DISCUSSION PAPERS IN ECONOMICS

Working Paper No. 21-03

Do Redistricting Commissions Reduce Partisan Gerrymandering? Evidence from Arizona

Loren Kruschke
University of Colorado Boulder

October 27, 2021
Revised November 5, 2021
Revised February 11, 2022

Department of Economics

University of Colorado Boulder
Boulder, Colorado 80309

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 affected 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 4, respectively. Still, a brief overview is useful here as a primer. Arizona contained five congressional districts in 1982, steadily increasing to nine districts in 2012. Congressional elections occur every two years, and so this data set spans 18 elections; ten prior to the AIRC's implementation and eight thereafter. This study uses publicly available data from 1982 - 2016 to match and forecast gerrymandering outcomes in Arizona's congressional elections.

In this regard, I employ a broad array of demographic, economic, and political variables that all from all plausibly impact election outcomes. These variables include: race, sex, age, birthplace, and education, state congressional seat count, percentage of elected congressional candidates that were incumbents (both in total and by party), state house vote share by party, state house seat share by party, measured gerrymandering by state, state unemployment rate, per capita disposable income, and industry composition by state. I remain agnostic about the degree to which any variable contributes to the predicted outcomes, instead utilizing lasso regression as an objective method to select variables.

More specifically, I utilize a variant of the synthetic control method, synthetic control using lasso regression, or SCUL, in order to conduct this analysis (Hollingsworth and Wing, 2020). The standard synthetic control method (Abadie and Gardeazabal, 2003) is heavily utilized by economists for causal inference, and has been described by Athey and Imbens (2017) as "arguably the most important innovation in the policy evaluation literature in the last 15 years." The synthetic control method is typically used when researchers observe time series data on a treated unit and many untreated units. A weighted combination of untreated units are used to construct a synthetic version of the treated group in which the synthetic control group matches the treatment group as closely as possible during the pre-treatment period. This is meant to generate a counterfactual for the treated group, where none naturally exists, in order to facilitate casual analysis. In this regard, only states that do not uti-

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 specifics, identification concerns, data specifics, and estimation technique. Section 5 introduces baseline results, including the estimated average treatment effect, statistical inference, and construction of the synthetic control group. Section 6 details a series of robustness checks as alternatives to baseline results. Section 7 concludes.

2 Related Literature

This study contributes to our understanding of the role redistricting institutions play in gerrymandering outcomes. It applies statistical techniques common in economics to a phenomenon typically scrutinized by legal scholars and political scientists. Given its interdisciplinary placement, this paper primarily operates at the intersection of three bodies of literature.

First, this paper is tied to studies measuring gerrymandering (Stephanopoulos and McGhee, 2015; Warrington, 2018; Cain et al., 2017). There are many potential ways gerrymandering could be measured, and this study uses a variant of the *efficiency gap* (Stephanopoulos and McGhee, 2015). The *efficiency gap* is a recently popularized metric that is tailored to measuring partisan gerrymandering, specifically (as opposed to, say, racial or incumbency gerrymandering).

This study prioritizes a metric known as the *cracking differential* (Kruschke, forthcoming). The *cracking differential* maintains most of the core features of the *efficiency gap*, and so is qualitatively similar in most cases. However, it is robust to criticisms about the *efficiency gap*'s inability to consistently measure gerrymandering – particularly in the face of extreme partisan vote splits. In Arizona's case, it is generally less volatile than the *efficiency gap*, and provides a more consistent measurement across states and election years. This is particularly useful in the current setting, where the SCUL method will potentially use *cracking differential* outcomes from some states to predict *cracking differential* outcomes in others. Still, a full discussion of the two metrics is beyond the scope of this section; metric specifics are discussed in section 4.1.

Second, this paper is linked to literature analyzing the consequences of redistricting commissions. A significant body of work exists in this field, with generally mixed results. Most find that independent commission redistricting results in more competitive U.S. house elections (Carson and Crespin, 2004; Stephanopoulos, 2013a; Lindgren and Southwell, 2013), though some dissenting scholars find the opposite to be true (Masket et al., 2012). Still, concerns about how competition is measured

– and its link to partisan gerrymandering – persist. To be sure, a noteworthy amount of academic work endorses the usage of independent political commissions, in theory (Kubin, 1996; McDonald, 2004; Issacharoff, 2002). Nonetheless, little work has been done to substantiate a causal link between partisan gerrymandering (as opposed to competition) and redistricting methods at an empirical level.

This paper's results largely align with the existing body of work in this field, but it attempts a more rigorous causal analysis. Most of the studies examining the effects of redistricting methods utilized a pre-post research design. That is, they evaluate some measure of competitiveness or gerrymandering before before and after implementation of a commission redistricting system (Moncrief et al., 2011). While this is a useful endeavor, it is vulnerable to criticisms that unobserved factors may confound results. One strategy scholars use to resolve this is to search for a valid counterfactual to use for comparison. In this context, it means searching for a group or region that represents what would have happened in Arizona had it not enacted the its independent redistricting commission. Clearly, this is a difficult task to accomplish.

Third, this paper relies on empirical techniques used for assessing causality when no counterfactual exists in nature. The synthetic control method (first introduced by Abadie and Gardeazabal, 2003) is at this point well established as a tool to conduct causal analysis in the economic literature (see Abadie et al., 2010 and Abadie et al., 2015 for example on implementation and usefulness of the synthetic control method). In essence, it creates a counterfactual for the group of interest by creating a synthetic composite of untreated groups. In our context, this means creating a "synthetic Arizona" out of weighted a combination of other states. The synthetic control method is potentially quite useful to scholars in this field. In theory, it can be applied to any region-specific policy (e.g., state legislation) that requires causal analysis – provided that identification requirements are met. Synthetic control method specifics are discussed in section Appendix A.

This study utilizes a recent innovation in the synthetic control method that allows scholars to utilize time series data with trends that "mirror" the target series (Hollingsworth and Wing, 2020). As before, method specifics are discussed in Appendix A. For now, it is enough to note that this is useful because the *cracking differential* can take positive or negative values, depending on which party is gerrymandered. For example, states with Democratic majorities may have levels of gerrymandering similar to Arizona, but trends that are opposite in sign.[4] These opposing trends potentially provide

[4]In general, I remain agnostic about how any state's predictors align with gerrymandering outcomes in Arizona,

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

regardless of partisan majority. Instead, I defer to the SCUL method's results when making these evaluations.

[5]Information on the details regarding the formation and function of Arizona's independent redistricting committee can be found on its website: https://irc.az.gov/.

[6]Other states that, in principle, utilized redistricting commissions to draw congressional districts starting in 2002 include: Hawaii, Idaho, Washington, Montana, and New Jersey. Commission type and utilization differ by state. To understand differences in commission utilization and composition, two examples are useful. First, Montana had only one congressional district in 2000, and so did not redraw congressional districts in that year. Since congressional redistricting did not occur in Montana in this year, they did not utilize their commission. Second, unlike Arizona, New Jersey utilizes a politician commission – meaning it may be comprised of politicians. This may cause some to question how removed New Jersey's commission's motives are from political interests. On the surface, then, Arizona's independent commission is likely a more ideal candidate for case study than New Jersey's politician commission – at least where partisan gerrymandering is concerned.

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 2014. Courts have repeatedly upheld its constitutionality, culminating in a 2015 Supreme Court case that rebuked the legislature's efforts (Arizona State Legislature v. Arizona Independent Redistricting Commission).

Beyond the legal concerns laid to rest by courts, there is little evidence of partisan gerrymandering following the AIRC's implementation. Why, then, would the legislative and executive branches exert so much effort to resist its influence? At a minimum, Arizona's majority party seems displeased with this institutional change. This suggests that the AIRC may have successfully affected the balance of political power within Arizona, making the state an ideal candidate for empirical investigation.

4 Measurement, Identification, Data, And Estimation

4.1 Measurement

Despite gerrymandering's pronounced role in American politics, there is still no consensus on how it should be measured. One recent and popular measure is the *efficiency gap* (Stephanopoulos and McGhee, 2015). The *efficiency gap* is a useful indicator variable, and has even been featured prominently in *Gill v. Whitford*, a Supreme Court case regarding partisan gerrymandering in Wisconsin. In spite of this, it has garnered considerable criticism for its inability to consistently measure gerrymandering – particularly in states with few congressional seats or extreme partisan vote splits (Cho, 2017). Kruschke (forthcoming) proposes a modified version of the *efficiency gap* that is robust to these concerns. It is named the "cracking differential," and constitutes the dependent variable in the succeeding analysis. Given this framework, a brief description of the two metrics is in order.

Both the *efficiency gap* and *cracking differential* are based on the notion that partisan gerrymandering occurs through two mechanisms: "packing" and "cracking." Packing entails concentrating one party's voters in a few districts, so that they win only a few elections by extreme margins. Cracking entails diluting a party's voters over many districts, so that they lose many elections by narrow margins. Packing results in what are known as "surplus votes"; votes in excess of the 50% margin necessary for victory. Cracking results in "lost votes"; votes cast for a losing candidate. Since both surplus and lost votes do not contribute to a political victory, they are considered inefficient, or wasted. The *efficiency gap* and *cracking differential* both measure the difference in wasted votes between parties; the larger and more enduring the imbalance, the more evidence there is of gerrymandering.

Nonetheless, there are important differences between the two metrics. In the *efficiency gap*, the amount of measured packing and cracking depends on the share of the statewide vote each party receives. It turns out that definitions of surplus and lost waste imply that the majority party in every state *must* attribute some of its vote to surplus waste.[7] Moreover, this minimum amount of surplus waste increases with the vote share of the majority party. The *efficiency gap* evaluates the difference in total wasted votes (both surplus and lost) between parties, and does not account for the majority party's

[7] To see this, note that only the first 50% of the vote attributed to the winning candidate in any election contribute to their victory. The remaining 50% of the vote is attributed to lost and surplus waste. Summed over all elections, 50% of the total vote in a state must be attributed to waste. Because the minority party receives less than 50% of the statewide vote, it cannot account for the full amount of wasted votes. The remainder of wasted votes must be attributed to the majority party.

minimum surplus waste. This has the potential to induce misleading results, and obfuscates interpretation. Just as the majority party's vote share differs across states and years, so too does the range of potential values the *efficiency gap* can take. This makes comparisons across states and years difficult.

