The Effects of Foreclosures on Homeowners

Samuel G.Z. Asher

Machine Learning and Causal Inference

Spring 2021

Introduction

Understanding the costs of foreclosure is essential to properly calibrating housing policy. Unfortunately, most estimates focus exclusively on financial costs[1]. To make headway, Diamond, Guren, and Tan (2020, henceforth DGT) compile new data on the universe of foreclosure filings in Cook County, Illinois between 2005 and 2012, merging in comprehensive microdata on non-financial outcomes[2]. For identification, the authors rely on random assignment of foreclosure cases to judges in the Circuit Court of Cook County and monotonicity in judges' ranking of case merits to construct a leniency design instrument, as well as a conditional parallel trends assumption to justify propensity-score-matched linear event study regressions. Where the two methods – IV and PSM – disagree, DGT argue the cause is treatment effect heterogeneity, with compliers in the IV case having different treatment effects to the average treated unit. To justify this, they show that a non-PSM OLS event study on sub-samples of foreclosure cases can recover point estimates equivalent to their IV estimates.

I make three contributions to the study of the causal effect of foreclosure on homeowners. Firstly, DGT construct propensity scores with a linear regression of treatment on an ad-hoc selection of predictors. I examine methods to predict probability of treatment with more principled variable selection and with less restrictive functional forms, and then probe robustness of DGT's main specification to these alternative propensity score methods. Secondly, I construct non-parametric estimators of treatment effects. These estimators target more desirable estimands than OLS, relax linearity assumptions, and are robust to nuisance parameter misspecification. Finally, I conduct a treatment effect heterogeneity analysis with fewer researcher

[1] See e.g. U.S. Department of Housing and Urban Development (2010) for an influential example.

[2] Due to space constraints, I do not elaborate on institutional details or descriptive statistics here. Please refer to Diamond et al. (2020) and Appendix A.1.

degrees of freedom, using causal forests (Athey et al., 2018) to allow for data-driven discovery. This constitutes the only data-driven exploration of heterogeneity in the literature on foreclosure[3].

Due to space constraints, there are many implementation details and some auxiliary analyses that I can not present in the main text, but that I believe are important to the rigour of the paper. As such, Appendix A.1 contains some analyses and Appendix A.2 contains some implementation details. Code for this project can be found at: `https://github.com/sgzasher/MI-CL-Final-Project`.

Propensity Score Prediction

The main DGT specification regresses outcomes Y on treatment indicator F denoting foreclosure and a vector of fixed effects $\mathbf{\Gamma}$:

$$Y_{i,k,s,t} = \beta_s F_k + \mathbf{\Gamma}_{i,k,s,t} + \epsilon_{i,k,s,t} \tag{1}$$

where i indexes individuals, k cases, s event-year, and t year. In particular, the suite of fixed effects includes (exhaustively): event-time, treatment, and individual fixed effects; as well as zipcode-year and date-of-filing fixed effects interacted with propensity score deciles. Identification in this setting relies on parallel pre-trends and a lack of anticipation effects (Borusyak et al., 2021) and standard errors are clustered at the case level. To construct ex-ante propensity scores, DGT first regress unrestricted the treatment indicator on 20 predictor variables and zipcode-year and date-of-filing fixed effects with data three years pre-treatment. They then take the five variables with the highest t-statistics and conduct the same regression with those predictors, forming propensity scores from the predicted values.

I present three alternatives to the above method. Firstly, we might want a more principled method for variable selection within the DGT framework: I implement a post-LASSO algorithm (Belloni and Chernozhukov, 2013). Secondly, the propensity score model will be mis-specified if the covariates and fixed effects do not enter the true data generating process in a linear, additive manner. To relax this assumption, I implement random tree and random forest algorithms to predict propensity scores from the same 20 predictors. The tree would be appropriate were the data-generating process a step function, while the forest averages trees for flexible non-parametric prediction[4].

[3] DGT describe: *"Our findings about neighborhood quality and divorce for owners in particular highlight the importance of considering heterogeneous treatment effects in evaluating foreclosure policy. We are not aware of prior research that shows evidence of significant treatment effect heterogeneity."*.

[4] I follow a cluster-robust cross-fitting method, extended from Athey et al. (2018), in all machine learning algorithms mentioned. See Appendix A.2.2 for details on implementation for all of these methods.

Figure 1 plots histograms of the resulting propensity score predictions for comparison. Notice that the DGT and post-LASSO propensity scores look almost identical; the fixed effects do not leave much variation for variable selection to work with, and the two methods select very similar predictors[5]. The forest and tree predictions look similar also, with a more skewed distribution than the linear methods, although the tree appears to be "lumpier" due to its lack of smoothing. All of the methods recover propensity score predictions with good overlap properties, although the linear models have small densities at zero. I return to this issue below.

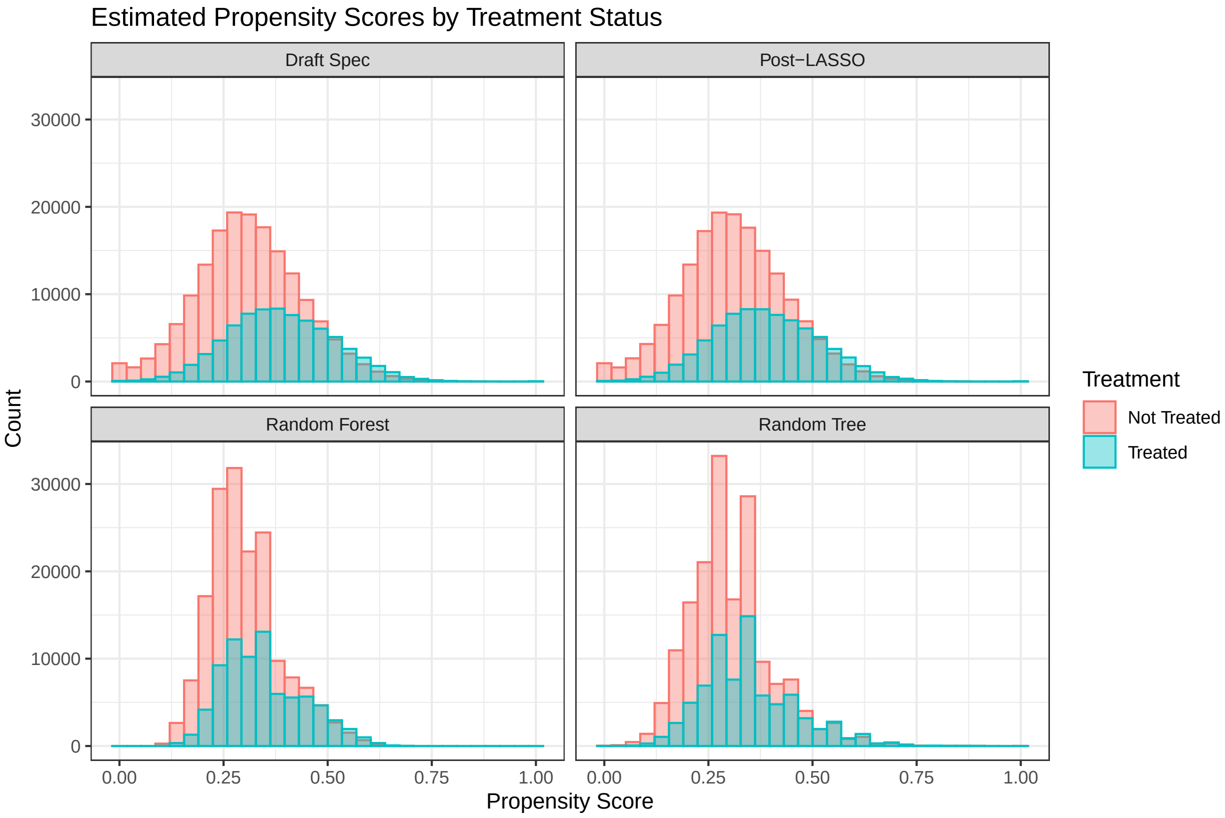

Figure 1: Estimated Propensity Scores

To assess model performance, Figure 2 plots standardised absolute mean differences between treatment and control units one year pre-treatment for a few covariates. Notice that although the unadjusted covariates are not balanced, they only differ by around five percentage points across groups. Worryingly, for quality of local schooling, zip-code income, home square-footage, age, and credit score, weighting by the linear model propensity scores appears to *worsen* pre-treatment balance. If we believe that unconfoundedness is satisfied, we should be worried that the linear models are mis-specified. The tree and forest algorithms both improve balance across

[5] The DGT variable selection includes an ownership indicator, log square footage, credit scores, number of mortgages with 90+ loans, and number of open mortages. This is a strict subset of the post-LASSO variable selection, which also selects average log zipcode income, middle school test score index, cumulative number of divorces, and number of vehicular loans. Of course, it is unwise to over-interpret these variable selections.

the board, with the forest out-performing the tree. Neither achieves perfect weighted balance, but the forest brings the difference down to around one percentage point. See Appendix A.1.2 for supplementary analyses including decile balance and the (strong) robustness of the DGT specification to these alternative methods.

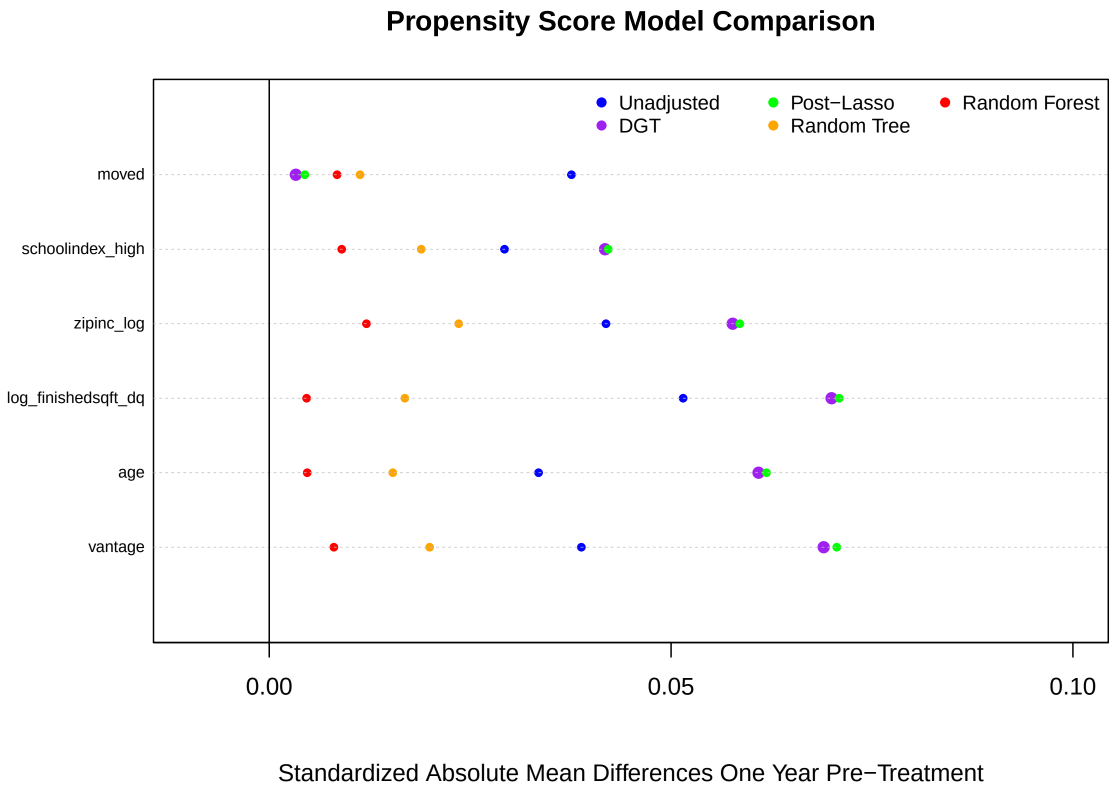

Figure 2: Propensity Score Balance Comparison

Event Study Design

Event study designs target estimands of the form $\tau_w = \sum_{it\in\Omega_1} w_{it}\tau_{it} \equiv w_1'\tau$, where Ω_1 is the set of treated units, w_{it} are observation weights and τ_{it} are unit treatment effects (Borusyak et al., 2021). In the DGT case, we are interested with the above in which Ω_1 is the set of treated units in a balanced panel and averaged at a specific event time[6]. Unfortunately, OLS recovers an opaque weighting scheme on the unit treatment effects that is 'improper' in that weights can potentially be negative due to OLS performing "forbidden comparisons" of later-treated to earlier-treated units in staggered designs (de Chaisemartin and D'Haultfœuille, 2020). My goal in this section is to construct cohort-pooled event study estimates of the ATT in event-year four that incorporate prediction methods without performing forbidden comparisons. I offer a brief elaboration here, with more details in Appendix A.2.3.

