It's a (Coarsened Exact) Match!

Non-Parametric Imputation of European Abstainers' Vote

Supplementary material

Damien Bol*

Marco Giani†

August 28, 2019

Abstract

There is a long tradition of imputation studies looking at how abstainers would vote if they had to. This is crucial for democracies because when abstainers and voters have different preferences, the electoral outcome ceases to reflect the will of the people. In this paper, we apply a non-parametric method to revisit old evidence. We impute the vote of abstainers in 15 European countries using Coarsened Exact Matching (CEM). While traditional imputation methods rely on the choice of voters that are on average like abstainers, and simulate full turnout, CEM only imputes the vote of the abstainers that are similar to voters, and allows to simulate an electoral outcome under varying levels of turnout, including levels that credibly simulate compulsory voting. We find that higher turnout would benefit social democratic parties while imposing substantial losses to extreme left and green parties.

Keywords: Turnout, Elections, Imputation, Coarsened Exact Matching, Western Europe

*Department of Political Economy, King's College London, damien.bol@kcl.ac.uk.

†Department of Political Economy, King's College London, marco.giani@kcl.ac.uk.

Appendix A1: Data

This paper relies on data from the European Election Study (ESS). The ESS regularly conducts face-to-face surveys on representative national samples in European countries. For the sake of comparability, we restrict our analysis to the 15 countries that were members of the European Union before the enlargement of 2004. In each of them, we analyze the two latest national elections available in the data until the release of the 7th round of the ESS. In total, we analyze 30 elections, two per country. It is important to mention that we only include national elections because: (1) the ESS lacks of systematic data for other elections, and (2) electoral behavior at the regional and European level tends to follow different logic due to the second-order nature of the elections. Also, as to minimize memory issues, we only analyze the surveys collected right after to each national election. In total, our analysis includes $56,037$ respondents, among which $11,137$ report an abstention.[1] The table below provides detailed information about each of the election that we use in the analysis. We report, for each country belonging to EU15, the year of the two elections that we analyze. For each of this election, we use the closest available round of the ESS. We also report, for each election, the real turnout and the sample turnout.

	1st Election				2nd Election			
	Year	N Obs.	Real	Sample	Year	N Obs.	Real	Sample
Austria (**AT**)	2008	2115	78.8	76.6	2013	1678	74.9	76.8
Belgium (**BE**)	2007	1560	89.3	91.9	2014	1585	88.5	90.3
Germany (**DE**)	2008	2734	70.8	81.0	2013	2813	71.5	83.3
Denmark (**DK**)	2007	1506	86.5	94.2	2011	1498	87.7	93.8
Spain (**ES**)	2008	2295	73.9	80.3	2011	1782	68.9	75.5
Finland (**FI**)	2007	2034	67.9	78.8	2011	2044	70.4	84.5
France (**FR**)	2007	1867	60.2	76.6	2012	1707	57.2	67.9
United Kingdom (**GB**)	2005	2275	61.4	72.0	2010	2324	65.1	71.9
Greece (**GR**)	2007	1947	74.1	86.9	2009	2528	70.9	78.4
Ireland (**IE**)	2007	1633	67.0	79.5	2011	2462	70.0	73.8
Italy (**IT**)	2001	1159	81.4	88.9	2013	933	75.2	79.5
Luxembourg(**LU**)	1999	1249	86.5	67.3	2004	1384	91.4	75.9
Netherlands (**NE**)	2010	1761	74.7	83.2	2012	1785	74.3	83.5
Portugal (**PT**)	2009	2053	59.7	74.2	2011	2010	58.1	68.2
Sweden (**SE**)	2010	1894	84.6	90.8	2014	1677	85.8	91.5

[1] We also exclude non eligible voters.

One advantage of the ESS is that interviewers make strong efforts not to have a sample skewed towards politically-interested individuals. Politically-interested individuals and voters are often over-represented in surveys that are specifically about political issues like the American National Election Study. The table above shows that self-reported sample turnout follows closely actual turnout. Although the reported turnout rates are often higher than the actual turnout rates, the differences between the two are small. Even in Belgium that uses compulsory voting, the proportion of abstainers in the survey is substantial, and very much in line with the proportion of abstainers in reality. All in all, the issue of turnout over-reporting is not severe in our sample.

Looking at each country separately, we observe that the sample turnout often exceeds the real one. This excess sample turnout, however, differs from country to country. In particular, the latter is negligible in Belgium and Sweden. In Austria, instead, sample turnout is lower than real turnout, though the two are very close. Note that Luxembourg is a strong outlier, as the actual turnout rate is higher than the one reported in the survey. This probably due to the high proportion of non-citizens living in the country. Note that Luxembourg is a strong outlier, as the actual turnout rate is higher than the one reported in the survey. This is probably due to the high proportion of non-citizens living in the country.