The *cracking differential* avoids this issue. It boils down to the difference in each party's relative amount of lost votes, *after* accounting for both parties' minimum waste.[8] This ensures that the *cracking differential* takes a consistent range of values in all elections. For this reason, the *cracking differential* allows for clear comparisons of election outcomes across states and years, and is the preferred metric for this study. A detailed description of the cracking metric can be found in Kruschke (forthcoming).

The *cracking differential* takes a value between -1 and 1. Negative values indicate Republican bias, positive values indicate Democrat bias, and values close to zero indicate a lack of bias for either party.[9] It is undefined when states have only one district, or very few districts and an extreme partisan vote split. This is a desirable feature of the metric. When a state contains only one congressional district, it spans the entire state; congressional redistricting does not occur, and gerrymandering is impossible. Likewise, when a state's vote split is extremely unbalanced and in favor of one political party, only that party's candidates can be elected. This occurs regardless of how political maps are drawn, rendering partisan gerrymandering meaningless. For example, in 2016 Hawaii had two congressional districts and a partisan vote split of 79%-21% in favor of Democrats. Because of this, the two democrat candidates were guaranteed victory no matter how political districts were drawn; gerrymandering could not have impacted election results.[10]

[8] In its full form, the *cracking differential* calculates each party's relative quantity of votes attributed to *both* surplus and lost waste in excess of their minimum values. This is done separately for each type of waste in order to form measures of packing and cracking, individually. These measures are then summed for each party, forming a measure of total vote waste for each party. The *cracking differential* is the difference between partisan vote waste metrics. It turns out that both parties' packing metric must be identical, and so what remains when evaluating the difference in partisan waste is the relative difference in lost votes between parties (hence the name "*cracking differential*"). For a more detailed explanation, see Kruschke (forthcoming).

[9] The relation between the *cracking differential*'s sign and partisan bias is arbitrary; one could construct it so that negative values indicate Democrat bias and positive values indicate Republican bias. So long as it is constructed consistently across elections, results are qualitatively unchanged. Note, however, that specifications that do not compute the *efficiency gap* as the difference between the two primary political parties will likely lead to an inconsistent or misleading metric. For example, if one were to define the cracking differential as the difference between majority and minority party efficiency, the sign could indicate bias in favor of different parties from one year to the next as partisan vote shares change. This would necessitate knowledge of which party was in the majority in any given year in any given state, and obfuscate metric interpretation.

[10] To see this, note that both districts have approximately equal voting populations, equal to about 50% of the statewide vote. In either district, candidates require $25\% + \epsilon$ of the statewide vote to secure victory. Even if one Republican candidate had received the party's full 21% vote share, they would not have been able to win their election.

On a more intuitive level, the *cracking differential* measures the extent to which election outcomes deviate from representation proportional to voting outcomes. That is, the *cracking differential* 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 efficiently than the competing party. Because any effective gerrymander must result in one party translating their votes into a disproportionately large seat share, large and enduring *cracking differential* values indicate that a state is effectively gerrymandered.[11]

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

First, prior to the AIRC's implementation, the *cracking differential* generally takes negative values, indicating that election outcomes were biased in favors of Republicans. During the map cycle spanning the 1980s, five states had *cracking differentials* larger in magnitude than Arizona, on average. During the map cycle spanning the 1990s, seven states did.[13] Thus, the magnitude of the *cracking differential* 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 flip typically Republican voters. First, Bill Clinton ran for office amid a national wave of Democrat support. Of 42 states with a defined *cracking differential* during to 1992 - 2000 map cycle, 33 had *cracking differentials* more favorable for democrats in 1992 than their average *cracking differential* 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.

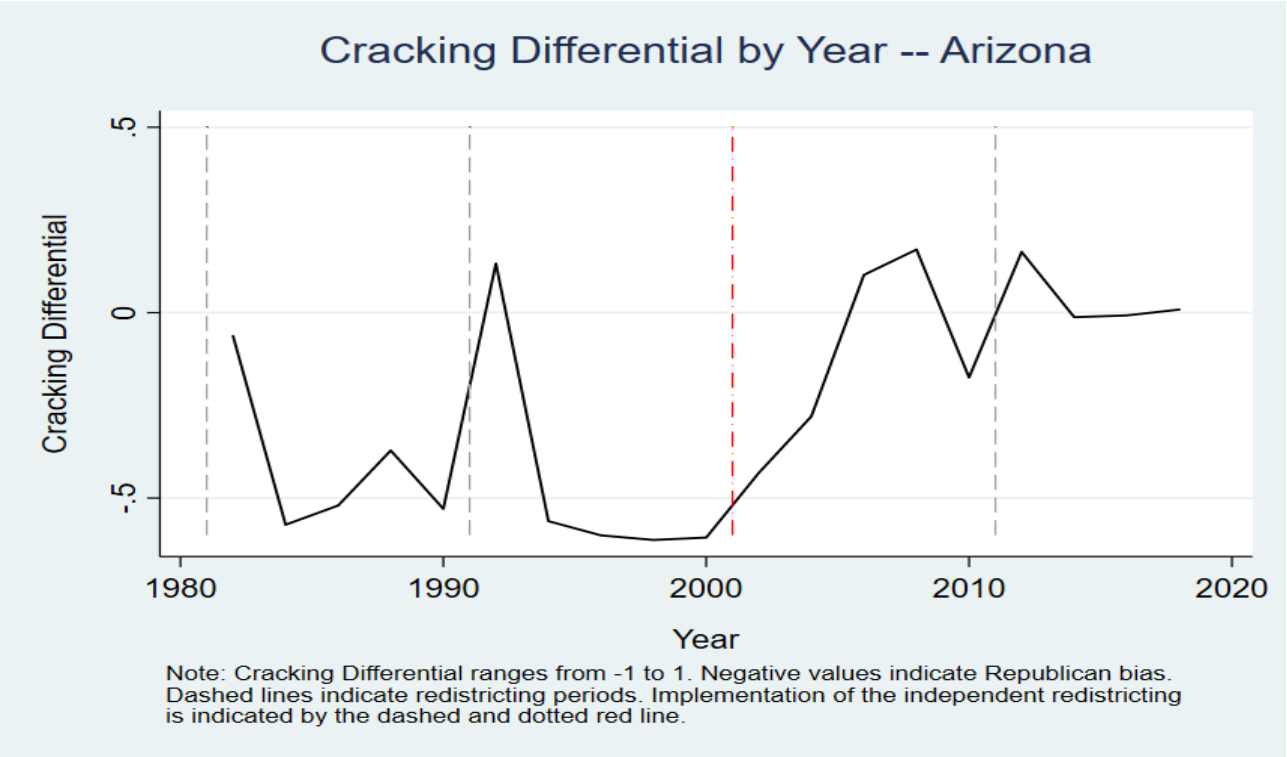
[11] It is worthwhile to note that a state can be effectively gerrymandered even if unintended at the time of redistricting.

[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 final year of one map cycle and the first 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.

Second, after the AIRC's implementation, the *cracking differential* moves toward 0 and fluctuates around the value for the remainder of the study. During this time, the average cracking differential value is about -0.06, which indicates a general lack of measured gerrymandering. More to the point, during the map cycle spanning the 2000s, 25 states had *cracking differentials* larger in magnitude than Arizona, on average. During the map cycle spanning the 2010s, Arizona had the smallest *cracking differential* of all states for which the metric was defined. Of course, this is merely a descriptive exercise; the remainder of this study evaluates whether this change in trends can be causally attributed to the AIRC's implementation.

Figure 1: Arizona Cracking Differential, 1982 - 2018



As previously mentioned, the *cracking differential* can be viewed as a measure of proportionality between partisan representation from partisan voting. To add context to Figure 1, Arizona's *cracking differential* values indicate that prior to the AIRC's implementation, Arizona Republican congressional representation was nearly halfway between their proportional vote share and a complete sweep of the state. After the AIRC's implementation, Republican congressional representation was nearly in line with its proportional vote share.[14]

[14]Specifically, the Republican party received over 57% of the bipartisan vote during the pre-treatment period, but won about 74% of Arizona's congressional seats in during the same time frame. After the AIRC was implemented, the Republican party received just under 53% of the bipartisan vote and won just over 53% of Arizona's congressional seats.

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 variables, describing the data used in this analysis is crucial. This study utilizes a number of political, demographic and economic variables to predict the election outcomes in Arizona.

Political controls include state congressional seat count, percentage of elected candidates that were incumbents (both in total and by party), state house vote share by party, and state house seat share by party.[15] Congressional seat count may be linked to strategic partisan choices, like allocation of party funds for campaign purposes. Incumbency is a well-known predictor of election outcomes; Congressional incumbents seeking reelection are generally successful over 90% of the time. State house vote share and seat share serve to indicate election outcomes in state elections, which likely carry predictive power for federal Congressional elections.

Demographic controls include race, sex, age, birthplace, and education.[16] These are all likely to affect election outcomes in generally understood ways. Women, minorities, immigrants, and younger individuals are all more likely to vote for Democrat candidates than their peers. Of course, these patterns differ by state. The SCUL method is designed to select these characteristics from states that have the

[15] Congressional election data is publicly available from the Harvard dataverse. State election data is publicly available at klarnerpolitics.org.

