DISCUSSION PAPERS IN ECONOMICS

Working Paper No. 21-03

! "#$%&'()*'+)',-#."/ /'(('",(#$%&0+%#12*)'(2,#
3%**4/2,&%*',-5#67'&%,+%#8*"/#9*':",2#

Do Redistricting Commissions Reduce Partisan Gerrymandering? Evidence from Arizona

Loren Kruschke[*]

University of Colorado Boulder

Updated: February 11, 2022

Abstract

A growing number of states have implemented commissions in order to design political districts, in large part as a response to concerns about partisan gerrymandering. While a significant amount of work endorses the use of independent redistricting commissions in theory, very little research has analyzed the causal effects of implementing redistricting commissions. In this paper, I contribute to our understanding of the role redistricting institutions play in gerrymandering outcomes by evaluating how Arizona's independent redistricting commission affected gerrymandering outcomes in congressional elections. To this end, I examine election outcomes in Arizona between the years of 1982 and 2016; two full redistricting cycles before the commission was implemented, and over one and a half redistricting cycles afterward. I use a novel variant of the synthetic control method, a recently popularized empirical tool for generating plausible control groups when none naturally exist, to facilitate this analysis. I find some suggestive evidence that commission-based redistricting in Arizona may have reduced partisan gerrymandering. While my baseline results fall short of full statistical significance, there is also no evidence that Arizona's redistricting commission made partisan gerrymandering outcomes worse; at a minimum, it seems to have done no harm where gerrymandering is concerned.

Keywords: Reapportionment; Voting; Efficiency Gap; Synthetic Control

JEL: H70, K16, Y40

[*] I thank Murat Iyigun, Martin Boileau, Miles Kimball, and Taylor Jaworski for their guidance, insights, and support. I also thank Evelyn Skoy, Brachel Champion, Lauren Schechter, Kyle Butts, and participants of the University of Colorado Department of Economics Applied Microeconomics and Graduate Student seminars for their comments and feedback.

1 Introduction

As the decade begins in earnest, so too will a process central to American democracy: redistricting. During this procedure, states will leverage census data to determine how the boundaries that govern election districts should be drawn. Fundamentally, this is meant to ensure that citizens are afforded relatively equal voting power { though this is often untrue in practice. In most states, politicians draw and enact the maps that govern elections. As one might expect, this conflict of interest often results in maps meant to benefit some individuals at the expense of others (Levitt, 2008; Issacharoff, 2002; McDonald, 2004).[1] This process of strategically redrawing political districts is known as gerrymandering, and has been a fixture in the American political landscape since at least the early nineteenth century (Griffith, 1907).

Although gerrymandering is clearly at odds with normative ideals of equal representation central to the constitution, only some variants are explicitly illegal. For example, racial gerrymandering { which entails redrawing political boundaries to systemically disadvantage racial minorities { is prohibited by law. By contrast, partisan gerrymandering, which systematically advantages one political party at the expense of another, is not. In fact, the Supreme Court's 2018 decision in *Rucho v. Common Cause* explicitly recognizes that gerrymandering for the purposes of systemically disadvantaging political parties is outside the purview of federal courts. As such, partisan gerrymandering promises to continue to be a source of controversy for years to come.

Generally, state legislatures both draw and ratify the maps that govern their own elections. This results in clear conflicts of interest, and has led to hyper-partisan congressional political maps.[2] To combat this, scholars have suggested that states implement redistricting commissions to draw maps in place of the legislature (Kubin, 1996; Issacharoff, 2002). A growing number of states have responded to these concerns, and adopted some type of commission-based redistricting process. However, relatively little work has analyzed the causal effects of commissions on gerrymandering outcomes.

[1] In general, this might mean advantaging incumbents, certain demographics, etc. In this paper, I specifically evaluate how political maps might be drawn to benefit one American political party at the expense of another.

[2] For example, North Carolina state representative David Lewis (Rep.) endorsed constructing a political map \I think electing Republicans is better than electing Democrats...I propose that we draw the map to give partisan advantage to 10 republican and 3 democrats because I do not believe it's possible to draw a map with 11 republicans and 2 democrats." North Carolina has a nearly equal share of votes cast for republican and democrat congressional candidates. Of the thirteen congressional districts located in North Carolina, at least nine were won by republican candidates each election cycle from 2012 and 2018.

This paper investigates the link between the method by which states enact redistricting and gerrymandering outcomes in congressional elections, using Arizona as a case study. Arizona amended their constitution to enact redistricting through an independent commission in the year 2000. This a ected the way in which future political maps were constructed, starting in 2002. Prior to this change, maps were constructed and enacted by the Arizona state legislature. If the Arizona Independent Redistricting Commission (AIRC) functioned as intended, one would expect to see a decline in partisan gerrymandering beginning with the political maps constructed in the 2002-2010 redistricting cycle.

Relevant institutional details and data are detailed in sections 3 and lize commission-based redistricting are used to forecast counterfactual voting outcomes in Arizona.[3] A detailed description of the synthetic control method { and the SCUL variant { can be found in Appendix A.

Robustness checks re-run this analysis in a variety of settings. First, I restrict the variety of economic covariates used as potential components of the synthetic counterfactual. This is meant to address concerns that I might be including variables that are spuriously correlated with election outcomes, leading to biased results. Second, I truncate the post-treatment period to reflect only the map cycle immediately following treatment. This check is meant to address concerns about the method's ability to forecast results in the post-treatment period, given the number of pre-treatment observations available in the data. Third, I re-run the analysis using an alternative metric for partisan gerrymandering. This addresses concerns that partisan gerrymandering may be measured inappropriately. Results are qualitatively consistent across all robustness checks. The totality of this analysis finds marginally statistically significant evidence that the AIRC reduced partisan gerrymandering outcomes in Arizona. Still, because it does not obtain full statistical significance, some may not find this evidence compelling. In either case, it appears the AIRC did no harm where partisan gerrymandering is concerned.

Beyond evaluating gerrymandering outcomes in Arizona, this paper serves as a demonstration of how to implement the SCUL method and interpret its results. While the standard synthetic control method is well established within economics, neither it nor its variant, SCUL, have widespread application evaluating redistricting outcomes. Because of this, showcasing their application to political scientists and legal scholars may help proliferate a useful empirical tool across academic fields.

The SCUL method is particularly useful with regard to studies regarding state-level redistricting institutions, where most studies are descriptive. It may therefore be of use to scholars analyzing any consequence of redistricting commissions, be it gerrymandering or otherwise. Furthermore, the SCUL method { and, more generally, synthetic control { can potentially be applied to analyze any state-level policy. It is therefore likely of interest to legal scholars and political scientists at large.