[6] This section presents estimates pooled for event time four i.e. treatment effects four years post-treatment.

I estimate the ATT separately for cohort-years (i.e. cross-sectionally at event-years for each cohort) using two methods, and then pool together for event-year estimates. Firstly, I implement the Imbens-Rubin subclassification (IRS) estimator (Imbens and Rubin, 2015) using the algorithms from the previous section. I also implement AIPW estimates of the treatment effect, estimating nuisance parameters using the above algorithms as well as causal forests. Moving to cross-sectional estimates, the identifying assumption changes from conditionally parallel trends and no anticipation effects to cross-sectional unconfoundedness; also, we directly compare only units simultaneously treated or never-treated, avoiding 'forbidden comparisons'. All estimates are on panels balanced over the 9-year horizon, and I trim the samples for propensity scores between 0.1 and 0.9, following Crump et al. (2009).

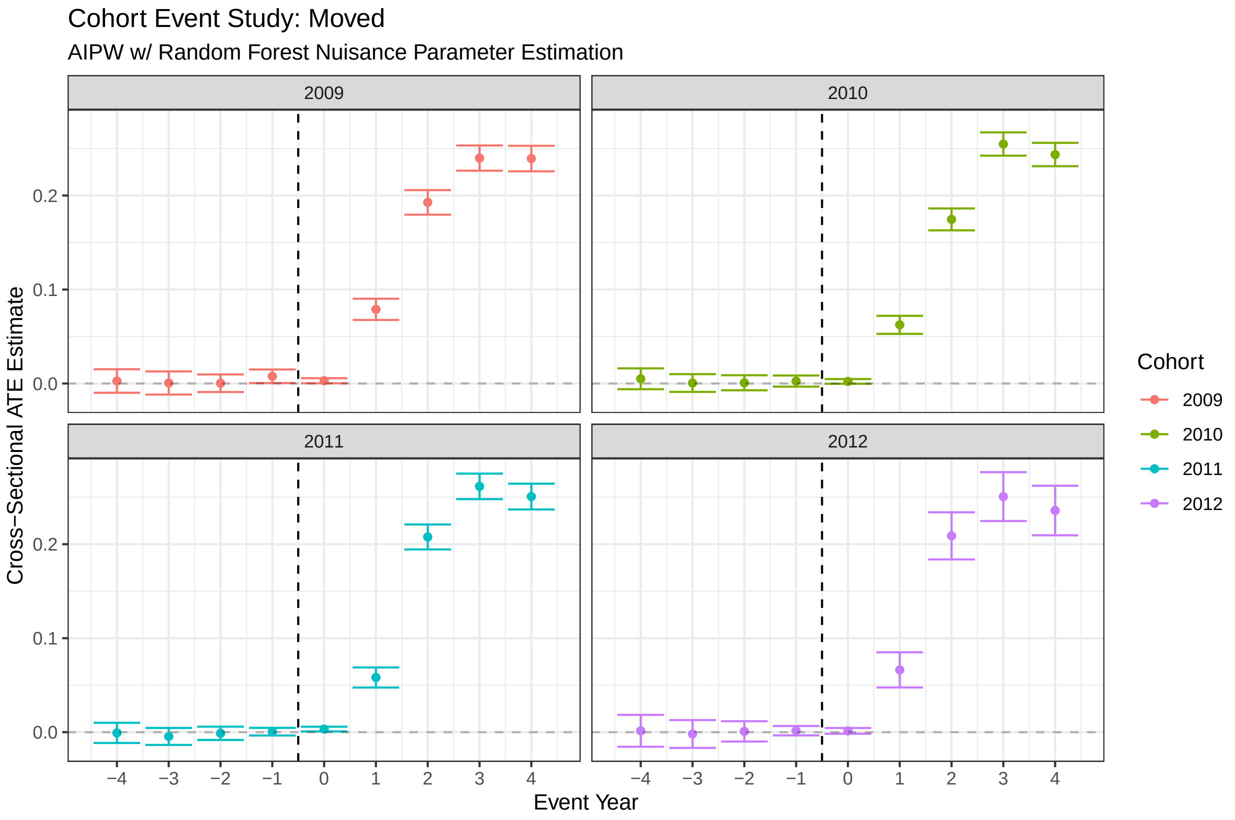

Figure 3: Forest-AIPW Pooled Event Study: Moved From Foreclosure Address

Table 2 reports the results[7]. Panel A reports the original DGT specificaiton results. Panel B reports the pooled IRS estimates; this estimator relaxes the linearity assumptions of the original specification but is not robust to mis-specification in the propensity scores. Panel C reports the pooled AIPW estimates; this estimator relaxes the linearity assumptions and is also potentially

[7] Comparison to the original specification is not straightforward, as I make several simultaneous adjustments in each specification. In particular, since the estimands recovered in Panels B and C target slightly different populations to the DGT specification, I compare between those panels and then suggest the implications for how we should assess the DGT specification.

robust to mis-specification in the propensity scores. Figures 4, 10, 11, and 12 display event studies using the pooled forest-IRS estimator; Figures 3, 13, 14, and 15 display the equivalent plots for the pooled forest-AIPW estimator. From them we may check pre-treatment estimated ATTs as placebo checks for our methods[8]. The AIPW estimates are centered more closely to zero in pre-treatment periods and are more precise than the IRS estimates. Considering this – as well as that the finite-sample performance of AIPW estimators, directly correcting for bias due to errors in propensity score estimates, is potentially better than stratification estimators – I prefer the pooled AIPW estimates to the IRS estimates when there is disagreement.

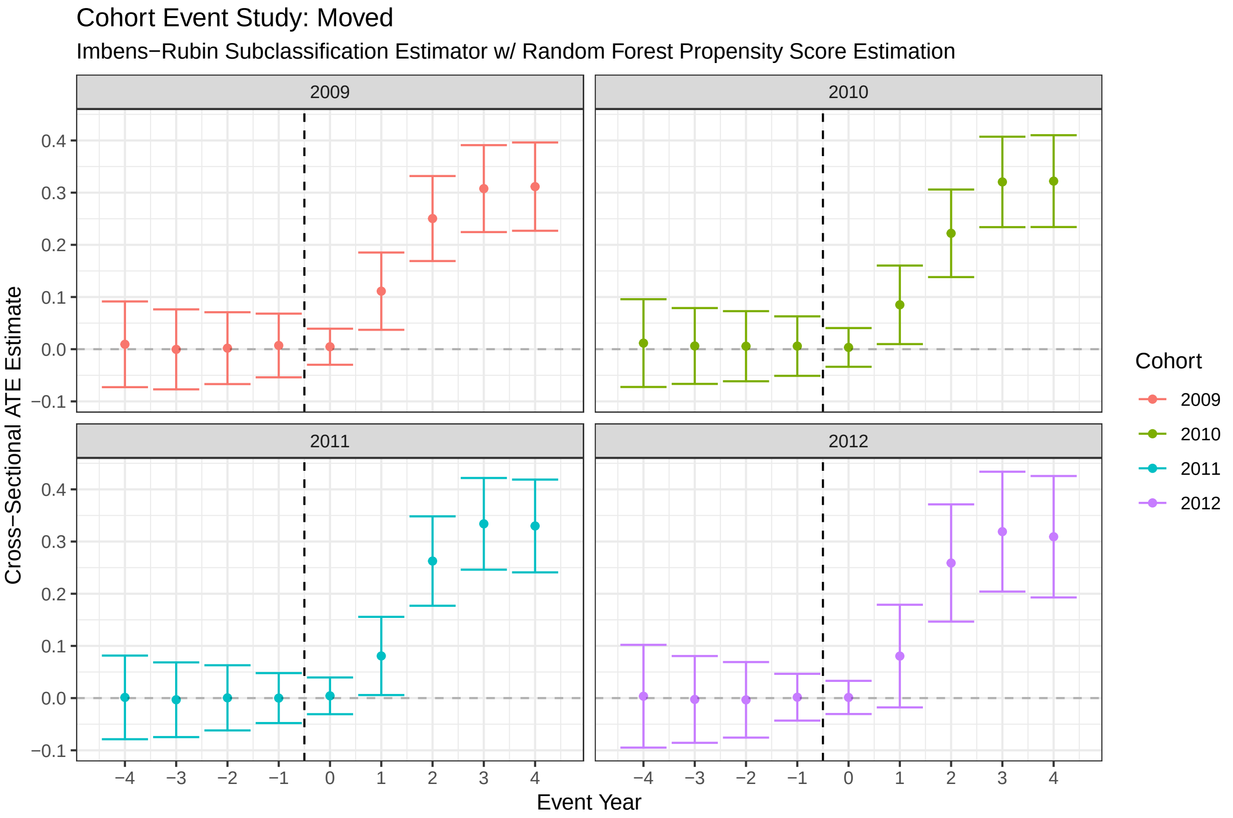

Figure 4: Forest-IRS Pooled Event Study: Moved From Foreclosure Address

I return now to Table 2. Ignoring the causal forest results (see Appendix A.1.3), we see that for the indicator for moving home and number of unpaid debt collections, the IRS estimates all agree with one another in magnitude, implying that the propensity score data generating process for those outcomes can be well approximated as linear. The AIPW results similarly are in agreement, suggesting the same[9]. For the cumulative number of divorces, the IRS and AIPW estimates are all centred on approximately the same near-zero value, though the AIPW estimates are much more precise and can reject the null hypothesis of zero effects. For the

[8] We can also learn about the dynamics of our estimated treatment effects; I leave this to later work.

[9] If we believe the propensity scores are well-specified here, then we cannot draw conclusions about the outcome DGP from the AIPW since it is robust to mis-specification in outcomes in that case.

quality of nearby schooling, the linear and non-linear IRS estimates disagree. Since the method is not robust to mis-specification in the propensity scores, the data generating process for the propensity scores satisfying unconfoundedness for this outcome may not be well approximated as linear. Similarly, the linear and non-linear AIPW estimates disagree with one another (though their confidence intervals overlap); this may suggest that the outcome data generating process also can not be well approximated as linear. There is a very large disagreement between the non-linear IRS and AIPW estimates; as discussed above, I am more inclined to trust the AIPW estimates in this circumstance, and this may be evidence that our propensity score models are particularly poorly calibrated for this outcome.

Heterogeneous Treatment Effects

Where the instrumental variables and PSM-event-study methods recover different point estimates in Diamond et al. (2020), the authors ascribe it to treatment effect heterogeneity. To explore this, DGT show that non-propensity-matched OLS event studies on study sub-samples can recover similar point estimates to their IV. The primary risk with this method is that, by selecting on confirmatory results, one might spuriously conclude heterogeneity exists where there is none. A more preferable strategy would let us take an agnostic approach to heterogeneity and test for its presence. To that end, I train causal forest (Athey et al., 2018) treatment effect estimates in the pooled event study setup described in the previous section[10].