[16] Demographic data is pulled from the IPUMS database. When annual data was not available, it is linearly imputed from the two closest neighboring years.

highest 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 different 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 briefly 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. Specifically, the synthetic control, y_t^*, is constructed as follows:

$$y_t^* = Y_{Dt}^{'} W_{SCUL}$$

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 differential* is defined for the study's entire time period.[19] SCUL method weights, W_{SCUL}, are lasso regression coefficients selected to minimize the difference between the observed time series of interest and its synthetic control. Specifically, weights are computed according to the following objective function:

$$W_{SCUL} = \arg\min_W \left(\sum_{t=1}^{T_{pre}} (y_{0t} - Y_{Dt}^{'} W)^2 + \lambda |W|_1 \right)$$

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 classifications 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] In total, 31 states are retained as potential donors for the synthetic control.

is similar to the one used for OLS regression, and constitutes the equivalent of a "main regression specification" in this setting. However, lasso regression adds a penalty term, λ, that increases with the vector representing the sum of the absolute values of coefficients, $|W|_1$. While a full description of this process is beyond this section's scope, a few major points are important.

First, the penalty term, λ, functions to attenuate the magnitude of all coefficients relative to their OLS counterparts. At one extreme, λ may be large enough to attenuate all coefficients to 0. At the other extreme, $\lambda = 0$ does not attenuate coefficients at all, and the process is equivalent to OLS. In general, most covariates are assigned a weight of 0, and others may have either positive or negative values. Variables with coefficients of 0 do not contribute to the synthetic control. This provides an objective method for selecting which variables are used to construct "synthetic Arizona," which is quite useful given the large number of potential donor series elements in this study.

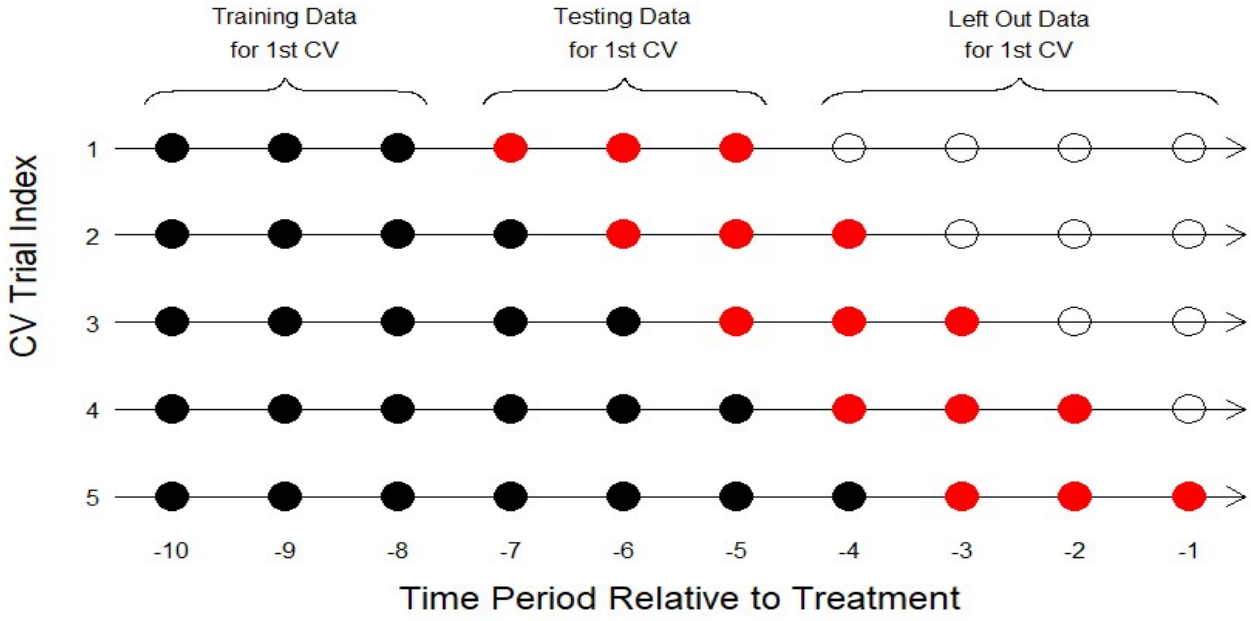
Second, and more specifically, the penalty parameter, λ, is chosen through a process known as rolling-origin cross-validation. This process partitions the focal time series' pre-treatment period into two main components: a training period and testing period. Lasso regression is performed on the training data, and the optimal λ value is determined by fitting predicted outcomes to observed data over the testing period.[20] This is iterated several times, As allowed by the lengths of the training and testing periods, and the pre-treatment period itself. Figure 2 illustrates this process by indicating which election years in the pre-treatment period are used for the training and testing data. Note that, because congressional elections are held every other year, there are 10 elections between 1982 and 2000 used for the rolling-origin cross-validation procedure.

Each cross-validation trial generates a different potential λ value. Of these the median value is used to determine W_{SCUL}. This rewards models that correctly predict outcomes during the testing period. It also provides a natural mechanism for balancing concerns about overfitting the data when λ is small, and eliminating valuable predictive variables when λ is too large.

Lastly, it should be noted that changing the testing period length entails a tradeoff: longer testing periods yield increased confidence about forecasting accuracy, but decrease the number of cross-validation trials that can be run. With fewer cross-validation runs, the median λ value may result in poor fit.

[20] Specifically, λ is chosen to minimize mean squared error between the observed and synthetic time series during the testing period.

Figure 2: Rolling-Origin Cross-Validation Visualization



These attributes are useful for several reasons. First, unlike the classic synthetic control method, SCUL allows for negative weights. This means that covariates with trends that "mirror" the focal group can be emphasized in the SCUL method's synthetic control group. This is particularly useful in the current context, because covariates that influence gerrymandering may have differing coefficient signs across states, depending on which party is benefited. Second, driving most weights to zero allows for coefficient estimation when the number of covariates is larger than the number of observations. Third, this process ensures that a sparse number of covariates contribute to the synthetic counterfactual, and avoids overfitting the data.

4.4 Identification

In order to identify the causal effect of the AIRC, it is important to confront factors that could confound analysis. Chief among these are issues which might result in endogeneity. These issues typically follow from two types of concerns: first, that unobserved factors are correlated with both treatment status and dependent variable; second, that there exists a simultaneity issue between the dependent variable and treatment status. I address these issues in order below.

By their nature, one cannot directly test for unobserved confounding factors. To completely eliminate concerns that unobserved factors are causing endogeneity, one would need a convincing instrumental

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]

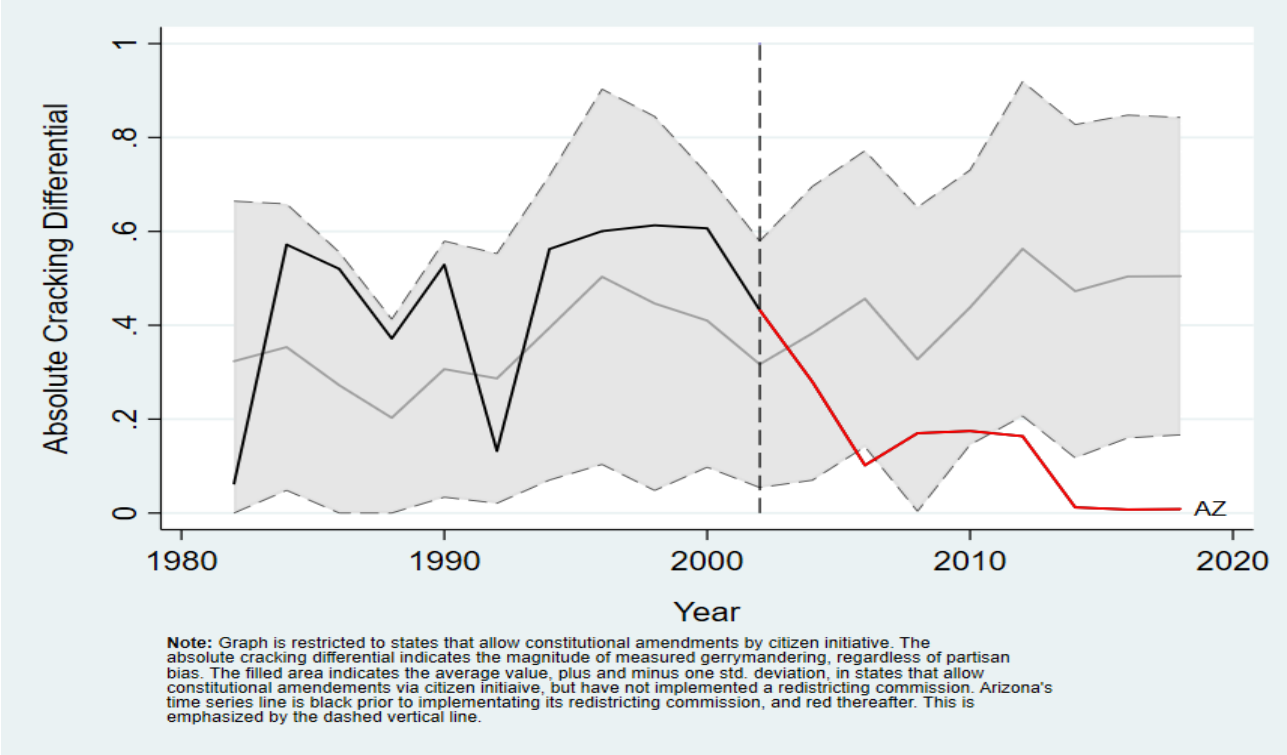
I now confront the potential that there exist simultaneity issues between partisan gerrymandering and AIRC implementation. Typically, these concerns follow two tracks. First, readers may be concerned that only states with low levels of gerrymandering are likely to enact commission-based redistricting 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: $y(0)_{st} = \delta_t \alpha_s + \epsilon_{st}$. Here, $y(0)_{st}$ are the synthetic outcomes for group s in period t, δ_t is a $1 \times K$ vector of period-specific unmeasured variables, and α_s is a $K \times 1$ vector of group-specific coefficients. If the observed outcomes for the treated group are generated by $y(0)_{0t} = \delta_t \alpha_0 + \epsilon_{0t}$, then the synthetic control method will match these outcomes in the pre-treatment period by selecting comparison units with values of α_s that are a close match for α_0. Since α_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 α_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.