The rest of this paper is organized as follows. Section 2 details related literature and this analysis' placement therein. Section 3 motivates Arizona's use as a case study for redistricting reform. Section

[3] This is done to ensure that predicted results are in no way impacted by redistricting commissions. Within the time period I analyze, California, Hawaii, Idaho, Montana, New Jersey, and Washington all implemented redistricting commissions, and so are not used to construct Arizona's synthetic control.

4 describes metric speci cs, identi cation concerns, data speci cs, and estimation technique. Section
5

useful predictive information about Arizona's counterfactual outcome.

3 Why Study Partisan Gerrymandering in Arizona?

America is unique among modern democracies in that it generally provides state legislatures authority over the redistricting process. Virtually every other democratic nation that enacts redistricting does so through the use of independent commissions (Stephanopoulos, 2013b). This is not merely an institutional oddity; power over state redistricting processes can determine the fortunes of political parties for an entire decade. Still, in 2000, Arizona amended its constitution via citizen initiative to enact redistricting through a commission of five non-politician members.[5] The Arizona Independent Redistricting Commission (AIRC) designs both state legislative and congressional districts, and is meant to prevent conflicts of interest that might arise from politicians designing the districts in which they are elected.

At the time the redistricting commission was implemented, Arizona was among six states which enacted redistricting of congressional maps through a commission.[6] That number has since grown to eleven states, as California, Colorado, Michigan, New York, and Virginia have passed similar measures in the last two decades. Given the increasing prevalence of commission-based redistricting reforms { and their stated objective of curbing political power { it is worthwhile to investigate their efficacy at deterring partisan gerrymandering. In this regard, Arizona represents an ideal case study for several reasons.

First, the timing with which Arizona passed its redistricting legislation enables researchers to evaluate gerrymandering outcomes in Arizona over the lifetime of several sets of political maps. This study examines election outcomes in Arizona between the years of 1982 and 2016; two full redistricting cycles before the commission was implemented, and nearly two full redistricting cycles afterward. This allows one to clearly determine post-treatment trends for a potentially noisy outcome variable, and runs in contrast to states which passed their legislation later. For example, California's redistricting commission first drew congressional maps that went into effect in 2012; available data would allow for analysis of less than one full life cycle of political maps following the commission's implementation.

Second, Arizona has contained a substantial number of congressional districts throughout the time period of this study. States with very few congressional districts tend to have noisy measures of partisan gerrymandering. At an extreme, states with one district have no defined gerrymandering metric, since redistricting does not take place in these states. These concerns most notably apply to Hawaii, Idaho, and Montana; all three states have commission-based redistricting systems, and two or fewer congressional districts during the lifespan of this study. In contrast, Arizona has contained an average of almost seven congressional districts throughout the time period of this study — and never fewer than five. This mitigates measurement concerns related to district quantity.

Third, Arizona's commission has been the target of backlash from state's majority party. Given that its implementation was due to a majority vote of the citizenry, and that the judicial branch has upheld its legality, this may suggest the majority party perceives it has affected election outcomes. Specifically, the AIRC chair was impeached in 2011 by the Republican-held governor's office. Removal from office was confirmed by a two-thirds vote in the state senate, where Republicans held 70% of the seats. Nonetheless, the Arizona Supreme Court ruled that the impeachment was improper, and reinstated the chair. Furthermore, the Arizona state legislature unsuccessfully sought to dissolve the AIRC in

On a more intuitive level, the *cracking di erential* measures the extent to which election outcomes deviate from representation proportional to voting outcomes. That is, the *cracking di erential* will favor a political party that wins a larger portion of congressional seats than their portion of the statewide vote. The logic underlying this is straightforward; if a party wins disproportionately more seats than votes, it must have distributed its votes more e ciently than the competing party. Because any e ective gerrymander must result in one party translating their votes into a disproportionately large seat share, large and enduring *cracking di erential* values indicate that a state is e ectively gerrymandered.[11]

Figure 1 shows how Arizona's *cracking di erential* has evolved over time. Vertical lines indicate political map life cylces, and the red vertical line indicates when the AIRC took e ect.[12] Here, there are two general trends that stand out.

First, prior to the AIRC's implementation, the *cracking di erential* generally takes negative values, indicating that election outcomes were biased in favors of Republicans. During the map cycle spanning the 1980s, ve states had *cracking di erentials* larger in magnitude than Arizona, on average. During the map cycle spanning the 1990s, seven states did.[13] Thus, the magnitude of the *cracking di erential* during the time period prior to the AIRC's implementation is suggestive.

There is one major exception to this trend in 1992, when two events coincided to ip typically Republican voters. First, Bill Clinton ran for o ce amid a national wave of Democrat support. Of 42 states with a de ned *cracking di erential* during to 1992 - 2000 map cycle, 33 had *cracking di erentials* more favorable for democrats in 1992 than their average *cracking di erential* over that decade. Second, Arizona gained a sixth Congressional seat in 1992, following redistricting. National pro-Democrat sentiment and a lack of a Republican incumbent competitor helped the Democratic candidate win this district. Following 1992, Republicans controlled this district for the remainder of the map cycle.

[11] It is worthwhile to note that a state can be e ectively gerrymandered even if unintended at the time of redistricting.

[12] Political map cycles begin in the second year of every decade (1982, 1992, etc.) and end on census years. Vertical lines are drawn in between the nal year of one map cycle and the rst year of the next. This is meant to avoid confusion that could arise if vertical lines coincided with the year values; it would not be obvious whether lines indicated the beginning or end of political maps cycles.

[13] During the 1980s, these states were: Georgia, Massachusetts, Nebraska, Utah, West Virginia. During the 1990s, these states were: Idaho, Iowa, Massachusetts, Nebraska, New Hampshire, Oklahoma, Rhode Island.

The *cracking differential*'s volatility is not inherently a shortcoming. Rather, it indicates that differences in partisan efficiency can shift in the face of changing political headwinds. Figure 1 shows that the Republican party consistently received a larger portion of political representation than votes prior to implementing the AIRC. However, as indicated by the spike in 1992, this advantage was not ironclad.

Lastly, it should be noted that the *cracking differential* is tailored to measure partisan gerrymandering, specifically. Other types of gerrymandering may not strictly follow partisan voting behavior, and so may not be captured by this metric. This is not to say the metric is flawed; rather, it is specialized. Because researchers must always make choices about how best to measure their outcome of interest, it is useful in the current context. Still, researchers should be careful about applying the *cracking differential* to measure other types types of gerrymandering.

4.2 Data

Because the synthetic counterfactual is constructed as a combination of relevant independent varihighest predictive power for election outcomes in Arizona.

Economic controls include state unemployment rate, per capita disposable income, and industry composition by state.[17] State industry controls are divided into 20 categories designed to match BEA industry employment reports. Each of these are likely to impact election outcomes in di erent ways, and may be contextually linked to individual states. As with other controls, I remain agnostic about the relationship between each economic control and election outcomes a priori, preferring instead to allow the SCUL method to make the determination empirically.

