Political Science Research and Methods (2021), **9**, 445–450
doi:10.1017/psrm.2019.44

RESEARCH NOTE

It's a (coarsened exact) match! Non-parametric imputation of European abstainers' vote

Damien Bol* and Marco Giani

Department of Political Economy, Quantitative Political Economy Group, King's College London, London, United Kingdom
*Corresponding author. E-mail: damien.bol@kcl.ac.uk

(Received 14 November 2018; revised 19 March 2019; accepted 15 May 2019; first published online 30 October 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

Abstention can lead to a major problem of political representation. If abstainers and voters do not have the same preferences, the electoral outcome ceases to represent the will of the people, and only represents the will of the voters. This well-known problem has pushed prominent political scientists to make big claims in favor of compulsory voting (Lijphart, 1997). Many studies evaluate the extent to which abstention constitutes a normative issue for political representation. They usually rely on survey analysis in order to infer the likely voting choice of abstainers. In particular, they estimate regression models predicting the party choice of voters based on a set of covariates available in the surveys, and reconstruct the likely choice of abstainers on the basis of the model's prediction (e.g., Bernhagen and Marsh, 2007). Hereafter, we call this strategy a *standard parametric imputation*.

In this paper, we apply a new non-parametric method in order to revisit old evidence. We use Coarsened Exact Matching (CEM) and survey data from 15 European democracies and 30 elections between 1998 and 2014 in order to match abstainers with voters. The voting choice of abstainers is then imputed from the voting choice of the voters with whom they are matched.[1] In line with the conventional wisdom regarding abstainers' political preference, we look at how the score of left-wing parties would change if abstainers voted, in distinguishing between social democratic, extreme left and green parties.

Using CEM as an imputation strategy has two key advantages. First, CEM uses an exact matching algorithm. Therefore, it imputes the voting choice of abstainers from the voting choice of voters that are *exactly* equal to them on a set of (possibly coarsened) covariates, rather than imputing from the voting choice of voters that are equal to them *on average* (like with standard

[1]The identification strategy is logically equivalent to comparing observations randomly assigned to a treatment (here: voting) in an experiment. For another study using similar identification strategies, though in other contexts and using different matching methods, see Dehejia and Wahba (2002).

parametric imputation). This is an important advantage in terms of internal validity, as the traditional determinants of electoral behavior predict voting choices not only directly, but also in interactions with each other. Second, CEM permits to simulate compulsory voting under varying levels of turnout, whereas standard parametric imputation automatically simulates full turnout. This is another important advantage, since, in reality, turnout is never full, as some individuals always decide to abstain. In Belgium, where compulsory has been used for more than a century, turnout varies between 88 and 95 percent. Moreover, studies show that citizens with a lower level of education attainment tend to vote less, even in countries where voting is compulsory (Katz and Levin, 2018). In other words, several covariates continue to predict turnout even in these situations. Simulating what would the electoral outcome under a high level of turnout that credibly approaches compulsory voting, instead of full turnout, is thus empirically relevant. Along this line, some recent studies exploit the historical abolishing of compulsory voting in a few European countries to estimate what would have been the voting choice of abstainers (Fewerda, 2014; Bechtel *et al.*, 2015; Miller and Dassonneville, 2016).[2] Our simulation is similar to them in spirit, as we are aiming for a similar counterfactual electoral outcome.

Most post-2000 standard parametric imputation studies find that, in Europe and the United States, abstainers and voters have similar political preferences, and thus conclude that abstention does not threaten political representation (Citrin *et al.*, 2003; Brunell and DiNardo, 2004; Bernhagen and Marsh, 2007).[3] Instead, we find that the imputed voting choice of abstainers substantially differs from the one of voters. Social democratic parties would gain from higher turnout, whereas extreme left and green parties would be worse off. Hence, our result implies that left-wing partisan preferences are differently misrepresented in final electoral outcomes. Counterintuitively, it is the moderate left, rather than the extreme one, that pays the price of low turnout.