Figure 3: Partisan Gerrymandering in States that Allow Constitutional Amendment by Citizen Initiative



represents the average, plus and minus one standard deviation, of the absolute *cracking differential* for the remaining 14 states that have not implemented redistricting commissions. Arizona did not experience an unusually high level of gerrymandering relative to other states that had the ability to implement redistricting reform through similar mechanisms.[22]

5 Baseline Results

This section presents the AIRC's estimated effect on gerrymandering outcomes. The outcome variable of interest is the *cracking differential*. As a reminder, the SCUL procedure's goal is to estimate the counterfactual path the *cracking differential* would have taken, had the AIRC not been implemented. Optimal weights are computed according to the procedure outlined in section 4.4.[23] The counterfac-

[22]In unreported results, I repeat this exercise with the remaining three states that have implemented commissions. Their measured gerrymandering also fall within one standard deviation of the average for states allowing constitutional amendments via citizen initiative, prior to redistricting. While this suggests there is not a systematic relationship between the magnitude of gerrymandering and commission implementation *across* states, it is not required for identifying the causal effect of redistricting reform *within* Arizona.

[23]This procedure is recounted in more detail in Appendix A

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.

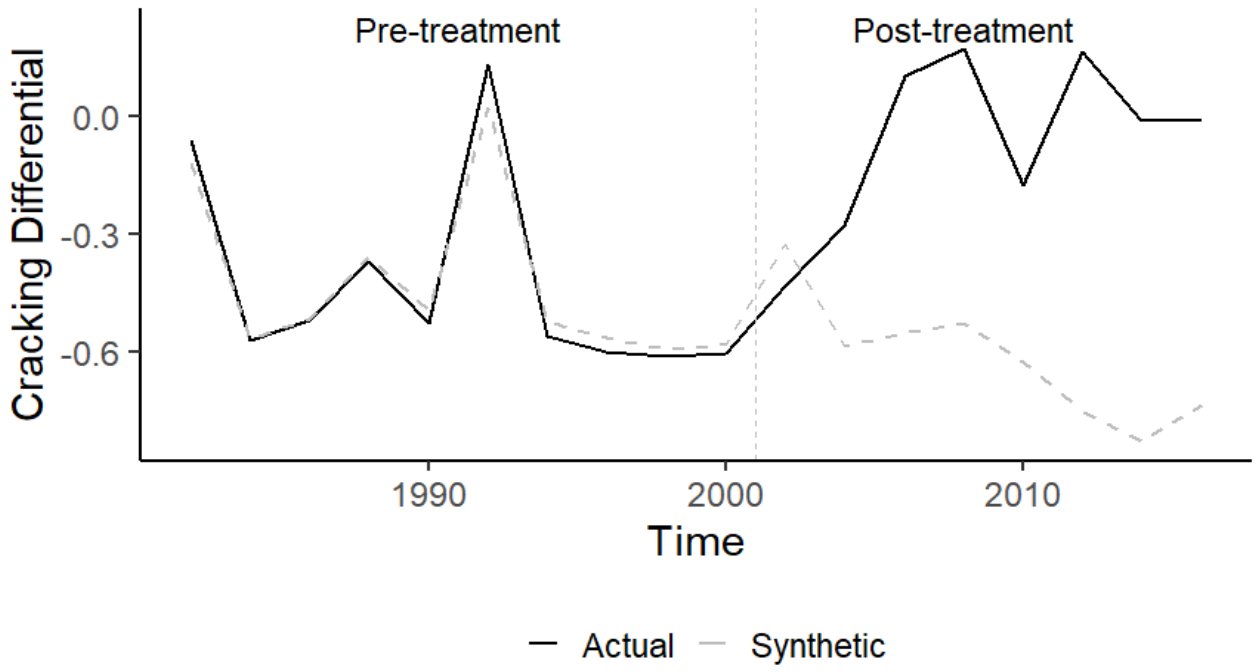
Baseline results present my findings when using the full set of variables in my dataset, and following guidelines for model fit suggested in the literature. I will show that this leads to concerns about the synthetic control's composition and statistical power, and address them in robustness checks. 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 Effect Estimates

Figure 4 depicts Arizona's observed *cracking differential* 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 fit is measured in terms of a modified version of Cohen's D. They suggest using a threshold of .25 for model fit, 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 fit.

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 observed *cracking differential* takes an average value of -0.06 after treatment, while its synthetic counterpart takes an average value of -0.62 after treatment. By comparison, the average *cracking differential* for both groups is -0.43 prior to treatment. All told, this constitutes a treatment effect equivalent to a 90% decrease in measured gerrymandering. While this is a quite large treatment effect at first glance, Section 5.3 will show that it does not guarantee statistical significance.

Figure 4: Arizona and its Synthetic Counterfactual



5.2 The Composition of the Synthetic Control

Given the preceding discussion on the effect of AIRC implementation, it is prudent to examine the donor elements used to construct Arizona's synthetic control. Depending on which variables are included in the synthetic control, it may gain or lose credibility as a counterfactual. Specifically, if any donor elements are included due to spurious correlation, the synthetic control's value as a counterfacutal is degraded; the synthetic control is matched to observed outcomes based partially on noise, and forecasted outcomes may be inaccurate. In this case, the SCUL method generally selects donor units that can reasonably be expected to hold predictive power for congressional election outcomes in Arizona. Still, a few included variables may be suspect. This is discussed below, and further addressed through a robustness check in Section 6.1.

Figure 5 relays the composition of Arizona's synthetic control. Aside from the intercept, the SCUL method places non-zero weight on five variables, each from various states. These are the share of the statewide vote received by Republicans in the state house, the share of seats received by Republicans in the state house,[24], the percent of seats held by reelected incumbents, the unemployment rate, and the state employment share in various industries. Industry categories are constructed to match BEA

[24]I name these variables "Republican State House Vote Share" and "Republican State House Seat Share" in reported tables, respectively.

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

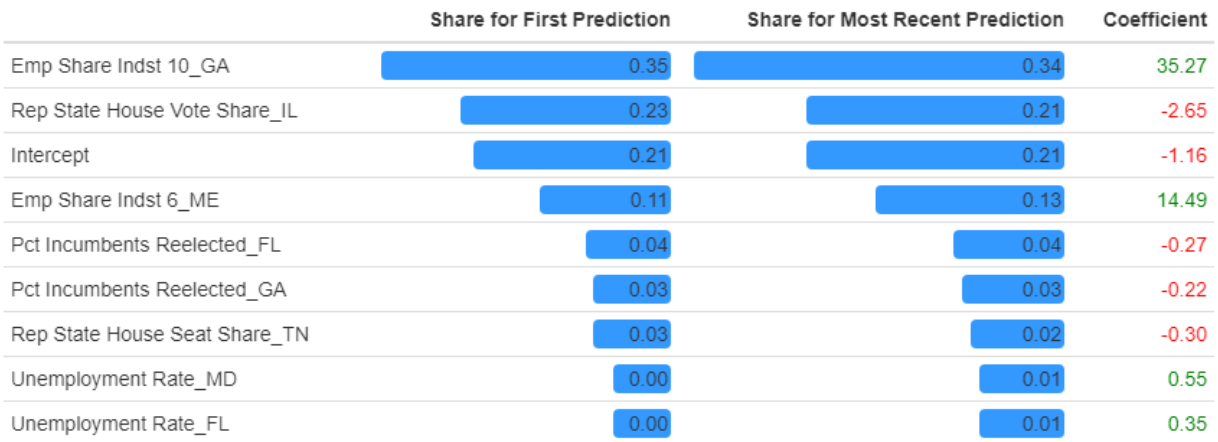


Table 1: Industry Employment Categories

Group	Industry	Group	Industry
01	Farm employment	12	Professional, scientific,
02	Mining, quarrying, oil		technical services
	and gas extraction	13	Enterprise management
03	Utilities	14	Administrative and support
04	Construction		and waste management
05	Manufacturing		and remediation services
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.
11	Real Estate, Rental,		and govt. enterprises)
	Leasing	20	Government, govt. enterprises

[25]This is noteworthy insofar as a synthetic control whose components' shares fluctuate significantly may be suspect; if donor elements comprise vastly different 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 null distribution. The two-sided p-value is therefore determined by comparing the absolute values of the estimated pseudo-effects against the absolute value of the estimated treatment effect. Specifically, I follow the advice of Hollingsworth and Wing (2020) when there are relatively few elements included in the null distribution, and report the p-value as the range of percentile values the treatment effect's rank may take when it is included in the null distribution.[27]

Figure 6 depicts a "smoke plot" which illustrates Arizona's estimated treatment effect against the null distribution of outcomes. Notably, the null distribution widens as forecasts are further removed from treatment; this indicates the extent to which one should expect model fit to deteriorate over time. Arizona's estimated treatment effect is indicated by the green line, and generally runs near the upper boundary of the null distribution throughout the post-treatment time period. This visually reinforces that it is at most marginally statistically significant.