4.3 Estimating the Synthetic Control Group

Given the preceding discussion of data, it is prudent to brie y discuss how covariates are used to estimate the synthetic counterfactual. To avoid distracting from the research question at hand, I recount only the most important aspects of this process here. A more detailed explanation can be found in Appendix A.

The SCUL method operates by assigning a weight to each covariate, which determines its contribution to the synthetic control group. Speci cally, the synthetic control, y_t , is constructed as follows:

$$y_t = Y_{Dt}^{0} W_{SCUL}$$

where Y_{Dt} represents the vector of observed outcomes for each covariate in time period t.[18] Covariates are restricted to states without commission-based redistricting systems, and for which the *cracking di erential* is de ned for the study's entire time period.[19] SCUL method weights, W_{SCUL}, are lasso regression coe cients selected to minimize the di erence between the observed time series of interest and its synthetic control. Speci cally, weights are computed according to the following objective function:

$$W_{SCUL} = \arg\min_W \left(\sum_{t=1}^{T_{pre}} (y_{0t} \quad Y_{Dt}^{0} W)^2 + \quad |W|_1 \right)$$

Here, y_{0t} indicates Arizona's observed outcomes in period t of the pre-treatment period. This process

[17] This data relies on the recent work of Eckert et al. (2020) to construct consistent industry classi cations for the sample time period. Unemployment and income data are compiled from reports made publicly available through the BLS and BEA, respectively.

[18] The full group of covariates that may contribute to the synthetic control is known as the \donor pool," and so the vector describing their outcomes is denoted with the subscript \D".

[19]

variable for treatment status. Failing this, however, the synthetic control method mitigates these concerns by attempting to implicitly match on unobserved factors. This intuition here is straightforward: to the extent that unobservable factors (e.g., culture) drive outcomes in Arizona elections, the SCUL method must select donor series elements that match on those same factors in order to recreate Arizona's outcomes prior to treatment. Figure 4 illustrates Arizona's observed and synthetic *cracking differential* over the lifespan of this study. Synthetic outcomes closely match their observed counterparts during the pre-treatment period, providing suggestive evidence that the SCUL method selects donor elements that match on relevant unobserved factors.[21]

I now confront the potential that there exist simultaneity issues between partisan gerrymandering and AIRC implementation. Typically, these concerns follow two tracks. First, readers may be concerned that only states with low levels of gerrymandering are likely to enact commission-based redistricting reform, since only un-gerrymandered legislatures will pass such legislation. Because Arizona passed its gerrymandering legislation as a constitutional amendment through citizen initiative, the legislature neither proposed nor ratified the AIRC. Thus, partisan attempts to block commission-based redistricting through the legislature are not a major concern in the present context.

Following this line of reasoning, some may then be concerned that Arizona may have only been motivated to implement its commission through citizen initiative given a sufficiently high level of gerrymandering. This does not appear to be the case. Figure 3 expounds on this point by plotting the absolute value of the *cracking differential* for Arizona over the lifespan of the study. The absolute value of the *cracking differential* is useful because it indicates the magnitude of measured gerrymandering, regardless of partisan bias. The line tracking the magnitude of Arizona's measured gerrymandering is black prior commission implementation, and red thereafter.

Of 18 states which allow constitutional amendments via citizen initiative, four have enacted redistricting commissions (Arizona, California, Colorado, and Montana). The gray-filled area in Figure 3

[21] To make this point more explicit, I follow Hollingsworth and Wing (2020) by considering a setting in which untreated counterfactual outcomes are generated by a simple interactive fixed effects model. Namely: $y(0)_{st} = {}_{t}\ {}_{s} + {}_{st}$. Here, $y(0)_{st}$ are the synthetic outcomes for group s in period t, ${}_{t}$ is a 1 K vector of period-specific unmeasured variables, and ${}_{s}$ is a K 1 vector of group-specific coefficients. If the observed outcomes for the treated group are generated by $y(0)_{0t} = {}_{t}\ {}_{0} + {}_{0t}$, then the synthetic control method will match these outcomes in the pre-treatment period by selecting comparison units with values of ${}_{s}$ that are a close match for ${}_{0}$. Since ${}_{s}$ values are unobserved, this matching procedure is implicit; two time series with closely matching values of $y(0)_{st}$ are likely to also have closely matching values of ${}_{s}$. Still, if this matching process is successful then the synthetic counterfactual will effectively control for relevant unobservable factors when estimating the effect of treatment.

tual outcome is then computed as the product of weights and donor unit values in the post-treatment period. The main analysis utilizes all state-level variables detailed in Section 4.2 over all the years in the dataset.

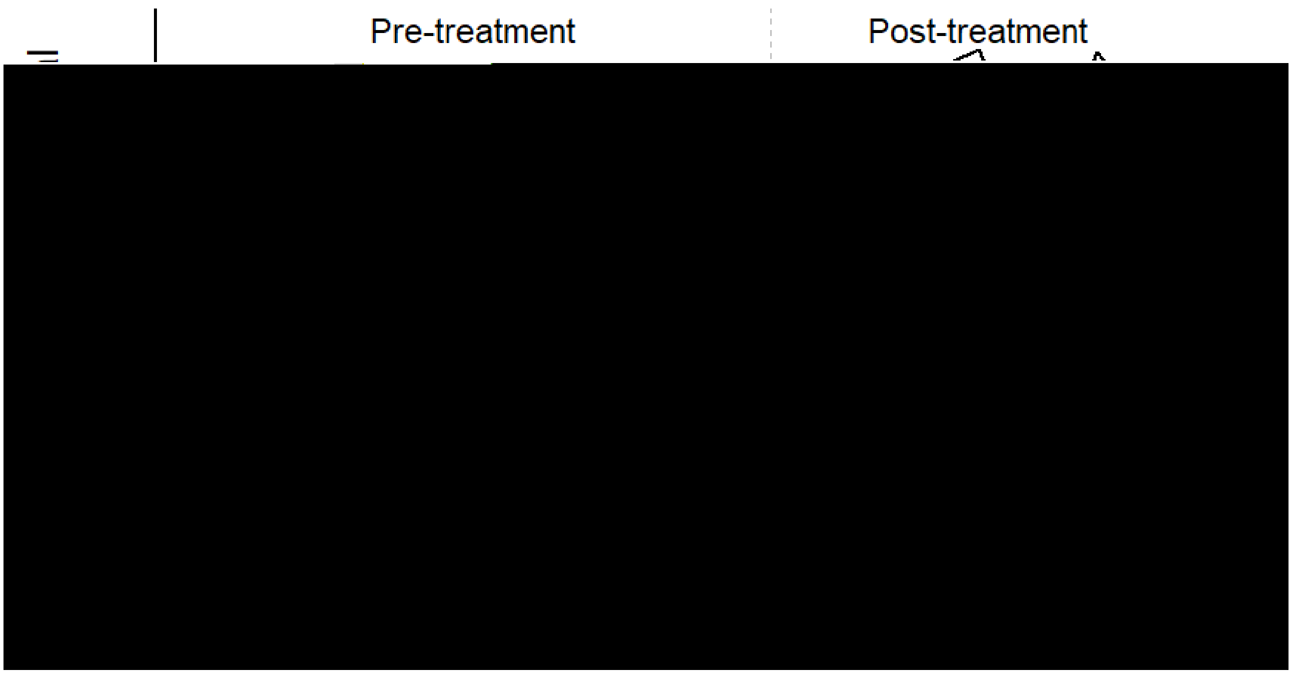
Baseline results present my ndings when using the full set of variables in my dataset, and following guidelines for model t suggested in the literature. I will show that this leads to concerns about the synthetic control's composition and statistical power, and address them in robustness checks. Still, presenting baseline results in this way emphasizes transparency. In robustness checks in Section 6, I diverge from standard practices only insofar as doing so enables me to address issues emphasized in this section.

