Estimating subnational preferences across the European Union

The European Political Science Association, 2018

Estimating subnational preferences across the European Union *

JANA LIPPS AND DOMINIK SCHRAFF

Subnational analyses of political preferences are substantively relevant and offer advantages for causal identification. Yet, our knowledge on regional political preferences across Europe is limited, not least because there is a lack of adequate data. The rich Eurobarometer data is a promising source for regional information across Europe. Yet, it is only representative for the national level. This paper compares state-of-the-art methods for estimating regional preferences from nationally representative Eurobarometer data, validating our estimations using regionally representative surveys. Our analysis highlights a number of challenges for estimating regional preferences across Europe, such as data availability, variable selection, and over-fitting. We find that predictions are best using a Bayesian additive regression tree with synthetic post-stratification.

*Jana Lipps is doctoral student at the European politics research group at ETH Zurich, Switzerland. Dominik Schraff is postdoctoral researcher at the European politics research group at ETH Zurich, Switzerland. We thank Thomas König, Lucas Leemann, Matthew Loveless, Santiago Olivella, and Fabio Wasserfallen for helpful comments and advice.

Subnational variation in political preferences across Europe has attracted increasing attention in recent years. [1] Elections in Europe have repeatedly demonstrated the divergences of political preferences between regions (e.g., Jones, Johnston, and Pattie 1992; Stockemer 2017). A prominent example is the British vote to leave the European Union, which was informed by strong geographical divides within the United Kingdom.[2] The substantial variation in living conditions across the territories of EU members states likely affects a wider range of political preferences, such as preferences for European integration or welfare state preferences (Beramendi 2012; Chalmers and Dellmuth 2015; Pittau, Zelli, and Gelman 2010). Moreover, regional variation offers great potential to improve causal identification in comparative politics, as the manifold con-founders on the national level can be held constant. Yet, we so far have very limited insights on subnational preferences in the European Union (EU). This is first and foremost due to a lack of reliable data.

Comparative social science research has seen an extensive mushrooming of surveys collecting information on all aspects of life in the European member states. Most prominent are probably the Eurobarometer (EB), the European Social Survey or the European Election Study. However, these surveys are not designed to provide representative samples on the subnational level. Regional information from these surveys is therefore noisy and potentially biased. Only a hand full of surveys have sampled public opinions on a regional level. A unique source covering the whole EU-28 is the Flash Eurobarometer on regional public opinion from 2015, surveying around 60,000 citizens. Certainly, there are not

[1]This paper uses the terms subnational and regional exchangeable. Both terms are used to describe geographical areas within EU member states.

[2]https://www.economist.com/news/britain/21701257- results-paint-picture-angry-country-divided-class-age-and-region-country-divided

more regional surveys due to the large financial resources required. Estimating regional preferences from nationally representative surveys, therefore, seems a more viable, less costly option (Lax and Phillips 2009). Recent studies on regional political preferences in the EU have been well aware of the noise in national surveys. So far, the usual strategy to tackle the noisiness of regional EB data was multi-level modeling (e.g., Lubbers and Scheepers 2010). However, we currently have no knowledge on the quality of regional estimates from standard EB surveys.[3]

This paper tries to tackle this issue using EB data - the data with the widest temporal and cross-sectional coverage of political preferences across the EU - to investigate regional preferences. First, we compare regionally disaggregated political preferences from standard EB surveys to averages from regionally representative surveys. We find that regional information from the standard EB is very noisy, hardly justifying using them as valid source for regional preferences. In a second step, we implement strategies to improve the estimation of regional preferences using EB data. Our goal is to develop a model that can be applied as a default method to estimate various regional political preferences from standard EB surveys. Our evaluation study finds that multi-level modeling with partial pooling and regional predictors leads to moderate improvements. Multi-level modeling with post-stratification (MrP) has great potential. Yet, the regression-based methods face a number of challenges with regard to model building. A Bayesian additive regression tree (BART) method provides the best predictions by requiring minimal researcher intervention. We therefore propose a BART model with synthetic post-stratification as a promising tool to improve the quality of regional information from standard EB surveys. We conclude

[3]Note that the paper by Toshkov (2015) applies the multi-level regression with post-stratification strategies to EB data. However, Toshkov (2015) estimates country-level preferences and disregards the subnational information in the data.

with a discussion on implications for future research on regional political preferences in the European Union.

Why the regional level matters

Citizens are nested in different jurisdictions of varying territorial scales. This has been described by scholars as multi-level governance (Hooghe and Marks 2001). Governance levels stretch from the global to the local level of communities. While the national political arena is clearly the most important one for our understanding of political processes, subnational levels have attracted more attention in the past decade. In Europe, regional governance structures have been shown to be a meaningful unit of analysis (Chalmers 2013; Marks, Hooghe, and Schakel 2008). Within the EU, regions have a political clout, acting strategically on the national and supranational level (Keating 1997, 2008; Tatham 2012). With their own representations and the Committee of Regions, regions are highly visible in Brussels. Although they lack the leverage of nation-states, they have increased the awareness for regional specificities. Due to the EU structural funds, regions have become direct beneficiaries of a large share of the EU's budget. Yet, EU citizen still face very different living conditions with regard to economic wealth, infrastructure, public services, etc. The EU's Eastern enlargement has increased these divides even further.

Therefore, analyzing regional preferences within the EU seems highly relevant to our understanding of politics in Europe. Not least because the subnational level moves closer to peoples' actual living context. A number of studies in recent years, many of which relied on the Eurobarometer data, have demonstrated the relevance of the regional level for political preferences in Europe (Chalmers and Dellmuth 2015; Dellmuth and Chalmers 2017; Lubbers and Scheepers 2010; Pittau, Zelli, and Gelman 2010; Schraff 2017). For

instance, radical right-wing voting does follow distinct regional patterns (Stockemer 2017). Moreover, support for European integration is affected by regional characteristics. Lubbers and Scheepers (2010), for example, show that changes in regional economic wealth affect Euroscepticism. Another example are Dellmuth and Chalmers (2017), who demonstrate that regional EU spending affects support for European integration. Furthermore, Beramendi (2012) shows that regional preferences on redistribution shape the design of national and supranational fiscal institutions. We could further improve our understanding of these processes with reliable data on political preferences across European regions.