1. Method

In this paper, we revisit existing evidence with the help of a new method. We use the well-known European Social Survey (ESS) dataset. For a description, see Appendix A1. Our goal is to impute the voting choice of abstainers, and compare it with the one of voters. We do so by relying on CEM, and follow the step-by-step procedure of Iacus *et al.* (2012). CEM can be used within an observational study to mimic experimental methods. The basic idea is simple and powerful: matching untreated observations that are exactly similar to treated observations on relevant covariates, thus forming strata of observations, and calculating the average treatment effect across these strata. In this paper, we use CEM as an imputation tool.

The first step is to select a set of covariates. We select socioeconomic variables and political interest that are known to be strong predictors of electoral behavior (Blais, 2000). We organize them into three matching specifications. In the basic specification, we match voters and abstainers according to their age, gender, ethnic group, household status, highest education attainment, and the subjective feeling of income insecurity.[4] In an augmented specification, we add the main source of income of the household and unemployment status. We acknowledge that the socioeconomic status is not the only determinant of electoral behavior. Consistently, in the full specification, we add self-reported political interest.[5] In the full specification, the combination of

[2]Note that these studies find effects pointing at different directions, compulsory voting giving an advantage to left-wing or right-wing parties.

[3]Note that Brunell and DiNardo (2004) use a propensity-score matching method to impute the voting choice of abstainers. Although based on a matching algorithm like CEM, this method is parametric and functions like a standard parametric imputation.

[4]The ESS also provides information about objective income decile, but this variable has many more missing data.

[5]We acknowledge that electoral behavior has other determinants, like personal and professional networks (Bond *et al.*, 2012). In Appendix A7, we report additional simulations using extra covariates like social capital and institutional trust. The results of these additional simulations are very similar to those reported in the main text, suggesting that the set of

covariates gives rise to 32,584 strata, 3,167 of which include observations. Details about the covariates can be found in Appendix A2, where we show that abstainers are systematically different from voters on most of the selected covariates. They are less educated, younger, more likely to be unemployed, more likely to be from a minority background, feel less secured economically, and are less interested in politics.

The second step is to match, for each election and each country, voters and abstainers that have equal scores on the selected categorical covariates and similar values on the selected continuous covariates, in the spirit of CEM. Concretely, consider the basic matching specification presented above. Abstainers are matched with voters with equal gender, ethnic background, and household status. Matching requires abstainers and voters to also share a close value on continuous covariates. Age is coarsened by working age categories (15–24; 25–34…). Education attainment and subjective feeling of income insecurity, which range respectively from 1 to 5 and 1 to 4, are coarsened using the Scott–Break method that maximizes the trade-off between homogeneous and sufficiently populated strata (Blackwell *et al.*, 2009).

In applying CEM, we identify three groups of individuals: (1) the *certain voters* whose covariates are such that they are not matched with any of the abstainers, (2) the *certain abstainers* whose covariates are such that they are not matched with any of the voters, and (3) the *marginal voters*.[6] This last group is of key importance for our analysis: it includes voters and abstainers who share equal, or close to equal in the case of coarsened continuous variables, scores on the relevant covariates. We think of marginal voters as those who sometimes vote, sometimes not. As such, we are treating their turnout decision as conditionally random: in a world where electoral behavior is determined by these covariates, marginal voters have equal *ex-ante* turnout probability.

In Figure 1, we show the distribution of subjective feeling of income insecurity of marginal voters (i.e., matched individuals) for each matching specification (basic, augmented, and full). For the sake of comparison, we also show the distribution of these variables for the entire sample. It reveals that marginal voters always have a lower income security and education level than the entire sample. Further, it also shows that this difference increases when we match on more covariates (from basic to full specifications). In other words, Figure 1 shows that the profile of marginal voters become closer to the one of abstainers when we add enough covariates.