5.1 Treatment E ect Estimates

Figure 4 depicts Arizona's observedcracking di erential and its synthetic counterpart. Encouragingly, the synthetic counterfactual produced by SCUL matches Arizona's observed outcomes well in the pre-treatment period. Per Hollingsworth and Wing (2020), model t is measured in terms of a modi ed version of Cohen's D. They suggest using a threshold of .25 for model t, meaning that only synthetic control groups with outcomes within a quarter of a standard deviation of the observed time series are used for analysis. Here, Cohen's D is .13 over the pre-treatment period, which is well within the threshold for model t.

Given the SCUL method's ability to accurately predict pre-treatment outcomes, the divergence between synthetic and observed outcomes in the post-treatment period is striking. The observed873h08(unit)--287(w

Figure 4: Arizona and its Synthetic Counterfactual



5.2 The Composition of the Synthetic Control

Given the preceding discussion on the e ect of AIRC implementation, it is prudent to examine the reports; Table 1 relays category composition along with their corresponding codes.

In general, these are variables one would expect to have significant impact on election outcomes; incumbency and unemployment rate effects have a long tradition of being used in related literature (see, for example, Lepper 1974, Hibbs Jr 1977 regarding unemployment; Abramowitz 1975, Krehbiel and Wright 1983 regarding incumbency). It also seems intuitive that Republican state house vote and seat share values in some states might have some predictive power for *cracking differential* outcomes in Arizona; national trends and coordinated partisan activity are likely to cause correlation in these outcomes.

The SCUL method presents an objective procedure for selecting variables that contribute to the synthetic control, and is preferable to alternatives that rely on researchers' subjective evaluations. Still, some may find the inclusion of industry employment shares questionable. Specifically, the SCUL method selects employment in Georgia's finance and insurance industry and employment in Maine's wholesale trade industry as holding predictive value for election outcomes in Arizona. On their face, these are not the most intuitive variables to select – though one can easily rationalize why they might be. For example, because Atlanta is a large financial hub it could very well be that employment in the finance and insurance correlates with national economic and political trends. Nonetheless, skeptics may not be convinced by ex-post rationalizations for these variables. To address this, I re-run this analysis while excluding state industry employment shares in Section 6.1. Specifics regarding this robustness check are relegated to Section 6.1; for now, it is enough to note that results are qualitatively unchanged.

Lastly, I examine the extent to which each included variable contributes to Arizona's synthetic control. Because the synthetic control is constructed using the product of the coefficients and corresponding characteristic levels, the share of the synthetic control that each characteristic comprises can vary from one time period to another. Coefficient values are reported in the right-most column, and reflect SCUL method weights (W_{SCUL}), as described in section 4.4. Figure 5 shows the share of the synthetic counterfactual comprised by each characteristic in the first and final prediction, which is meant to indicate how the synthetic control group's composition varies over time. In each column, shares sum to one. Each characteristic's relative importance and contribution the to synthetic control are generally stable between the first and final prediction. This means that each donor element seems to provide relatively stable predictive power within the synthetic control over time.[25]

Figure 5: Synthetic Arizona Composition

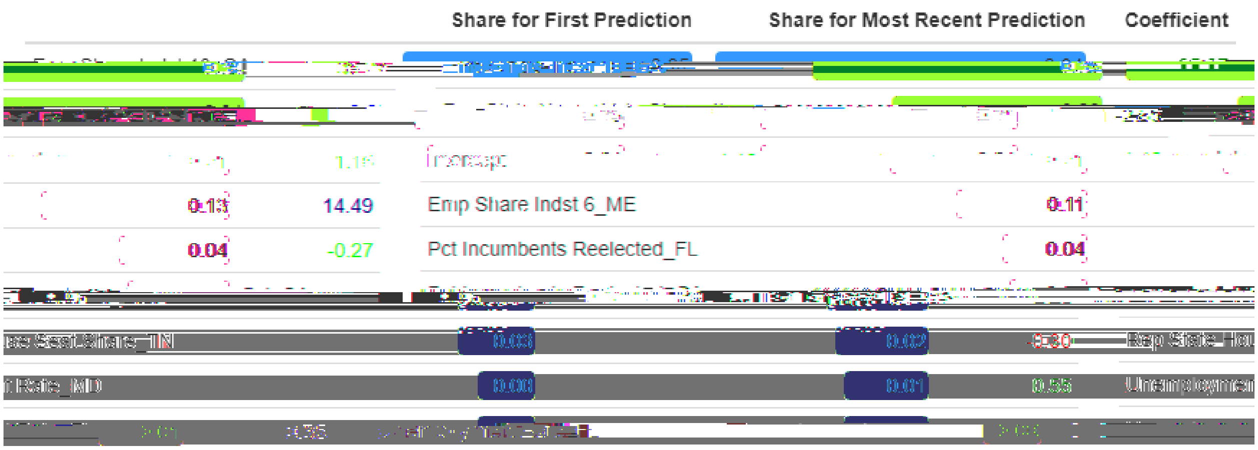


Table 1: Industry Employment Categories

Group	Industry	Group	Industry
01	Farm employment	12	Professional, scienti c, technical services
02	Mining, quarrying, oil and gas extraction	13	Enterprise management
03	Utilities	14	Administrative and support and waste management and remediation services
04	Construction		
05	Manufacturing		
06	Wholesale Trade	15	Educational Services
07	Retail Trade	16	Healthcare, social assistance
08	Transportation and warehousing	17	Arts, entertainment, recreation
09	Information	18	Accommodation, food services
10	Finance and Insurance	19	Other services (except govt. and govt. enterprises)
11	Real Estate, Rental, Leasing	20	Government, govt. enterprises

[25] This is noteworthy insofar as a synthetic control whose components' shares uctuate signi cantly may be suspect; if donor elements comprise vastly di erent shares of the synthetic control over time, one would need to provide a rationalization at the very least.