Further, regional-level data offers great potential to strengthen causal inference in comparative politics (cf, Rodden 2010). Usually, country-level data shows limited variability and suffers from many potential con-founders. For example, the effects of electoral system features on political preferences can be identified more credibly on the regional level. This, of course, requires reliable data for subnational districts. Comparative research across the EU is limited by the number of member states (currently 28). Yet, the number of regions in the EU is already around 200, even when not using the most fine-grained administrative grid. Research focusing on the explanatory power of living conditions across Europe, such as economic or social contexts, could benefit greatly from subnational analyses, as conditions vary strongly across space (cf, Pittau, Zelli, and Gelman 2010).

Regional information in the EB data

The EB data has a number of properties that provide great advantages for comparative social science research in Europe. It is highly standardized and has a large temporal

and spatial coverage. It provides long time-series and covers all EU member states (and some more). Moreover, standard EB surveys record the region of residence of respondents, which usually refers to the NUTS2 level. [4] The standard EB samples around 1,000 respondents per member state, which are representative for the country as a whole. The sample size and the sampling procedure are not designed to provide valid regional preferences. This is a serious challenge for research interested at the regional context. The number of respondents per region in the EB is usually rather low and the regional samples are not representative for the regional population. With EB data, it can easily happen that we only observe a hand full of respondents for sparsely populated areas. As a result, regional information from standard EB surveys is likely to be very noisy. This makes it much harder to identify the true parameter values in statistical analyses and potentially leads to biased results. [5]

Of course, recent studies have been well aware of this challenge. So far, the standard account to tackle the noisiness of regional EB data is multi-level modeling. Under this approach, a set of good predictors and partial pooling (also called global smoothing) should improve the estimation of regional averages. Partial pooling should shrink regional estimates towards the country mean, therefore addressing imprecisions in regions with very few respondents. Further, predictors that are informative of regional differences in political preferences should improve estimates (Hanretty, Lauderdale, and Vivyan 2016).

[4]The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU (http://ec.europa.eu/eurostat/web/nuts/background).

[5]It is important to note that the Eurobarometer data comes with some other issues regarding methodological transparency, the quality of field work, and measurement validity. Other surveys, such as the European Social Survey, are stronger in this regard. Yet, the EB still has the widest geographical and temporal scope, which provides unique added value for researchers.

However, we so far lack empirical evidence on how well this approach works. In fact, even though the EB is a highly prominent data source, we currently have no knowledge on the quality of regional estimates from standard EB surveys. To address this gap, we provide a first investigation of the validity of regional information from EB data. Moreover, we also compare the performance of different strategies to derive regional estimates from standard EB data. Before we turn to the more advanced estimation procedures, we use the next section to provide an overview on how noisy regional information from the EB is.

The challenge of heterogeneity

Even though the Eurobarometer data provides regional information since decades, research on regional public opinion in the EU is limited. We suspect that this is also due to concerns with the data quality of regional information from the EB. Recent advances in statistical modeling might allow us to mitigate some of these concerns. Yet, we so far lack systematic evidence on how good (or bad) regional information from EB surveys are. There are good reasons for this research gap. An evaluation of the validity of regional EB data requires a benchmark. In this paper, we use unique regionally representative public opinion surveys for Europe to evaluate regional estimates from standard EB data.

Primarily, we rely on a 2015 Flash EB (Flash Eurobarometer 427) on regional public opinion. The 2015 Flash EB was designed to capture regional political preferences across the EU-28, therefore sampling 300 respondents per region. [6] While the Flash EB 2015 is able to provide us with valid regional opinions ("true" values), it only covers a very small

[6]Flash Eurobarometer are special surveys conducted by the EU Commission. They are distinct from the standard EB surveys as they have no core set of questions and are not run annually.

set of political preferences. The only political preference that is congruent with items in standard EB surveys is a dichotomous measure of EU trust (1 - trust EU; 0 - not trust EU). We therefore focus our evaluation study on estimating regional EU trust. EU trust is a prominent dependent variable, as it captures diffuse support for EU integration (Foster and Frieden 2017; Harteveld, Meer, and Vries 2013; Muñoz, Torcal, and Bonet 2011). Such diffuse support is central for the legitimacy of a political system.

We use two further datasets that provide valid regional information. We can compare results to a 2012 Flash EB (Flash Eurobarometer 356) that is very similar to the 2015 Flash EB, but covers less regions. Moreover, we use the raw survey data from the Quality of Government (QoG) project (Teorell et al. 2017). The European QoG data are based on two regionally representative surveys from 2010 and 2013. Even though the QoG data does not cover the whole EU and has many idiosyncratic questions that are not comparable to EB surveys, the QoG provides a general left-right self-placement measure that can be compared to the EB's left-right scale.

In a first step, we check how noisy regional estimates from standard EB surveys are. We only should be interested in the issue of regional preferences estimation if the correspondence between the raw standard EB data and the true values is weak. Table 1 presents errors and correlations (Pearson's Rho) between the true Flash EB values and weighted regional averages from standard EB surveys. We simply disaggregate EU trust by region in the standard EB 83.3 and compare it to the Flash EB 2015 values. [7]. The fieldwork of the two surveys was conducted around the same time. The Flash EB 2015 data was collected in March 2015 and the Standard EB 83.3 data was collected in May 2015. The number of cases per region in the Standard EB 83.3 is 124 on average. Yet,

[7]All regional averages are weighted with the corresponding post-stratification weight.

TABLE 1 *Performance of Disaggregation*

	Flash EB 2012: EU Trust				Flash EB 2015: EU Trust				QoG 2013: Left Share			
	cor	cov	MAE	RMSE	cor	cov	MAE	RMSE	cor	cov	MAE	RMSE
Eurobarometer 77.3	0.498	0.222	0.142	0.180			–				–	
Eurobarometer 83.3			–		0.428	0.332	0.120	0.149			–	
EES 2014			–		0.407	0.357	0.106	0.136			–	
Eurobarometer 79.5			–				–		0.338	0.348	0.084	0.109

Note: cor = Pearson's Rho; cov = number of disaggregated means falling within confidence intervals of true values; MAE = mean absolute error; RMSE = root-mean-square error

there are sparsely populated regions with only 7 or 8 respondents.

