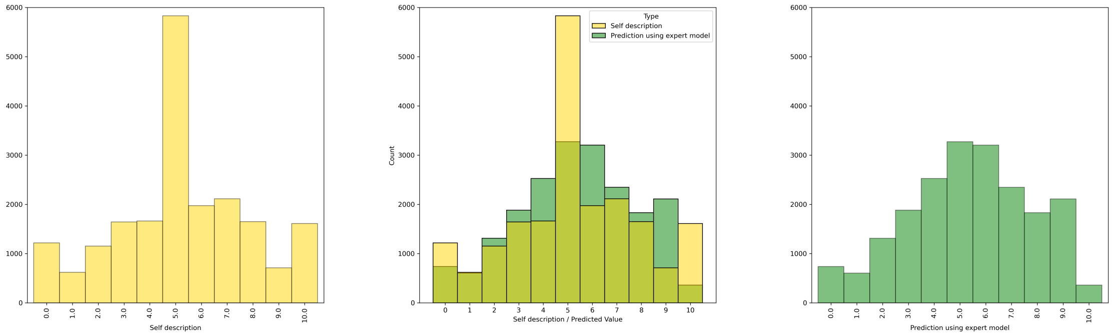


Figure 5: Self-described and predicted values for voters based on the LRGEN model (11-point Likert scale)

In a second step, we predict the economic left-right ideology of the whole population. Since the EES 2019 does not ask voters to evaluate themselves on this scale, we cannot compare our results with self-assessment data. Figure 6 shows that the predicted distribution exhibits three modes: One each at the center, center-left and center-right.

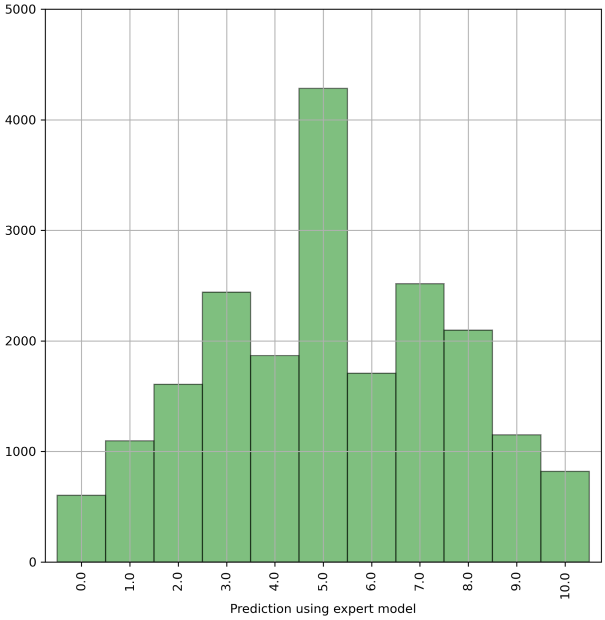


Figure 6: Predicted values using EES based on the LRECON model (11-point Likert scale)

Thirdly, we predict the distribution of the 'GAL-TAN' scale. Figure 7 shows that this prediction is markedly different from the other two ideological scales, as we have four modes and two of them are located at the extreme values of the distribution. Hence, we are confronted with a high level of political polarization along this axis on the European level. This result seems to be highly plausible in the light of discussions surrounding, e.g., immigration policy, but is masked when looking at a general left-right scale.

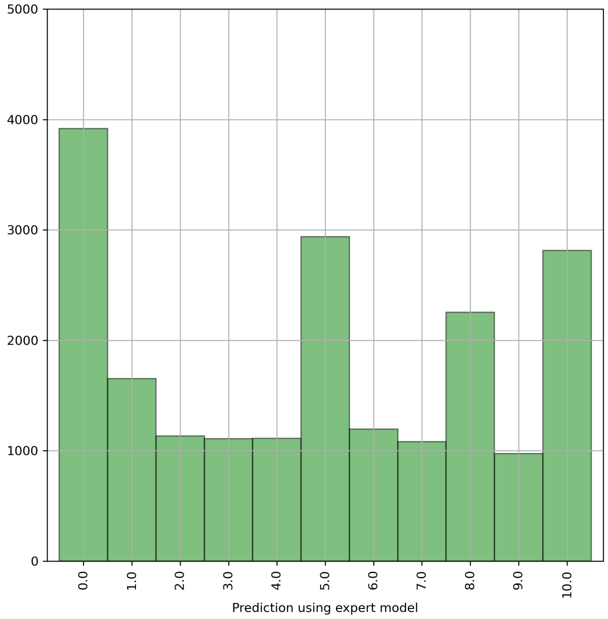


Figure 7: Predicted values using EES based on the GAL-TAN model (11-point Likert scale)

3.3.2 Mapping LRECON and GALTAN on a reduced scale

As a next step, we predict the economic left-right ideology and the GAL-TAN ideology as three-point scales. This is particularly useful if we want to take a look at how individuals are distributed along the two-dimensional ideology space covering 'social' and 'economic' values. Figure 8 and Table 7 show the results of this exercise, which suggests the existence of two distinct modes/poles at the European level, namely one being authoritarian and the other being libertarian. Economically, the authoritarian pole is mostly center or right-wing, whereas the libertarian pole is mostly center or left-wing. A position which is at the center of each ideological scale is only shared by about 13% of the survey respondents. This result again supports the suspicion that it is far more useful to describe voter ideology in two dimensions than in one. Figure 19 in the Appendix shows respective country-level results for all countries covered by the EES 2019.

3.3.3 Cross-country results

In a further analysis, we want to study how voter ideology is distributed along the three dimensions among the respondents of the EES 2019 on a country-level in order to shed light on cross-national heterogeneity and to study how the predictions relate to the self-described ideology in each country. This is not only interesting in its own right, but also important to understand the challenges that the European Union may face in devising policies in the face of heterogeneous preferences.

Figure 9 shows the mean general left-right ideology according to the self description of survey respondents and according to our predictions. Our random forest model predicts the

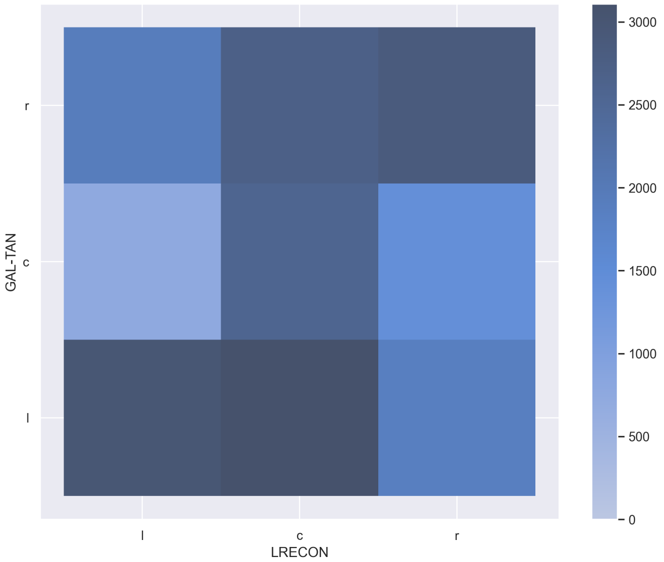


Figure 8: Political compass using EES-based predictions of LRECON and GALTAN

GALTAN	LRECON	Total	Share
l	c	3109	15.40%
l	l	2982	14.77%
r	r	2848	14.11%
r	c	2698	13.37%
c	c	2586	12.81%
r	l	1930	9.56%
l	r	1862	9.22%
c	r	1422	7.04%
c	l	749	3.71%

Table 7: Political compass: total / share of EES-based predicted values of LRECON and GALTAN

Maltese respondents to be on average the most left-wing and survey respondents from the Baltics to be the most right-wing in Europe. While this trend exists in the self-described ideologies, it is much more pronounced in our predictions. We can further see that the survey respondents in many central European countries are on average predicted to be more right-wing than they describe themselves (e.g. Germany, Austria, Denmark etc.). On the other hand, respondents from Romania and Bulgaria consider themselves to be much more right-wing than experts are predicted to place them along this ideology.

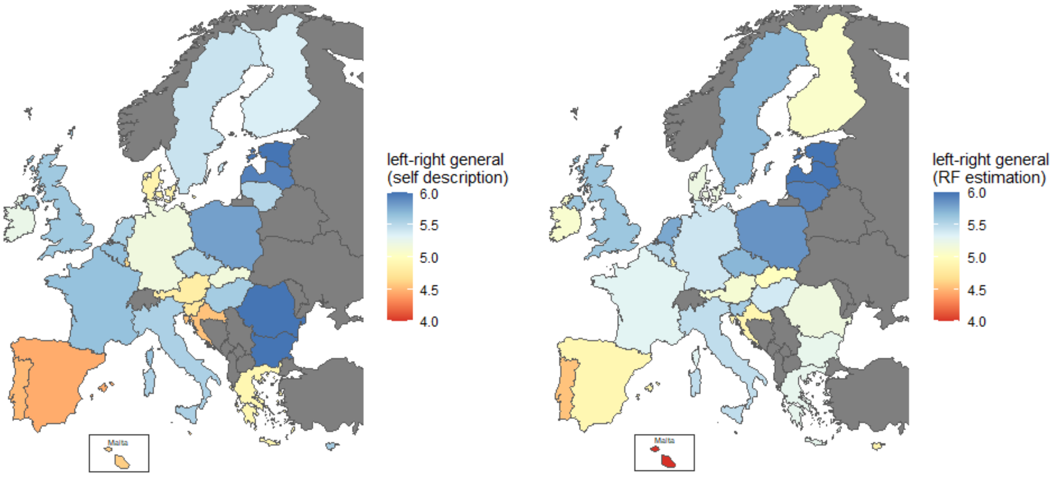


Figure 9: Mean value of general left-right ideology according to self-description (left) and random forest predictions (right) across Europe. The results are bounded to be between 4 (the average Maltese respondent was predicted to be more left-wing) and 6 (respondents from the Baltics were predicted to be more right-wing on average)

Figure 10 shows predicted mean values of the economic left-right ideology and the 'GAL-TAN' ideology of survey respondents of each country in the European Election Studies 2019. While there are some countries where survey respondents were on average both economically left-wing and socially libertarian (in particular in south Western Europe) and some where they are both economically right-wing and socially authoritarian (in particular in the Baltics), we can see that the mean respondent is ideologically more diverse in many other countries. A particular interesting case in this regard is given by Cyprus, where respondents are economically very left-wing, but socially very conservative/authoritarian.[4]

These results again emphasize that a multi-dimensional approach to ideology is crucial (at least as a complementary analysis), as significant ideological heterogeneity is masked in a unidimensional approach. Furthermore, our results suggest significant ideological polarization across European countries, as the mean respondent in many Eastern European countries (including countries that never belonged to the Eastern bloc, such as Cyprus and Greece) are socially more authoritarian/conservative than the rest of Europe, whereas respondents from Southern Europe, France, Luxemburg, Ireland and Finland are on average economically more left-wing than respondents from other countries.