5.3 Statistical Inference

To determine whether the estimated treatment effect is statistically significant, it is compared to the estimated pseudo-treatment effects for all untreated placebo units. In this setting, placebo units are the *cracking differential* outcomes for all states included in this study.[26] In turn, the pseudo-treatment effects are used to construct the null distribution of outcomes one could expect to observe due to random chance, under the null hypothesis that implementing a redistricting commission has no effect. A statistically significant effect should be larger in magnitude than the pseudo-treatment effects in the measured in standard deviations during the pre-treatment period.[28] Of 31 potential placebo units, 11 survive for this analysis. One placebo has a larger estimated effect over the post-treatment period than Arizona, resulting in a p-value range of $(.08, .17]$. This contains the $.1$ threshold for marginal statistical significance. While this is clearly outside the $.05$ threshold required for full statistical significance, Arizona's rank as the second largest effect is suggestive.

Figure 6: Smoke Plot of Estimated Treatment and Pseudo-Treatment Effects

5.4 Statistical Power

Some of the pseudo-treatment effects shown in Figure 6 are quite large. This raises concerns about statistical power; it could be that forecasted results in untreated states are so noisy that I am unable to detect a true effect of AIRC implementation, if it exists. Because there are relatively few pseudo-treatment effects included in the null distribution, Arizona would need to be the largest effect in order taking an average value of 0.16 over that time span; Arizona would need to have election outcomes biased in favor of Democrats in order to register a statistically significant effect. Since the AIRC is intended to produce fair and balanced elections, we should not expect to observe election outcomes biased in favor of either party after its implementation, assuming it is performing effectively.

Table 2: Smoke Plot Effect Sizes and Fit

State	Post-Treatment Effect Size	Pre-Treatment Fit
Maryland	3.02	0.11
Arizona	2.26	0.13
Alabama	1.88	0.07
Tennessee	1.79	0.20
Iowa	0.91	0.02
Kansas	0.73	0.03
Indiana	0.57	0.02
South Carolina	0.47	0.07

employment share variables from the donor pool in the hope that it will improve the accuracy of post-treatment forecasts. In turn, this mitigates the magnitude of pseudo-treatment effects, allowing me to detect smaller treatment effects. This entails a trade-off: while post-treatment forecast accuracy may be improved, match quality during the pre-treatment period may also be degraded. This can result in some states being dropped from the analysis if their pre-treatment fit exceeds the .25 standard deviation threshold for model fit. In general, the f35(74 Td osomeetaten)-3ar8(d)1(n)-3inclnithold n analyral, thee

6.1 Excluding State Industry Composition

The rst robustness check restricts the set donor pool variables to exclude state industry composition. Figure 7 depicts Arizona's observed *cracking di erential* and its synthetic counterpart. Here, model t is improved during the pre-treatment period, and there is a slightly smaller divergence in post-treatment outcomes than in Figure 4. The synthetic control's post-treatment average *cracking di erential* is -0.52, leading to an estimated treatment e ect of 0.46. This would constitute a 88% decrease in measured gerrymandering over the post-treatment period. As before, while this e ect seems large at rst glance, it does not guarantee statistical signi cance.

Figuredepi-etpTJ/F30 34.3462 Tf 79.0.726 Td [(isizona'sdep26nd)-3

Figure 9: Smoke Plot of Estimated Treatment and Pseudo-Treatment E ects

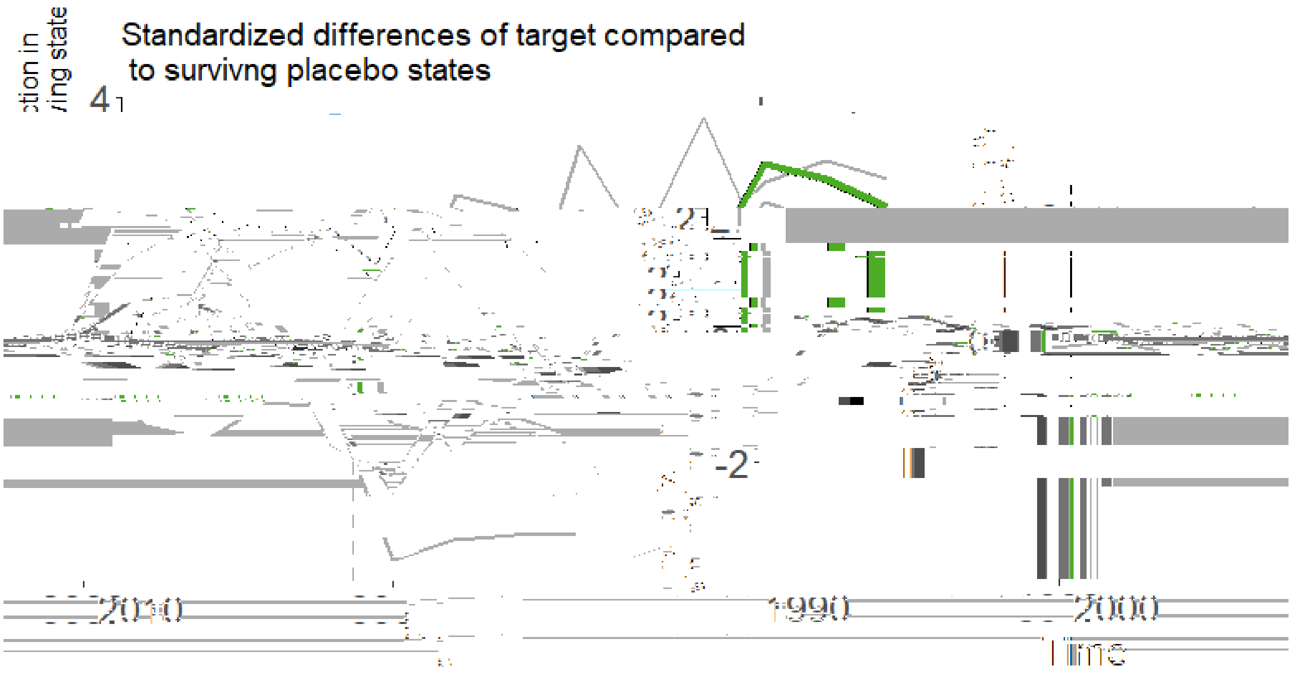


Table 3: Smoke Plot E ect Sizes and Fit

State	Post-Treatment E ect Size	Pre-Treatment Fit
Maryland	2.15	0.14
Arizona	1.80	0.04
Tennessee	1.75	0.24
Florida	1.27	0.16
Iowa	1.24	0.24
Oregon	1.16	0.16
Alabama	0.98	0.20
Louisiana	0.89	0.24
Georgia	0.33	0.15
Kansas	0.29	0.03

Note: E ect size and t are measured in terms of each state's pre-treatment standard deviation. Only states with Pre-treatment ts smaller than 0.25 are retained for the smoke plot.