Table 1 shows that regional EU trust from the standard EB is only correlated by 0.43 to the true regional EU trust values. This suggests that regional information from standard EBs is very noisy. Indeed, the mean absolute error (MAE) of the disaggregated standard EB values is 0.12 and the residual mean squared error (RMSE) is at 0.15. Hence, on average, disaggregated regional EU trust from the standard EB is off by 15 percentage points. The coverage measure (cov) reports how many of the disaggregated values lie within the confidence intervals of the true values. This is a useful measure as the Flash EB values are estimates and have a random error. Table 1 shows that only about 30 percent of all regions fall within the confidence intervals of the true values using simple disaggregation.

Figure 1 provides a graphical representation of the relationship between the two EU trust measures. Each row in the figure corresponds to a region, and the graph plots the means for each region from both datasets, including confidence intervals for the Flash EB means. The graph clearly shows that the standard EB data hardly aligns with the true values from the Flash EB.

To make sure that this low correspondence between disaggregated EU trust from the standard EB and the true values is not incidental, we investigate the performance of disaggregation for 2012 as well. Here, we compare actual regional EU trust from the 2012 Flash EB with the disaggregated EU trust from the 77.3 Standard Eurobarometer.

Correlation looks slightly better, but coverage is worse and errors are even larger. We can also apply disaggregation to an identical EU trust measure included in the European Election Study 2014 and compare it to the 2015 Flash EB. The correspondence is similarly weak, with correlation around 0.4 and an RMSE of 0.14.

Finally, we check that the issue of heterogeneity in regional EB data is not unique for EU trust. Using the regionally representative QoG data from 2013 and the standard EB 79.5 from the same year, we investigate the performance for the well-known left-right self-placement variable. We calculate a regional left share, which measures the share of respondents placing themselves on the left of the political spectrum. Again, as Table 1 shows, we see weak associations between the standard EB data and the regionally representative information.

In sum, we see substantial deviations between the disaggregated regional information from the standard EB surveys and the regionally representative surveys. The deviations between national and regional representative surveys are independent of the survey data and the variable under consideration. Frequently, disaggregated regional EU trust values are off by 15 percentage points or more. Even worse, they differ so strongly that substantive interpretations can lead scholars in complete opposite directions. Consider an example where standard EB data shows a rather high EU trust of 55 percent, suggesting that a majority of people trusts the EU. Accounting for noise, actual EU trust can easily be around 40 percent, suggesting that a majority clearly distrusts the EU.

Predicting regional preferences

The previous section has demonstrated that regional information from standard EB data is very noisy. This makes the endeavor of improving regional estimates with statistical

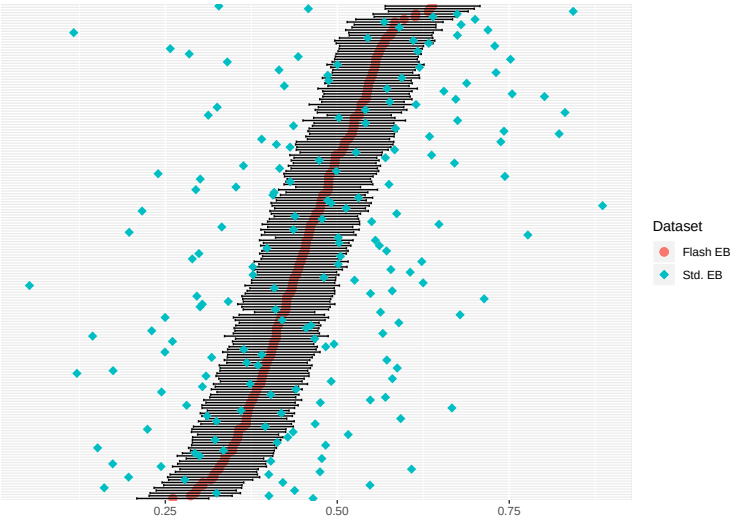

Figure 1. Noise in regional EU trust

methods even more relevant. The remaining part of the paper will evaluate different strategies to more adequately predict regional preferences with standard EB surveys. Below, we shortly introduce the available strategies for estimating subnational preferences. Since, we evaluate a number of alternative strategies, we have limited space to provide detailed explanations of the methods. For a more extensive discussion of the methods we want to refer readers to Hanretty, Lauderdale, and Vivyan (2016), who provide a very structured and clear overview of regression-based methods. We refer to Leemann and Wasserfallen (2017) for more details on synthetic post-stratification. Finally, Montgomery and Olivella (2018) provide a great introduction to tree-based methods for the prediction of subnational preferences.

Global smoothing in multi-level regression

Current approaches to deal with the noisiness of regional information in the standard EB data apply multi-level regression (Mr) (Chalmers and Dellmuth 2015; Dellmuth and Chalmers 2017; Lubbers and Scheepers 2010; Pittau, Zelli, and Gelman 2010). The hierarchical structure in the standard EB gives rise to a two-level model, with random effects for country and region. Regional estimates should improve due to partial-pooling, as sparsely sampled regions more strongly rely on their respective country average. Additionally, good regional predictors should improve predictions by making regions more comparable (Hanretty, Lauderdale, and Vivyan 2016). Regional predictors are also an important extension to the more complicated multi-level approaches introduced below. However, a major short-coming of these methods refers to the selection of regional co-variates. Eurostat, for example, offers a rich set of variables on the regional level. Selecting a regional predictor frequently seems arbitrary and brings the risk of over-fitting the data. In this paper, we want to develop general models that minimize the threat of over-fitting. This is crucial so that we confidently can apply any model to new, unseen data. We therefore work with a parsimonious set of regional predictors in all multi-level modeling applications.

Classical MrP

An extension of simple Mr is multi-level regression with post-stratification (MrP). This method gained increasing attention in the past decade and is currently the standard approach for estimating subnational preferences from nationally representative data (Buttice and Highton 2013; Hanretty, Lauderdale, and Vivyan 2016; Lax and Phillips 2009; Leemann and Wasserfallen 2017; Reilly, Gelman, and Katz 2001; Rodden 2010; Toshkov 2015;

Warshaw and Rodden 2012). MrP comes with the advantages of simple Mr, but additionally allows to correct for systematic differences between the estimation sample and the true population (Hanretty, Lauderdale, and Vivyan 2016; Lax and Phillips 2009). Multi-level models are estimated with a set of demographic characteristics (e.g., gender, education, age). Predictions from the multi-level model are then weighted with census data on the true population structure of the demographics.