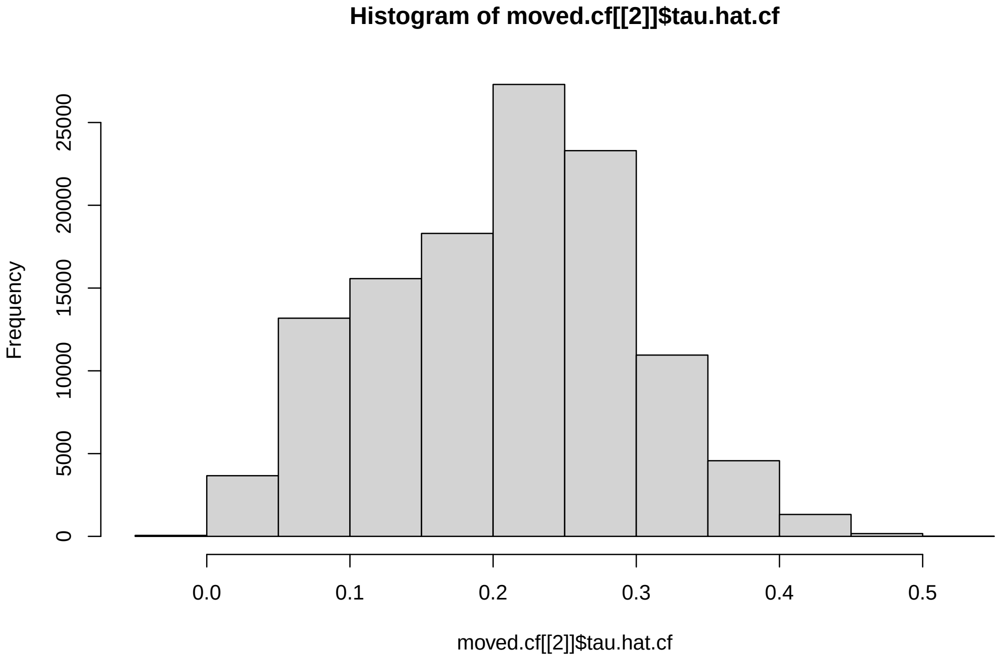

Figure 5: Histogram of Estimated CATE Scores: Moved From Foreclosure Address

[10] See Appendix A.2.4

For initial inspection, pooling all CATE estimates in a given event-year, Figures 5, 16, 17, and 18 plot histograms of the estimated CATEs for each outcome. For each of the outcomes, we can see some dispersion of CATE estimates; for cumulative number of divorces there is a reasonably symmetric distribution around 0.01, for number of unpaid collections there is a distribution above zero with a notable lack of corresponding density below zero, for the schooling index there is a large density between -2 and 2, and for the moving home indicator there is a large density of CATEs between 0 and 0.4. However, this does not necessarily mean that the CATE is a better unit treatment effect estimate than a simple ATT. I conduct two tests to assess whether there is evidence of treatment effect heterogeneity in the data. First, I regress synthetic predictors on out-of-bag CATE estimates to implement the "best linear predictor" calibration test (Athey and Wager, 2019). Secondly, I examine whether the causal forests can reliably identify sub-groups of the data in which treatment effects are different. The implementation for both of these tests required significant adaptations for the setting; please see Appendix A.2.4[11].

Table 3 presents the estimates of the linear predictor calibration test. For all of the outcome models, a one-sided heteroskedasticity-robust hypothesis test for the presence of heterogeneity rejects the null of no heterogeneity at any conventional significance level. Inspecting the slope coefficients, we see that the model for movement outcomes is exceptionally well calibrated, with a coefficient very close to one. For the other outcomes, the forest CATE models appear to be over-stating the extent of heterogeneity in the data; in expectation, a one-unit deviation in estimated CATE implies only between a 0.3 and 0.5 unit deviation in true CATE. Nevertheless, finding evidence of heterogeneity for all outcomes is encouraging.

Figures 6, 19, 20, and 21 plot the ATT estimates for the subset of observations in each quintile of the CATE scores. For the movement outcome, we can see again how well the causal forest has performed; it has clearly identified five separate subgroups that can be ordered with monotonically increasing treatment effects. This is very strong evidence that there is treatment effect heterogeneity in who is moving home four years after foreclosing. For the cumulative number of divorces, the evidence is weaker; it appears that the forest may have successfully identified two subgroups for whom treatment effects are distinct. For the number of unpaid collections and high school test index, the causal forests appear to be less well calibrated, with the lowest identified CATE group recovering higher ATT point estimates than the next lowest group in both cases (for the number of unpaid collections, the confidence intervals do

[11] For the sub-group analysis, the implementation is different enough that I am slightly unsure of the validity of my confidence intervals, and refrain from conducting inference across groups. I am interpreting this test as high-quality qualitative evidence of the ability of the causal forests to identify subgroups.

not overlap for the two groups)[12]. However, for both outcomes the forest does at least seem to have successfully identified a group of reliably high-CATE individuals. There is certainly good evidence of the causal forests successfully identifying subgroups with different treatment effects, without researcher input, for all of the outcomes.

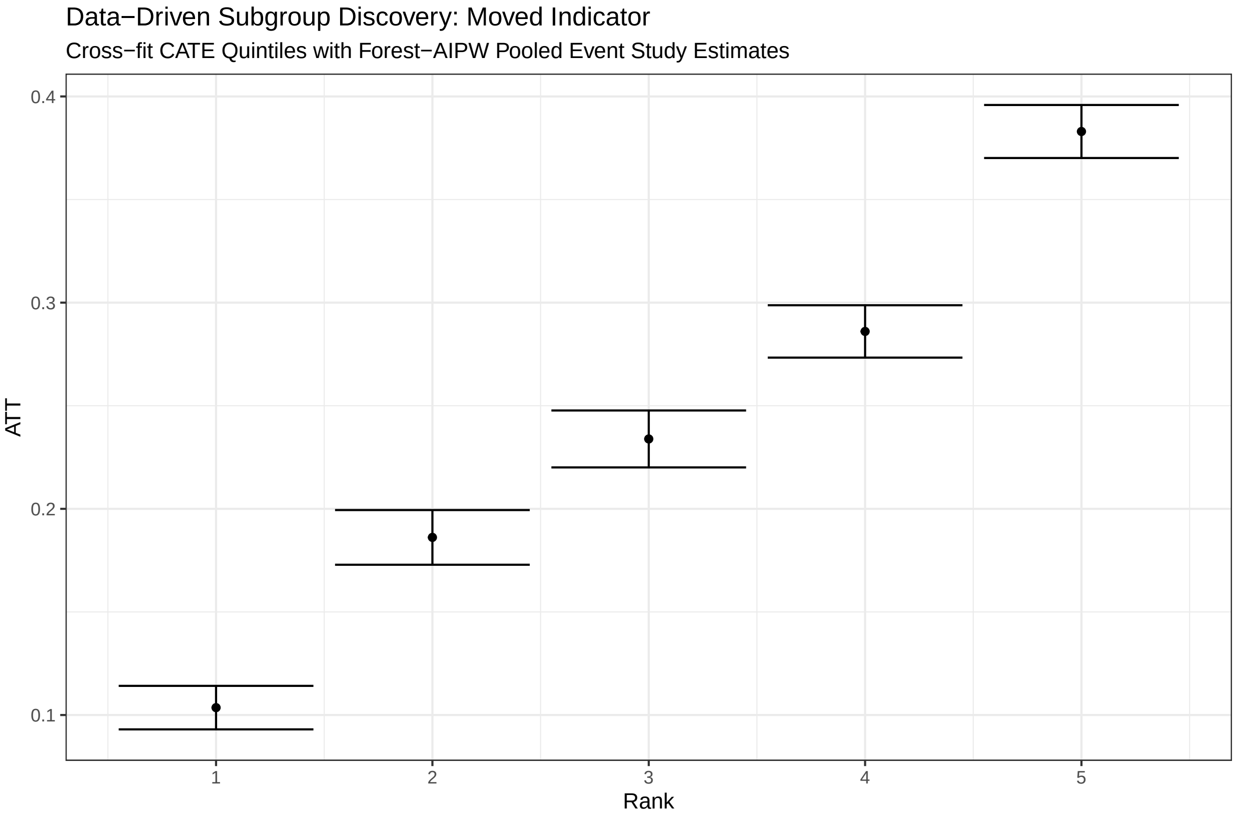

Figure 6: Forest-AIPW Pooled Estimates Within CF-Identified Subgroups

A logical next step would be to compare the heterogeneity results to the IV estimates, and to compare units in different CATE groups to the identified complier space in Diamond et al. (2020). In the interest of time, space, and scope, I leave this to future efforts.

[12] Note that the quintiles are stratified by cohort due to the estimation procedure; there are an approximately equal number of 2009 and 2012 cohort units in each quintile, for example. If there is strong time-specific heterogeneity, then the apparent miscalibration could be due to e.g. a cluster of units in the lowest quintile with a relatively low CATE for their cohort, but a relatively high CATE for the panel. Indeed, notice in Figures 14 and 15 that there is across-cohort heterogeneity in the fourth year post-treatment.

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017, November). When Should You Adjust Standard Errors for Clustering? Technical Report w24003, National Bureau of Economic Research.

Athey, S., J. Tibshirani, and S. Wager (2018, April). Generalized Random Forests. *arXiv:1610.01271 [econ, stat]*. arXiv: 1610.01271.

Athey, S. and S. Wager (2019, February). Estimating Treatment Effects with Causal Forests: An Application. *arXiv:1902.07409 [stat]*. arXiv: 1902.07409.

Belloni, A. and V. Chernozhukov (2013, May). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2). arXiv: 1001.0188.

Borusyak, K., X. Jaravel, and J. Spiess (2021, May). Revisiting Event Study Design: Robust and Efficient Estimation. Working Paper.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017, January). Double/Debiased/Neyman Machine Learning of Treatment Effects. *arXiv:1701.08687 [stat]*. arXiv: 1701.08687.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009, March). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

de Chaisemartin, C. and X. D'Haultfœuille (2020, September). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review 110*(9), 2964–2996.

Diamond, R., A. Guren, and R. Tan (2020, June). The Effect of Foreclosures on Homeowners, Tenants, and Landlords. Technical Report w27358, National Bureau of Economic Research, Cambridge, MA.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Knittel, C. R. and S. Stolper (2019, December). Using Machine Learning to Target Treatment: The Case of Household Energy Use. Technical Report w26531, National Bureau of Economic Research.

U.S. Department of Housing and Urban Development (2010). “Economic Impact Analysis of the FHA Refinance Program for Borrowers in Negative Equity Positions. Technical report, Washington DC.

Tables and Figures

	Moved	Divorces	Schooling	Collections
Panel A: Propensity Score Deciles				
DGT	0.282	0.001	1.001	0.256
	(0.003)	(0.001)	(0.126)	(0.027)
	[0.000]	[0.352]	[0.000]	[0.000]
Post-LASSO	0.282	0.001	1.031	0.257
	(0.003)	(0.001)	(0.126)	(0.027)
	[0.000]	[0.489]	[0.000]	[0.000]
Forest	0.284	0.001	0.992	0.258
	(0.003)	(0.001)	(0.126)	(0.027)
	[0.000]	[0.164]	[0.000]	[0.000]
Tree	0.285	0.001	1.002	0.264
	(0.003)	(0.001)	(0.127)	(0.027)
	[0.000]	[0.172]	[0.000]	[0.000]
Panel B: Imbens-Rubin Stratification				
DGT	0.290	0.001	1.026	0.232
	(0.001)	(0.001)	(0.161)	(0.036)
	[0.000]	[0.539]	[0.000]	[0.000]
Post-LASSO	0.288	0.001	1.114	0.227
	(0.003)	(0.001)	(0.154)	(0.034)
	[0.000]	[0.503]	[0.000]	[0.000]
Forest	0.281	0.001	0.936	0.255
	(0.001)	(0.001)	(0.151)	(0.034)
	[0.000]	[0.225]	[0.000]	[0.000]
Tree	0.289	0.001	0.979	0.290
	(0.003)	(0.001)	(0.152)	(0.034)
	[0.000]	[0.139]	[0.000]	[0.000]

Standard errors in parentheses, p-values in brackets. Standard errors clustered at the case level. “Moved” refers to an indicator variable for having changed address in the given year. “Divorces” refers to the cumulative number of divorces for a given individual at a given year. “Schooling” refers to the relative test score rank of an individual’s closest high school. “Collections” refers to an individual’s number of unpaid debt collections in a given year. All models are OLS event study regressions with pooled event-year 3 and 4 coefficients. Propensity stratification bins enter as fixed effects interacted with zip-code-year and date of initial filing fixed effects.