Arizona again has the second largest effect in the smoke plot, which contains a total of 10 states. As such, its p-value falls in the range $(.1, .2]$. As before, Maryland has the largest effect, with a pseudo-effect of 2.15 pre-treatment standard deviations. This allows me to detect statistical significance for an effect size 29% smaller than in baseline results. Given that the synthetic control takes an average value of -0.52 during the pre-treatment period, Arizona's observed outcomes would need to take an average value of 0.01 during the post-treatment period to reach statistical significance.[30] This would indicate a lack of bias in favor of either party, and is close to what is actually observed in Arizona during the post-treatment period. This indicates that statistical power is not so lacking that detecting statistical significance would require an impossibly large treatment effect.

Still, because relatively few states are contained in the smoke plot, only the largest measured effect can be measured as even marginally statistically significant; any rank lower than 1/10 results in a p-value

Table 4: Smoke Plot Effect Sizes and Fit

State	Post-Treatment Effect Size	Pre-Treatment Fit
Maryland	2.15	0.14
Arizona	1.80	0.04
Tennessee	1.75	0.24
Florida	1.27	0.16
Iowa	1.24	0.24
Minnesota	1.14	0.38
Mississippi	1.10	0.31
Oregon	1.16	0.16
Alabama	0.98	0.20
Louisiana	0.89	0.24
Georgia	0.33	0.15
Kansas	0.29	0.03
Kentucky	0.20	0.29
Massachusetts	0.12	0.29
Oklahoma	0.02	0.40

6.2 Truncating the Post-Treatment Period

The second robustness check truncates the post-treatment period so that it ends in 2006. This means that the SCUL method need only forecast 3 time periods of election outcomes, equivalent to just over half a redistricting cycle. Moreover, the testing and forecasting periods are balanced, which is in line with recommendations made by Hollingsworth and Wing (2020). This improves confidence in forecasted outcomes, but entails a trade off: if the effect of AIRC implementation grows over time, truncating the post treatment period may impede my ability to capture its entire effect. Given that Arizona's cracking differential takes a few election cycles to move towards zero after AIRC implementation, this concern is relevant.[33] Still, it is useful to determine whether a detectable treatment effect

[33] For example, it could be that Republican representatives benefited from incumbency advantages in the early 2000s, which dissipated as they retired or voter sentiments changed. This would bias election results in favor of Republicans even if congressional districts were drawn in an unbiased way, leading to a treatment effect that grows over time.

of post-treatment data than examined here), in order to match the four pre-treatment periods accurately predicted by the SCUL method. In this case, Arizona is again the second largest e ect measured.

Figure 13: Smoke Plot of Estimated Treatment and Pseudo-Treatment E ects

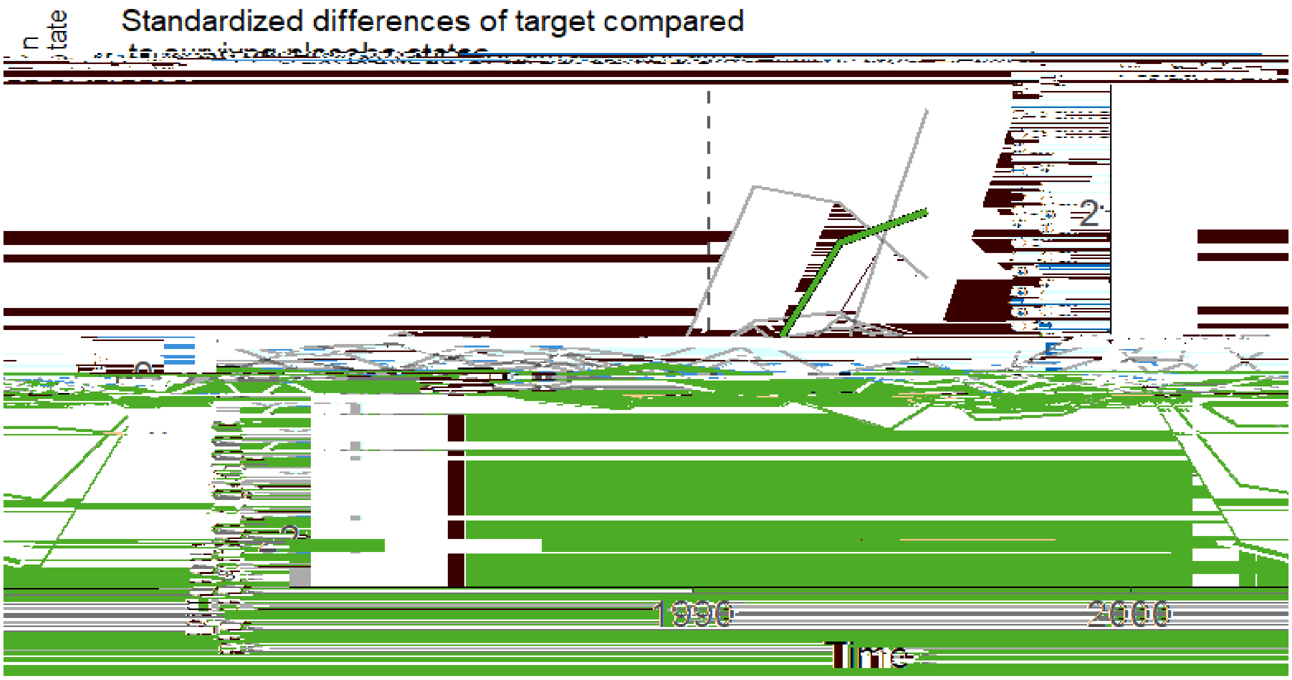


Table 5: Smoke Plot E ect Sizes and Fit

State	Post-Treatment E ect Size	Pre-Treatment Fit
Maryland	1.88	0.15
Minnesota	1.68	0.24
Arizona	1.26	0.04
Florida	1.11	0.03
Iowa	1.05	0.24
Alabama	0.63	0.20
Georgia	0.45	0.15
Kansas	0.44	0.03
Oklahoma	0.30	0.10
Louisiana	0.19	0.24
Tennessee	0.18	0.24

The totality of this robustness check is generally aligned with previous results. Arizona is among the larger treatment effects estimated, but is not statistically significant. Treatment effect estimates are more credible over the shorter time period examined, but may mitigate the magnitude of the estimated treatment effect if it grows over time.