Finally, we want to look at the ideological fragmentation/polarization with regard to each ideological scale by depicting the standard deviation of our predictions, as well as the self-assessed left-right ideology. Again, this analysis (shown in Figure 11) is not only interesting for its own sake, but also important because high degrees of ideological heterogeneity within

[4] The country-level analysis presented in the Appendix in Figure 19 shows that only few respondents from Western Europe are predicted to be left-authoritarian. This relates to the finding by Hillen and Steiner (2020), who analyze that Western European parties The exact causality is unclear, as the literature suggests that parties' policy stances are influenced by voters, but voters are also influenced by parties (see Mellacher, 2020)

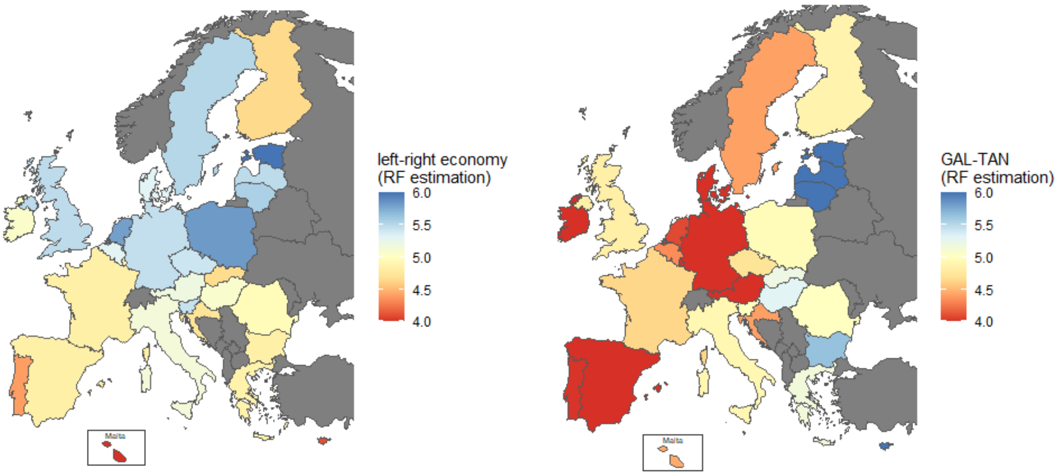


Figure 10: Mean values of the economic left-right ideology (left) and the GAL-TAN ideology (right) according to our random forest predictions across Europe. The results are bounded to be between 4 and 6.

countries likely fuels social conflict (see, e.g., Esteban and Schneider, 2008).

Our analysis produces some interesting results. First, the predicted level of polarization in every ideological measure is generally much higher than the polarization based on self-described left-right ideology. Second, the ranking between countries may vary drastically between polarization based on self-described ideology and polarization based on predicted ideology. For instance, Maltese respondents exhibit the highest polarization in self-described ideology, but relatively low levels of polarization with regard to our predictions. On the other hand, our predictions suggest, e.g., that the Baltic countries are highly polarized, although their polarization with regard to self-described left-right ideology is rather low.

3.4 Application to a spatial voting model

In our final analysis, we test whether our predictions can improve a simple spatial voting model that aims to predict the vote choice of individuals based on their (perceived and/or predicted) ideological distance to the political parties.

To this end, we use the stacked data matrix of the European Election Studies 2019 provided by Carteny et al. (2022) which matches every respondent of the voter study to every national party and provides various metrics to describe the relationship between them. We then link this data set with our predictions, as well as the Chapel Hill Expert Survey 2019 using and amending a 'translation table' provided by the ParlGov database (Döring et al., 2022) which links, among others, some of the party identification codes used the EES with those party codes used by the CHES. Due to the emergence (or rebranding) of parties, we had to manually add 62 party codes. In total, our new 'translation table' allows to link 161 political parties covered by both the EES 2019 and the CHES 2019.

In our analysis, we use five variables given by the stacked data matrix of the European Election Studies 2019:

i) The propensity to ever vote for a specific party as given by a scale from 0 (respondent has a very low propensity to vote for the stack party) to 1 (respondent has a very high propensity to vote for the stack party).

ii) The vote choice of individuals at the last national elections as a binary variable which is 1 if the respondent voted for the respective party at this election and 0 otherwise.

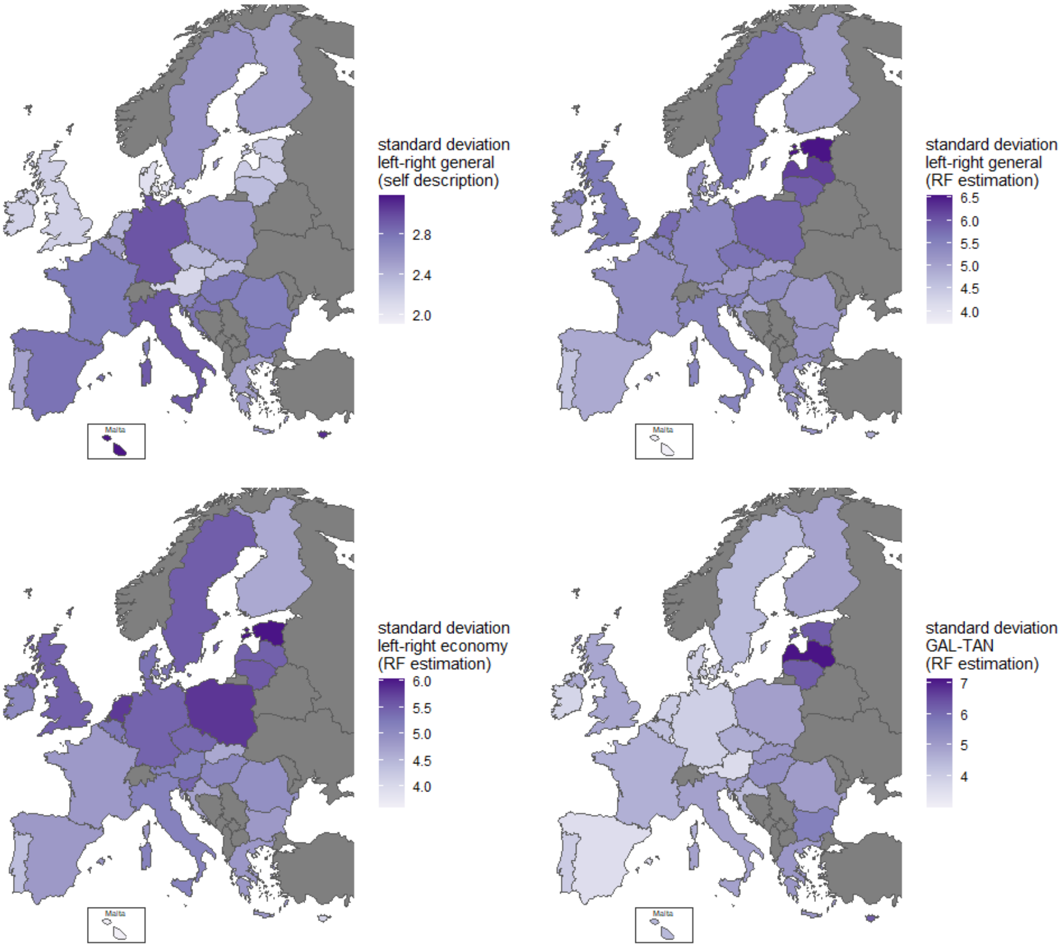


Figure 11: Standard deviation of different measures of ideology. Please note that the color scale is not uniform!

iii) The vote choice of individuals at the elections to the European parliament as a binary variable which is 1 if the respondent voted for the respective party at this election and 0 otherwise.

iv) Self-described voter ideology as a left-right scale between 0 and 10 (this variable is identical to the one used in the analysis in the previous subsections).

v) The perceived ideological distance between the voter and the party: This is originally a value from 0 to 1 that is one tenth of the absolute distance between the self-described voter ideology and the party ideology as described by the voter. We multiply the variable by 10 in order to make it quantitatively commensurable with our predictions.

In our spatial voting models, the variables i), ii) and iii) are our dependent variables. We use variables iv) and v), as well as data from the Chapel Hill Expert Survey 2019 (Bakker et al., 2020a) and our predictions to construct four metrics which will act as independent variables in the models.

1.) The absolute difference between the self-described ideology and the general left-right party ideology as perceived by the mean expert of the Chapel Hill Expert Survey (the latter of which is a continuous variable between 0 and 10). This metric is useful to assess whether linking ideological data from the Chapel Hill Expert Survey and the European Election Studies can improve the spatial voting model on their own.

2.) The distance on the general left-right scale between the predicted voter ideology according to our Random Forest classification model and the party ideology as perceived by the mean expert of the Chapel Hill Expert Survey (i.e. their absolute difference). This variable can be interpreted as 'objective' ideological distance on the left-right scale as predicted by our model.

3.) The distance on the economic left-right scale between the predicted voter ideology according to our Random Forest classification model and the party ideology as perceived by the mean expert of the Chapel Hill Expert Survey. This variable can be interpreted as 'objective' ideological distance on the economic left-right scale as predicted by our model.

4.) The distance on the GAL-TAN (Green, Alternative, Libertarian - Traditional, Authoritarian, Nationalist) scale between the predicted voter ideology according to our Random Forest classification model and the party ideology as perceived by the mean expert of the Chapel Hill Expert Survey. This variable can be interpreted as 'objective' ideological distance on the GAL-TAN scale as predicted by our model.

We use fixed effects models to test the predictive power of our model results (variables 2-4) compared to simply linking the EES and CHES data sets (variable 1). We conduct a linear regression for the propensity to ever vote for a specific party (variable i), and probit regressions for the vote choice at the last national elections (variable ii) and European parliament elections 2019 (variable iii) using the fixest package (Bergé, 2018) for R (R Core Team, 2022). The models are following a specification described in Equation 5:

$$y_{i,j} = x_{i,j}^T \beta + \alpha_j + \epsilon_{i,j} \tag{5}$$

In each model, $y_{i,j}$ is the dependent variable, i.e. the propensity of voter i to vote for the party j, or a binary variable describing whether i voted for j at the last national/European elections, $x_{i,j}^T$ is a vector of variables measuring the ideological distance between the voter and the party and β a vector of the respective coefficients. α_j are party fixed effects that account for factors that increase the voting probability across the ideological spectrum, such as charisma of a candidate, and $\epsilon_{i,j}$ is an error term.

Table 8 shows the results of the OLS estimates of the propensity to ever vote for a specific party, Tables 9 and 10 show the results of probit regressions of the probability to have voted for a specific party at the last national/European elections respectively. Our estimates across the three dependent variables are highly similar in terms of the statistical significance of the coefficients. This is particularly true if we only consider a single ideological dimension, namely the general left-right scale.

The most important predictor is the subjective ideological proximity as measured by the distance between the left-right ideologies of the voter and the party as measured by the respective voter. Naturally, this measure avoids some of the problems of ideological self-assessment: If, for instance, a right-wing voter believes that her views are left-wing, she will likely also place right-wing parties to the left. A similar argument could be made about voters who, for some reason, perceive a skewed distribution of the left-right scale. This coefficient is statistically significant and negative in any of our models. what implies that voters are less likely to vote for a party if they perceive it to be ideologically distant.

After accounting for the subjective ideological proximity as described above, we do not find a significant effect of adding the distance between the self-described voter ideology and the party ideology as perceived by the experts of the CHES 2019, i.e. some measure of objective party ideology. Hence, simply linking the ideological measures of the two data sets does not improve the outcome of these voting models.

Our next coefficient is the predicted ideological distance on the general left-right scale measured as the difference between our random forest classification predictions and the party ideology as perceived by the mean expert in the CHES 2019, which is negative and significant for all uni-dimensional models. This implies that voters are less likely to vote for parties that are predicted to be ideologically more distant—even after accounting for the subjective ideological distance. Hence, our predictions are indeed able to improve a simple spatial voting model.

We then turn to the multi-dimensional analysis, as we add the distance between the predicted voter ideology and the party ideology according to the mean expert in the CHES 2019 along the economic left-right axis. This coefficient is negative and significant at the 5% level for the propensity to ever vote for a specific party, as well as for the probability to have ever voted for a party, but not significant for the probability to have voted for a specific party at the European elections of 2019. Furthermore, this coefficient is only significant if we do not also include the general left-right ideology in the model. This indicates that respondents from the EES 2019 placed less emphasis on economic policy with regard to the vote choice.

Our final coefficient is the difference between the predicted voter ideology and the mean party ideology according to the CHES 2019 along the 'GAL-TAN' axis. This coefficient is negative, significant and sizeable in any regression estimate—even in those models that also include the general left-right ideology as an additional independent variable. This result points to the fact that the 'GAL-TAN' axis is highly salient with regard to the vote choice of individuals, and again supports our conclusion that a two-dimensional approach to political ideology is highly useful (and inferable from our approach!).

In Appendix A.6, we conduct robustness checks where we estimate the same models, but include respondent fixed effects. Doing so may help to account for factors that induce individuals to be more (or less) likely to vote for *any* party, such as trust in the political process (or lack thereof), but is expensive in terms of the degrees of freedom, as there are only a few observations (one for each party) per respondent. However, the results are highly similar in terms of significance of the coefficients, except for the fact that the economic left-right dimension seems to be more important in these models.