Table 1: OLS Event Study Specifications with Propensity Stratification Interactive Fixed Effects

	Moved	Divorces	Schooling	Collections
Panel A: Baseline DGT Specification				
DGT	0.282 (0.003) [0.000]	0.001 (0.001) [0.352]	1.001 (0.126) [0.000]	0.256 (0.027) [0.000]
Panel B: Imbens-Rubin Subclassification Estimator				
DGT	0.315 (0.023) [0.000]	0.008 (0.014) [0.540]	0.609 (0.179) [0.000]	0.182 (0.065) [0.005]
Post-LASSO	0.314 (0.023) [0.000]	0.004 (0.014) [0.746]	0.584 (0.185) [0.002]	0.176 (0.064) [0.006]
Forest	0.320 (0.024) [0.000]	0.007 (0.015) [0.616]	1.254 (0.195) [0.000]	0.169 (0.066) [0.010]
Tree	0.321 (0.025) [0.000]	0.007 (0.016) [0.649]	1.239 (0.183) [0.000]	0.160 (0.070) [0.022]
Panel C: AIPW Estimator				
Post-LASSO	0.259 (0.003) [0.000]	0.004 (0.002) [0.011]	0.470 (0.131) [0.000]	0.255 (0.028) [0.000]
Forest	0.244 (0.004) [0.000]	0.005 (0.002) [0.003]	0.286 (0.110) [0.009]	0.184 (0.027) [0.000]
Tree	0.266 (0.004) [0.000]	0.007 (0.002) [0.000]	0.211 (0.144) [0.139]	0.247 (0.028) [0.000]
Causal Forest	0.208 (0.003) [0.000]	0.006 (0.001) [0.000]	0.321 (0.146) [0.027]	0.418 (0.033) [0.000]

Standard errors in parentheses, p-values in brackets. Standard errors reply on cohort-year estimates being independent. “Moved” refers to an indicator variable for having changed address in the given year. “Divorces” refers to the cumulative number of divorces for a given individual at a given year. “Schooling” refers to the relative test score rank of an individual’s closest high school. “Collections” refers to an individual’s number of unpaid debt collections in a given year.

Table 2: Non-Parametric Pooled Event Study Specifications

	Estimate	HC3	t -stat	P -value
Panel A: Moved From Foreclosure Address				
Intecept	0.982	0.012	85.10	0.000
Slope	0.999	0.024	40.88	0.000
Panel B: Cumulative Number of Divorces				
Intercept	0.947	0.258	3.675	0.000
Slope	0.365	0.095	3.819	0.000
Panel C: High School Test Score Index				
Intercept	1.205	0.336	3.586	0.000
Slope	0.308	0.089	3.453	0.001
Panel D: Number of Unpaid Collections				
Intercept	0.983	0.073	13.42	0.000
Slope	0.543	0.091	5.956	0.000

"Moved" refers to an indicator variable for having changed address in the given year. "Divorces" refers to the cumulative number of divorces for a given individual at a given year. "Schooling" refers to the relative test score rank of an individual's closest high school. "Collections" refers to an individual's number of unpaid debt collections in a given year. Heuristically, the intercept in the 'best linear predictor' calibration test absorbs the average treatment effect and the slope coefficient measures the quality of treatment heterogeneity estimates. A one-sided t-test on the slope coefficient being positive can be interpreted as a formal hypothesis test of the null that there is no deviation in treatment effects (i.e. that there is no treatment effect heterogeneity.)