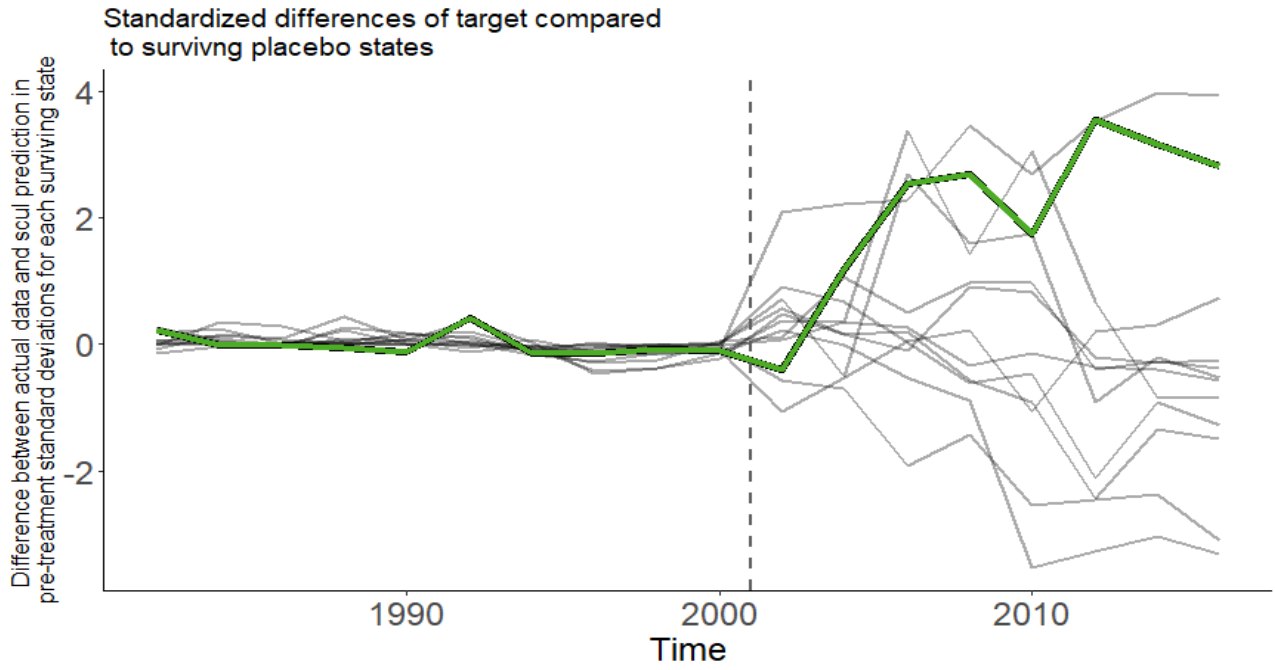
More specifically, the smoke plot depicts the standardized differences in synthetic and observed time series for Arizona on placebo units. Only placebo units which have a pre-treatment fit within the modified Cohen's D threshold of 0.25 are depicted, as suggested by Hollingsworth and Wing (2020). Cohen's D reflects the average difference between the observed outcome and its synthetic counterpart,

[26] 31 placebo states are included. This omits states which have commission-based redistricting systems at any point, or which have an undefined *cracking differential* at any point.

[27] This is done to account for a lack of granularity in potential p-values when using only the rank of the estimated treatment effect within the null distribution. For example, if 9 states comprise the null distribution of pseudo-treatment effects, the treated group's smallest rank percentile possible is 1/10, or 0.1. This reflects the fact that the treatment effect is larger than 90% of elements in the null distribution, and constitutes the upper bound on the range of p-values its effect may take. This reporting convention makes sense when there are a large number of elements in the null distribution; the range of potential values the treatment effect may take is small, and conservatively reporting p-values is responsible. In this example, however, it is equally true that no pseudo-treatment effects in the null distribution are larger than the estimated treatment effect, and so the lower bound for its p-value is arbitrarily close to 0. In this hypothetical, the treatment effect's p-value falls somewhere in the broad range $(0, .1]$, due to the small number of elements included in the null distribution. Instead of reporting a specific value within that range, it is most transparent to simply report the range itself.

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 to obtain full statistical significance; instead, it is the second largest effect. Table 2 shows all states included in Figure 6, including the magnitude and fit of all effects shown (measured in pre-treatment standard deviations). In order to be the largest effect measured, Arizona would need to have an effect size of at least 3.02 pre-treatment standard deviations. Given that its synthetic control takes an average value of -0.62 during the post-treatment period, this would necessitate observed outcomes

[28] Explicitly, $D_s = \sum_{t=1}^{T_{pre}} \left| \frac{y_{st} - y^*_{st}}{\sigma_s} \right|$, $\sigma_s = \sqrt{\frac{1}{T_{pre}} \sum_{t=1}^{T_{pre}} (y_{st} - y_s)^2}$

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
Oklahoma	0.32	0.13
Kentucky	0.19	0.23
Florida	0.14	0.06
Georgia	0.10	0.15

Note: Effect size and fit are measured in terms of each state's pre-treatment standard deviation. Only states with Pre-treatment fits smaller than 0.25 are retained for the smoke plot.

With this in mind, it seems likely that this analysis is indeed underpowered. This problem occurs for two reasons. First, in at least some cases, post-treatment forecasts are very inaccurate. This results in pseudo-treatment effects that are so large that Arizona's estimated treatment effect would have to be unreasonably large in order to attain statistical significance. Second, there are relatively few pseudo-treatment effects included in the null distribution. Because of this, if even one is larger in magnitude than Arizona's observed treatment effect, Arizona's treatment effect cannot attain statistical significance.

In the next section, I take steps to address each of these issues. First, I remove state industry

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 fewer states are included in the analysis, the lower its statistical power.

Second, I relax the threshold required for model fit. Hollingsworth and Wing (2020) suggest enforcing a Cohen's D threshold of 0.25 to determine whether the SCUL estimates fit observed data well, though they note this threshold is somewhat arbitrary. When combined with removing state industry employment share variables from the data, this provides a sizeable boost to statistical power.

All told, baseline results fall within the range of outcomes necessary for marginal statistical significance. Given that lack of statistical power likely prevents me from correctly identifying a statistically significant effect, if one exists, marginal statistical significance is suggestive. Still, some may find this line of reasoning unconvincing. While results are merely suggestive, it bears noting that there is also no evidence that AIRC implementation made partisan gerrymandering outcomes worse. At a minimum, AIRC implementation seems to have been a lateral move where gerrymandering is concerned.

6 Robustness Tests

The preceding analysis presented my findings when using all covariates included in my data set. It is meant to constitute a transparent first step towards measuring the causal effect of AIRC implementation on partisan gerrymandering outcomes. Still, there are alternative ways the synthetic control could be constructed, lengths of time over which I could forecast election outcome, or metrics that could be used to measure gerrymandering. In this section, I repeat the analysis from section 5 in a variety of different settings in order to address each of these concerns. Rationale for each is given below.

First, I address concerns about statistical power and synthetic control construction. In Section 6.1, I re-run the preceding analysis while omitting state industry employment variables. This ensures that potentially problematic variables included in the synthetic control in Section 5.2 can no longer be used. This is beneficial for two reasons: first, eliminating questionable variables improves the syn-

thetic control's credibility as a counterfactual; second, this generally improves the synthetic control's forecasting accuracy, improving statistical power.

As mentioned earlier, this modification comes at a cost. By eliminating variables from the donor pool, I reduce the accuracy with which synthetic controls match their observed counterparts in some states. In turn, this eliminates some states from the null distribution. In response, I relax the threshold required to model fit in order to include states with model fit just outside the traditional benchmark. Nonetheless, Arizona remains the second largest effect in the distribution; results are qualitatively aligned with those in Section 5.

Second, some may be concerned about the SCUL method's ability to accurately forecast outcomes in this setting. Because there are 10 pre-treatment time periods and 8 post-treatment time periods, the testing period for this study is short relative to the post-treatment period. This study uses a three period testing length for the cross-validation period used to generate λ values. This constitutes just over a third of the post-treatment period, which may cause concern about the model's ability to accurately predict the entire post-treatment period. Unfortunately, data limitations prevent addressing this by extending the pre-treatment period.

I address this in section 6.2. Because I cannot add data to the pre-treatment period, I instead truncate the post-treatment period after 2006. This means that the SCUL method need only forecast three time periods of outcomes after AIRC implementation. This ensures that the testing period and post-treatment period lengths are balanced, and is meant to mitigate concerns that treatment effect estimates are extrapolate too far beyond the testing period length to be considered credible. This modification also entails a trade-off: if the effect of AIRC implementation grows over time, truncating the post-treatment period after 2006 may prevent me from fully measuring its effect.

Third, some may be skeptical of the *cracking differential's* ability to accurately measure gerrymandering. In response, I re-run analyses using an alternative measure of gerrymandering in sections 6.3: the *efficiency gap* created by Eric McGhee, EG_{McGhee} (McGhee, 2014). The SCUL method does not match or forecast *efficiency gap* outcomes well, and generally does not provide credible results. This emphasizes the value of using the *cracking differential* to measure gerrymandering in this context.

6.1 Excluding State Industry Composition

The first robustness check restricts the set donor pool variables to exclude state industry composition. Figure 7 depicts Arizona's observed *cracking differential* and its synthetic counterpart. Here, model fit 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 differential* is -0.52, leading to an estimated treatment effect of 0.46. This would constitute a 88% decrease in measured gerrymandering over the post-treatment period. As before, while this effect seems large at first glance, it does not guarantee statistical significance.