The third step is to impute the voting choice of abstainers based on the voting choice of their matches.[7] Concretely, the imputation looks like this: suppose that we match one abstainer and three voters, two voting for Party A, and one voting for Party B; we impute the vote choice of the abstainer as being Party A with a probability of 2/3, and Party B with a probability of 1/3. In turn, this means that the vote share of each party remains the same within each stratum of marginal voters. We thus have a vote distribution for abstainers that we can compare to one ofthe voters. In an attempt to keep results synthetic, we show the difference of vote distribution between voters and abstainers for three groups of parties: social democratic, extreme left, and green parties. The classification of the parties in the countries covered in this paper can be found in Appendix A3.

In total, we are able to match between 62.0 and 81.9 percent of the abstainers, depending on the specification. It is important to note that the voting choice of certain abstainers (i.e., unmatched abstainers) is not imputed. These certain abstainers can be seen as individuals who would not vote even if voting was compulsory. Related to this, the proportion of imputed observations is almost nil in countries where voting is already compulsory (Belgium and Luxembourg)

covariates of the full specification already contains most important determinants of electoral behavior, and that the omitted variable bias is limited. Note that it is up to the researcher to decide upon the number of covariates to include in a simulation using CEM knowing that there is a trade-off between accuracy and the simulated turnout rate. As the number of covariates increases, the number of individuals that CEM matches decreases, and so does the level of simulated turnout, see below.

[6]We borrow the expression "marginal voter" to Fowler (2015). In this paper, marginal voters are those who sometimes vote, sometimes not, depending on the circumstance (e.g., if it rains).

[7]Note that we do not restrict matching to be one-to-one: abstainers can be matched with more than one voter, and *vice-versa*. Also we do not match individuals from different countries or elections. Only nationals of the same country are matched together.

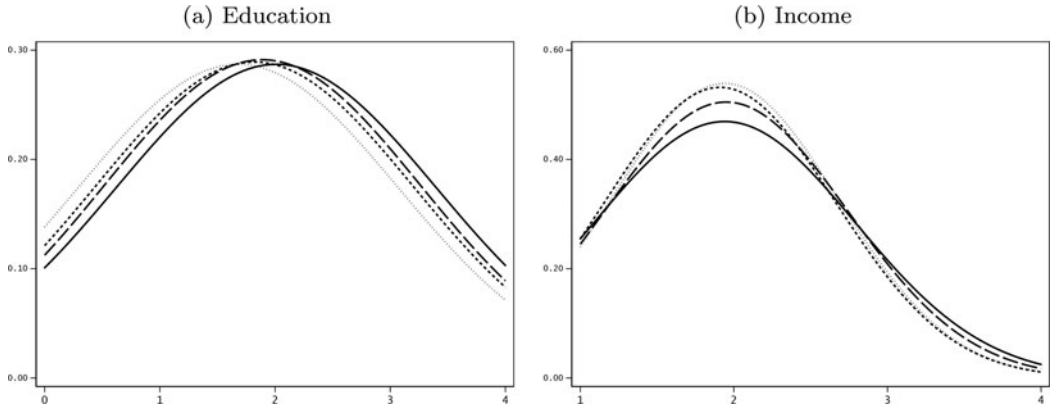

Figure 1. Imbalance in education and income, per matching specification.
Note: The lines represent the distribution of the variables education for the entire sample (solid line), for individuals matched individuals per the basic specification (long-dashed), augmented specification (dashed), and full specification (dotted).

under the full matching specification (3 and 7 percent respectively). The size of the group of certain abstainers depends on the number of covariates upon which we require matching. CEM is a flexible method that permits to set different matching specifications that, in turn, yield different levels of simulated turnout. The more covariates we use for matching voters with abstainers, the lower the number of matched abstainers. However, it is important to keep in mind that when the number of covariates is small, the strata of matched individuals are less homogeneous, which threatens the validity of the imputation (see below). Finally, the higher the number of covariates, the stronger the reduction in imbalance obtained through CEM.[8]