Table 3: Best Linear Projection Calibration Test

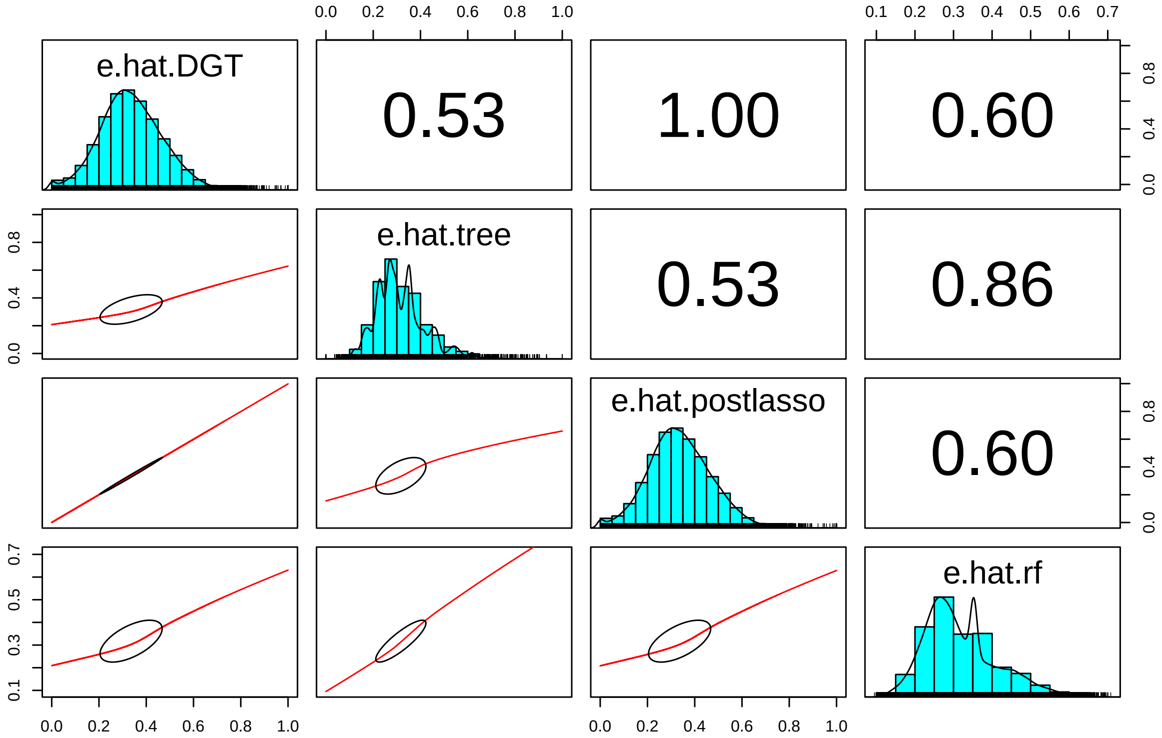

Figure 7: Propensity Score Correlations

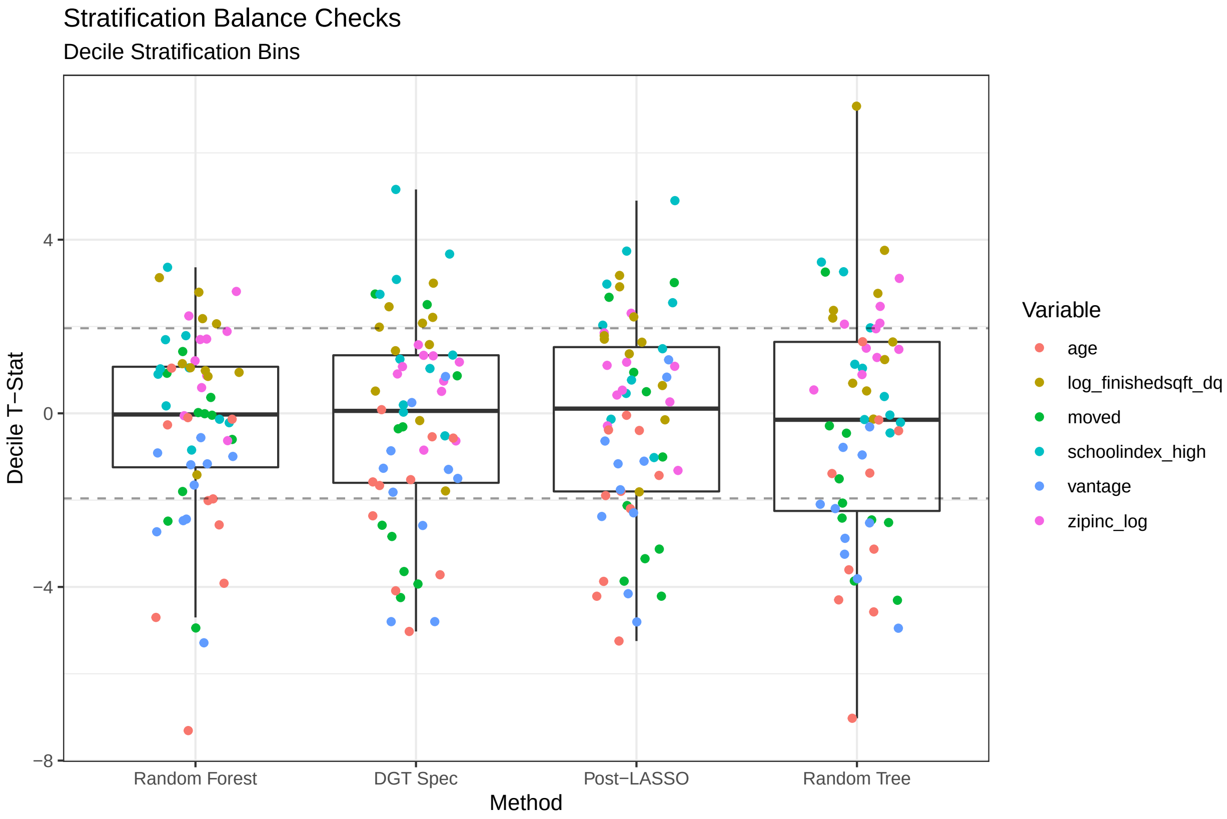

Figure 8: Propensity Score Decile Balance Comparison

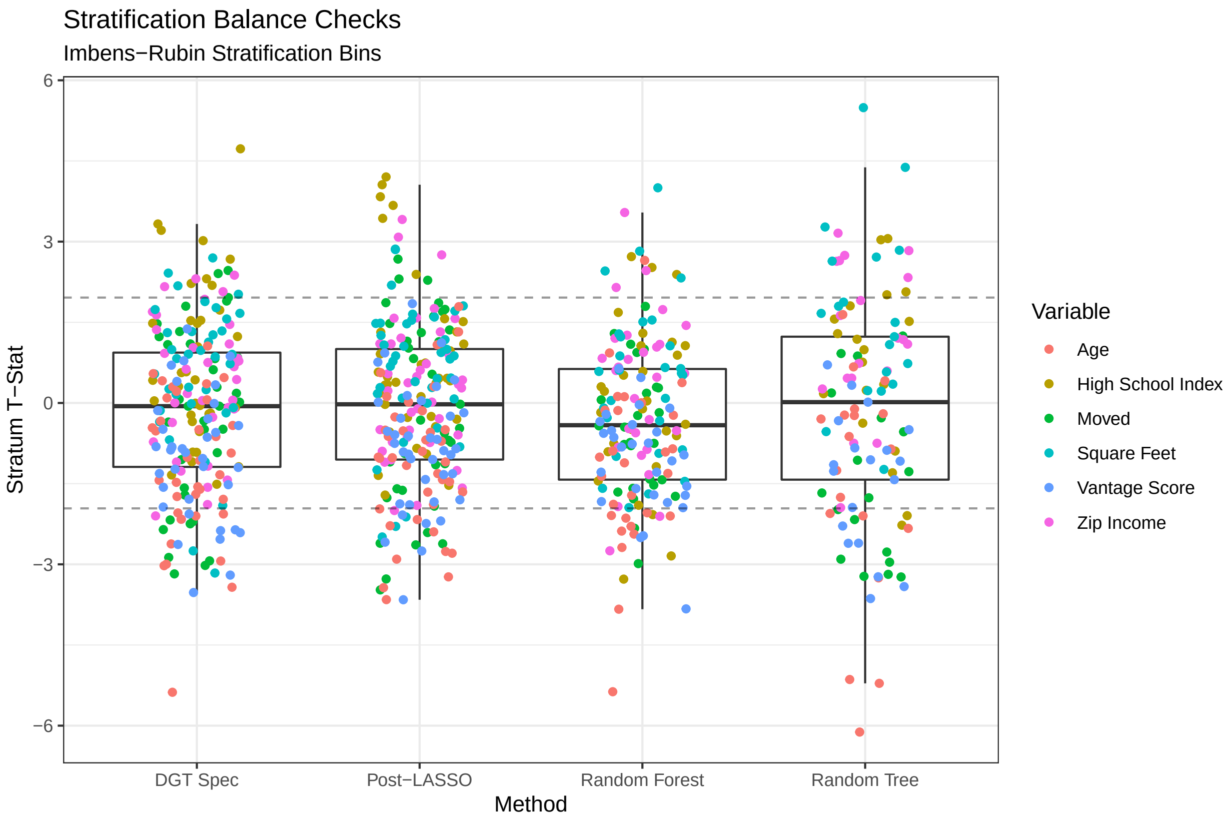

Figure 9: Propensity Score Imbens-Rubin Bin Balance Comparison

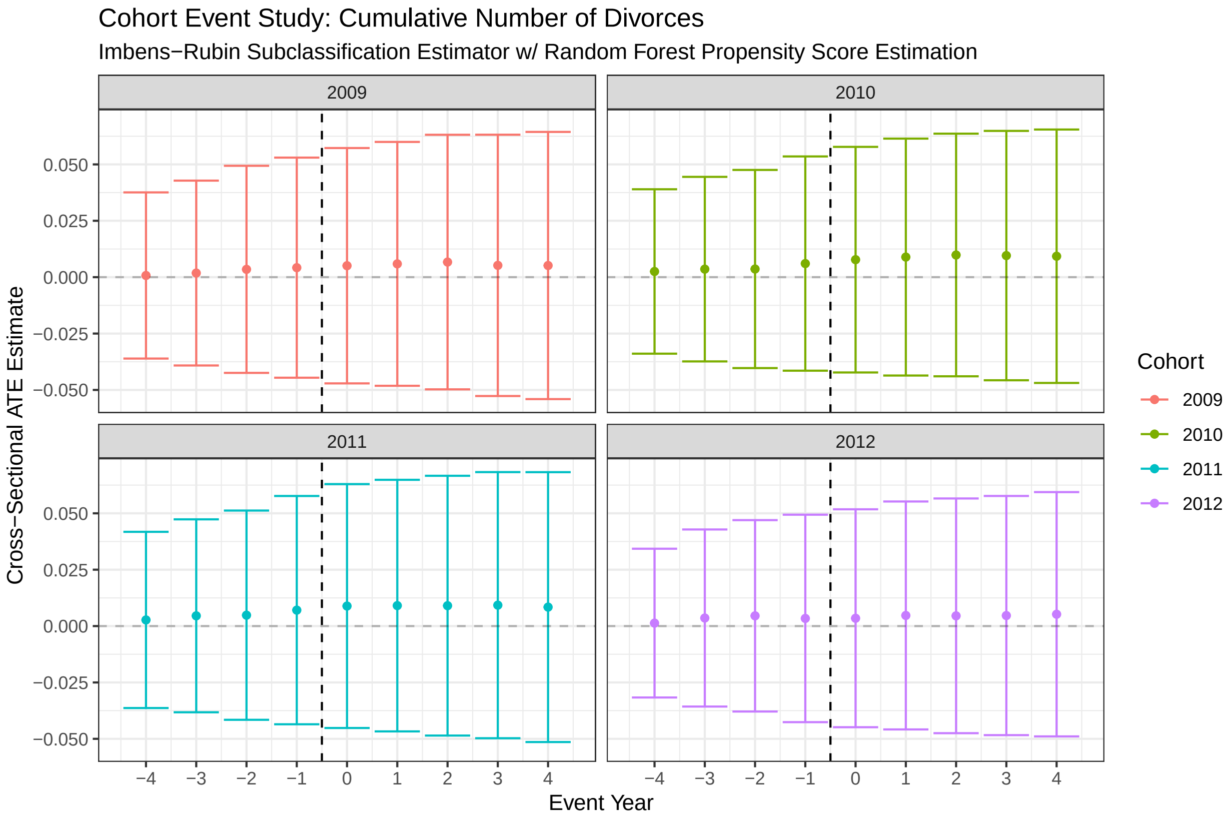

Figure 10: Forest-IRS Pooled Event Study: Cumulative Number of Divorces

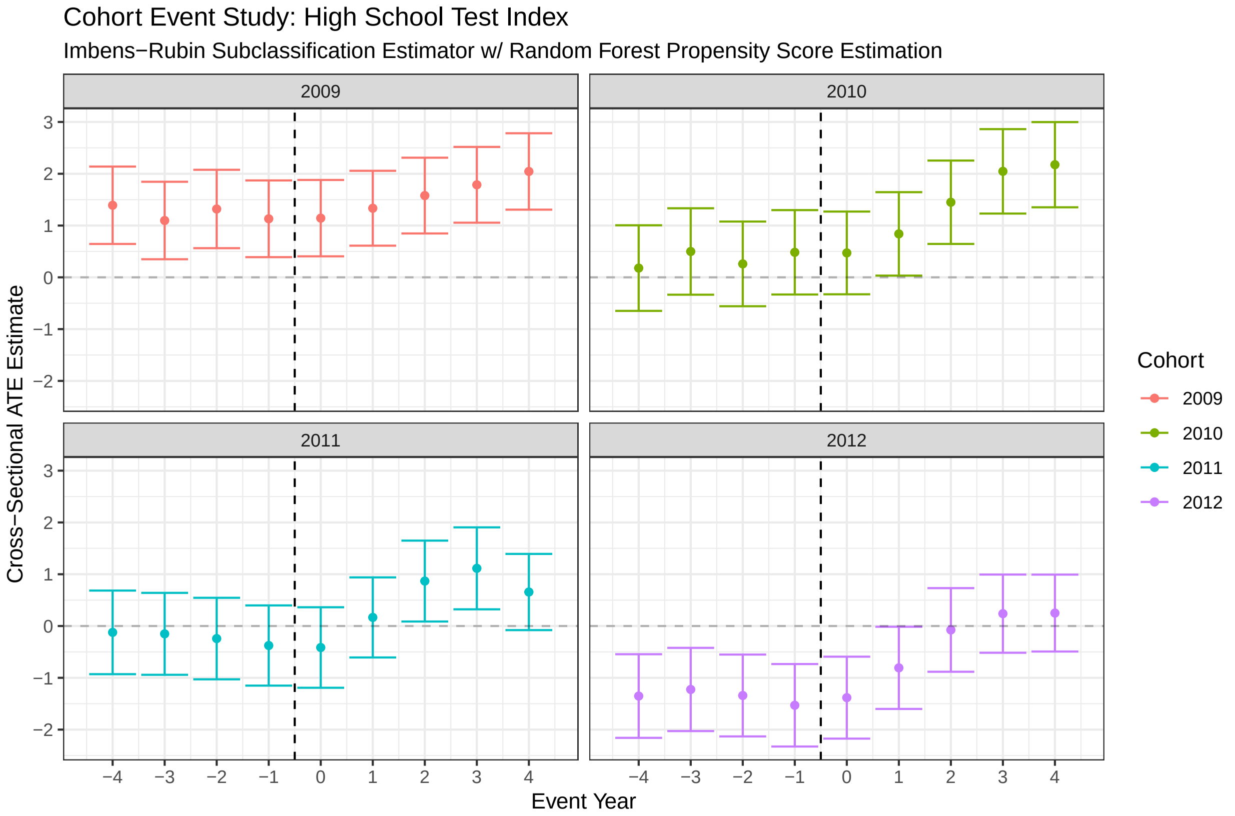

Figure 11: Forest-IRS Pooled Event Study: High School Test Score Index