Appendix A2: Descriptive statistics of covariates

The table below summarizes the descriptive statistics of covariates. For the purpose of evaluating imbalance, we provide separate descriptive statistics for voters and abstainers. From a demographic perspective, abstainers are on average younger, more likely to be women and to have ethnic minority background. From a socioeconomic perspective, abstainers are poorer, less educated, more likely to be unemployed, and feel less secure economically speaking. Finally, abstainers' level of political interest is lower.

	Voters			Abstainers		
	N	mean	sd	*N*	mean	sd
Education	44703	2.09	1.40	11020	1.69	1.31
Age	44798	51.62	17.34	11097	43.86	18.83
Sex	44888	1.53	0.50	11135	1.55	0.50
Children living at home	44871	1.63	0.48	11122	1.65	0.48
Belong to minority ethnic group	44598	1.97	0.16	10975	1.94	0.24
Feeling about income	44664	1.88	0.83	10984	2.19	0.88
Ever unemployed	44736	1.74	0.44	11062	1.65	0.48
Wage	44900	0.54	0.50	11137	0.57	0.49
Pension	44900	0.28	0.45	11137	0.20	0.40
Self-employed	44900	0.05	0.21	11137	0.11	0.32
Others sources	44426	0.00	0.00	10918	0.00	0.00
How interested in politics	44796	2.44	0.91	11081	3.01	0.89

Appendix A3: Classification of parties

In the analysis, we only include parties that have a parliamentary representation. The *social democratic* parties are: SPÖ (Austria), PS and SPa (Belgium), SD (Denmark), SDP (Finland), PS (France), SPD (Germany), Pasok (Greece), Labour (Ireland), PD (Italy), LSAP (Luxembourg), PvdA (the Netherlands), PS (Portugal), PSOE (Spain), SSDP (Sweden), and Labour (Great Britain).

The *extreme left* parties included in the analysis are: PTB (Belgium), SF (Denmark), VAS (Finland), PCF and FG (France), Linke (Germany), KKE and Syriza (Greece), Rifondazione Comunista (Italy), Déi Lénk (Luxembourg), SP (the Netherlands), BE (Portugal), IU (Spain), V (Sweden).

The *green* parties are: Die Grunen (Austria), Ecolo and Groen (Belgium), Enhlø(Denmark), VIHR (Finland), EELV (France), Die Grunen (Germany), Green Party (Ireland), Girasole (Italy), Déi Gréng (Luxembourg), GL (the Netherlands), MP (Sweden), Green Party (Great Britain).

Appendix A4: Statistical significance

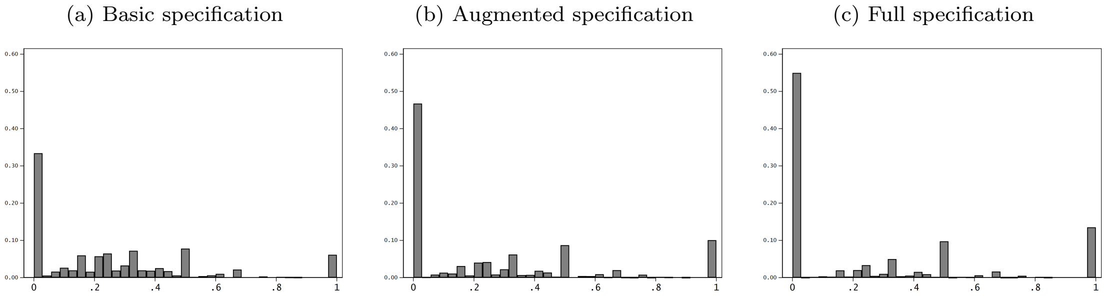
In the table below, we test whether the electoral behavior of marginal voters (matched voters/abstainers) differs significantly from those of certain voters (unmatched voters), using a standard two sided $z-$test. The reported coefficients refer to the difference between the score of each party in the two groups. For instance, the number 0.018 in the first entry means that the score of the social democratic party among marginal voters is higher than the one among certain voters by 1.8%-points.

	Basic	Augmented	Full
Δ Social democratic parties			
Matched - Unmatched	.018***	.019***	.022***
Standard error	(.004)	(.004)	(.005)
Δ Extreme left parties			
Matched - Unmatched	-.019***	-.014***	-.017***
Standard error	(.002)	(.002)	(.002)
Δ Green parties			
Matched - Unmatched	-.015***	-.012***	-.017***
Standard error	(.002)	(.002)	(.002)

Covariates in Basic specification: education (1-5), age (15-110), gender (0-1), household status (0-1), minority status (0-1), feeling of income insecurity (1-4). Augmented: add source of income (categorical variable), and unemployment (0-1). Full: add political interest (1-4). For CEM, we match units within each election. Age is coarsened according to standard age categories, with intervals of 10 years. We require exact matching on all dummy and categorical variables including country/election. For income and education, coarsening is based on the Scott-break algorithm provided by CEM. For parametric imputation, we use binary logit regressions with the same covariates than the basic, augmented and full specification, including country/election fixed effects. *** $p < .01$.