Data requirements for classical MrP are quite demanding. Researchers need regional census data on all individual level predictors included in the multi-level model. This comes with a number of challenges. First, census data needs to be congruent with items in the survey data. Education, for example, is measured in the Eurobarometer as the age when respondents finished full-time education. This is not comparable to regional census data from Eurostat, which reports education based on the international standard classification of education. Second, regional census data is only available for a limited set of variables, which are not necessarily good predictors of certain political preferences. Yet, the gains of post-stratification depend on the extent to which individual level predictors explain regional differences (Leemann and Wasserfallen 2017).

Synthetic MrP (MrsP)

A recent advancement in the methodological literature is multi-level regression with synthetic post-stratification (MrsP). MrsP addresses some of the weaknesses of classical MrP. More specifically, MrsP allows researchers to use individual level variables that are powerful predictors, but are not available from census datasets. MrsP weights multi-level estimates with so-called synthetic joint probabilities (Leemann and Wasserfallen 2017). Synthetic joints are created by simply multiplying a number of marginal probabilities from

the regional level.

Consider an example where one specifies a model using gender, education, and occupational status as individual level predictors. Classical MrP requires regional census data on gender, education, and occupation for post-stratification: meaning, true joint probabilities. For instance, MrP needs to identify for each region how many citizens are female, tertiary educated, and employed (continuing this for each logical combination of the categories of individual level predictors). Only very detailed census data can provide these information.

MrsP does not need such detailed data. For MrsP, one collects regional data on the relative gender, education and occupation shares (regional margins). Joints are then created by multiplying these margins for each logical combination of the categories of individual level predictors. These are called synthetic joints as they are not representing the true probability distributions within each region. This is because multiplication assumes that individual level predictors are independent from each other. However, recent research has found that synthetic joints work very well, even if the zero-correlation assumption is violated (Leemann and Wasserfallen 2017).

The great advantage of MrsP is that it allows models with individual level predictors that are not included in census surveys. Leemann and Wasserfallen (2017), for example, use regional margins on party identification to improve estimates of regional political behavior. The inclusion of such powerful predictors can greatly enhance predictions of regional preferences. A major restriction to MrsP is - again - model building. For example, a good individual-level predictor of EU trust is not necessarily a good predictor of redistributive preferences. To build general models, it therefore is important to select variables for synthetic joints that are potentially informative for many political preferences. Otherwise, one risks building idiosyncratic models that do not travel well to other settings.

However, as with the selection of regional predictors in multi-level regression models, one cannot validate these model building steps. As a results, there is a large degree of researcher judgment involved in the selection of individual and regional level predictors.

Post-stratified BART

A recently proposed alternative to the regression-based methods introduced above is a Bayesian additive regression tree (BART) (Chipman, George, and McCulloch 2012; Montgomery and Olivella 2018). There are two major advantages of this tree-based method. First, it is not fitting a global model, but rather estimates highly interactive effects across partitions of the data (a classical tree feature). BART therefore implements deep interactions to predict regional preferences as envisioned by Ghitza and Gelman (2013). Second, BART is an ensemble method which retrieves average predictions across many trees using a Bayesian back-fitting algorithm. BART therefore selects the best predictors and their interactions in an iterative process, in which every tree is a weak learner. This allows researchers to provide a set of predictors and let the algorithm select and interact them in a non-parametric way, just as the data calls for them. As this is done across an ensemble of trees (e.g., 200 trees), we avoid over-fitting the data. BART therefore promises greater precision (truly deep interactions) as well as more generality due to less researcher intervention in the modeling process (ensemble method). BART predictions can also be post-stratified with classical or synthetic joints, just as in the MrP and MrsP approaches. This provides a further safeguard against bad predictions.

Of course, there are also shortcomings to BART. First, it is a black box method which is more difficult to interpret compared to multi-level regression models. Second, the Bayesian back-fitting algorithm is computationally more costly. Moderately complex BART models

can easily take 20 to 30 minutes to compute and computation time increases exponentially with more individual level predictors. Finally, a few hyper-parameters have to be set by the researcher before running the BART model. Details on the specific design of the BART model and the hyper-parameters can be found in Chipman, George, and McCulloch (2012). One has to define the number of trees the BART model should estimate. Here, Chipman, George, and McCulloch (2012) have shown that 200 trees provide reliable predictions. Further, researchers can set parameters for the priors to control the influence of individual trees on the overall fit. We have chosen parameters to ensure a conservative fit by limiting the influence of individual trees.[8] However, we have varied all hyper-parameters and the resulting predictions remain very similar. Generally, BART has well defined default settings and works good as an off-the-shelf method (Montgomery and Olivella 2018).

Data

We are using the Standard Eurobarometer 83.3 to predict regional EU trust and compare the results with the regionally representative data from the 2015 Flash EB. After cleaning and aligning the two datasets, we are able to analyze EU trust across 184 EU regions. However, depending on the co-variates used for post-stratification, the number of regions decreases to 154. Mostly, we analyze NUTS2 regions. The NUTS2 level is the regional level on which the EU regional policy is applied. It is defined as regions with a population size between 800,000 and 3 million. However, for Germany, Belgium, Greece, and the United Kingdom, Eurobarometer does only report the larger NUTS1 regions. The intra-regional correlation (regional ICC) for EU trust in the EB 83.3 is 0.11, meaning that 11 percent of

[8]For this, we chose a high value of ten for the "k" parameter (Chipman, George, and McCulloch 2012).

the total variation in EU trust is due to regional differences.

We add regional-level predictors to improve the model fit. From Eurostat, we retrieve complete regional data on GDP per capita, area size, population density, unemployment rate, tertiary education shares, manufacturing shares, and median household income. We use all regional predictors in the BART model, because the Bayesian back-fitting algorithm does select on the most relevant predictors. However, we have to select manually some variables for the multi-level regressions to avoid over-fitting. Therefore, we choose a parsimonious set of predictors for the regression-based methods. Here, we account for the effect of economic conditions on EU trust (Foster and Frieden 2017), adding the regional gross domestic product (per capita) recorded by Eurostat. We suspect that regional wealth divides shape a number of political preferences across Europe. Moreover, we add a variable measuring whether a region lies in eastern, western, northern or southern Europe ("grand region").[9]

For post-stratification, we implement two strategies. First, we apply classical post-stratification using regional demographics from the 2011 census conducted by Eurostat (classical joints). The 2011 EU census provides a limited set of demographics for all EU regions. All census data can be subset by gender and age groups, variables that are also included in the standard EB. Another demographic variable that is congruent between the census and the EB is marital status. Hence, we use census data on gender, age, and marital status. Multi-level regression results are therefore post-stratified by all logical

[9]Note that we compare models that do not have the same set of predictors. One might suggest that a "fair" comparison would require identical inputs. However, the different estimation strategies vary in their ability to capitalize on the available data. As our goal is to develop general models that provide best predictions, we provide each strategy with the adequate data. This also highlights the advantages and disadvantages of the different strategies.

combinations of two gender, six age (ten year groups from 15 years onward, with the last category being "65 years and older"), and four marital status (single, widow, (re-)married, divorced or separated) categories.