Jointly, our results indicate that i) ideology indeed plays a role in the vote choice of individuals, even if voters are not fully aware of it, and ii) that a multi-dimensional approach to ideology is better suited than a uni-dimensional one to describe the behavior of voters. Hence,

Table 8: OLS estimates for the propensity to ever vote for a specific party

	(1)	(2)	(3)	(4)	(5)
Distance between self-placement and subjective party placement	-0.049***	-0.046***	-0.049***	-0.046***	-0.045***
	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)
Distance between self-placement and party placement by mean expert	0.0008		0.005		
	(0.003)		(0.003)		
Distance between predicted ideology and party placement by mean expert (left-right general)		-0.011***	-0.012***		-0.007***
		(0.002)	(0.002)		(0.002)
Distance between predicted ideology and party placement by mean expert (left right economy)				-0.004*	-0.0008
				(0.002)	(0.002)
Distance between predicted ideology and party placement by mean expert (GAL-TAN)				-0.010***	-0.009***
				(0.001)	(0.001)
Num.Obs.	120154	120154	120154	120154	120154
R2	0.135	0.139	0.139	0.141	0.142
R2 Adj.	0.134	0.138	0.138	0.140	0.141
R2 Within	0.070	0.074	0.074	0.077	0.078
R2 Within Adj.	0.070	0.074	0.074	0.077	0.078
AIC	59891.1	59389.2	59331.3	59041.4	58910.1
BIC	61481.3	60979.4	60931.2	60641.3	60519.8
RMSE	0.31	0.31	0.31	0.31	0.31
Std.Errors	by: party	by: party	by: party	by: party	by: party
FE: party	X	X	X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: Probit regression estimates for the probability to have voted for a specific party at the last national elections

	(1)	(2)	(3)	(4)	(5)
Distance between self-placement and subjective party placement	-0.023***	-0.024***	-0.021***	-0.024***	-0.023***
	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)
Distance between self-placement and party placement by mean expert	-0.003		-0.002		
	(0.003)		(0.003)		
Distance between predicted ideology and party placement by mean expert (left-right general)		-0.005***	-0.004***		-0.003*
		(0.001)	(0.001)		(0.002)
Distance between predicted ideology and party placement by mean expert (left right economy)				-0.003*	-0.001
				(0.001)	(0.001)
Distance between predicted ideology and party placement by mean expert (GAL-TAN)				-0.004***	-0.004***
				(0.001)	(0.001)
Num.Obs.	115425	115425	117962	115425	115425
McFadden's R2	0.191	0.192	0.194	0.194	0.194
McFadden's R2 Adj.	0.186	0.187	0.187	0.188	0.189
McFadden's R2 Within	0.048	0.049	0.063	0.051	0.051
McFadden's R2 Within Adj.	0.047	0.049	0.063	0.050	0.051
AIC	50459.9	50369.0	37497.6	50300.0	50275.9
BIC	52043.6	51952.7	39094.5	51893.3	51878.8
RMSE	0.30	0.30	0.28	0.30	0.30
Std.Errors	by: party	by: party	by: party	by: party	by: party
FE: party	X	X	X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: Probit regression estimates for the probability to have voted for a specific party at the elections to the European parliament 2019

	(1)	(2)	(3)	(4)	(5)
Distance between self-placement and subjective party placement	-0.021***	-0.022***	-0.021***	-0.022***	-0.022***
	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)
Distance between self-placement and party placement by mean expert	-0.003		-0.002		
	(0.003)		(0.003)		
Distance between predicted ideology and party placement by mean expert (left-right general)		-0.005***	-0.004***		-0.002
		(0.001)	(0.001)		(0.001)
Distance between predicted ideology and party placement by mean expert (left right economy)				-0.001	-0.0001
				(0.001)	(0.001)
Distance between predicted ideology and party placement by mean expert (GAL-TAN)				-0.006***	-0.005***
				(0.001)	(0.0009)
Num.Obs.	117962	117962	117962	117962	117962
McFadden's R2	0.192	0.194	0.194	0.198	0.198
McFadden's R2 Adj.	0.185	0.187	0.187	0.191	0.191
McFadden's R2 Within	0.062	0.063	0.063	0.068	0.068
McFadden's R2 Within Adj.	0.062	0.063	0.063	0.068	0.068
AIC	37568.9	37506.2	37497.6	37318.3	37304.7
BIC	39156.1	39093.4	39094.5	38915.2	38911.3
RMSE	0.28	0.28	0.28	0.28	0.28
Std.Errors	by: party	by: party	by: party	by: party	by: party
FE: party	X	X	X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

these results further validate and bolster our approach.

4 Limitations

Our analysis presupposes that individuals are aware of their true policy stances. The first limitation is hence given by the fact that we do not know whether the survey respondents truly 'understand' each survey item on average, in particular whether they understand it in the same way as political experts do. A detailed analysis of the data suggests, for instance, that a fraction of German speakers in the European Election Studies seem to misunderstand the question about their opinion on immigration due to low sophistication. This suspicion is fueled by the fact that i) the question wording is 'difficult', ii) the items are ordered in the opposite way than other survey items (i.e. the left-most position is coded as 10 instead of 0), iii) a high share of voters for German-speaking right-wing populist parties (AfD in Germany and FPÖ in Austria) stated that they were fully opposed to a restrictive immigration policy—this is in stark contrast to the voters of other right-wing populist parties in Europe.

The second limitation given by the data set is that we do not know how salient each particular topic is. Consider a worker who is economically left-wing and holds an anti-immigration view. Whether she considers immigration to be the most or least important factor surely makes an enormous difference in her vote choice and perhaps ultimately also her worldview. Enhancing future voter studies by including items that are able to capture this kind of information could hence further improve the predictive power of our approach.

Thirdly, as usual in quantitative empirical research, more data would increase the validity of our approach. Typically, one way to increase the size of the data sets is to change the handling of missing values in data: instead of deleting data rows with missing data, imputation algorithms could predict missing values by identifying similarities across the observations. This approach was not successful in this study. Thus, it would be very useful to be able to rely on a larger sample size in the first place. This is particularly true for the cross-national analysis, as the sample size of the EES, which is 1000 participants per country (except for Cyprus, Luxembourg and Malta where it is 500) may not be statistically representative for every country.

Finally, we could extend our search for the optimal parameter settings of the models by transferring the computations to high-performance computers, hence possibly (marginally) improving the predictive power of our models. Nevertheless, we expect that increased data availability have a much greater impact on model performance than excessive model tuning.

5 Conclusion

Since the French revolution, people use 'left' and 'right' to describe political ideologies. Even though the left-right concept is inherently relative (see, e.g., Bobbio, 1996, pp. 56-57), survey studies usually assume that respondents who are political non-experts are able to envision the same left-right scale and are able to correctly place themselves on it.

In the present work, we confronted this assumption in a novel way. Building on Downs' 1957 assertion that political ideologies represent bundles of policy stances and validating it using data-driven approaches, we study how political experts would place voters ideologically using the voters' policy stances. Our approach can help to i) 'objectively' place voters along various ideological axes, ii) understand how many dimensions of ideology are appropriate to describe a voters' policy preferences, and iii) to understand if political ideology has any effect on the behavior of people, e.g. during elections.

In our study, we use machine learning algorithms that are well-known to exhibit high predictive power on two well-established publicly available survey data sets: The Chapel Hill Expert Survey (CHES) 2019 (Bakker et al., 2020a), which covers policy stances and various measures of political ideology of parties as perceived by political experts of their country, and the voter study of the European Election Studies (EES) 2019 (Schmitt et al., 2022), which covers—among others—policy views, vote choice and self-described left-right ideology of voters. The two data sets share six policy-related survey items and are hence ideally suited for our approach.

As a first step, we explored in a data-driven way whether experts perceive a clear relationship between latent ideological scales and policy stances by using exploratory factor analysis and principal component analysis (PCA). Exploratory factor analysis suggested that the six policy stances that link the two data sets—but also the complete set of 15 policy items included in the CHES 2019—are well-described by two ideological scales which correspond to the well-known axes that are already included in the CHES 2019: economic left-right and 'GAL-TAN', capturing social, cultural and ecological values. In addition to that, principal component analysis suggested that the parties are clustered according to their general left-right ideology (as perceived by the experts) along the two principal components (as created by PCA based on the the policy items). These results suggest that political experts indeed view a strong link between ideologies and policies, complementing other research on 'policy interconnections' (Wagner, 2009).

As a second step, we tested which machine learning algorithm is best able to predict how experts evaluate a party's ideology given their policy stances. In order to do so, we first split the CHES 2019 data set in training, test and validation data sets. For each algorithm, we then conducted a parameter search where we optimized the parameter configuration of the algorithms based on how well models (which are estimated using the training data set) are able to predict the test data set. Finally, we compared the different optimized machine learning algorithms based on how well they are able to predict the validation data set.

In all of the three ideological dimensions that we consider (general left-right, economic left-right, and 'GAL-TAN', representing social/cultural and ecological values), random forest models exhibit the highest predictive power. Our models predict exactly how a particular expert will place a party on each 11-point scale in about 40% of the cases included in the (completely separate) test data set. In addition to that, more than 40% predictions miss the empirical value only by a single point.

Emboldened by the strong performance of our models, we use them to predict how experts would have evaluated the ideologies of respondents in the voter study of the EES 2019. Aggregated European data suggests i) the existence of a sizeable 'center bias', i.e. survey respondents are much more likely to place themselves at the ideological center than experts are predicted to do so, and ii) that Europeans are much more polarized in the GAL-TAN dimension than in the economic or in the general left-right dimension.

We also showed that country-specific predictions vary remarkably across Europe both with regard to mean ideologies, but also with regard to the level of political polarization based on the standard deviation of the predictions. Since our approach allows us to predict voter ideology along the same scale(s), it is well-suited to be used in trans-national studies, in particular compared to the self-described ideology which may have a different 'meaning' depending on the specific country. Indeed, our results show that 'subjective', i.e. self-described, left-right ideology may vary remarkably from 'objective', i.e. predicted, left-right ideology.

Finally, we used our predictions to test whether they are able to improve a simple spatial voting model based on probit regressions. We find that the predicted ideological distance between a voter and a party has a significant effect on the vote choice even after accounting for the subjective ideological distance in the left-right dimension, and that a two-dimensional approach fares better than a uni-dimensional approach. In particular, the 'GAL-TAN' axis seems to be highly important for the vote choice of individuals. Hence, this axis seems not only to be more polarized, but also more salient than the economic or the general left-right axis. These findings again validate our approach and show that it can produce interesting insights.

More generally, we showed that machine learning models are able to make meaningful predictions based on—even a limited set—of survey data of political experts and voters. We are confident that these methods can be fruitfully applied in other fields covered by political science as, e.g., modelling the vote choice of individuals.

Our approach can be used to study i) spatial political competition in the tradition of Downs (1957) in more detail, and ii) how behavior is shaped by political ideology more generally. In order to foster the latter use of our approach, we encourage researchers who conduct surveys on political behavior to include the six items used in this study in their surveys.

Acknowledgements

PM has received funding from the Austrian Science Fund (FWF) [P 35228]. We thank the participants of the Austrian Political Science Day 2022 (in particular Dominik Duell) and the participants of the research seminar of the School of Business, Economics and Social Sciences of the University of Graz for their valuable suggestions and comments. Any errors are ours.

References

R. Bakker, L. Hooghe, S. Jolly, G. Marks, J. Polk, J. Rovny, and M. Steenbergen. 2019 Chapel Hill Expert Survey. Version 2019.3. Available on chesdata.eu. Chapel Hill, NC: University of North Carolina, Chapel Hill, 2020a.

R. Bakker, S. Jolly, and J. Polk. Multidimensional incongruence, political disaffection, and support for anti-establishment parties. *Journal of European Public Policy*, 27(2):292–309, 2020b. doi: 10.1080/13501763.2019.1701534. URL https://doi.org/10.1080/13501763.2019.1701534.

P. C. Bauer, P. Barberá, K. Ackermann, and A. Venetz. Is the Left-Right Scale a Valid Measure of Ideology? *Political Behavior*, 39(3):553–583, Sept. 2017. ISSN 1573-6687. doi: 10.1007/s11109-016-9368-2. URL https://doi.org/10.1007/s11109-016-9368-2.