2. Results

Table 1 summarizes our results. We compare the party scores among all voters in the survey with the party scores among matched abstainers, across the three matching specifications. In addition, we contrast our results with those obtained when we impute the vote of abstainers from two standard parametric methods. For the first one (*logit*), we estimate a logit regression predicting the choice of voters for social democratic, extreme left, and green parties with the same covariates than in the three matching specifications. Then, we simulate the vote of abstainers using the predicted probabilities as given by the coefficient estimates. For the second one (*logit+*), we estimate two logit regressions. We first predict turnout with the specified covariates (basic, augmented, and full), and then prune abstainers that have low predicted turnout probability (<50 percent). In the second logit regression, we predict voting choice with the same covariates, just like we do in the *logit* estimation, but in excluding the abstainers that we prune after the first regression. This second parametric imputation, inspired by the two-step procedure of Ho *et al.* (2007), is an interesting benchmark given that, similar to CEM, it only imputes the voting choice of abstainers that have a reasonable chances of voting.[9] Unsurprisingly, simulations based on the *logit+* method are half way between CEM and *logit*.

First, we find evidence that abstainers would be more supportive of the social democratic parties in the 30 elections covered in the data using CEM. The difference between voters and abstainers goes from 0.6 to 1.7 percent-points in the full specification. Interestingly, the exact

[8]CEM provides a measure of multivariate imbalance which is the absolute difference over all the cell values. In our data, it goes from 0.95 (0.77) before matching to 0.66 (0.63) after matching in the full (basic) specification.

[9]A series of Hosmer–Lemeshow's goodness-of-fit tests reveals that, without surprise, the augmented and full specification are always better than the basic one. More importantly, it reveals that the second regression of the *logit+* is most of the time better than the one of the *logit* method (regardless of the party). This suggests that the *logit+* method is superior.

Table 1. Voting choice of voters and abstainers

	Basic	Augmented	Full
% Social democratic parties			
Voters	25.1		
Abstainers (CEM)	25.7	26.2	26.8
Abstainers (logit)	26.0	26.4	25.8
Abstainers (logit+)	25.8	26.0	26.2
% Extreme left parties			
Voters	5.5		
Abstainers (CEM)	4.9	4.7	4.2
Abstainers (logit)	6.8	7.3	6.7
Abstainers (logit+)	6.6	6.8	5.7
% Green parties			
Voters	5.5		
Abstainers (CEM)	5.0	4.9	4.2
Abstainers (logit)	6.3	6.4	6.0
Abstainers (logit+)	6.1	6.1	5.3
% Turnout			
Sample turnout	80.1		
Compulsory (CEM)	96.4	92.4	88.3
Compulsory (logit)	100	100	100
Compulsory (logit+)	96.4	92.4	88.3

Note: 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 country/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.

opposite happens for extreme left and green parties. Abstainers would vote substantially less for these parties (from 0.5 to 1.3 percent-points). Higher turnout would thus increase the score of social democratic parties, at the expenses of other left-wing parties. In Appendix A4, we show that these differences are statistically significant at a level of $p<0.01$. By contrast, when we impute the voting choice of abstainers with the standard parametric imputation and full specification, we find that, consistently with the literature, abstainers are almost as likely to support a social democratic party than voters. Also, Table 1 reveals that the results for the *logit+* method gives different results than CEM, which suggests that the difference is not only due to the selective imputation on abstainers that are likely to vote, typically in a situation of compulsory voting.

Second, we observe differences between matching specifications. The more covariates we use to match voters and abstainers, the larger the difference in party scores. This result holds true for every party. Intuitively, this means that increasing the number of covariates makes the group of marginal voters smaller, and more homogeneous, which decreases the simulated turnout rate, from 88.3 percent with the full specification to 96.4 percent with the basic one. This is an important advantage as the researcher can construct a simulation with varying turnout levels. For example, the turnout rate with the full specification can be seen as credible simulation of compulsory voting. In contrast, the simulated turnout is always the same when we use the parametric imputation method (*logit*). It is automatically 100 percent (full turnout).