Second, we implement synthetic post-stratification by adding a promising individual-level predictor. We were looking for a predictor that potentially speaks to many political preferences. This is to ensure a certain generality of the model and avoid over-fitting. We chose the general left-right measure, which is surveyed in the EB regularly. We group respondents into the three categories "left", "center", and "right", and assume that political preferences are informed by the divides across these ideological camps. Margins on left-right self-placement are taken from the regionally representative 2013 QoG survey. Combining this with regional margins on gender, age, and marital status from the census data allows us to create a table of synthetic joints. Table A1 in the Appendix provides an overview of the data sources and coding schemes for the level-1 predictors.

Table 2 provides an overview of the different strategies and predictors used in the analysis. We evaluate four strategies. For each strategy, we calculate the regional predictions with and without post-stratification. This results in an overall set of eight prediction outcomes. Note that the multi-level models are estimated with a reduced set of regional predictors. Contrarily, the BART models include the full set of regional predictors. Here, all regional predictors are country-mean centered to capture the within-country, between-region variation. Further, to account for variation between countries, we add the country mean of each regional predictor. This results in a set of 14 level-2 predictors for the BART models.

TABLE 2 *Overview of predictors and strategies*

Strategy	Level-1 Predictors	Level-2 Predictors
MrP	gender, age, marital status	region, grand region, country, gdp
MrsP	gender, age, marital status, left-right	region, grand region, country, gdp
BART CJ	gender, age, marital status	gdp, population density, area size, tertiary education rate, unemployment rate, manufacturing share, median household income
BART SJ	gender, age, marital status, left-right	gdp, population density, area size, tertiary education rate, unemployment rate, manufacturing share, median household income

Note: Regional predictors for the BART models enter the algorithm group-mean centered and as country mean. BART CJ = BART with classical joints; BART SJ = BART with synthetic joints

Evaluation results

Figure 2 presents the results of the different estimation strategies. Note that the results are based on the 154 regions for which we have complete data on all variables. This is to ensure comparability between the model estimates. We later also report the results for the method that is able to cover the full sample. Our baseline result is simple disaggregation of the standard EB 83.3 (see Table 1). We therefore present the gains of our estimations relative to disaggregation. Panel a of Figure 2 plots the reduction in RMSE compared to disaggregation. Panel b plots the reduction in MAE. Overall, we find that every method improves upon simple disaggregation. Put differently, simple disaggregation is never the preferred method for estimating regional EU trust. But which strategy provides the best predictions?

A simple BART model only including the covariates for the classical joints (gender, age, martial status), plus regional predictors, provides the smallest gains ("bart.p.cj" and "bart.cj"). Note that the BART model is provided with all the seven regional predictors we could get complete data on. Even though BART does select on the most relevant predictors, we do not see substantial gains by adding regional covariates. Most likely, the available regional covariates are not very informative of the outcome EU trust. Mainly due to the individual level predictors, we get a reduction in RMSE of around -0.025 and

in MAE of a bit more than -0.02. These are moderate gains, as, for example, an error in regional EU trust of 10 percent would merely decrease to 8 percent. We also see that post-stratification does not make a difference. The raw predictions ("bart.cj"") perform identical to the post-stratified predictions ("bart.p.cj").

In the mid-range of our estimation gains are the simple multi-level model predictions ("ml.sj" and "ml.cj"). The predicted values from a multi-level model improve RMSE of around -0.035 and MAE of around -0.03. Here, the multi-level model with the covariates for classical joints performs even better than the model including left-right categories. These differences even increase if we post-stratify the Mr results. We see the post-stratified classical joints model ("post.cj") approaching an RMSE reduction of -0.04, while the post-stratified synthetic joints model drops in RMSE reduction to -0.03.

The best predictions are reported by the post-stratified BART using synthetic joints. RMSE decreases by -0.05. This means that - on average - a region with a 10 percent deviation of EU trust under disaggregation, improves to a 5 percent error after synthetic BART. While the MAE was 0.12 under disaggregation, synthetic BART decreases this noise by 40 percent to 0.07. This is a substantial gain in precision. Again, we see that post-stratification does not reduce prediction error substantially. The results for the raw predictions ("bart.sj") and the post-stratified ones ("bart.p.sj") are nearly identical.[10]

A number of insights emerge from our evaluation study. Previous research frequently relied on Mr to analyze regional EB data. However, our Mr results show moderate gains.

[10]There are many potential explanations on why post-stratification does not improve predictions substantially in this case. Most likely, the representativeness of the EB is fairly good with regard to the level-1 predictors used. This result is not unique to our analysis. The post-stratified BART application presented in Montgomery and Olivella (2018) does also not gain a lot from re-weighting. Still, post-stratification is a good safeguard against bad predictions.

Mr is definitely better than running a regression on the disaggregated regional means. However, the reduction in prediction error is below synthetic BART. Most importantly, the gains of Mr will always depend on the specific modeling strategy. It could well be that there are much better Mr models than the ones we estimated, for example using individual-level predictors that are more closely related to the outcome variable. These, however, run the risk of being idiosyncratically tuned towards the specific outcome and data under investigation.

The same reservation applies to the MrsP model. There might be other individual level predictors that provide substantive gains in post-stratification. For instance, one could use trust in national political institutions as predictor of EU trust, since these variable have been shown to be substantially related (Muñoz, Torcal, and Bonet 2011). Indeed, we can tune a MrsP model including martial status, age, left-right, and national parliament trust to approach the gains reported under the synthetic BART model. Figure A1 in the appendix presents the results for this extended MrsP model. The model reaches similar gains to the synthetic BART model, with a MAE of 0.07. This, however, comes at the cost of building a more idiosyncratic model. Developing this model, we had to re-run many MrsP analyses, throw out gender and play around with different sets of regional predictors. This means that there was a lot of data-fitting and researcher judgment involved. We therefore are very skeptical that this performance will hold in new, unseen settings. Not least because national parliament trust is a level-1 predictor selected specifically for the outcome EU trust.