6.3 Measuring Gerrymandering Using the Standard efficiency gap, EG_{McGhee}

The third robustness check re-runs the primary analysis in section 4 using the standard efficiency gap, EG_{McGhee} (McGhee, 2014; Stephanopoulos and McGhee, 2015). The *cracking differential* is this study's preferred metric because it provides consistent measures for gerrymandering, even when partisan vote shares are highly imbalanced. The more partisan vote shares are imbalanced, the more EG_{McGhee} will favor the majority party; at an extreme, EG_{McGhee} will always find a party which receives more than 75% of the statewide vote to be the *victim* of gerrymandering. In Arizona's case, congressional vote shares were typically most skewed in favor of Republicans during the 80s and 90s. During these decades, the Republican party typically received between 55% and 60% of the bipartisan vote, and on average more than 58%. This imbalance has the potential to skew measured gerrymandering in favor of democrats during the time period in question. Still, many may find it valuable to approach this issue using a more established metric than the *cracking differential*.

As a reminder, the SCUL method chooses which donor variables are assigned non-zero weight by using rolling-origin cross-validation to select a value. Unfortunately, the cross-validated results in poor model fit; Cohen's D during the pre-treatment period is larger than the 0.25 threshold for model fit. As before, the SCUL method is modified to iteratively select the next lowest value from the pool of generated values until the synthetic control group meets the Cohen's D threshold for model fit, or all values are exhausted. In this case, the lowest value out of the pool of generated values induces model fit during the pre-treatment period (Cohen's D = 0.05). Again, a warning is in order: this has the potential to overfit the data. Nonetheless, evaluating a suspect robustness check is likely preferable to having no robustness check at all.

Figure 14 depicts Arizona's observed value for EG_{McGhee} alongside its synthetic counterpart, given a sufficiently small value. Post-treatment, there is again an estimated reduction in gerrymandering, as measured by EG_{McGhee}. However, further analysis suggests that the model is indeed fitting on noise. Analysis of Figures 16 and 15 expounds on this point.

Figure 15 displays the structure of the synthetic control group in this robustness check. As with

Figure 15: Synthetic Arizona Composition

	Share for First Prediction	Share for Most Recent Prediction	Coefficient
Intercept	0.44	0.46	0.64
Pct Incumbents Reelected_FL	0.27	0.25	-0.48
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
0.06 -0.17 Rep State House Vote Share_NV	0.06		
-0.16 Standard Efficiency Gap_CO	0.02		0.01
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
0.00 -0.06 Pct Effectively Uncontested_MS	0.00		
0.03 0.53 Pct Effectively Uncontested_NY	0.00		

synthetic control does indeed t the observed trend based on noise; the inclusion of extra donor

References

Alberto Abadie and Javier Gardeazabal. The economic costs of con ict: A case study of the basque country. *American economic review*, 93(1):113{132, 2003.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the e ect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493{505, 2010.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495{510, 2015.

Alan I Abramowitz. Name familiarity, reputation, and the incumbency e ect in a congressional election. *Western Political Quarterly*, 28(4):668{684, 1975.

Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3{32, 2017.

Marianne Bertrand, Esther Du o, and Sendhil Mullainathan. How much should we trust di erences-in-di erences estimates? *The Quarterly journal of economics*, 119(1):249{275, 2004.

Bruce E Cain, Wendy K Tam Cho, Yan Y Liu, and Emily R Zhang. A reasonable bias approach to gerrymandering: Using automated plan generation to evaluate redistricting proposals. *Wm. & Mary L. Rev.*, 59:1521, 2017.

Jamie L Carson and Michael H Crespin. The e ect of state redistricting methods on electoral competition in united states house of representatives races. *State Politics & Policy Quarterly*, 4(4):455{469, 2004.

Wendy K Tam Cho. Measuring partisan fairness: How well does the e ciency gap guard against sophisticated as well as simple-minded modes of partisan discrimination. *U. Pa. L. Rev. Online*, 166:17, 2017.

Arindrajit Dube and Ben Zipperer. Pooling multiple case studies using synthetic controls: An application to minimum wage policies. 2015.

Fabian Eckert, Teresa C Fort, Peter K Schott, and Natalie J Yang. Imputing missing values in the us census bureau's county business patterns. Technical report, National Bureau of Economic Research, 2020.

Elmer Cummings Gri th. *The rise and development of the gerrymander*. Scott, Foresman, 1907.

Douglas A Hibbs Jr. Political parties and macroeconomic policy. *The American political science review*, pages 1467{1487, 1977.

Alex Hollingsworth and Coady Wing. Tactics for design and inference in synthetic control studies: An applied example using high-dimensional data. *Available at SSRN 3592088*, 2020.

Samuel Issacharo . Gerrymandering and political cartels. *Harv. L. Rev.*, 116:593, 2002.

Keith Krehbiel and John R Wright. The incumbency e ect in congressional elections: A test of two explanations. *American Journal of Political Science*, pages 140{157, 1983.

Loren Dean Kruschke. Measuring partisan e ciency in redistricting. forthcoming.

Je rey C Kubin. Case for redistricting commissions. *Tex. L. Rev.*, 75:837, 1996.

Susan J Lepper. Voting behavior and aggregate policy targets. *Public Choice*, 18(1):67{81, 1974.

Justin Levitt. A citizen's guide to redistricting. *Available at SSRN 1647221*, 2008.

Eric Lindgren and Priscilla Southwell. The e ect of redistricting commissions on electoral competitiveness in us house elections, 2002-2010. *J. Pol. & L.*, 6:13, 2013.

Seth E Masket, Jonathan Winburn, and Gerald C Wright. The gerrymanderers are coming! legislative redistricting won't a ect competition or polarization much, no matter who does it. *PS: Political Science and Politics*, pages 39{43, 2012.

Michael P McDonald. A comparative analysis of redistricting institutions in the united states, 2001{02. *State Politics & Policy Quarterly*, 4(4):371{395, 2004.

Eric McGhee. Measuring partisan bias in single-member district electoral systems. *Legislative Studies Quarterly*, 39(1):55{85, 2014.

Gary F Moncrief, Barbara Norrander, and Jay Wendland. *Reapportionment and Redistricting in the West*. Lexington Books, 2011.

Nicholas O Stephanopoulos. The consequences of consequentialist criteria. *UC Irvine L. Rev.*, 3:669, 2013a.

Nicholas O Stephanopoulos and Eric M McGhee. Partisan gerrymandering and the e ciency gap. *U. Chi. L. Rev.*, 82:831, 2015.

Arizona State Legislature v. Arizona Independent Redistricting Commission. 576 U.S. 35 (2015).

Gregory S Warrington. Quantifying gerrymandering using the vote distribution. *Election Law Journal*, 17(1):39{57, 2018.