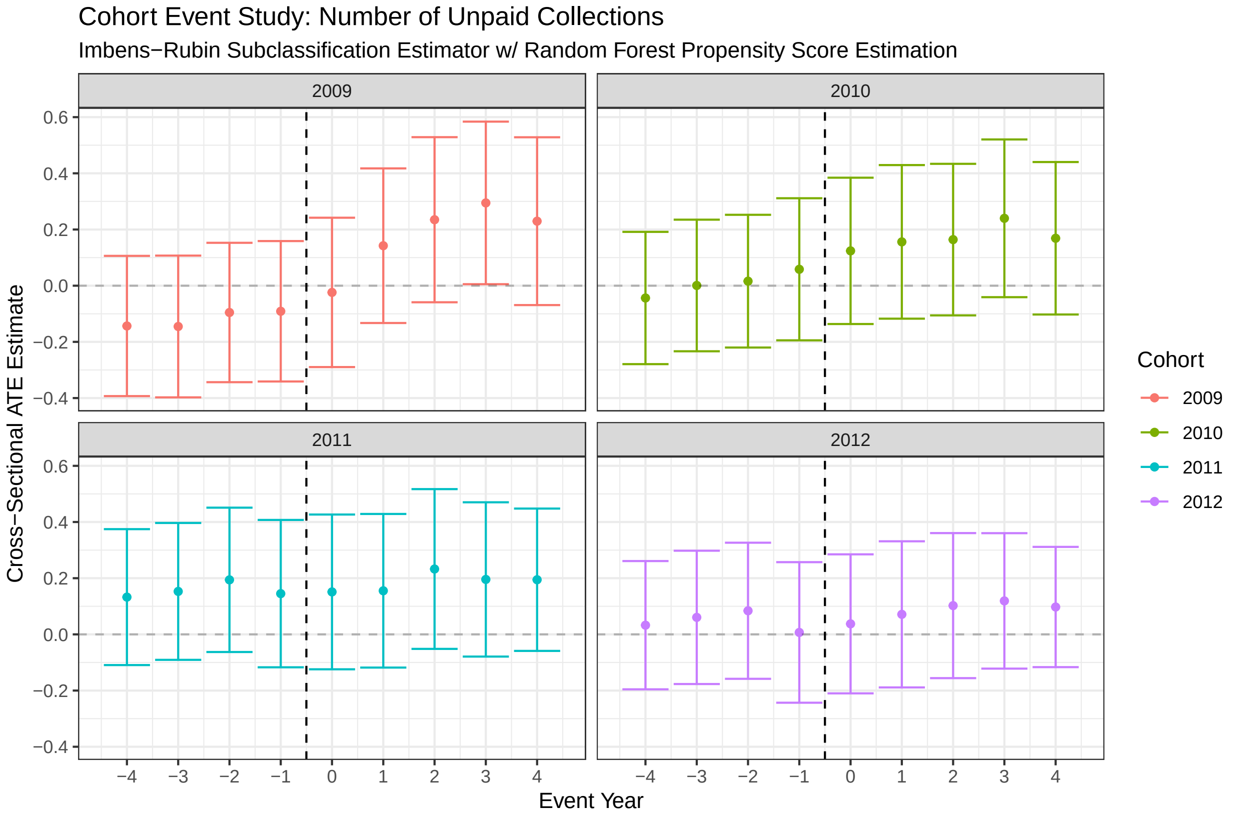

Figure 12: Forest-IRS Pooled Event Study: Number of Unpaid Debt Collections

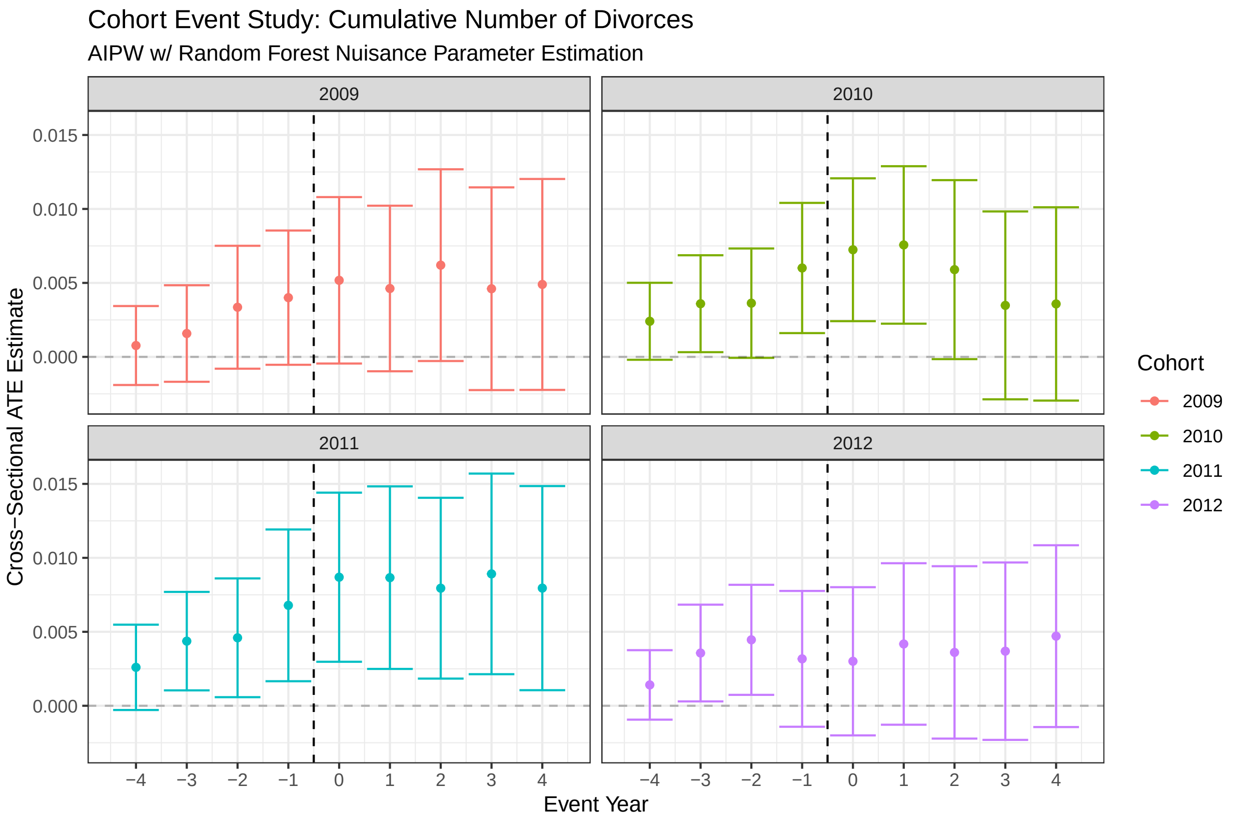

Figure 13: Forest-AIPW Pooled Event Study: Cumulative Number of Divorces

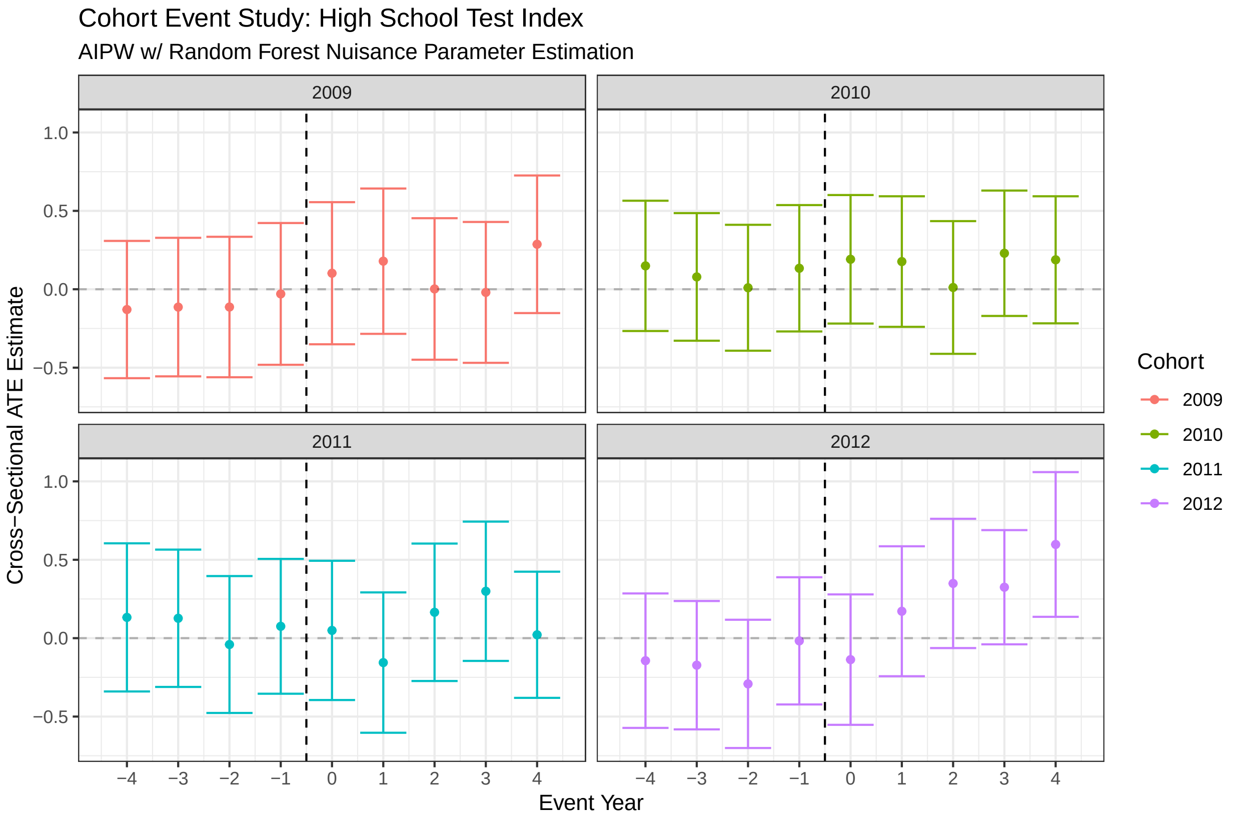

Figure 14: Forest-AIPW Pooled Event Study: High School Test Score Index

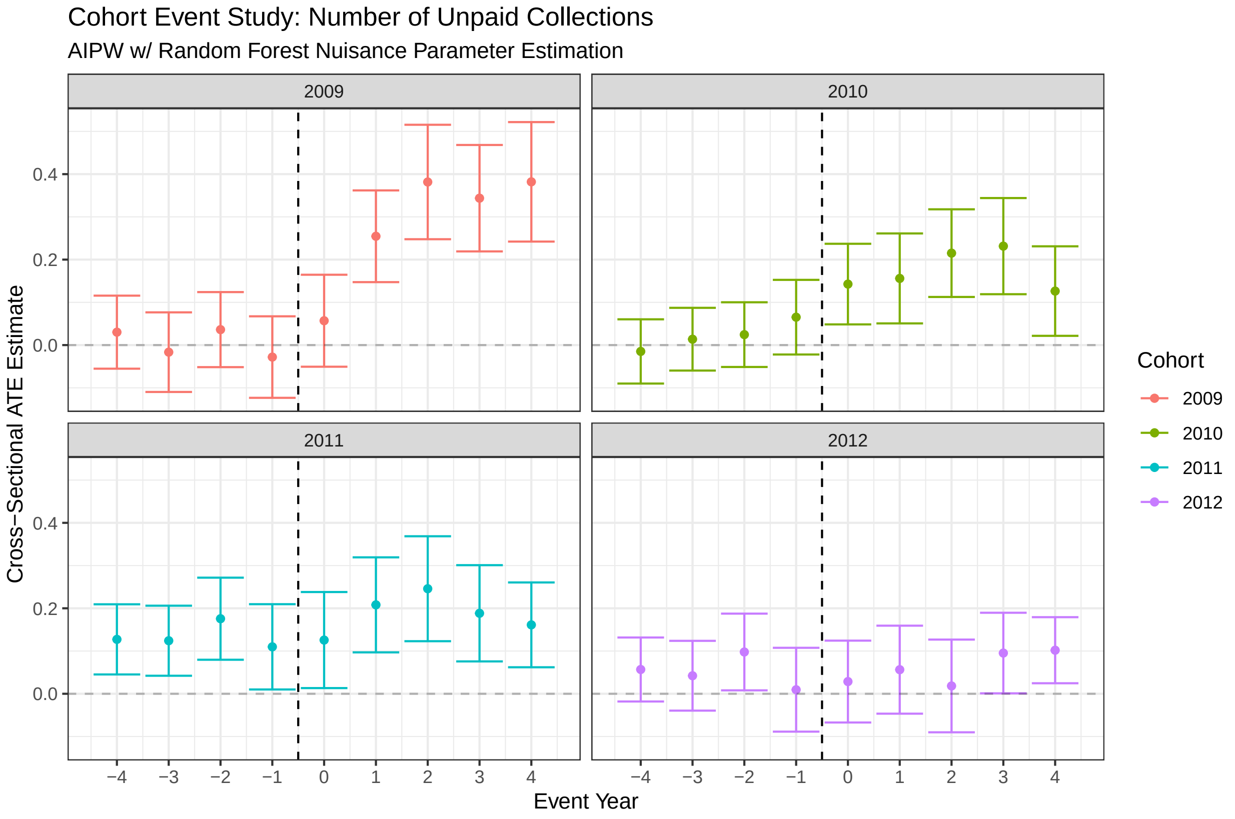

Figure 15: Forest-AIPW Pooled Event Study: Number of Unpaid Debt Collections

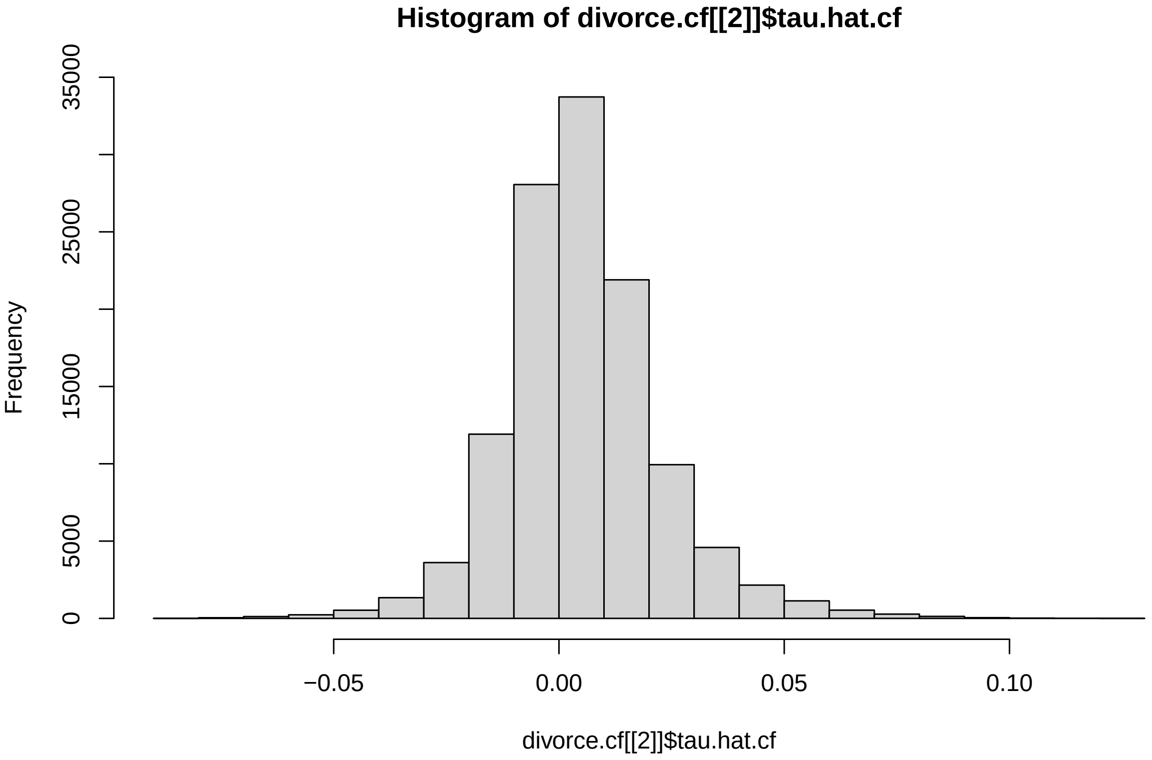