There is less of a concern regarding over-fitting with the classical MrP model. There is not much of a choice in the selection of individual-level predictors for post-stratification with classical joints. For the whole EU, the only available data source is the EU census. Here, only age, gender, and marital status are compatible to items in the EB. There still is a

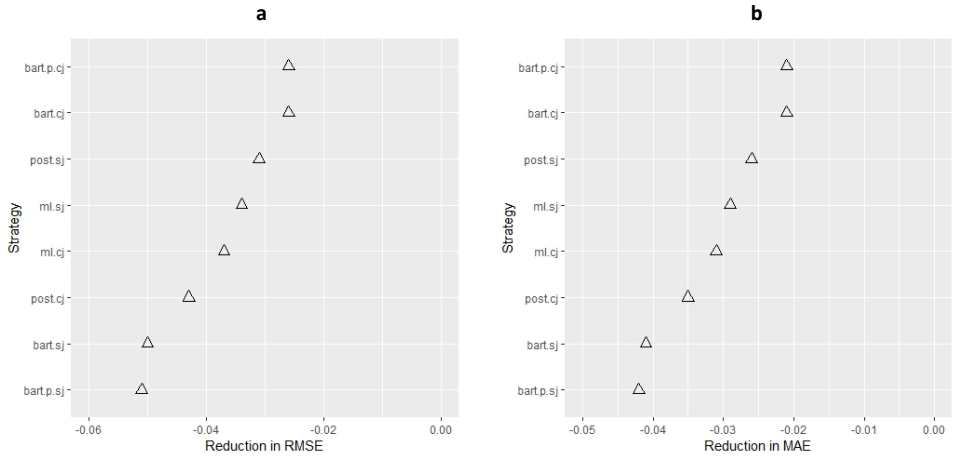

Figure 2. The Performance of Prediction Strategies

risk of over-fitting coming from the selection of regional predictors. But if researchers keep the set of regional predictors parsimonious, MrP is a viable option. A major advantage of classical MrP is that it is able to predict regional values for the complete case universe. Approaches involving the left-right measure are based on the synthetic joints created with the help of the QoG data. This introduces missing data and reduces the sample to 154 regions. We therefore recommend using classical MrP for applications requiring complete data. For the complete sample of 184 regions, the reduction in RMSE of the MrP model is 0.044.

For applications that can accept a smaller sample, synthetic BART is most promising. It automatically bases its predictions on the most relevant predictors, allowing researchers to provide it with the available data and let BART take it from there. The minimized researcher interference makes us put most confidence in BART to provide valid estimates in new, unseen settings. In addition, our synthetic BART model includes the general left-right measures, which we expect to be a useful predictor for many political preferences

besides EU trust. BART also accounts for deep interactions within the data, but only includes them if they improve fit (Montgomery and Olivella 2018). Still, a few tuning parameters of the BART model require researcher judgment. These, however, refer to a limited set of parameters. Also, we have not seen major differences in our predictions when varying them.

Application: Estimating regional social trust in Spain

The goal of our evaluation study above was to develop a general model that seems promising for predicting regional preferences from standard EB data across the EU. Since our strategies are evaluated on a large and heterogeneous set of European regions, they promise to provide reasonable estimates across Europe. Of course, single country analyses can be improved by using more detailed national census data or more meaningful regional predictors from national statistical offices. Yet, we still wanted to know how our preferred strategy - the synthetic BART model - performs in a new, unseen setting. Further, we want to demonstrate how our models can easily be applied to other EB data and single country analyses.

For this, we use the standard Eurobarometer 74.1 from 2010 to predict regional social trust in Spain. Social trust is measured by a dummy variable, coded as "1 - trust people" and "0 - do not trust people". The intra-regional correlation in social trust is 0.13. We chose social trust, because we can evaluate our predictions against a regionally representative measure of social trust from the QoG dataset. Spain is a good test case as it has a relatively large number of 16 NUTS2 regions. Moreover, disaggregated regional social trust is moderately high correlated with the true values (Pearson's Rho is 0.67). Yet, with a RMSE of 0.16 and a MAE of 0.13, the disaggregated regional means from the EB 74.1 are quite

noisy.

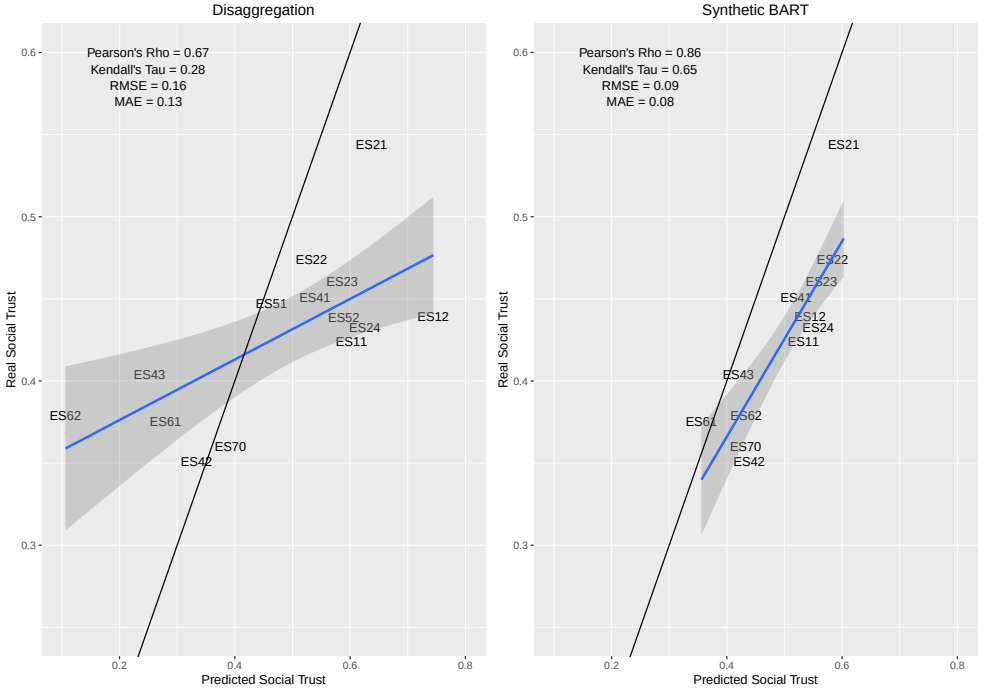

Figure 3. Predicting Social Trust across Spanish Regions

We simply apply our BART model with synthetic post-stratification to the Spanish EB data. The results are plotted in Figure 3. The plot on the left summarizes the performance of disaggregation, which - again - serves as a benchmark. We see a moderately high correlation wit a Pearson's Rho of 0.67. However, the rank correlation reported by Kendall's Tau is much weaker. Further, RMSE and MAE suggest that disaggregation provides very noisy estimates.