L. Bergé. Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers*, (13), 2018.

N. Bobbio. *Left and right: The significance of a political distinction.* University of Chicago Press, 1996.

M. Bolanowski, A. Paszkiewicz, and B. Rumak. Coarse Traffic Classification for High-Bandwidth Connections in a Computer Network Using Deep Learning Techniques. In L. Barolli, K. Yim, and T. Enokido, editors, *Complex, Intelligent and Software Intensive Systems*, volume 278 of *Lecture Notes in Networks and Systems*, pages 131–141. Springer International Publishing, 2021. ISBN 978-3-030-79724-9 978-3-030-79725-6. doi: 10.1007/978-3-030-79725-6_13. URL https://link.springer.com/10.1007/978-3-030-79725-6_13.

P. F. Brennan and S. Bakken. Nursing Needs Big Data and Big Data Needs Nursing: Nursing Needs Big Data. *Journal of Nursing Scholarship*, 47(5):477–484, 2015. ISSN 15276546. doi: 10.1111/jnu.12159. URL https://onlinelibrary.wiley.com/doi/10.1111/jnu.12159.

E. Bublitz. Misperceptions of income distributions: cross-country evidence from a randomized survey experiment. *Socio-Economic Review*, 20(2):435–462, 04 2022. ISSN 1475-1461. doi: 10.1093/ser/mwaa025. URL https://doi.org/10.1093/ser/mwaa025.

G. Carteny, H. Schmitt, W. Häußling, J. Leiser, and M. Körnig. 2019 european election studies (ees) stacked data matrix. GESIS, Cologne. ZA7890 Data file Version 1.0.0, https://doi.org/10.4232/1.13967, 2022.

R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1 (2):245–276, 1966. doi: 10.1207/s15327906mbr0102_10. URL https://doi.org/10.1207/s15327906mbr0102_10. PMID: 26828106.

D. Chicco. Ten quick tips for machine learning in computational biology. *BioData Mining*, 10 (1):35, 2017. ISSN 1756-0381. doi: 10.1186/s13040-017-0155-3. URL https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0155-3.

P. Crosetto, A. Filippin, P. Katuščák, and J. Smith. Central tendency bias in belief elicitation. *Journal of Economic Psychology*, 78:102273, 2020. ISSN 0167-4870. doi: https://doi.org/10.1016/j.joep.2020.102273. URL https://www.sciencedirect.com/science/article/pii/S0167487020300301.

G. Cruces, R. Perez-Truglia, and M. Tetaz. Biased perceptions of income distribution and preferences for redistribution: Evidence from a survey experiment. *Journal of Public Economics*, 98:100–112, 2013. ISSN 0047-2727. doi: https://doi.org/10.1016/j.jpubeco.2012.10.009. URL https://www.sciencedirect.com/science/article/pii/S004727271200117X.

T. W. David, H. Anizelli, T. J. Jacobsson, C. Gray, W. Teahan, and J. Kettle. Enhancing the stability of organic photovoltaics through machine learning. *Nano Energy*, 78:105342, 2020. ISSN 22112855. doi: 10.1016/j.nanoen.2020.105342. URL https://linkinghub.elsevier.com/retrieve/pii/S2211285520309198.

C. E. De Vries, A. Hakhverdian, and B. Lancee. The dynamics of voters' left/right identification: The role of economic and cultural attitudes. *Political Science Research and Methods*, 1(2): 223–238, 2013.

H. Döring, C. Huber, and P. Manow. Parliaments and governments database (parlgov): Information on parties, elections and cabinets in established democracies. *Development version*, 2022.

I. Douven. A bayesian perspective on likert scales and central tendency. *Psychonomic bulletin & review*, 25(3):1203–1211, 2018.

A. Downs. *An economic theory of democracy.* Harper & Row, 1957.

L. Esser. OSEMN is AWESOME, 2022. URL https://medium.com/analytics-vidhya/osemn-is-awesome-3c9e42c3067d.

J. Esteban and G. Schneider. Polarization and conflict: Theoretical and empirical issues. *Journal of Peace Research*, 45(2):131–141, 2008. doi: 10.1177/0022343307087168. URL https://doi.org/10.1177/0022343307087168.

J. Fernández-Albertos and A. Kuo. Income perception, information, and progressive taxation: Evidence from a survey experiment. *Political Science Research and Methods*, 6(1):83–110, 2018. doi: 10.1017/psrm.2015.73.

M. Funke, M. Schularick, and C. Trebesch. Going to extremes: Politics after financial crises, 1870–2014. *European Economic Review*, 88:227–260, 2016. ISSN 0014-2921. doi: https://doi.org/10.1016/j.euroecorev.2016.03.006. URL https://www.sciencedirect.com/science/article/pii/S0014292116300587. SI: The Post-Crisis Slump.

A. Gelman, S. Goel, D. Rivers, and D. Rothschild. The Mythical Swing Voter. *Quarterly Journal of Political Science*, 11(1):103–130, 2016. ISSN 1554-0626. doi: 10.1561/100.00015031. URL http://dx.doi.org/10.1561/100.00015031.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(null):723–773, mar 2012. ISSN 1532-4435.

S. Hillen and N. D. Steiner. The consequences of supply gaps in two-dimensional policy spaces for voter turnout and political support: The case of economically left-wing and culturally right-wing citizens in western europe. *European Journal of Political Research*, 59(2):331–353, 2020. doi: https://doi.org/10.1111/1475-6765.12348. URL https://ejpr.onlinelibrary.wiley.com/doi/abs/10.1111/1475-6765.12348.

H. L. Hollingworth. The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 7(17):461–469, 1910.

L. Hooghe, G. Marks, and C. J. Wilson. Does left/right structure party positions on european integration? *Comparative Political Studies*, 35(8):965–989, 2002. doi: 10.1177/001041402236310. URL https://doi.org/10.1177/001041402236310.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, 2021. ISBN 978-1-07-161417-4 978-1-07-161418-1. doi: 10.1007/978-1-0716-1418-1. URL https://link.springer.com/10.1007/978-1-0716-1418-1.

S. A. Jessee. Partisan bias, political information and spatial voting in the 2008 presidential election. *The Journal of Politics*, 72(2):327–340, 2010. doi: 10.1017/S0022381609990764. URL https://doi.org/10.1017/S0022381609990764.

H. F. Kaiser. The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151, 1960.

H. Kim and R. C. Fording. Voter Ideology, the Economy, and the International Environment in Western Democracies, 1952–1989. *Political Behavior*, 23(1):53–73, Mar. 2001. ISSN 1573-6687. doi: 10.1023/A:1017669614814. URL https://doi.org/10.1023/A:1017669614814.

O. Knutsen. EUROPEANS MOVE TOWARDS THE CENTER: A COMPARATIVE LONGITUDINAL STUDY OF LEFT–RIGHT SELF-PLACEMENT IN WESTERN EUROPE. *International Journal of Public Opinion Research*, 10(4):292–316, Dec. 1998. ISSN 0954-2892. doi: 10.1093/ijpor/10.4.292. URL https://doi.org/10.1093/ijpor/10.4.292.

M. Kumar. Project Management in Data Science using OSEMN, 2022. URL https://medium.com/international-school-of-ai-data-science/project-management-in-data-science-using-osemn-50e46f95eec7.

M. D. Laméris, R. Jong-A-Pin, and H. Garretsen. On the measurement of voter ideology. *European Journal of Political Economy*, 55:417–432, 2018. ISSN 0176-2680. doi: https://doi.org/10.1016/j.ejpoleco.2018.03.003. URL https://www.sciencedirect.com/science/article/pii/S017626801730277X.

H. Mason and C. Wiggins. A taxonomy of data science, 2010. URL https://introdatasci.dlilab.com/pdf/A_Taxonomy_of_Data_Science.pdf.

M. Mayo. Frameworks for Approaching the Machine Learning Process, 2022. URL https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html.

P. Mellacher. The impact of corona populism: Empirical evidence from austria and theory, 2020. URL https://arxiv.org/abs/2012.14962.

S. Merrill. A unified theory of voting : directional and proximity spatial models, 1999.

T. M. Meyer and M. Wagner. Perceptions of parties' left-right positions: The impact of salience strategies. *Party Politics*, 26(5):664–674, Sept. 2020. ISSN 1354-0688. doi: 10.1177/1354068818806679. URL https://doi.org/10.1177/1354068818806679. Publisher: SAGE Publications Ltd.

T. R. Palfrey and K. T. Poole. The Relationship between Information, Ideology, and Voting Behavior. *American Journal of Political Science*, 31(3):511–530, 1987. ISSN 00925853, 15405907. doi: 10.2307/2111281. URL http://www.jstor.org/stable/2111281. Publisher: [Midwest Political Science Association, Wiley].

T. pandas development team. Pandas-dev/pandas: Pandas, 2020. URL https://doi.org/10.5281/zenodo.3509134.

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

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

M. M. Rahman and Z. Govindarajulu. A modification of the test of shapiro and wilk for normality. *Journal of Applied Statistics*, 24(2):219–236, 1997.

W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University, Evanston, Illinois, 2022. URL https://CRAN.R-project.org/package=psych. R package version 2.2.9.

W. Revelle and T. Rocklin. Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, 14(4):403–414, 1979.

J. Rosset and C. Stecker. How well are citizens represented by their governments? Issue congruence and inequality in Europe. *European Political Science Review*, 11(2):145–160, 2019. ISSN 1755-7739. doi: 10.1017/S1755773919000043. URL https://www.cambridge.org/core/article/how-well-are-citizens-represented-by-their-governments-issue-congruence-and-inequality-in-europe/4FCE7A0C8503A85A496B17428C096E78. Edition: 2019/06/06 Publisher: Cambridge University Press.

J. S. Saltz and I. Krasteva. Current approaches for executing big data science projects—a systematic literature review. *PeerJ Computer Science*, 8:e862, 2022. ISSN 2376-5992. doi: 10.7717/peerj-cs.862. URL https://peerj.com/articles/cs-862.

H. Schmitt, S. B. Hobolt, W. v. d. Brug, and S. A. Popa. European Parliament Election Study 2019, Voter Study. GESIS, Köln. ZA7581 Datenfile Version 2.0.1, https://doi.org/10.4232/1.13846, 2022.

C. Schröer, F. Kruse, and J. M. Gómez. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181:526–534, 2021. ISSN 18770509. doi: 10.1016/j.procs.2021.01.199. URL https://linkinghub.elsevier.com/retrieve/pii/S1877050921002416.

B. Shamir. Ideological position, leaders' charisma, and voting preferences: Personal vs. partisan elections. *Political Behavior*, 16(2):265–287, June 1994. ISSN 1573-6687. URL https://doi.org/10.1007/BF01498880.

S. Stevens and H. B. Greenbaum. Regression effect in psychophysical judgment. *Perception & Psychophysics*, 1(5):439–446, 1966.

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R, 2011. URL https://www.jstatsoft.org/v45/i03/.

W. F. Velicer. Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3):321–327, 1976.

M. Wagner. *Policy interconnections in party competition: Issue linkages in 23 countries*. London School of Economics and Political Science (United Kingdom), 2009.

M. Wagner and S. Kritzinger. Ideological dimensions and vote choice: Age group differences in austria. *Electoral studies*, 31(2):285–296, 2012.

A. Zafiroglu and Y.-N. Chang. Scale, Nuance, and New Expectations in Ethnographic Observation and Sensemaking. *Ethnographic Praxis in Industry Conference Proceedings*, 2018(1):663–690, 2018. ISSN 1559890X. doi: 10.1111/1559-8918.2018.01228. URL https://onlinelibrary.wiley.com/doi/10.1111/1559-8918.2018.01228.