Figure 7: Arizona and its Synthetic Counterfactual

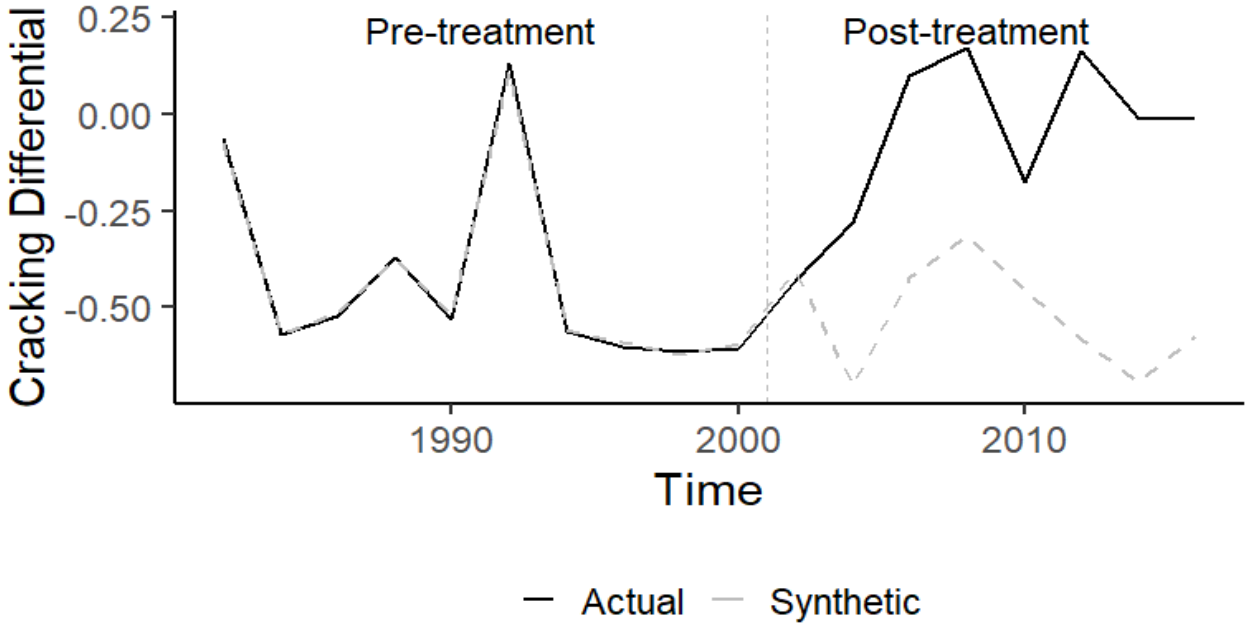


Figure 8 displays the structure of the synthetic control group in this robustness check. Aside from the intercept, the SCUL method places non-zero weight on four variables, each from various states. These are the share of the statewide vote received by Republicans in the state house, the percent of seats held by reelected incumbents, the unemployment rate, and the *cracking differential*.

These variables are generally aligned with those selected in baseline results, though a few changes are noteworthy. Industry employment shares are now omitted, and cracking differential outcomes in Michigan contribute modestly to Arizona's synthetic control. As before, variables that overlap with those discussed in Section 5.2 are all factors one would expect to have significant impact on election

outcomes. It also seems intuitive that *cracking differential* values in some states might have some predictive power for *cracking differential* outcomes in Arizona; national trends are likely to ensure their correlation.

Figure 8: Synthetic Arizona Composition

	Share for First Prediction	Share for Most Recent Prediction	Coefficient
Intercept	0.44	0.46	2.45
Rep State House Vote Share_IL	0.23	0.21	-2.66
Rep State House Vote Share_KY	0.11	0.09	-1.43
Pct Incumbents Reelected_FL	0.09	0.09	-0.63
Rep State House Vote Share_NM	0.07	0.07	-0.90
Pct Incumbents Reelected_PA	0.02	0.02	-0.14
Pct Incumbents Reelected_GA	0.02	0.03	-0.20
Unemployment Rate_MD	0.02	0.04	2.34
Cracking Differential_MI	0.00	0.00	0.02

All told, the synthetic control seems to be constructed in a way that performs at least as well as baseline results. Industry employment shares, which may have been suspect, no longer contribute to the synthetic control. Variables that are selected by the SCUL method are unlikely to be controversial. Furthermore, Figure 7 is qualitatively similar to Figure 4, but is fit slightly better by the synthetic control during the pre-treatment period.[29] Taken together, this suggests that removing industry employment shares from the pool of donor variables results in a slightly more accurate synthetic control.

Having confronted concerns about the construction of the synthetic control, I now address questions regarding statistical inference and power. Figure 9 reports the standardized differences between observed and synthetic controls for Arizona and surviving placebo states. Pseudo-treatment effects take a more narrow range of values than in Section 5.3, indicating that forecasted outcomes are generally more accurate for untreated states. However, the removal of industry employment shares results in fewer states reaching the threshold required for model fit; here, 9 states states survive for placebo analysis. Table 3 makes explicit which states are reflected in Figure 9, along with their pre-treatment fit and post-treatment pseudo-effect sizes.

[29]Specifically, Cohen's D is 0.04 in this specification and 0.13 in baseline results.

Figure 9: Smoke Plot of Estimated Treatment and Pseudo-Treatment Effects

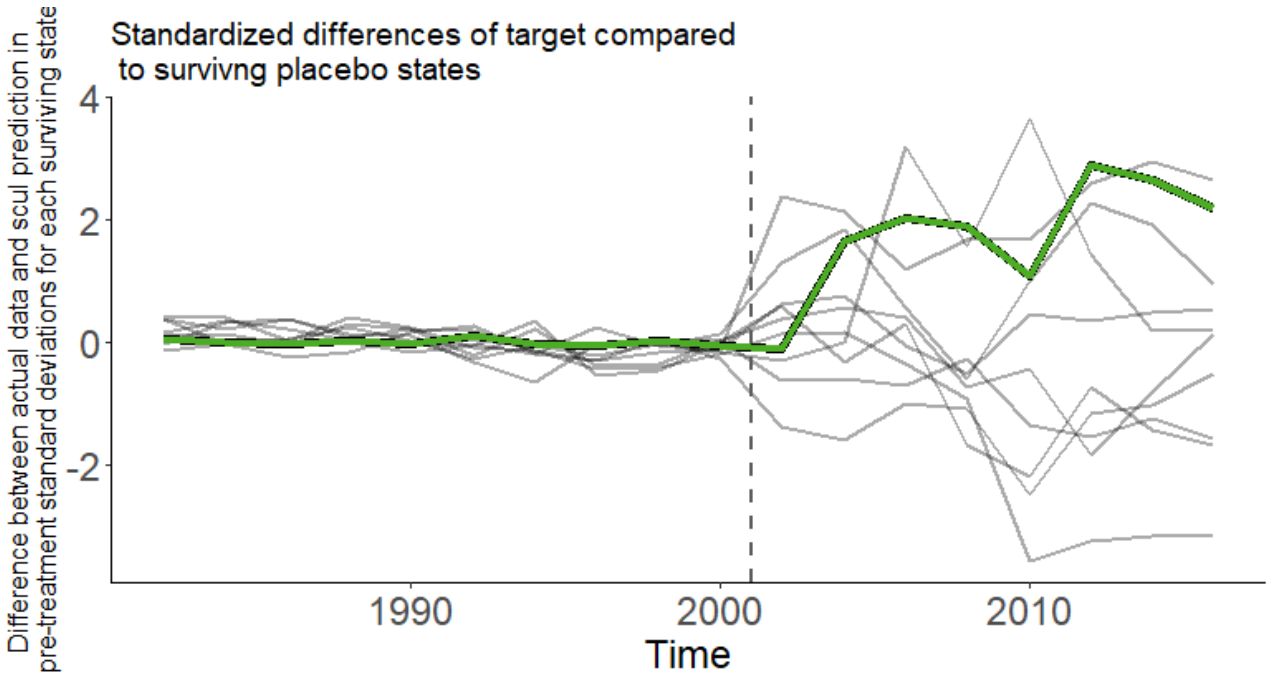


Table 3: Smoke Plot Effect Sizes and Fit

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
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: Effect size and fit are measured in terms of each state's pre-treatment standard deviation. Only states with Pre-treatment fits 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 greater than .1. This lack of granularity is also an issue for statistical power, insofar as it distorts our ability to distinguish extreme outcomes from mild ones.

In response to this, I relax the threshold for model fit beyond the 0.25 mark suggested by Hollingsworth and Wing (2020). While the traditional threshold of 0.25 standard deviations is somewhat arbitrary, the logic underlying it is sound: the worse the synthetic control fits its observed counterpart during the pre-treatment period, the less credibility its forecasts have during the post-treatment period.[31] Thus, this procedure entails a trade off: relaxing the threshold for model fit allow for more states to be included in the analysis, but reduces our confidence in forecasted outcomes for newly included states.

In order to balance these concerns, I report results when relaxing the threshold to 0.4. This is not vastly beyond the traditional benchmark, and is small enough that included states have reasonable synthetic control composition.[32] This allows five additional states to be used for analysis. An updated smoke plot and corresponding table of effect sizes are reported in Figure 10 and Table 4, respectively.

[30] Alternatively, given that Arizona's observed outcomes take an average value of -0.06 during the post-treatment period, the synthetic control would need to predict that the cracking differential take an average vale of -0.59 during the same time frame to attain statistical significance.

[31] Of course, this is no longer true if synthetic controls are matched to observed time series using spuriously correlated variables; in this case, pre-treatment fit might be excellent, but forecasted outcomes would be completely inaccurate. Indeed, this concern is precisely what led to this robustness check.

[32] In unreported results, I relax the threshold to 0.75 – three times larger than the traditional benchmark. In general, states with a fit worse than 0.5 pre-treatment standard deviations have very few variables in their synthetic control, and potentially use only the lasso regression intercept. In these states, the SCUL method does not produce a convincing match for observed outcomes. Thus, the maximum threshold I use is 0.4 pre-treatment standard deviations; large enough to include more states in the analysis, but below the point at which fit is so poor that results completely lose credibility.

Arizona remains the second largest effect measured, placing its p-value in the range $(.7^*, .13]$. This suggests that Arizona's rank within the null distribution is robust to changes in the threshold for model fit, and may affect statistical significance.

Results here are qualitatively aligned with those in Section 5, but address concerns regarding synthetic control composition and statistical power. Donor elements used to construct the synthetic control are less controversial than those used in Section 5.2, which lends credibility to the synthetic control's use as a counterfactual. Additionally, psuedo-treatment effects are no longer so large that I would need to observe an unreasonably large treatment effects to attain statistical significance. Arizona's treatment effect is the second largest measured, even when relaxing the standard threshold for model fit. While results do not obtain full statistical significance, this provides suggestive evidence that AIRC implementation reduced partisan gerrymandering. As before, no matter one's stance on the strength of these results, there is no indication that AIRC implementation made gerrymandering outcomes worse. At a minimum, AIRC implementation seems to have done no harm when it comes to partisan gerrymandering.