The plot on the right presents the results for the BART model with synthetic post-stratification. Visually, we already see that the linear regression line moves much closer to the 45 degree line, indicating a good fit. Indeed, Pearson's Rho increases to 0.86 and

Kendall's Tau rises to 0.65. RMSE drops by -0.07 and MAE by -0.05. Overall, BART somewhat over-predicts regional social trust, as most predictions deviate to the right of the 45 degree line. However, the BART predictions align much closer with the true values and reductions in RMSE and MAE are around 40 percent. Also note that the rank order correlation improves substantially. The estimation procedure was therefore very successful in correcting the order of regions into less and more trusting areas. Such relative differences can be of substantive interest for social scientist and the disaggregated values have been highly misleading here.

The results of our short application gives us confidence that the synthetic BART model offers great potential to improve regional estimates from standard EB data. We have also compared the performance of synthetic BART to predictions from the classical MrP model. Figure A2 in the Appendix presents the results of MrP estimates for regional social trust in Spain. Similar to our evaluation results, MrP estimates provide substantial gains as well. Reductions in MAE and RMSE are equally large. Yet, Person's Rho and Kendall's Tau improve much less. The rank order correlation is only at 0.42, compared to 0.65 under synthetic BART. This suggests that MrP has more problems in recovering the adequate sorting of social trust across regions. This demonstrates that MrP is a viable alternative, which, however, is outperformed by synthetic BART.

Discussion and outlook

This paper set out to evaluate the quality of regional information from standard Eurobarometer data. Across several datasets and for different variables, we have shown that regional EB data is very noisy. Therefore, we have implemented a number of strategies to improve regional estimates from standard EB surveys. Our evaluation study indicates

that BART with synthetic post-stratification performs best, substantially increasing the correspondence between regional estimates and the true values. In our analysis, most gains come from the inclusion of powerful individual-level predictors. Contrarily, regional predictors provide limited gains. This is not due to our selection of predictors. BART allows us to work with all seven regional predictors we could get from Eurostat. Rather, we suspect that the moderate level of intra-regional correlation (ICC=0.11) in EU trust limits the role of regional predictors. It is reasonable to expect that BART will perform even better with outcomes that are more strongly clustered spatially.

MrP with classical joints from the EU census also works well and allows researchers to estimate regional preferences for the complete case universe. We also show that MrsP has great potential, but increases the risk of over-fitting. Here, the selection of predictors remains a challenge, asking for careful judgment by researchers.

This paper develops and evaluates models for estimating regional preferences across the EU using Eurobarometer data. As the analysis demonstrated, a lot of choices in the modeling process are dictated by data availability. There is a limited set of variables in the EU census, the regionally representative surveys, and the EBs. And, of course, there are only a few variables that can be measured consistently across the different data sources. A contribution of our study is that it explores various possibilities for predicting regional preferences under these restrictive conditions.

A lot of the restrictions in our application comes from the fact that we attempt to estimate regional preferences across the whole EU. This is a highly relevant endeavor for comparative research. However, it should be noted that there is more potential in estimating regional preferences for single countries, either using national Eurobarometer samples or other national surveys. Researchers interested in regional preferences of a single country might get better predictions from using more detailed national census

data, other regional margins for synthetic joints, and different regional predictors from national sources. However, our short application to estimate regional social trust in Spain demonstrates that the synthetic BART model can already provide substantial gains.

Considering our results, one might ask whether the presented gains are 'good enough'? This is a difficult question. Our analysis makes clear that the raw EB data is sub-optimal and that we can improve data quality substantially using estimation procedures. The gains of our analysis are within the range of gains reported in previous evaluation studies (Hanretty, Lauderdale, and Vivyan 2016). Also, we have seen that every estimation strategy improved on disaggregation. So implementing MrP, MrsP, or BART will not harm and ideally brings some gains for the quality of regional estimates. The high level of noise in the EB data has two implications. First, it makes estimation procedures for regional preferences even more relevant. Here, simple multi-level modeling does not seem sufficient. Second, estimation results from MrP, MrsP, or BART models using EB data should come with a footnote on the remaining noise in the data.

Even though the strong heterogeneity in EB data remains a challenge, we are optimistic that scholars will be able to use the EB's rich regional information. We have not pushed the methods to their limits yet. Local smoothing strategies, for example, could provide additional gains (Selb and Munzert 2011). We hope that our analysis lays the ground for the estimation of regional preferences across the EU, enabling more disaggregated comparative research, as well as investigations of regional change across time. Scholars can use our replication file to estimate regional preferences for any standard EB, simply requiring some preprocessing of the respective EB data. Along with our code, we provide some convenience functions for the preprocessing, as well as the tables for classical and synthetic post-stratification.

References

Beramendi, Pablo. 2012. *The Political Geography of Inequality. Regions and Redistribution.* New York: Cambridge University Press.

Buttice, Matthew K., and Benjamin Highton. 2013. "How does multilevel regression and poststratification perform with conventional national surveys?" *Political Analysis* 21 (4): 449–467. ISSN: 14764989. doi:10.1093/pan/mpt017.

Chalmers, Adam William. 2013. "Regional authority, transnational lobbying and the allocation of structural funds in the European union." *Journal of Common Market Studies* 51 (5): 815–831. ISSN: 00219886. doi:10.1111/jcms.12038.

Chalmers, Adam William, and Lisa Maria Dellmuth. 2015. "Fiscal redistribution and public support for European integration." *European Union Politics* 16 (3): 386–407. ISSN: 1465-1165. doi:10.1177/1465116515581201. http://eup.sagepub.com/cgi/doi/10.1177/1465116515581201.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2012. "BART: Bayesian additive regression trees." *Annals of Applied Statistics* 6 (1): 266–298. ISSN: 19326157. doi:10.1214/09-AOAS285. arXiv: 0806.3286.