Appendix A Appendix

A.1 Ideology scales and policy items in CHES 2019 and EES 2019

Variable name	CHES 2019	EES 2019
LRGEN	**Position of the party in 2019 in terms of its overall ideological stance.** 0 = Extreme left 5 = Center 10 = Extreme right	**In political matters people talk of ”the left” and ”the right”. What is your position? Please indicate your views using any number on an 11-point-scale. On this scale, where 0 means ”left” and 10 means ”right,” which number best describes your position?** 0 = left 10 = right
LRECON	**Position of the party in 2019 in terms of its ideological stance on economic issues. Parties can be classified in terms of their stance on economic issues such as privatization, taxes, regulation, government spending, and the welfare state. Parties on the economic left want government to play an active role in the economy. Parties on the economic right want a reduced role for government.** 0 = Extreme left 5 = Center 10 = Extreme right	-

GAL-TAN	**Position of the party in 2019 in terms of their views on social and cultural values. “Libertarian” or “postmaterialist” parties favor expanded personal freedoms, for example, abortion rights, divorce, and same-sex marriage. “Traditional” or “authoritarian” parties reject these ideas in favor of order, tradition, and stability, believing that the government should be a firm moral authority on social and cultural issues.** 0 = Libertarian/Postmaterialist 5 = Center 10 = Traditional/Authoritarian	-
IMMIGRATE_POLICY	**Position on immigration policy** 0 = Strongly favors a liberal policy on immigration 10 = Strongly favors a restrictive policy on immigration	**Immigration** 0 = You fully favour a restrictive policy on immigration 10 = You fully oppose a restrictive policy on immigration (we recoded this variable such that 0 reflects a strong preference for a liberal policy in order to correctly use our models)
REDISTRIBUTION	**Position on redistribution of wealth from the rich to the poor.** 0 = Strongly favors redistribution 10 = Strongly opposes redistribution	**Redistribution of wealth** 0 You fully favour redistribution from the rich to the poor in [country] 10 You fully oppose redistribution of wealth from the rich to the poor in [country]
ENVIRONMENT	**Position towards environmental sustainability.** 0 = Strongly supports environmental protection even at the cost of economic growth 10 = Strongly supports economic growth even at the cost of environmental protection	**Environment** 0 Environmental protection should take priority even at the cost of economic growth 10 Economic growth should takenpriority even at the cost of environmental protection

ECON_INTERVEN	**Position on state intervention in the economy.** 0 = Fully in favor of state intervention 10 = Fully opposed to state intervention	**What do you think of state regulation and control of the economy** 0 You fully favour state intervention in the economy 10 You fully oppose state intervention in the economy
CIVLIB_LAWORDER	**Position on civil liberties vs. law and order.** 0 = Strongly favors civil liberties 10 = Strongly favors tough measures to fight crime	**Civil liberties** 0 You fully support privacy rights even if they hinder efforts to combat crime 10 You fully support restricting privacy rights in order to combat crime
SOCIALLIFESTYLE	**Position on social lifestyle (e.g. rights for homosexuals, gender equality).** 0 = Strongly supports liberal policies 10 = Strongly opposes liberal policies	**Same-sex marriage** 0 You fully favour same sex marriage 10 You fully oppose same sex marriages
MULTICULTURALISM	**Position on integration of immigrants andasylum seekers (multiculturalism vs. assimilation).** 0 = Strongly favors multiculturalism 10 = Strongly favors assimilation	-
SPENDVTAX	**Position on improving public services vs. reducing taxes during 2019.** 0 = Strongly favors improving public services 10 = Strongly favors reducing taxes.	-
DEREGULATION	**Position on deregulation of markets.** 0 = Strongly opposes deregulation of markets 10 = Strongly favors deregulation of markets	-
RELIGIOUS_PRINCIPLES	**Position on role of religious principles in politics.** 0 = Strongly opposes religious principles in politics 10 = Strongly supports religious principles in politics	-

ETHNIC_MINORITIES	**Position towards ethnic minority rights.** 0 = Strongly favors more rights for ethnic minorities 10 = Strongly opposes more rights for ethnic minorities	-
NATIONALISM	**Position towards cosmopolitanism vs. nationalism.** 0 = Strongly promotes cosmopolitan conceptions of society 10 = Strongly promotes nationalist conceptions of society	-
URBAN_RURAL	**Position on urban/rural interests.** 0 = Strongly supports urban interests 10 = Strongly supports rural interests.	-
PROTECTIONISM	**Position towards trade liberalization/protectionism.** 0 = Strongly favors trade liberalization 10 = Strongly favors protection of domestic producers.	-
DECENTRALIZATION	**Position on political decentralization to regions/localities.** 0 = Strongly favors political decentralization. 10 = Strongly opposes political decentralization.	-

A.2 Descriptive statistics: overview of data

A.2.1 Descriptive statistics: CHES 2019

The initial data set consists of 3,823 observations and 63 columns. In line with our research objectives, we select nine relevant columns that cover 6 policy stances and 3 dimensions of ideology. As shown in the respective descriptive statistics (Table 12), all columns contain some missing values ('nan'). For five columns, more than 10% of all values are missing. Mean and median are all within the range of four to six.

Fig. 12 shows histograms of the variables extracted from the columns (where we ignore missing values). They reveal mode values of '10' for the variable *immigrate_policy* (indicating a preference for a highly restrictive immigration policy), and on the left for *sociallifestyle* (i.e. favoring equal rights for women, homosexuals etc.). While the distribution of *galtan* values appear as if they were drawn from a uniform distribution, the remaining six variables are distributed such that the mode is located at our near the center.

	econ_interven	environment	redistribution	civlib_laworder	immigrate_policy	sociallifestyle	lrgen	lrecon	galtan
count	3432	3262	3396	3405	3438	3461	3610	3554	3596
nan	391	561	427	418	385	362	213	269	227
mean	4.55	5.19	4.41	5.15	5.68	4.47	5.30	4.95	5.08
std	2.63	2.65	2.53	2.93	2.94	3.27	2.58	2.45	3.11
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	3.00	3.00	3.00	3.00	3.00	2.00	3.00	3.00	2.00
50%	5.00	5.00	4.00	5.00	6.00	4.00	5.00	5.00	5.00
75%	6.00	7.00	6.00	8.00	8.00	7.00	7.00	7.00	8.00
max	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00

Table 12: Descriptive statistics of CHES 2019

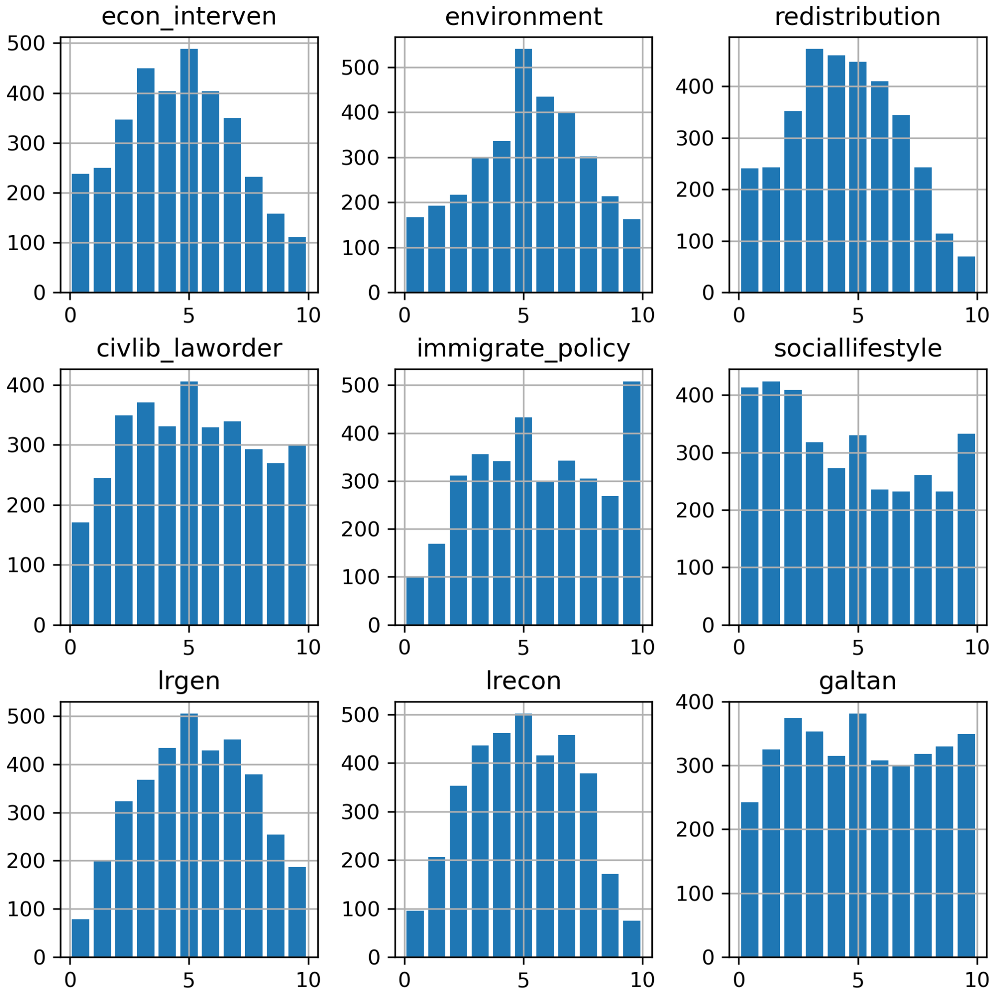


Figure 12: Distribution of variables of the CHES 2019 data set (without missing values)

A.2.2 Descriptive statistics: EES 2019

The EES 2019 data set contains 26,538 observations. Based on our research agenda, we extract 10 variables from the data and check the quality of the data. We describe the numeric data in Table 13 and omit categorical data (IDs of respondents, country of residence, country code).

	econ_interven	environment	redistribution	civlib_laworder	immigrate_policy	sociallifestyle	lrgen_selfdescription
count	23,954	25,686	24,907	25,204	25,124	25,543	22,826
nan	2584	852	1631	1334	1414	995	3712
mean	5.01	3.27	4.36	4.91	5.60	4.56	5.22
std	2.60	2.75	2.98	3.07	3.27	4.00	2.59
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	3.00	1.00	2.00	3.00	3.00	0.00	4.00
50%	5.00	3.00	5.00	5.00	5.00	5.00	5.00
75%	7.00	5.00	6.00	7.00	9.00	9.00	7.00
max	10.00	10.00	10.00	10.00	10.00	10.00	10.00

Table 13: Descriptive statistics of EES 2019

Fig. 13 shows histograms for the seven numeric variables (again omitting missing values) in the EES 2019 data set. Variables *econ_interven* and *lrgen_selfdescription* show a strong tendency towards the center, whereas most other values (except '1' and '9') appear to be similar to a uniform distribution. In contrast, all other variables are more differentiated, in particular with regard to the extreme values ('0' and '10'), even though local maxima at '5' exist for any variable. These results indicate that the electorate is more polarized about five out of the six policy stances than about their left-right self description and their stance on state intervention in the economy.

A.3 Details on handling of missing data of data set CHES 2019

We use the variables LRGEN, LRECON, and GAL-TAN as dependent variables in our models. Accordingly, we can only use observations that include the respective variable to train and validate our models. In order to maximize the number of observations available for each dependent variables, we analyze the variables separately and hence use observations to estimate, e.g. our model for the LRGEN variable, even if they do not include values for LRECON or GAL-TAN.

Following this procedure, we are left with 3,610 observations (out of 3,823 total) for LRGEN. Analogously, 3,554 observations include values for for LRECON and 3,596 observations include data about the GAL-TAN ideological scale.

A.3.1 Data preparation related to the target variables

In order to understand how this procedure affects the information available in the data, we present some statistics in Table 14. There are only minor differences between the data set including missing values and the data sets without missing values for LRGEN, LRECON and GAL-TAN respectively, as only a few observations are removed, and both mean and standard deviation (std) only differ marginally. Thus, we can proceed with our analysis.

In the next step, we have to exclude observations for which our independent variables are missing. The exclusion threshold n determines the number of independent variables that must be included in an observation for it to be used in our analysis. We set n to six, i.e. an observation must contain all independent variables for it to be used. As indicated in Table 15, this procedure reduces the number of observations from 3,823 to 2,912 (without considering the presence of independent variables).