Figure 10: Smoke Plot of Estimated Treatment and Pseudo-Treatment Effects

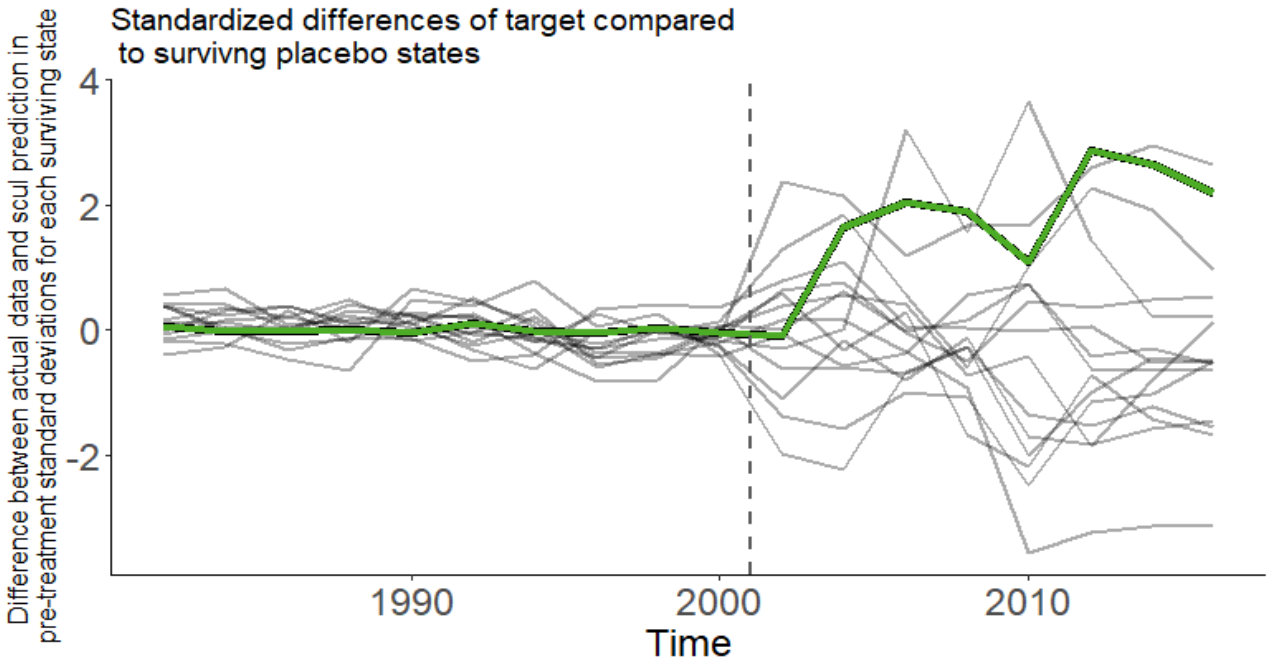


Table 4: Smoke Plot Effect Sizes and Fit

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.

exists over years in which we are most confident in SCUL method forecasts.

Because of concerns regarding synthetic control construction, I continue to report results when excluding state industry composition controls from the pool of donor variables. This guards against concerns investigated in the previous robustness check, and maintains narrative continuity. Still, in unreported results I re-run this robustness check while including state industry controls. Results are qualitatively unchanged.

Figure 11 depicts Arizona's observed *cracking differential* and its synthetic counterpart. Notably, this robustness check is equivalent to truncating baseline results in Section 5 after 2006. As before, both the observed and synthetic time series diverge following AIRC implementation. After treatment, the average *cracking differential* is -0.51 and -0.2 for the synthetic and observed time series, respectively. This leads to a smaller estimated treatment effect than in section 4, largely because the observed time series takes values smaller in magnitude during the portion of the post-treatment period that has been removed.

Figure 11: Arizona and its Synthetic Counterfactual

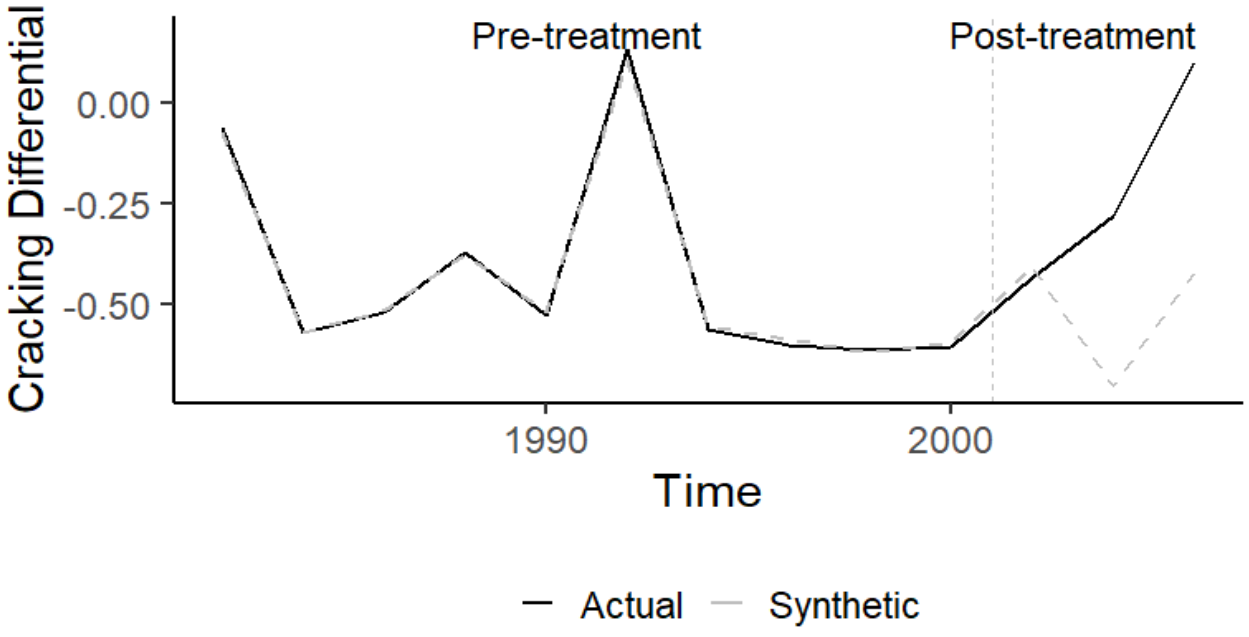


Figure 12 indicates the structure of the synthetic control group in this robustness check. Because the donor pool is identical to the used in Section 6.1, its construction is identical to Figure 5. This affirms that the SCUL method provides consistent results, and is again a reasonable way to construct the

synthetic control.

Figure 12: Smoke Plot of Estimated Treatment and Pseudo-Treatment Effects

	Share for First Prediction	Share for Most Recent Prediction	Coefficient
Intercept	0.44	0.46	2.45
Rep State House Vote Share_IL	0.23	0.21	-2.66
Rep State House Vote Share_KY	0.11	0.09	-1.43
Pct Incumbents Reelected_FL	0.09	0.09	-0.63
Rep State House Vote Share_NM	0.07	0.07	-0.90
Pct Incumbents Reelected_PA	0.02	0.02	-0.14
Pct Incumbents Reelected_GA	0.02	0.03	-0.20
Unemployment Rate_MD	0.02	0.04	2.34
Cracking Differential_MI	0.00	0.00	0.02

Figure 13 is a smoke plot of standardized differences between observed and synthetic controls for Arizona and surviving placebo states. As with previous iterations, Table 5 makes explicit which states are included and what their effect sizes are. Results again are similar to a truncated version of Figure 6. In this robustness check, the p-value range is $(.18, .27]$; of the 11 states in Figure 13, two have relative treatment effects larger than Arizona. Here, Arizona's observed outcomes are more biased in favor of Republicans during the truncated post-treatment period, and so Arizona's estimated treatment effect is mitigated somewhat. This leads to its ranking within the null distribution falling by one position, and results in statistical insignificance.

It bears reiterating that this robustness check is designed to balance the lengths of the testing and post-treatment periods. The intuition underlying this is that outcomes are forecast during the testing period onward; so long as these forecasts during the pre-treatment period fall within the 0.25 threshold for model fit, it provides confidence that the SCUL method can accurately forecast at least as many periods are included during the testing period. In this case, the rolling-origin cross-validation procedure chooses the λ value corresponding to the scenario in which training data spans 1982 - 1992, and testing period data spans 1994 - 1998. Given that the pre-treatment Cohen's D value is 0.04, this means the SCUL method accurately predicts 4 periods of pre-treatment data; the 3 elections included in the testing period (1994, 1996, 1998), and the election immediately following (2000). In unreported results, I restrict the post-treatment period to include four election cycles (one more year

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 effect measured.