Figure 16: Histogram of Estimated CATE Scores: Cumulative Number of Divorces

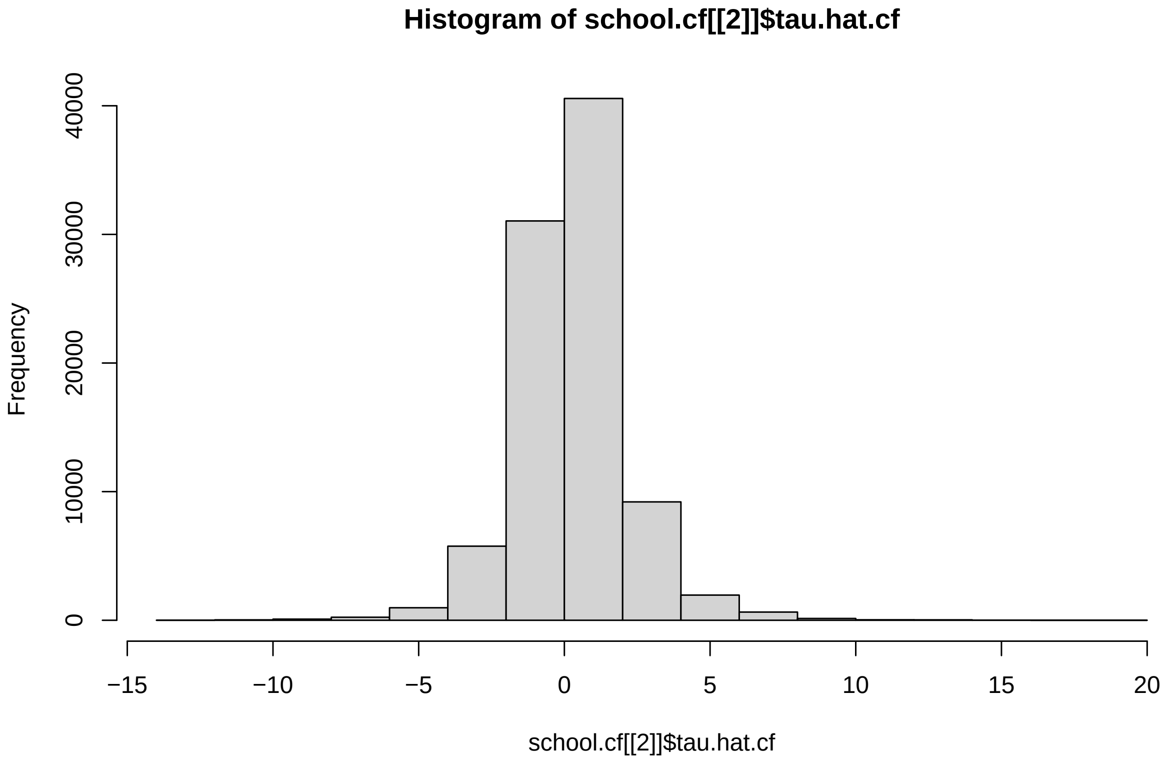

Figure 17: Histogram of Estimated CATE Scores: High School Test Score Index

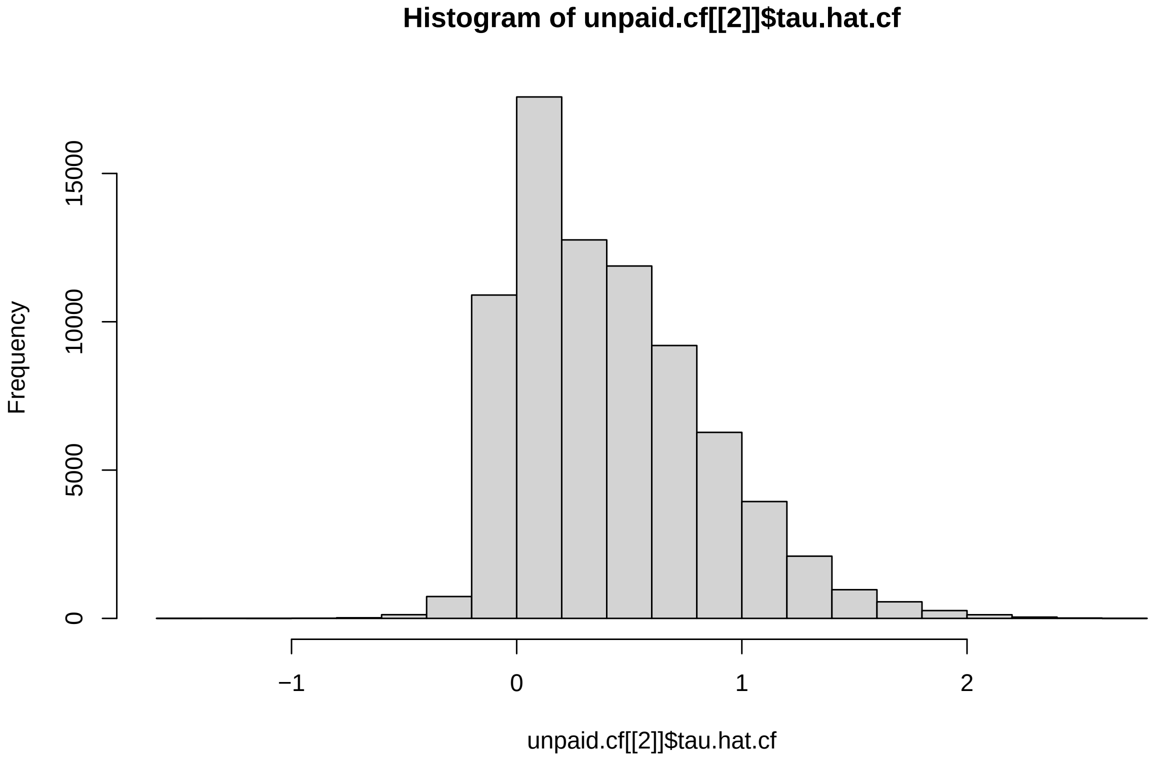

Figure 18: Histogram of Estimated CATE Scores: Number of Unpaid Debt Collections

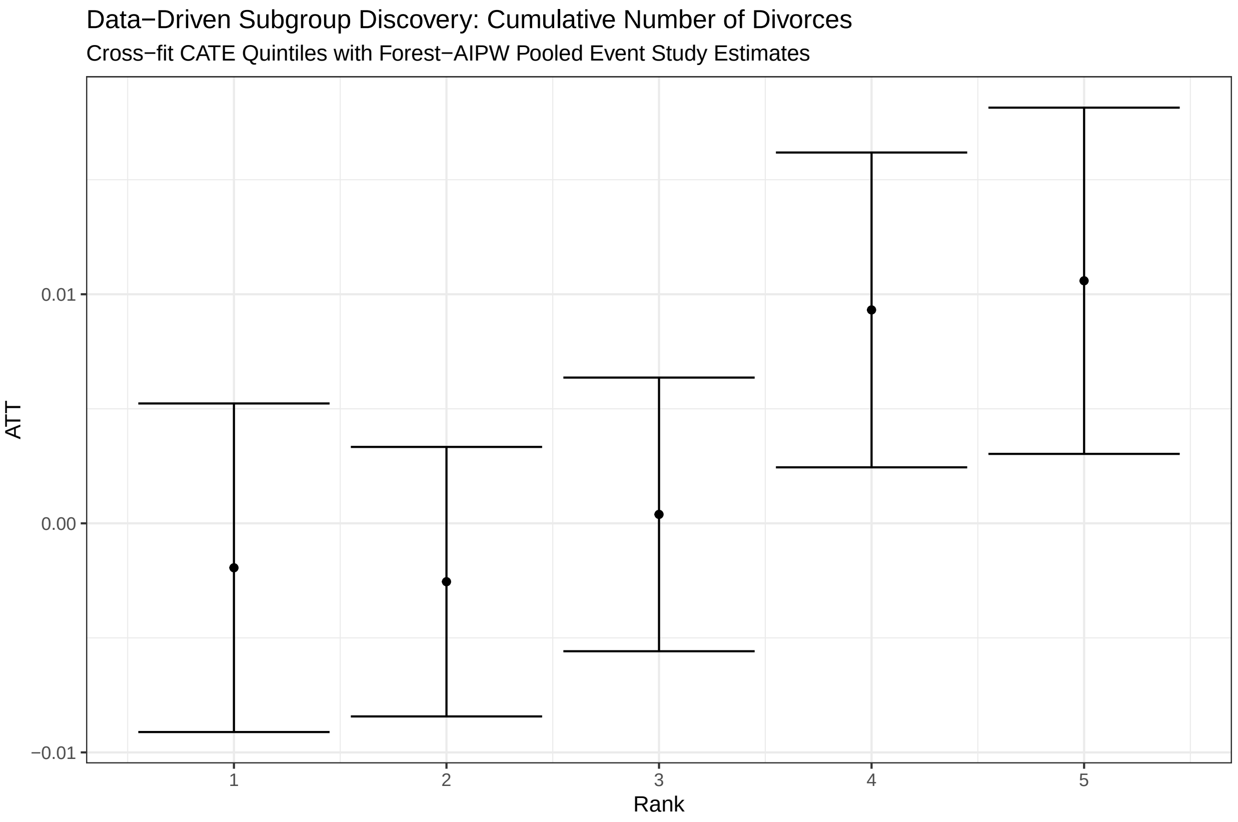

Figure 19: Forest-AIPW Pooled Estimates Within CF-Identified Subgroups: Divorce

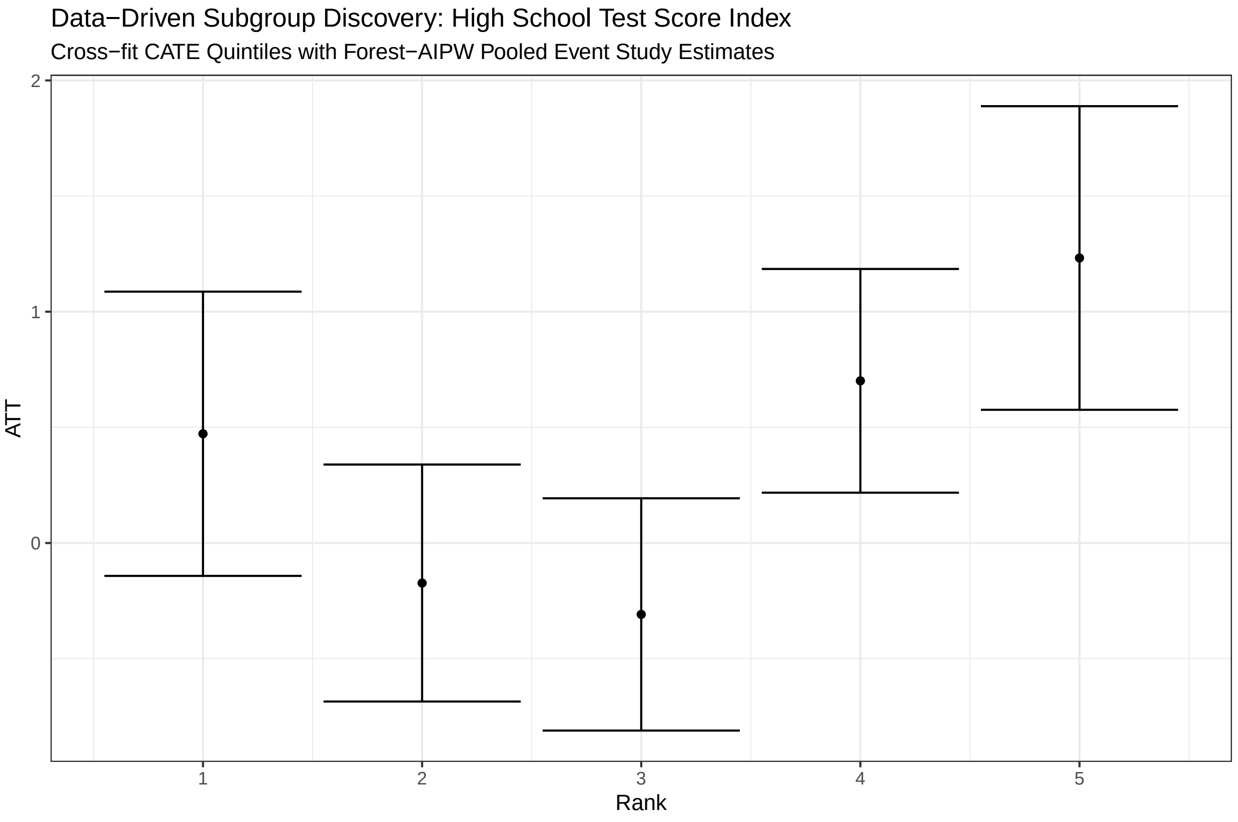

Figure 20: Forest-AIPW Pooled Estimates Within CF-Identified Subgroups: Schooling