Please note that we also experimented with imputation algorithms such as, e.g. the CART imputation algorithm in the R package mice (van Buuren and Groothuis-Oudshoorn, 2011). This approach allows to impute missing values based on similarities compared to other obser-

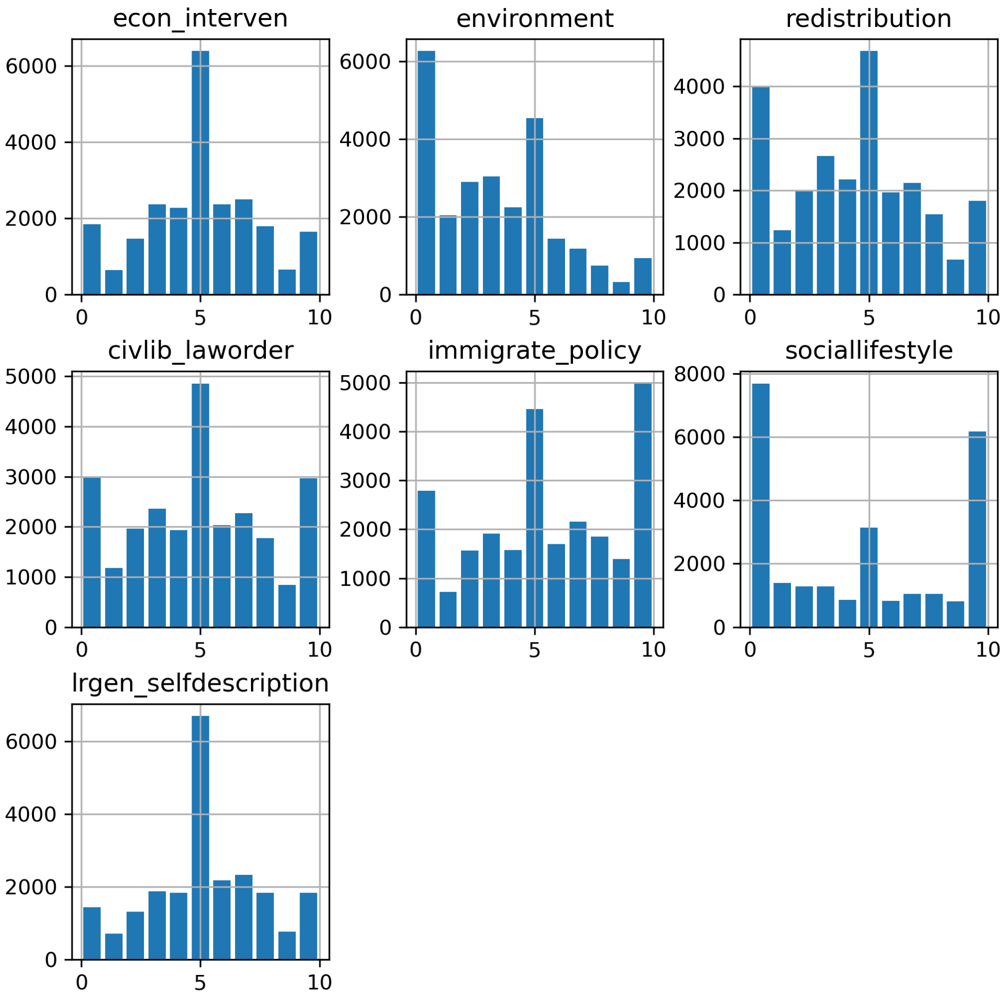


Figure 13: Distribution of variables of the EES 2019 data set (without missing values)

	LRGEN			LRECON			GAL-TAN		
	count	mean	std	count	mean	std	count	mean	std
econ_interven	-47	-0.007	0.000	-57	-0.003	0.006	-55	-0.004	0.004
environment	-41	-0.001	0.004	-55	-0.008	-0.002	-41	-0.001	0.006
redistribution	-47	-0.001	0.002	-56	-0.010	0.002	-61	-0.005	0.003
civlib_laworder	-55	0.000	0.004	-84	-0.016	-0.004	-49	-0.006	0.010
immigrate_policy	-46	-0.014	-0.001	-84	-0.036	-0.006	-47	-0.009	0.007
sociallifestyle	-57	-0.014	-0.002	-104	-0.046	-0.017	-51	-0.010	0.007
lrgen	0	0.000	0.000	-103	-0.025	-0.010	-72	0.007	0.008
lrecon	-47	-0.002	-0.003	0	0.000	0.000	-64	0.003	0.004
galtan	-58	-0.007	-0.001	-106	-0.038	-0.008	0	0.000	0.000

Table 14: Comparison between data sets with / without missing values

Exclusion threshold n	Remaining observations	Difference to total total number of observations
0	3,823	0
1	3,649	174
2	3,599	224
3	3,535	288
4	3,425	398
5	3,274	549
6	2,912	911

Table 15: Remaining observations dependent on exclusion threshold

vations. However, integrating the imputed observations did not yield any improvements with regard to our research objectives. Thus, we decided to only accept complete observations.

A.3.2 Data preparation related to the independent variables

Combining the cleaning of the scales and the exclusion threshold results in the basic data set for each scale. Due to the trade-off between available information and number of removed observations, an exclusion threshold of $n = 6$ is chosen, thus all observations with any missing data are removed from the data set.

Following data sets result from this step:

- Dataset *LRGEN:* 3,610 total entries without missing values for LRGEN
  1. 728 removed observations due to exclusion threshold $n = 6$
  2. Overview of independent variables:
     - econ_interven 2,882
     - environment 2,882
     - redistribution 2,882
     - civlib_laworder 2,882
     - immigrate_policy 2,882
     - sociallifestyle 2,882
- Dataset *LRECON:* 3,554 total entries without missing values for LRECON
  1. 676 removed observations due to exclusion threshold
  2. Overview of independent variables:
     - econ_interven 2,878
     - environment 2,878
     - redistribution 2,878
     - civlib_laworder 2,878
     - immigrate_policy 2,878
     - sociallifestyle 2,878
- Dataset *GAL-TAN:* 3,596 total entries
  1. 707 removed observations due to exclusion threshold
  2. Overview of independent variables:
     - econ_interven 2,889

- environment 2,889
- redistribution 2,889
- civlib_laworder 2,889
- immigrate_policy 2,889
- sociallifestyle 2,889

Although approaches for comparing multivariate distributions exist (see, e.g., Gretton et al., 2012), to the best of the authors' knowledge there is no method allowing to compare multivariate distributions with and without missing values. In order to provide some statistical insights for individual columns of the data set, we first confirm using the Shapiro-Wilk test that none of the variables are distributed according to a normal distribution. Accordingly, we use the Mann–Whitney U test to check the differences for each variable between the data sets with (1) all non-missing values for the individual columns, and (2) the reduced data set that includes all complete observations for LRGEN, LRECON and GAL-TAN respectively (i.e. all six independent and the respective dependent variable). Our tests shows that *sociallifestyle* is considered to be drawn from different populations in the case of LRGEN and GAL-TAN (p-value < 0.05). In addition to that, the distribution of *GAL-TAN* is also significantly different (p-value < 0.05).

Figure 14 plots overlapping histograms of the relative distributions of variables in order to help understanding where these changes come from. In particular, the relative distribution of the *immigrate_policy* and the *sociallifestyle* variables are shifted to the left from the data set containing all non-missing values for the individual columns (blue) to the reduced data set only including complete observations (i.e. $n = 6$). The distributions are highly similar for LRGEN, LRECON, and GAL-TAN.

Importantly, the changes do not seem to eliminate certain policy stances from our data set, but to reduce the number of observations for values that appear frequently. Therefore, even though some of the removed values do significantly impact the distribution of variables, we do not consider these results as problematic. This is supported by our experiments with imputed data, where we did not observe an increase in the quality of our models.

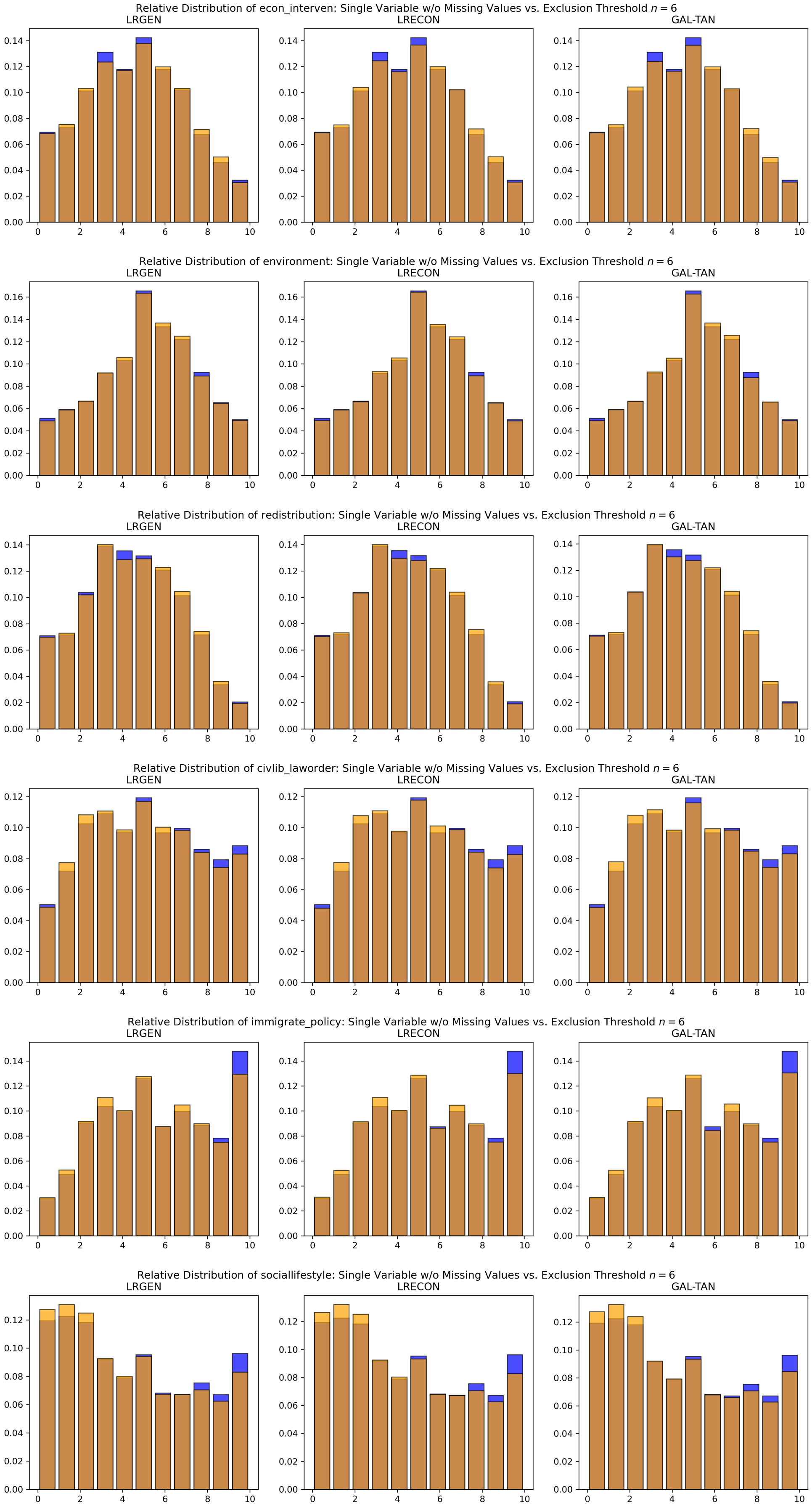
A.4 Details on handling of missing data of the voter study of the EES 2019

For the voter study of the European Election Study 2019, we again face the problem that some variables are missing (e.g. because respondents refused to answer a specific question). Since we use all six policy stances in our models and we want to consider our predictions with regard to general left-right ideology with the respondents' self-description, we can only consider complete observations. This leaves us with 20,186 observations (see Table 16).

	econ_interven	environment	redistribution	civlib_laworder	immigrate_policy	sociallifestyle	lrgen_selfdescription
count	20,186	20,186	20,186	20,186	20,186	20,186	20,186
mean	5.05	3.31	4.45	4.89	5.57	4.60	5.25
std	2.57	2.73	2.93	3.02	3.23	3.95	2.58
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	3.00	1.00	2.00	3.00	3.00	0.00	4.00
50%	5.00	3.00	5.00	5.00	5.00	5.00	5.00
75%	7.00	5.00	7.00	7.00	8.00	9.00	7.00
max	10.00	10.00	10.00	10.00	10.00	10.00	10.00