Dellmuth, Lisa Maria, and Adam William Chalmers. 2017. "All spending is not equal: European Union public spending, policy feedback, and citizens' support for the EU." *European Journal of Political Research* Forth.1–38.

Foster, Chase, and Jeffry Frieden. 2017. "Crisis of trust: Socio-economic determinants of Europeans' confidence in government." *European Union Politics* 18 (4): 511–535. ISSN: 17412757. doi:10.1177/1465116517723499.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups." *American Journal of Political Science* 57 (3): 762–776. ISSN: 15405907. doi:10.1111/ajps.12004.

Hanretty, Chris, Benjamin E. Lauderdale, and Nick Vivyan. 2016. "Comparing Strategies for Estimating Constituency Opinion from National Survey Samples." *Political Science Research and Methods:* 1–21. ISSN: 2049-8470. doi:10.1017/psrm.2015.79. http://www.journals.cambridge.org/abstract%7B%5C_%7DS2049847015000795.

Harteveld, Eelco, Tom van der Meer, and Catherine E. de Vries. 2013. "In Europe we trust? Exploring three logics of trust in the European Union." *European Union Politics* 14 (4): 542–565. ISSN: 14651165. doi:10.1177/1465116513491018. https://doi.org/10.1177/1465116513491018.

Hooghe, Liesbet, and Gary Marks. 2001. *Multi-Level Governance and European Integration.* Oxford: Rowman & Littlefield Publishers, INC.

Jones, K., R. J. Johnston, and C. J. Pattie. 1992. "People, Places and Regions: Exploring the Use of Multi-Level Modelling in the Analysis of Electoral Data." *British Journal of Political Science* 22 (3): 343–380. ISSN: 14692112. doi:10.1017/S0007123400006426.

Keating, Michael. 1997. "The invention of regions: political restructuring and territorial government in Western Europe" [in en]. *Environment and Planning C: Government and Policy* 15, no. 4 (December): 383–398. ISSN: 0263-774X. doi:10.1068/c150383. http://epc.sagepub.com/content/15/4/383.abstract.

Keating, Michael. 2008. "Thirty Years of Territorial Politics." *West European Politics* 31 (1-2): 60–81. ISSN: 0140-2382. doi:10.1080/01402380701833723.

Lax, Jeffrey R., and Justin H. Phillips. 2009. "How Should We Estimate Public Opinion." *American Journal of Political Science* 53 (1): 107–121. ISSN: 00925853. doi:10.1111/j.1540-5907.2008.00360.x.

Leemann, Lucas, and Fabio Wasserfallen. 2017. "Extending the Use and Prediction Precision of Subnational Public Opinion Estimation." *American Journal of Political Science* 00:1–20. ISSN: 00925853. doi:10.1111/ajps.12319.

Lubbers, Marcel, and Peer Scheepers. 2010. "Divergent trends of euroscepticism in countries and regions of the European Union." *European Journal of Political Research* 49 (6): 787–817. ISSN: 03044130. doi:10.1111/j.1475-6765.2010.01915.x.

Marks, Gary, Liesbet Hooghe, and Arjan H. Schakel. 2008. "Measuring regional authority." *Regional and Federal Studies* 18 (2-3): 111–121. ISSN: 13597566. doi:10.1080/13597560801979464.

Montgomery, Jacob M, and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 00 (00): 1–16. ISSN: 00925853. doi:10.1111/ajps.12361. http://dx.doi.org/10.7910/.

Muñoz, Jordi, Mariano Torcal, and Eduard Bonet. 2011. "Institutional trust and multilevel government in the European Union: Congruence or compensation?" *European Union Politics* 12 (4): 551–574. ISSN: 14651165. doi:10.1177/1465116511419250.

Pittau, M. Grazia, Roberto Zelli, and Andrew Gelman. 2010. "Economic disparities and life satisfaction in European regions." *Social Indicators Research* 96 (2): 339–361. ISSN: 03038300. doi:10.1007/s11205-009-9481-2.

Reilly, Cavan, Andrew Gelman, and Jonathan Katz. 2001. "Poststratification without population level information on the poststratifying variable with application to political polling." *Journal of the American Statistical Association* 96 (453): 1–11. ISSN: 1537274X. doi:10.1198/016214501750332640.

Rodden, Jonathan. 2010. "The Geographic Distribution of Political Preferences." *Annual Review of Political Science* 13 (1): 321–340. ISSN: 1094-2939. doi:10.1146/annurev.polisci.12.031607.092945. http://www.annualreviews.org/doi/10.1146/annurev.polisci.12.031607.092945.

Schraff, Dominik. 2017. "Regional redistribution and Eurosceptic voting." *Journal of European Public Policy* Forth.1–20. ISSN: 1350-1763. doi:10.1080/13501763.2017.1394901. https://doi.org/10.1080/13501763.2017.1394901.

Selb, Peter, and Simon Munzert. 2011. "Estimating constituency preferences from sparse survey data using auxiliary geographic information." *Political Analysis* 19 (4): 455–470. ISSN: 10471987. doi:10.1093/pan/mpr034.

Stockemer, Daniel. 2017. "The success of radical right-wing parties in Western European regions – new challenging findings." *Journal of Contemporary European Studies* 25 (1): 41–56. ISSN: 1478-2804. doi:10.1080/14782804.2016.1198691. https://www.tandfonline.com/doi/full/10.1080/14782804.2016.1198691.

Tatham, Michaël. 2012. "You do what you have to do? Salience and territorial interest representation in EU environmental affairs." *European Union Politics* 13 (3): 434–450. ISSN: 14651165. doi:10.1177/1465116512441505.

Teorell, Jan, Stefan Dahlberg, Sören Holmberg, Bo Rothstein, Anna Khomenko, and Richard Svensson. 2017. *The Quality of Government Standard Dataset.* Technical report. University of Gothenburg: The Quality of Government Institute.

Toshkov, Dimiter. 2015. "Exploring the performance of multilevel modeling and post-stratification with eurobarometer data." *Political Analysis* 23 (3): 455–460. ISSN: 14764989. doi:10.1093/pan/mpv009.

Warshaw, Christopher, and Jonathan Rodden. 2012. "How should We measure district-level public opinion on individual issues?" *Journal of Politics* 74 (1): 203–219. ISSN: 00223816. doi:10.1017/S0022381611001204.