Appendix A5: Number of covariates and validity

We calculate, in each stratum of matched voters/abstainers, the proportion of votes for social democratic parties (see Figure A6).[2] The proportions are either very close to 0 or 1. This means that matching on these covariates strongly discriminates supporters of these parties. Of all strata with observations, 42.9% are composed of either exactly 0 or 100% of social-democratic voters when we use the basic matching specification. Also, we show that the strata are increasingly homogeneous in vote shares depending on the number of covariates (from basic to full specification). In other words, the more covariates included, the more valid the imputation is. With the full specification, 67.7% of all strata are composed of either 0 or 100% of social democratic voters.


(a) Basic specification

(b) Augmented specification

(c) Full specification

[2] The results are similar for extreme left and green parties. However, since there are many less voters for these two groups of parties the figure is not as easy to read.

Appendix A6: Out-of-sample validation

In the table below, we show the results of some out-of sample validation test for the full specification. We focus on the respondents for which we have information regarding voting choice (i.e., we exclude abstainers), and evaluate how accurately CEM imputes a voting choice for them. To do so, we randomly split voters into two groups: a test group ($N = 3,000$) and a training group ($N = 56,037 - 3,000 = 53,037$). We remove the voting choice of the test group and treat it as a group of abstainers. We then perform the CEM imputation described in the main text, and compare the imputed voting choice of individuals of the test group to their actual voting choice.[3] In addition, we perform the same test using the two the standard parametric imputation methods described in the main text. As to assess model dependency, we also perform another parametric imputation. In the column *quadratic*, we do the same parametric imputation than for the column logit, except that we add a squared term to each continuous and discrete covariates.

	Real	CEM	Logit	Logit+	Quadratic
% Social democratic parties					
Test group complete random	**24.6**	26.7	25.5	25.5	25.4
Test group random within low SES	**28.1**	29.6	29.1	28.9	28.2
% Extreme left parties					
Test group complete random	**6.1**	4.8	7.7	7.7	7.7
Test group random within low SES	**6.6**	3.7	9.1	9.1	8.5
% Green parties					
Test group complete random	**4.9**	4.6	6.7	6.7	6.7
Test group random within low SES	**3.9**	1.9	4.2	4.5	4.1

Entries are real voting choices, or CEM and parametric imputation, in the test sample (N = 3,000). Based on the full specification, including education, age, gender, household status, minority status, feeling of income insecurity, source of income, unemployment, and political interest. The *logit*, and *logit+* methods are the same than in the Table 1 in the main text. The *quadratic* method adds squared terms for each continuous and discrete covariates. We repeat the analysis twice: one for a test group selected at complete random, and another with a test group selected at random within individuals with a low socio-economic status (SES).

We reproduce the out-of-sample validation tests twice: one in which the test group is selected at complete random out of the entire sample, and one in which it is selected randomly within the poorest and less educated of the sample ($> 2X$ on the variable income insecurity, and $< 4X$ on the variable education). The reason is that in reality voters and non-voters differ on many covariates. As we show in Figure 1 in the main text, the poorest and less educated are less likely to vote. Hence, excluding 3,000 voters at complete random is not representative of a real

[3] Note that we compare the simulation to the actual electoral behavior of *all* individuals of the test group, not only those who have an imputation.

situation. Excluding 3,000 voters in a group of individuals who are the least likely to vote among voters is more realistic.

From the table above, we observe that the difference between real and imputed vote shares is relatively small, especially for the social democratic parties.

Appendix A7: Potential omitted variable bias

In the table below, we replicate the main analysis presented in the main text (Table 1) by progressively adding further covariates that are known to affect electoral behavior: the amount of social capital of individuals, their level of institutional trust, and their self-reported ideology. We observe that that the results are very similar to those of the 'full specification' of Table 1 in the main text. This suggests that the full specification as reported in the main text already includes most of the important determinants of electoral behavior.

	Social	Institutional	Ideological
	capital	trust	placement
% Social democratic parties			
Voters	25.1		
Abstainers (CEM)	25.7	28.2	26.4
Abstainers (logit)	25.8	25.5	27.3
% Extreme left parties			
Voters	5.5		
Abstainers (CEM)	3.8	4.5	2.2
Abstainers (logit)	6.8	7.1	7.9
% Green parties			
Voters	5.5		
Abstainers (CEM)	3.9	3.9	2.9
Abstainers (logit)	6.0	6.1	6.6
% Turnout			
Sample Turnout	80.4		
Compulsory (CEM)	84.8	83.1	83.4
Compulsory (logit)	100	100	100

In each analysis, we include: education, age, gender, household status, minority status, feeling of income insecurity, source of income, unemployment, and political interest. In the first column, we also include social capital, proxied by the survey item "how often you socially meet". In the second column, we include Institutional trust, proxied by the self-reported level of trust in parliament. In the last column, we include ideological placement, proxied by the self-reported left-right position on a 0-10 scale.