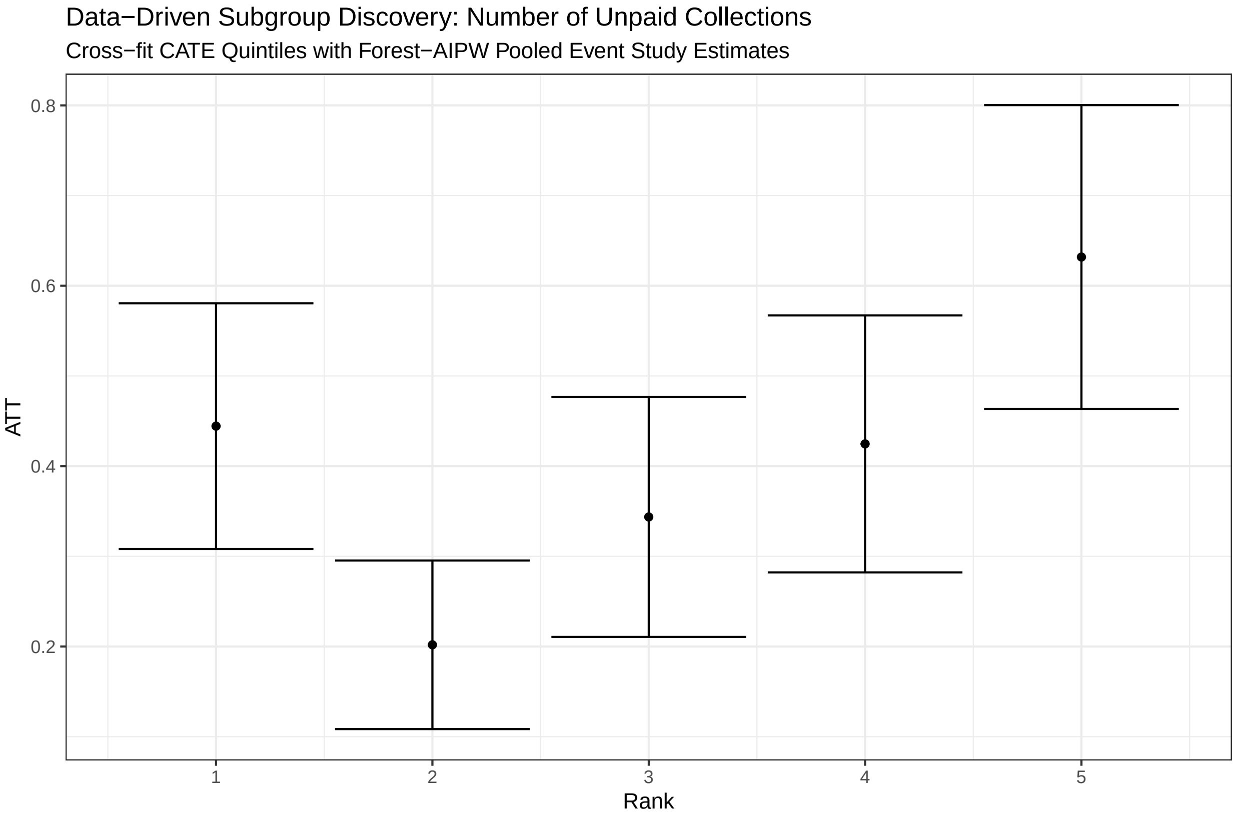

Figure 21: Forest-AIPW Pooled Estimates Within CF-Identified Subgroups: Debt Collections

Appendix A Supplementary Appendix

A.1 Auxiliary Analysis

A.1.1 Variable Description

There are twenty relevant variables used in the analysis (other than e.g. zip-code indicators).

1. (Indicator) Moved from Foreclosure Address
2. (Indicator) Owns Primary Residence
3. Log Square Footage of Living Space
4. Log Zip Code Average Income
5. Elementary School Test Score Rank
6. Middle School Test Score Rank
7. High School Test Score Rank
8. Cumulative Number of Divorces
9. Cumulative Number of Crimes Convicted
10. Cumulative Number of Bankruptcies
11. Cumulative Number of DUI Convictions
12. Vantage Credit Score
13. (Indicator) Death
14. Number of Foreclosures
15. Number of Unpaid Collections
16. Number of Auto Loans
17. Number of Mortgages 90+ DPDs
18. Number of Mortgages with Loan Mod
19. Number of Open Mortgages
20. Open Mortgage Balance

A.1.2 Propensity Score Models

Figure 7 reinforces the direct estimated propensity score inspections, plotting a matrix to visualise the model similarities, with overlapping rows and columns corresponding to correlation coefficients and plots of LOESS-smoothed bivariate scatter relationships. The linear models have a near-perfect correlation, while the tree and forest models have recovered similar, but distinct, predictions. All of the LOESS-scatters are monotone; each method appears to broadly agree on the ordinal ranking of propensity scores, which is encouraging for later stratification.

Figure 8 plots the t-statistic on a within-decile difference in mean test for the same covariates as in Figure 2, for each propensity score method. The average decile t-statistic is lowest for the random forest method (2.16), though for all methods the average t-stat is above 2. To improve on this situation, and to remove researcher degrees of freedom in stratification, I implement the Imbens and Rubin (2015) algorithm for propensity subclassification. Figure 9 plots the results; the average t-statistic is now lowest for the random forest (1.53) and post-LASSO methods (1.52), though the tree remains the poorest performer by this metric. The step-function nature of the tree prediction leaves it ill-suited to this test, since within-strata there is a distance guaranteed between predictions of different leaves. Of course, this may leave trees generally ill-suited to stratification methods. As seen in Figure 2, the tree is doing a good job at balancing covariates.

Table 1 documents for four outcomes the robustness of the original DGT specification to alternative prediction methods for the propensity score. Each row corresponds to regressions of form (1) with propensity scores intereacted fixed effects from different methods. Panel A uses decile stratification, while Panel B uses the Imbens-Rubin subclassification algorithm for stratification[13] (Imbens and Rubin, 2015). Evidently, the original specification is highly robust to alternative propensity score predictions and methods for stratification, with both alterations generally only changing point estimates at the third significant figure and inference conclusions unchanged. Once again, the explanation is that the fixed effects included in the regressions are absorbing most of the interesting variation (all of the regression results presented have an R^2 above 0.6). Moving forwards, I consider alternative methods to estimate the ATT.

A.1.3 Event Study Design

Turning to the pooled causal forest AIPW estimates in Table 2, note that they generally align in magnitude and standard error with the forest-AIPW estimates, with the exception of the unpaid collections outcome. One key difference between my forest-AIPW and the causal forest

[13] See Appendix A.2.2 for details.

algorithm is that the causal forest algorithm trains orthogonalized forests. Orthogonalization is usually important if there are features that are predictive of propensity scores but not very predictive of outcomes (Athey and Wager, 2019), which could be the explanation here. Given time, space, and computation constraints, I am unable to go back and recover variable importance estimates for the submission deadline on this paper. It would be an interesting and important avenue for future work on this project.

A.2 Implementation Appendix

I use this appendix for implementation details that could not be included in the main text due to space constraints, but which I believe are important to the rigour of the project.

A.2.1 General Notes

The setting demands some adaptations of out-of-the-box prediction methods and prediction-based causal estimators. Although outcomes are collected at the individual level, treatment is assigned at the case level; I implement prediction algorithms and standard errors that are robust to arbitrary within-cluster outcome correlation (Abadie et al., 2017; Athey and Wager, 2019). Also, several of the methods I implement rely on cross-sectional estimation. To make use of the data's panel structure, I redefine outcomes to be the difference between the period-specific outcome and the pre-treatment outcome mean, following Knittel and Stolper (2019).

A.2.2 Propensity Score Models

To ensure cluster-robustness and to avoid regularization bias, the predictions are cross-fit such that units are only considered out-of-bag if their cluster was not used in the training step (Abadie et al., 2017; Chernozhukov et al., 2017; Athey and Wager, 2019). This is implemented automatically in the forest algorithm thanks to the *grf* package (Athey et al., 2018). The forest algorithm includes a first variable selection phase following Athey and Wager (2019), in which a forest is trained and variables split on more than the average characteristic are considered selected. The second stage trains a forest using the selected covariates. All forests in the paper are trained with 1000 trees unless otherwise mentioned. For the tree algorithms, I divide the data into 5 folds, ensuring no cluster enters multiple folds, and predict across folds, to emulate the cluster-robust cross-fitting. I follow this method for cluster-robust cross-fitting in all later machine learning applications.

I implement a post-LASSO algorithm by substituting the unrestricted first-stage of propensity score prediction in the DGT approach with the LASSO (absent fixed effects), before running the unrestricted second-stage with the covariates that recovered non-zero coefficients from the LASSO and taking predicted values from that regression as the propensity scores.

I implement the Imbens-Rubin recursive stratification algorithm to create propensity score strata without researcher input. My implementation recursively splits strata for which a t-statistic on equal propensity score means between treatment and control groups is greater than 1.96, if and only if the resulting strata would contain (1) more than three treated and control units and (2) more than 23 units total ($K + 3$, where K is the number of covariates we are adjusting for [Imbens and Rubin 2015].

A.2.3 Event Study Design

First, consider the pooled Imbens-Rubin Subclassification version of my estimates. First, I compute a subclassification estimator in each given cohort-year, taking the stratum-size-weighted average of stratum difference in means estimates and computing cluster-robust standard errors. For the standard errors, first consider, within a given stratum, J_t clusters in the treatment group and J_c in the control group. The within-stratum variance estimate is then:

$$\hat{\sigma}^2_{s,q} = \frac{1}{J_t(J_t-1)}\sum_{j=1}^{J_t}\left(\bar{Y}_j - \bar{Y}_t\right)^2 + \frac{1}{J_c(J_c-1)}\sum_{j=1}^{J_c}\left(\bar{Y}_j - \bar{Y}_c\right)^2$$

where e.g. $\bar{Y}_j$ is the cluster average and $\bar{Y}_c$ is the average cluster average in the control group. The cohort-year variance estimate is therefore:

$$\hat{\sigma}^2_q = \sum_{s=1}^{S}\hat{\sigma}^2_s \cdot \left(\frac{J(s)}{J(q)}\right)^2$$

Given the above variance estimate as well as a resulting $\hat{\tau}_q$, we get the ATT and variance estimates pooled over Q cohorts as:

$$\hat{\tau} = \sum_{q=1}^{Q}\hat{\tau}_q \cdot \left(\frac{N(q)}{N}\right)$$

$$\hat{\sigma}^2 = \sum_{q=1}^{Q}\hat{\sigma}^2_q \cdot \left(\frac{J(q)}{J}\right)^2$$

Standard errors are computed completely analogously for AIPW estimates, substituting observed outcomes for AIPW scores:

$$\hat{\sigma}^2_q = \frac{1}{J(J-1)}\sum_{j=1}^{J}\left(\bar{\Gamma}_j - \bar{\Gamma}\right)^2$$

Treatment effects are also pooled analogously. The variances are valid under the assumption that, conditional on the propensity scores, each stratum treatment effect estimate is independent. There is an analogous argument for AIPW score independence. Each cohort-year estimate is then a consistent estimate of the cohort-year treatment effect τ_q, and the target estimand is a properly N-weighted average of cohort treatment effects. For all estimators, I construct Gaussian confidence intervals, appealing to theorems from class and from Imbens and Rubin (2015).

A.2.4 Heterogeneous Treatment Effects

Causal forests are trained at the cohort-year level and clustered at the case level, and variable selection is implemented as in my forest-AIPW estimators. Aggregation for the event-year level ATTs is as described in the previous section.

For the best linear predictor calibration test: I train causal forest estimates for a given cohort-year and then pool estimated CATEs across cohorts to get all estimated CATE scores in a given event-year. I recover also for each observation the estimated propensity score $\hat{e}(X_i)$ and outcome model $\hat{m}(X_i)$. I then regress:

$$Y_i - \hat{m}(X_i) = \bar{\tau}(W_i - \hat{e}(X_i)) + (\hat{\tau}(X_i - \bar{\tau})(W_i - \hat{e}(X_i)) + \epsilon_i$$

Where the above ignores cross-fitting superscripts (rest assured, all predictions are out-of-bag). Standard errors are HC3 heteroskedasticity-robust.

For the sub-group analysis: for a given cohort-year, I first divide the data into five folds such that no case appears in multiple folds. I then train a causal forest, clustered at the case level, on out-of-fold data, and use that model to make predictions on held-out folds (i.e. I train K causal forests for the K folds; to adjust for the computational intensity of this task, each forest is trained using 250 trees). On the held-out fold, I then assign the observations into quintiles of predicted CATEs. I then aggregate quintiles across cohorts. For each quintile, I then implement the pooled forest-AIPW estimator described in the previous section. The result is that I have subgroups identified by out-of-bag causal forest predictions, and for each subgroup an ATT estimate that is independent of the CATE scores used to identify the subgroups. I believe this is a valid approach, but am not certain. The implementation difficulty with respect to the tutorial (4.2.2.1) implementation is that, for clustered data, we cannot use the *causal_forest* cluster argument and must instead K -fold by hand; I then cannot recover out-of-bag propensity score and outcome predictions from the forests.