Table 16: Descriptive statistics of EES 2019 (excluding observations with missing values)

Since the number of observations exceeds the limit of 5,000 observations for the Shapiro-Wilk test (cf. Rahman and Govindarajulu, 1997), we apply the Kolmogorov–Smirnov test confirm that none of the variables are distributed normally. We then again apply the Mann–Whitney U test to investigate whether there are differences between the data sets (1) including all non-missing values for the individual columns, and (2) the reduced data set that only includes


Relative Distribution of econ_interven: Single Variable w/o Missing Values vs. Exclusion Threshold $n = 6$

Relative Distribution of environment: Single Variable w/o Missing Values vs. Exclusion Threshold $n = 6$

Relative Distribution of redistribution: Single Variable w/o Missing Values vs. Exclusion Threshold $n = 6$

Relative Distribution of civlib_laworder: Single Variable w/o Missing Values vs. Exclusion Threshold $n = 6$

Relative Distribution of immigrate_policy: Single Variable w/o Missing Values vs. Exclusion Threshold $n = 6$

Relative Distribution of sociallifestyle: Single Variable w/o Missing Values vs. Exclusion Threshold $n = 6$

Figure 14: Comparison of relative distributions: CHES 2019

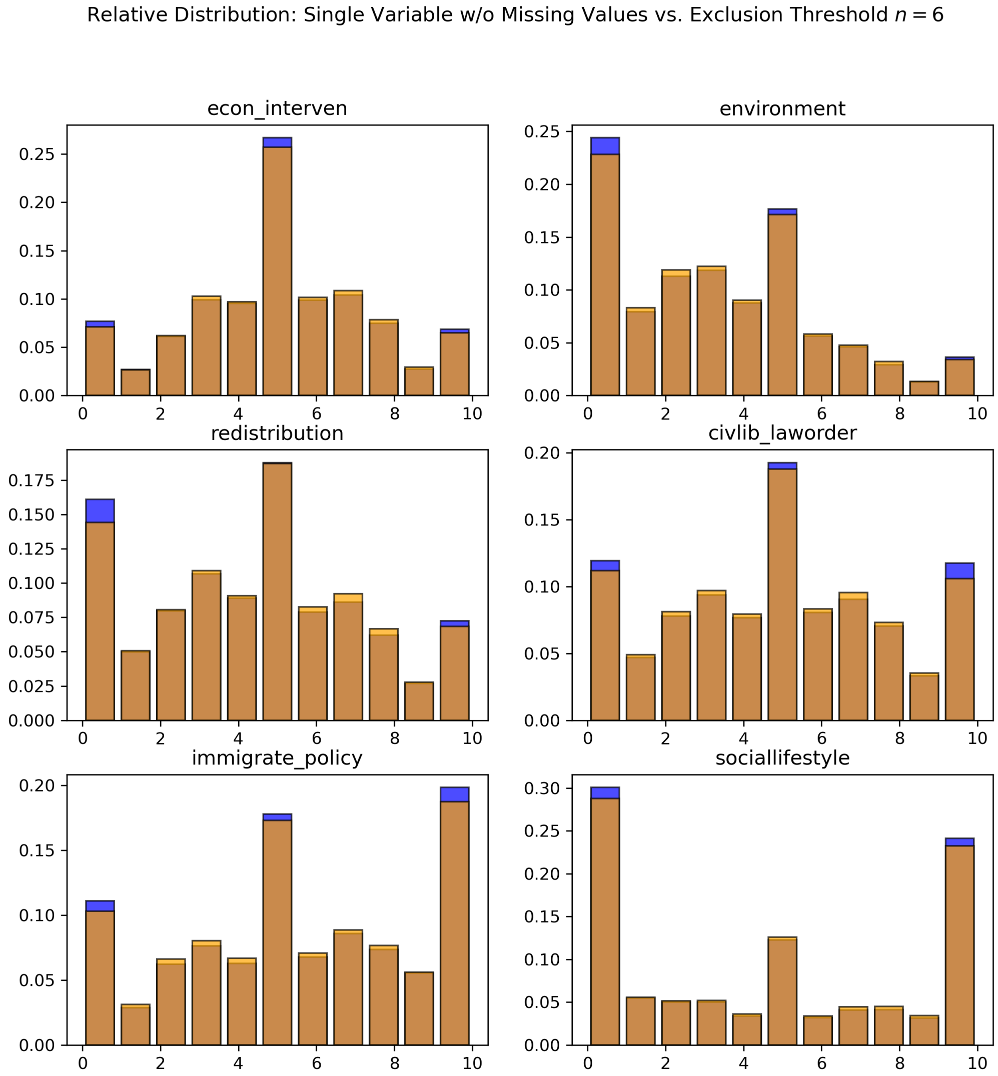


Figure 15: Comparison of relative distributions: EES 2019

complete observations (i.e. with an exclusion threshold of six and no missing values for *lrgen_selfdescription* suggests that the independent variables *environment* (p-value=0.04635) and *redistribution* (p-value=0.00105) are drawn from different distributions. Fig 15) gives an overview of the relative distributions of each variable, analogous to Fig. 14. It suggests that the changes skew the data a bit from the left to the right for these two variables, hence the electorate may be economically more left-wing and socially more green/alternative/libertarian than predicted. Furthermore, the values of the other variables are reduced a bit at the modes, in particular at the extreme ones. Hence, we may slightly underestimate the level of polarization. However, the magnitude of the changes are rather small and hence can plausibly be assumed to not have a large impact.

A.5 Country-level results

Figures 16-18 show how our predictions with regard to left-right ideology diverge from the self-described ideology in different countries.

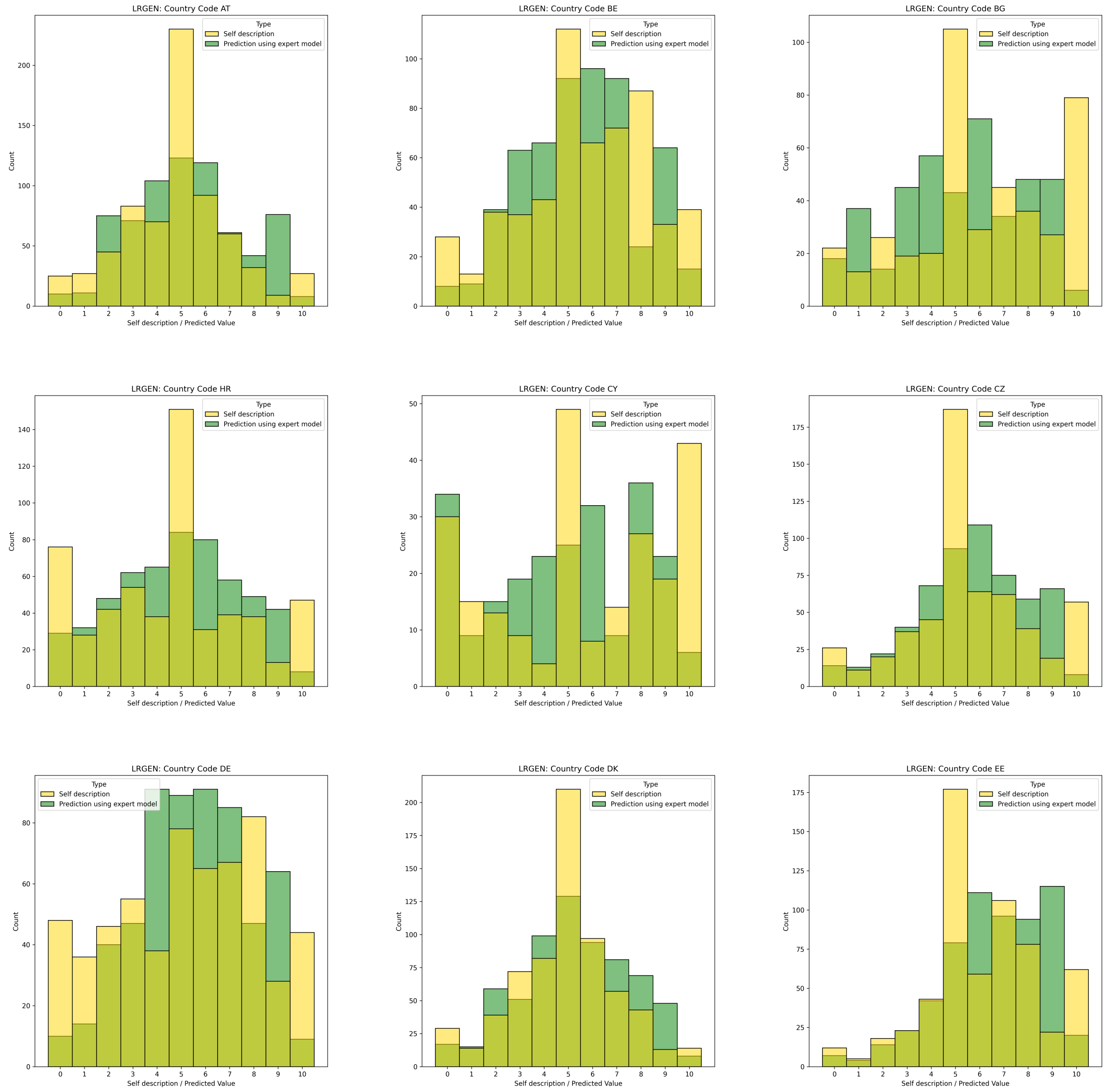


Figure 16: Self-described and predicted values for voters based on the LRGEN model in several countries (11-point Likert scale)