Appendix A8: Further matching and regression analyses

We compare imputation outcomes of the main analysis (see Table 1 in the main text) with those of other advanced methods. We use an alternative matching method (kernel matching), and an alternative regression method (kernel regression). Kernel matching works as a two-step procedure. Firstly, using the basic, augmented or full set of covariates, it predicts turnout propensity scores. Then, it uses on the propensity scores of actual voters to match them with abstainers, and hence predict their voting choice. Kernel regression works like a regression, in which the function of the basic, augmented, or full set of covariates is decided inductively as to fit the data.

Interestingly, we observe in the table below that kernel matching gives results similar to CEM (score increase for social democratic parties, lower score for green and extreme left parties), and that kernel regression gives results similar to the logit regression (similar for social democratic parties, slight increase for green and extreme left parties).This further analysis proves important to establish the robustness of our main analysis.

	Basic	Augmented	Full
% Social democratic parties			
Voters	25.1		
Abstainers (Kernel matching)	26.8	27.0	26.4
Abstainers (Kernel regression)	25.0	25.4	25.3
% Extreme left parties			
Voters	5.5		
Abstainers (Kernel matching)	6.4	6.8	4.5
Abstainers (Kernel regression)	5.3	5.2	5.2
% Green parties			
Voters	5.5		
Abstainers (Kernel matching)	5.4	5.6	4.5
Abstainers (Kernel regression)	6.0	6.0	6.0

Appendix A9: Outcomes by election

In the table below, we show the results of the main analysis (full specification) broken down by election using both CEM and logit. Specifically, we show the results for all left parties together (social democratic, extreme-left, and green parties), and for all incumbent parties. Echoing the result of the analysis in the main text (Table 1), the score of left and incumbent parties are relatively unaffected when we simulate how abstainers would have voted using a standard parametric method. However, we see greater discrepancies with CEM. It falls beyond the scope of this paper to explain between-countries heterogeneity.

Country	Year	All-Left			Incumbent			year	All-Left			Incumbent		
		Real	Matching	Logit	Real	Matching	Logit		Real	Matching	Logit	Real	Matching	Logit
AT	2008	39.7	37.6	40.6	55.2	48.4	45.9	2013	39.2	42.3	39.1	50.8	55.2	49.3
BE	2007	30.2	38.3	37.8	45.4	45.2	42.4	2014	32.8	40.5	41.2	46.6	63.8	54.4
DE	2008	45.6	40.9	45.0	72.3	52.9	48.0	2013	42.7	41.2	46.2	49.3	40.8	35.3
DK	2007	40.7	42.2	42.1	36.6	35.3	33.7	2011	40.2	48.1	43.8	33.9	24.7	27.3
ES	2008	48.2	43.8	43.5	44.4	41.9	40.1	2011	32.3	29.7	31.6	25.9	27.3	22.5
FI	2007	38.7	33.3	34.4	49.1	42.4	40.1	2011	34.5	31.5	32.2	47.8	41.8	46.2
FR	2007	29.2	34.4	40.3	45.9	29.8	26.7	2012	37.6	38.7	40.9	27.1	29.4	23.9
GB	2005	35.2	39.8	39.9	35.2	39.8	37.1	2010	29.0	31.9	29.6	29.0	31.9	26.4
GR	2007	51.3	46.7	47.2	41.8	25.8	22.4	2009	56.1	37.7	41.2	33.5	30.1	29.5
IE	2007	14.8	10.4	12.2	44.3	43.0	37.0	2011	19.5	14.6	16.6	17.5	17.1	16.7
IT	2001	25.4	18.8	22.1	34.9	22.5	26.9	2013	26.3	30.0	32.0	27.9	10.0	14.2
LU	1999	36.6	23.7	27.2	53.5	44.6	44.2	2004	37.0	24.6	25.1	40.4	54.6	49.5
NL	2010	36.1	28.4	35.4	41.1	40.1	32.0	2012	36.8	35.1	36.1	35.1	36.2	32.0
PT	2009	47.9	34.9	32.4	37.7	33.0	27.0	2011	34.6	28.1	29.7	29.2	25.8	23.9
SE	2010	43.6	34.4	38.9	49.3	48.6	45.0	2014	43.6	41.7	37.5	39.4	39.4	41.3