A Implementation and Inference Under Synthetic Control Using Lasso Regression

A.1 The Synthetic Control Method

This paper utilizes a variant of an established method in applied microeconomics, but not common to the literature surrounding gerrymandering. It is therefore important to provide an overview of both the standard synthetic control method (Abadie and Gardeazabal, 2003), and its more recent variant, the SCUL technique (Hollingsworth and Wing, 2020). The synthetic control technique is used for causal analysis when one (or a few) groups undergo a policy change, but no counterfactual exists in nature. It operates by creating a plausible counterfactual that \looks like" the treated group during the pre-treatment period. This is done by creating a weighted combination of untreated units such that the outcome value, and some set of predictive variables, closely match those of the treated group. Researchers can then determine whether the policy change was e ective by examining the extent to which synthetic and observed outcomes diverge, after it goes into e ect.

Abadie et al. (2015

group, and X_D represent the K N matrix of statistics of interest for each unit in the donor pool. In Abadie et al. (2015), there were ve statistics of interest and 16 OECD nations in the donor pool; thus, X_0 would be a 1 16 vector and X_D would be a 5 16 matrix in its context.

Given this setup, one must then de ne two sets of weights. First, one de nes weights for each donor characteristic. Then, one must de ne weights for each donor unit. For this purpose, let V be the K K positive semi-de nite matrix of characteristic weights.[34] Furthermore, let W be the N 1 vector of weights for units in the donor pool. Elements in W must be non-negative and sum to one. The synthetic control outcome is then computed for each time period, t, as:

$$y_t$$

without its drawbacks. Chief among these for our purposes is that its inability to assign negative weights means that untreated units with trends that "mirror" the treatment group are underweighted or omitted entirely from the synthetic control. This removes information from the synthetic control that might otherwise provide a more realistic counterfactual.

A.2 The SCUL Technique

Hollingsworth and Wing (2020) propose a variant of the standard synthetic control method that is adopted for this study. Because it is a recent innovation, this section will closely follow their own explanation of the method. The key difference between SCUL and the standard method is that SCUL provides an alternative method for choosing the weights on time series elements which comprise the synthetic controls. The primary benefit this method provides is that it allows for negative weights. Negative synthetic control weights are particularly useful in this context because factors that are negatively correlated with Republican gerrymandering are likely to be useful in constructing a synthetic counterfactual (i.e., factors that predict a positive, rather than negative, cracking differential). To achieve this, they suggest using a lasso regression framework to generate weights. This is dubbed "Synthetic Control Using Lasso" (SCUL).

Given this framework, a brief overview of lasso regression is in order. Lasso regression operates by minimizing the sum of squared residuals in the same way as OLS regression, but adds a penalty term that increases with the magnitude of coefficients. Specifically, SCUL computes weights as follows:

$$W_{SCUL} = \arg\min_W \left(\sum_{t=1}^{T_{pre}} (y_{0t} - Y_{Dt}^{0} W)^2 + \lambda |W|_1 \right) \tag{4}$$

where $|W|_1$ is the sum of the absolute values of the coefficients associated with each variable in the donor pool. The penalty parameter reduces the magnitude of all coefficients, and, at an extreme, will reduce them to zero. When the penalty parameter, λ, is zero, coefficients are unpenalized and lasso is analogous to OLS regression. At the other extreme, when $\lambda = \infty$ all coefficients are reduced to zero.[36] In general, lasso will reduce some coefficients to zero, while mitigating the magnitude of those that survive.

This is useful in several ways. First, because several coefficients may be set to zero, it allows for

[36] In general, λ need only be sufficiently large for this to be the case.

estimation even when the number of predictive variables exceeds the number of observations. Second, this method allows \the data to do the talking" when researchers are unsure which predictive variables

Figure A1: Rolling-Origin Cross-Validation Visualization

A.4 Synthetic Control Weights Using SCUL

Equation (1

treated group is $ATT = \frac{1}{(T - T_{pre} - 1)} \sum_{t=T_{pre}+1}^{T} (y_{st} - \hat{y}_{st})$, where T is the final time period in the post-treatment period.

The estimated treatment effect need not be estimated over the entirety of the post-treatment period. Because the synthetic control's predictive ability deteriorates as it becomes further removed from the onset of treatment, in some settings it may be preferable to restrict estimation to a subset of data closely following treatment. Alternatively, researchers may be interested in estimating the treatment effect in individual years throughout the post-treatment period. Decisions about how to best estimate treatment effects are largely contextual, and left to researchers' discretion. This study utilizes the entire post-treatment period for such calculations.

A.7 Statistical Inference

To test whether the ATT is statistically significant, one must ascertain whether it is likely to have occurred due to chance alone. To accomplish this, Hollingsworth and Wing (2020) utilize placebo tests, which are employed throughout the synthetic control literature and beyond (Abadie et al., 2010; Dube and Zipperer, 2015; Bertrand et al., 2004). Specifically, they compute a distribution of placebo ATT estimates from untreated states. These act as the distribution of outcomes one would expect to find if treatment had no effect. Given this null distribution, one compares the absolute value of the standardized ATT estimate to the absolute values of the standardized placebo ATT estimates. This constitutes a rank-based, two-sided test of statistical significance, where the p-value is the rank of the estimated ATT within the placebo distribution in fraction form. In tests with relatively few placebo units, it may be preferable to report the p-value as a range. For example, in tests with one treatment group and nine placebo units, when the treated unit has the largest estimated effect size its rank is 1/10. Transparency dictates that the p-value be reported as existing in the range $(0, .1]$ (as opposed to a single point). Following this logic, p-values are reported as a range of potential values in this study.

When constructing the distribution of placebo outcomes, researchers must carefully distinguish between variables included as donor series and variables included as placebos. In this study, each element in the pool of donor variables is a predictive variable for election outcomes (e.g., state racial composition). Notably, gerrymandering outcomes in some states are likely to have predictive value for gerrymandering outcomes in others, and so are included in the pool of donor variables. Meanwhile, the outcome variable of interest is the gerrymandering metric for the state of Arizona. Placebo effects should therefore only evaluate gerrymandering outcomes in other states; it would not make sense to compare the ATT for Arizona gerrymandering to a placebo e ect on other donor variables, like state racial composition in New Mexico. This illustrates that it is generally unwise to treat the entire pool of donor variables and placebo variables as interchangeable. In this setting, only the subset of donor variables that are directly comparable to the outcome variable have use as placebos. In general, there may be no overlap between placebo and donor variables whatsoever.[40]

After determining which variables should be included in the pool of potential placebos, one should determine whether these variables' synthetic estimates t observed outcomes su ciently well for use create the synthetic control runs counter to this goal, and confounds analysis. To protect against this, donor variables any state that implemented a redistricting commission are eliminated. In general, it is suggested that researchers pursue similar a similar strategy when estimating the ATT in their own work.