Why does CEM give different results than parametric imputation? CEM is a fully nonparametric method, and therefore is less sensitive to model selection, and to the choice of a function for the covariates. Parametric imputation can lead to inaccurate predictions when the imputed observations have extreme values on covariates (King and Zeng, 2006). To gauge this, we conduct several validation tests in the appendix. First, we show that the validity of the method increases with the number of covariates (A5). The strata are homogeneous in votes shares in the basic specification, but are even more so in the full specification. Second, we also perform some out-of-sample validation tests (A6). We compare CEM to the first (*logit*) and second parametric imputation methods described above (*logit+*), as well as a third one in which we use alternative

functions for the covariates (i.e., second-order polynomials). We find that CEM is as good as other methods at recollecting the known voting choice of a test sample. Finally, in Appendix A8, we show the results of our main analysis using alternative advanced imputation methods. We use kernel matching and kernel regression. This further analysis proves important to establish the robustness of our results: kernel matching gives results similar to CEM, and kernel regression gives results similar to standard parametric imputation. The imputation outcomes obtained with CEM may hence be generalized to other non-parametric matching methods.

3. Conclusion

Our contribution is threefold. First, from a normative point of view, we find that representation is problematic in European democracies: abstainers would vote differently than voters. This means that, without compulsory voting, the electoral outcome does not adequately represent the will of the people. Second, from a positive point of view, we document that compulsory voting does not change the overall score of left-wing parties, but would affect its composition, favoring social democratic parties at the expenses of extreme left and green parties. Third, from a methodological point of view, we demonstrate how CEM, and more generally non-parametric methods, can be used to impute missing values. To our knowledge, our study is the first one to use CEM for that purpose. This method is particularly appealing when missing data are, themselves, at the center of the research question, like in this paper.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/psrm.2019.44

References

Bechtel MM, Handgartner D and Schmid L (2015) Does compulsory voting increase support for leftist policy? *American Journal of Political Science* **60**, 752–767.

Bernhagen P and Marsh M (2007) The partisan effects of low turnout: analyzing vote abstention as a missing data problem. *Electoral Studies* **26**, 548–560.

Blackwell M, Iacus S, King G and Porro G (2009) CEM: coarsened exact matching in stata. *The Stata Journal* **9**, 524–546.

Blais A (2000) *To Vote or Not to Vote: The Merits and Limits of Rational Choice Theory*. Pittsburgh: University of Pittsburgh Press.

Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, Settle JE and Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* **489**, 295 EP.

Brunell TL and DiNardo J (2004) A propensity score reweighting approach to estimating the partisan effects of full turnout in american presidential elections. *Political Analysis* **12**, 28–45.

Citrin J, Schickler E and Sides J (2003) What if everyone voted? simulating the impact of increased turnout in senate elections. *American Journal of Political Science* **47**, 75–90.

Dehejia RH and Wahba S (2002) Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics* **84**, 151–161.

Fewerda J (2014) Electoral consequences of declining participation: a natural experiment in Austria. *Electoral Studies* **35**, 242–252.

Fowler A (2015) Regular voters, marginal voters and the electoral effects of turnout. *Political Science Research and Methods* **3**, 205–219.

Ho D, Imai K, King G and Stuart E (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15**, 199–236.

Iacus SM, King G and Porro G (2012) Causal inference without balance checking: coarsened exact matching. *Political Analysis* **15**, 1–24.

Katz G and Levin I (2018) A general model of abstention under compulsory voting. *Political Science Research and Methods* **6**, 489–508.

King G and Zeng L (2006) The dangers of extreme counterfactuals. *Political Analysis* **14**, 131–159.

Lijphart A (1997) Unequal participation: democracy's unresolved dilemma. *American Political Science Review* **91**, 1–14.

Miller P and Dassonneville R (2016) High turnout in the low countries: partisan effects of the abolition of compulsory voting in the Netherlands. *Electoral Studies* **44**, 132–143.

Cite this article: Bol D, Giani M (2021). It's a (coarsened exact) match! Non-parametric imputation of European abstainers' vote. *Political Science Research and Methods* **9**, 445–450. https://doi.org/10.1017/psrm.2019.44