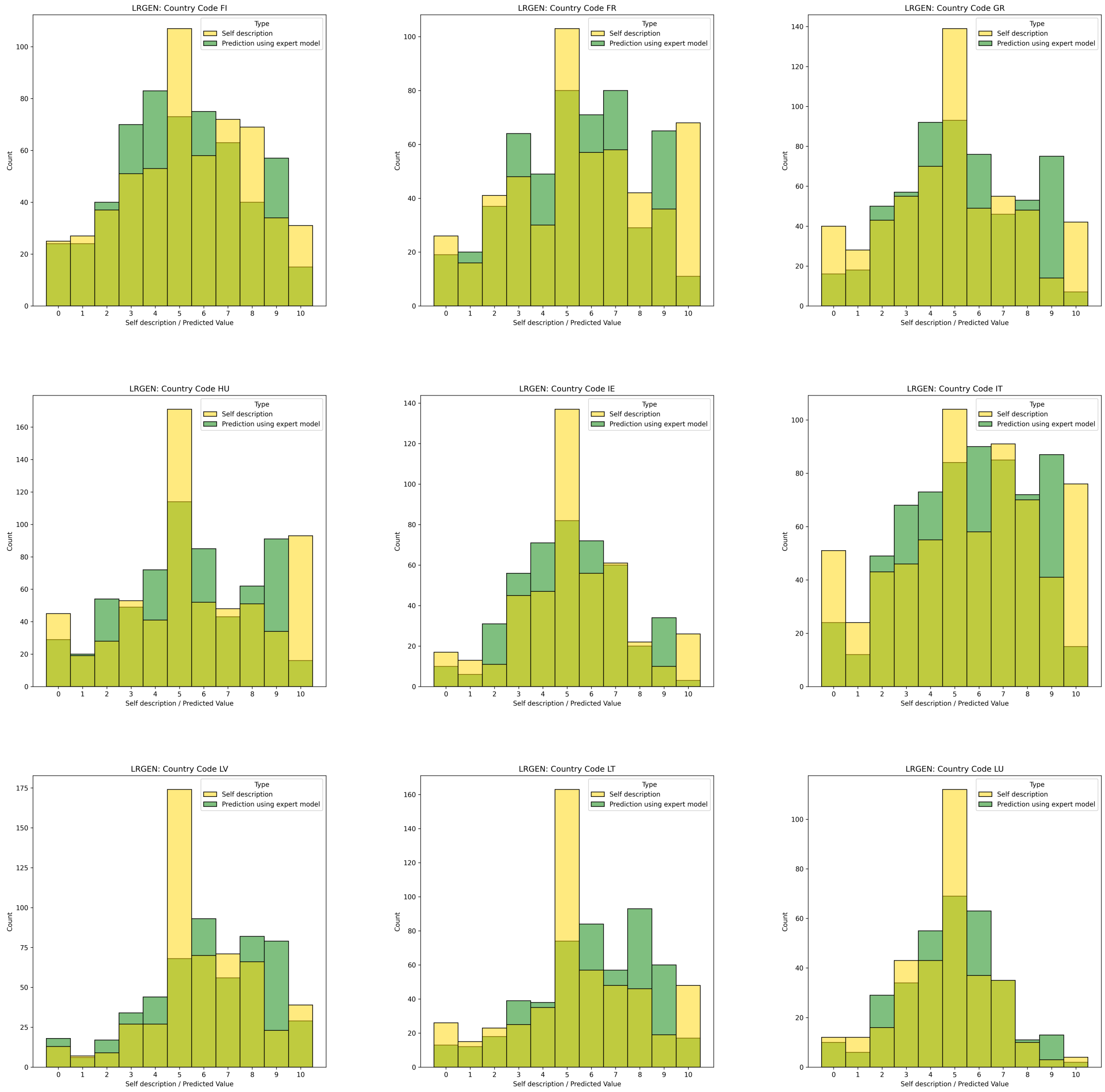
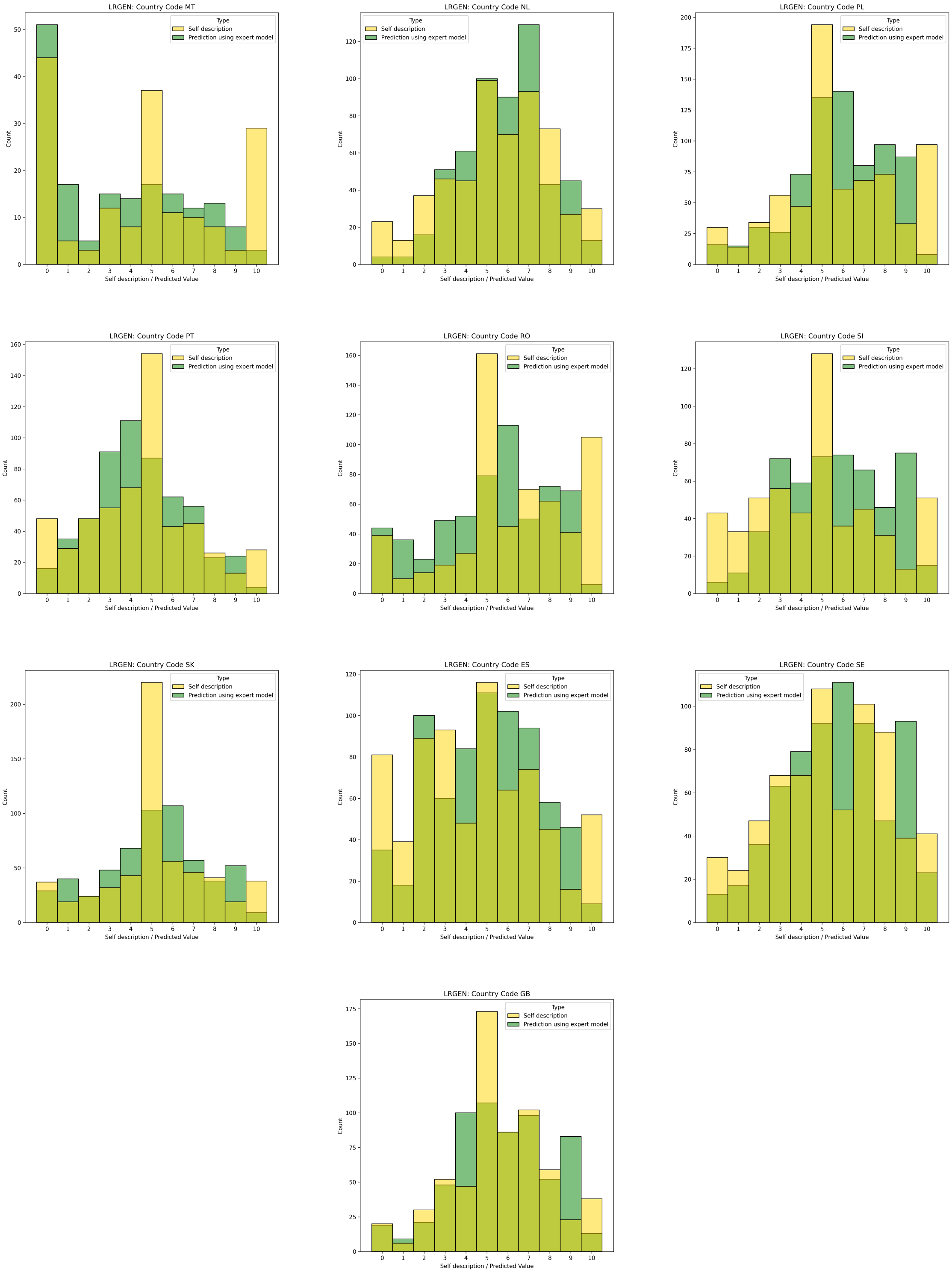


Figure 17: Self-described and predicted values for voters based on the LRGEN model in several countries (11-point Likert scale)



LRGEN: Country Code MT

LRGEN: Country Code NL

LRGEN: Country Code PL

LRGEN: Country Code PT

LRGEN: Country Code RO

LRGEN: Country Code SI

LRGEN: Country Code SK

LRGEN: Country Code ES

LRGEN: Country Code SE

LRGEN: Country Code GB

Figure 18: Self-described and predicted values for voters based on the LRGEN model in several countries (11-point Likert scale)

Fig. 19 plots heat maps of GAL-TAN and the economic left-right ideology that show how the respondents of each country surveyed by the European Election Study 2019 are predicted to be located on a 'political compass'.

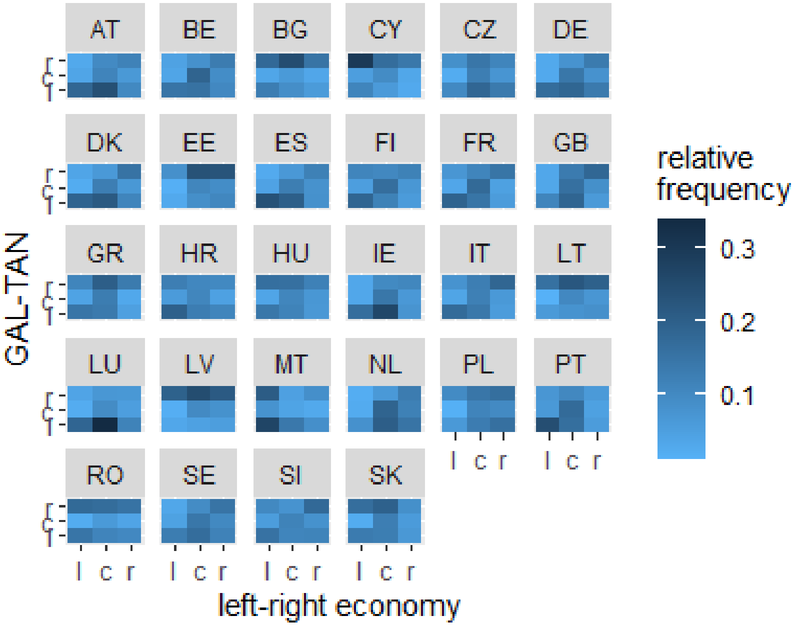


Figure 19: Heatmap of the reduced economic left-right and GAL-TAN predictions for each country

A.6 Robustness checks of the spatial voting models

Tables 17-19 shows regression estimates that do not only include party fixed effects, but also respondent fixed effects.

Table 17: OLS estimates for the propensity to ever vote for a specific party (two-way fixed effects)

	(1)	(2)	(3)	(4)	(5)
Distance between self-placement and subjective party placement	-0.083***	-0.083***	-0.083***	-0.082***	-0.082***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Distance between self-placement and party placement by mean expert	-0.002		0.001		
	(0.002)		(0.002)		
Distance between predicted ideology and party placement by mean expert (left-right general)		-0.010***	-0.010***		-0.004*
		(0.002)	(0.002)		(0.002)
Distance between predicted ideology and party placement by mean expert (left right economy)				-0.005*	-0.002
				(0.002)	(0.002)
Distance between predicted ideology and party placement by mean expert (GAL-TAN)				-0.010***	-0.009***
				(0.002)	(0.002)
Num.Obs.	120154	120154	120154	120154	120154
R2	0.403	0.405	0.405	0.407	0.408
R2 Adj.	0.285	0.287	0.287	0.290	0.291
R2 Within	0.148	0.151	0.151	0.154	0.155
R2 Within Adj.	0.148	0.151	0.151	0.154	0.155
AIC	54662.2	54297.0	54295.8	53803.3	53754.0
BIC	246692.3	246327.1	246335.6	245843.1	245803.5
RMSE	0.26	0.26	0.26	0.26	0.26
Std.Errors	by: party	by: party	by: party	by: party	by: party
FE: respondent	X	X	X	X	X
FE: party	X	X	X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 18: Probit regression estimates for the probability to have voted for a specific party at the last national elections (two-way fixed effects)

	(1)	(2)	(3)	(4)	(5)
Distance between self-placement and subjective party placement	-0.046***	-0.048***	-0.046***	-0.047***	-0.047***
	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)
Distance between self-placement and party placement by mean expert	-0.006		-0.004		
	(0.003)		(0.003)		
Distance between predicted ideology and party placement by mean expert (left-right general)		-0.006***	-0.005***		-0.002
		(0.002)	(0.001)		(0.002)
Distance between predicted ideology and party placement by mean expert (left right economy)				-0.005**	-0.003*
				(0.001)	(0.002)
Distance between predicted ideology and party placement by mean expert (GAL-TAN)				-0.005***	-0.005***
				(0.001)	(0.001)
Num.Obs.	115425	115425	115425	115425	115425
McFadden's R2	0.370	0.370	0.371	0.372	0.373
McFadden's R2 Adj.	-0.242	-0.242	-0.241	-0.240	-0.240
McFadden's R2 Within	0.123	0.124	0.125	0.127	0.127
McFadden's R2 Within Adj.	0.123	0.124	0.125	0.127	0.127
AIC	76977.7	76936.0	76903.6	76817.2	76807.7
BIC	260081.9	260040.2	260017.4	259931.1	259931.2
RMSE	0.29	0.29	0.29	0.29	0.29
Std.Errors	by: party	by: party	by: party	by: party	by: party
FE: respondent	X	X	X	X	X
FE: party	X	X	X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 19: Probit regression estimates for the probability to have voted for a specific party at the elections to the European parliament 2019 (two-way fixed effects)

	(1)	(2)	(3)	(4)	(5)
Distance between self-placement and subjective party placement	-0.043***	-0.046***	-0.043***	-0.045***	-0.045***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)
Distance between self-placement and party placement by mean expert	-0.006*		-0.004		
	(0.003)		(0.003)		
Distance between predicted ideology and party placement by mean expert (left-right general)		-0.005**	-0.004**		-0.0002
		(0.002)	(0.001)		(0.001)
Distance between predicted ideology and party placement by mean expert (left right economy)				-0.003*	-0.003*
				(0.001)	(0.002)
Distance between predicted ideology and party placement by mean expert (GAL-TAN)				-0.007***	-0.007***
				(0.001)	(0.001)
Num.Obs.	117962	117962	117962	117962	117962
McFadden's R2	0.487	0.487	0.488	0.493	0.493
McFadden's R2 Adj.	-0.354	-0.354	-0.353	-0.348	-0.348
McFadden's R2 Within	0.195	0.195	0.197	0.204	0.205
McFadden's R2 Within Adj.	0.195	0.195	0.196	0.204	0.204
AIC	62406.7	62415.6	62365.6	62135.6	62137.4
BIC	249929.9	249938.8	249898.5	249668.5	249680.0
RMSE	0.27	0.27	0.27	0.27	0.27
Std.Errors	by: party	by: party	by: party	by: party	by: party
FE: respondent	X	X	X	X	X
FE: party	X	X	X	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$