Figure 13: Smoke Plot of Estimated Treatment and Pseudo-Treatment Effects

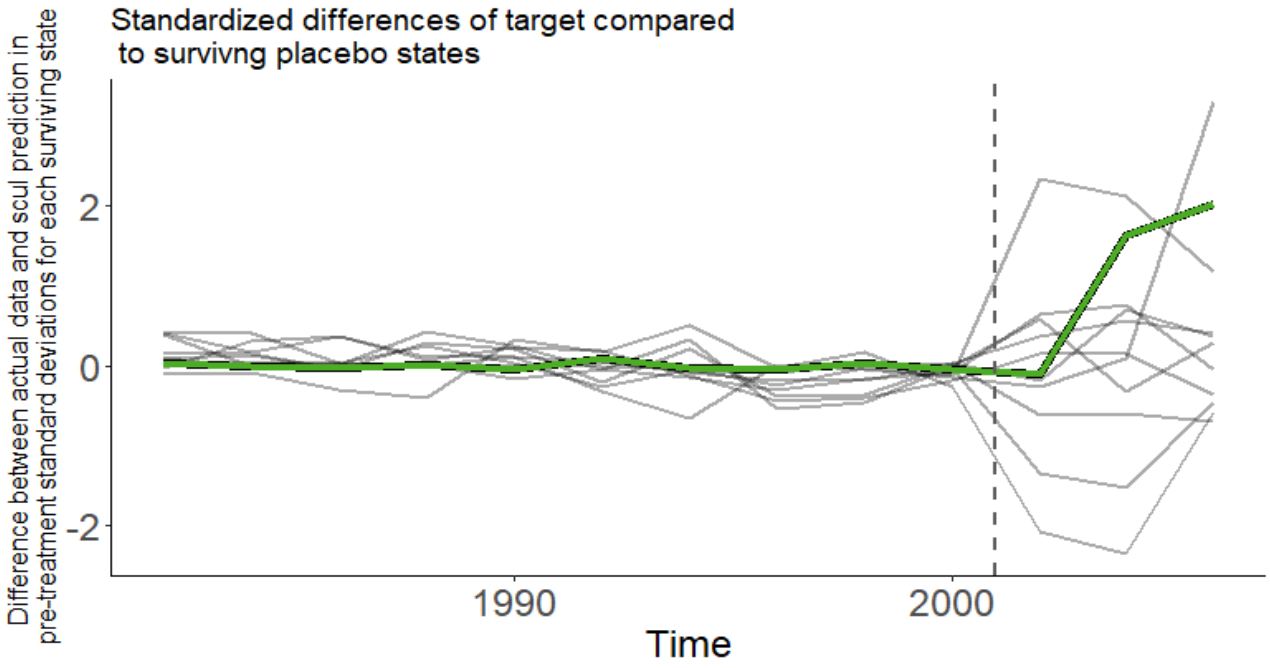


Table 5: Smoke Plot Effect Sizes and Fit

State	Post-Treatment Effect 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, $\mathbf{EG_{McGhee}}$

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 λ value. Unfortunately, the cross-validated λ 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 λ value from the pool of generated values until the synthetic control group meets the Cohen's D threshold for model fit, or all λ values are exhausted. In this case, the lowest λ 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 λ 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 previous iterations, the primary donor variables included are the percentage of seats held by reelected incumbents, the Republican state house vote share, and measures of gerrymandering from other states. The percentage of congressional candidates in Mississippi New York who were effectively uncontested (meaning they received at least 95% of their district's vote) also appears in Figure 15.

Figure 16 reports the standardized differences between observed and synthetic controls for Arizona and surviving placebo states. Results are discouraging. Few placebo states survive to be used for analysis, and post-treatment predictions are extremely volatile for all states. This suggests that the SCUL method does not predict EG_{McGhee} outcomes well for the vast majority of states in the pre-treatment period, and, even when it does, its post-treatment forecasts are so volatile they cannot be considered credible. The estimated effect for Arizona is the largest in Figure 16, and so its p-value is in the range $(0^{***}, 12]$. Still, it seems likely this result is due to noise, and is not credible. All told, the SCUL method does not appear to be able to reliably predict EG_{McGhee} outcomes, at least within the context of this study.

Figure 14: Arizona and its Synthetic Counterfactual

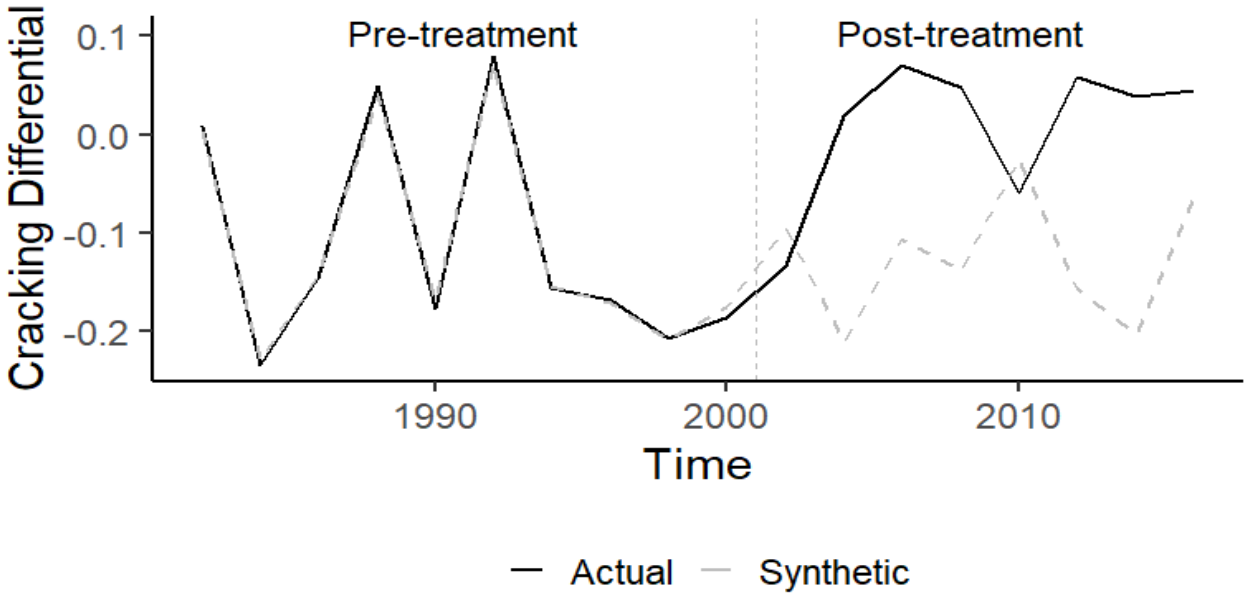
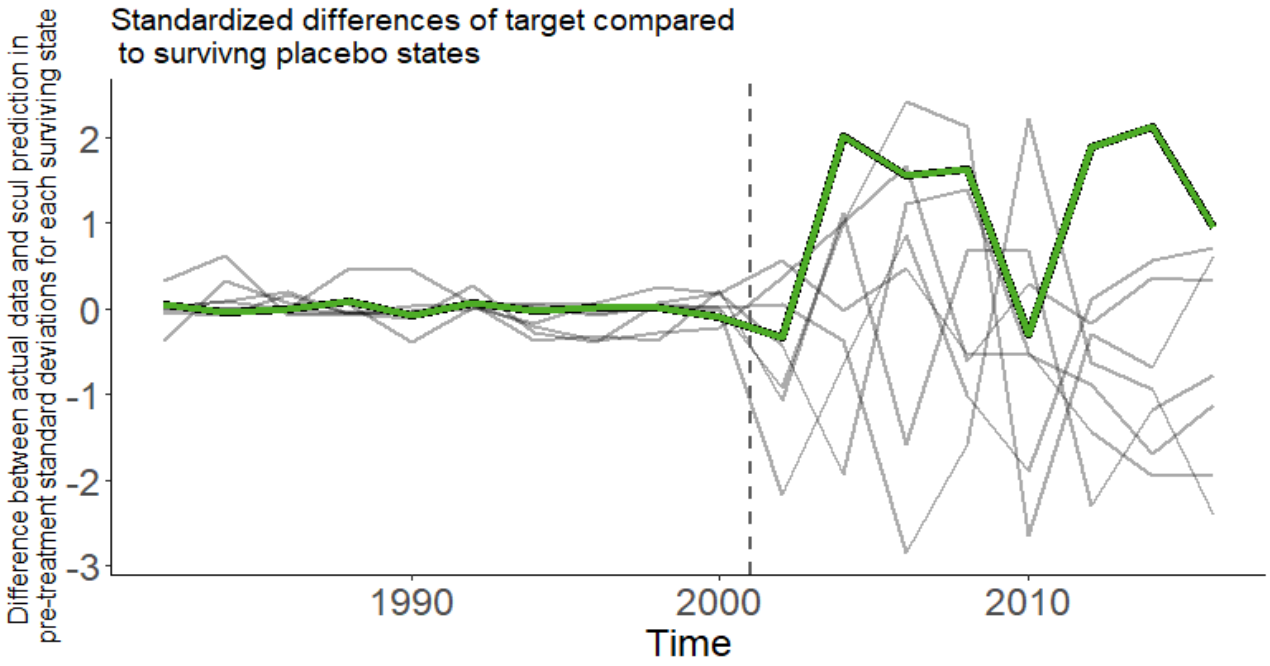


Figure 15: Synthetic Arizona Composition

	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
Rep State House Vote Share_KY	0.11	0.10	-0.41
Pct Incumbents Reelected_PA	0.08	0.07	-0.13
Rep State House Vote Share_NV	0.06	0.06	-0.17
Standard Efficiency Gap_CO	0.02	0.01	-0.16
Pct Incumbents Reelected_IL	0.00	0.00	-0.01
Standard Efficiency Gap_NC	0.00	0.01	0.06
Pct Effectively Uncontested_MS	0.00	0.00	-0.06
Pct Effectively Uncontested_NY	0.00	0.03	0.53

Figure 16: Smoke Plot of Estimated Treatment and Pseudo-Treatment Effects



Given the SCUL method's poor performance in this robustness check, it seems likely that at least some donor elements are included due to spurious correlation. This is likely because the smallest possible λ value was chosen for this test, in order to preserve the possibility of generating a counterfactual for observed EG_{McGhee} outcome in Arizona. Unfortunately, Figures 14 - 16 jointly suggest that the

synthetic control does indeed fit the observed trend based on noise; the inclusion of extra donor variables has improved fit over the pre-treatment period, but in a way that holds almost no predictive power during the post-treatment period. Taken together, this suggests the SCUL method does not match or forecast *efficiency gap* outcomes in Arizona well, and underscores the value of using the *cracking differential* to measure gerrymandering in this setting.

7 Conclusion

Redistricting is a process central to American democracy, and follows the census every decade. While it is meant to afford every person approximately equal representation in government, it is often used as an opportunity to align election outcomes with partisan goals. A growing number of states have implemented commission-based redistricting, in large part as a response to concerns about partisan gerrymandering. While there is considerable interest in the consequences of redistricting commissions, very little work has evaluated their casual effects.

This paper investigates the extent to which Arizona's independent redistricting commission reduced partisan gerrymandering outcomes in congressional elections. Arizona presents an ideal case study in this regard, largely because the timing with which it implemented its commission allows one to clearly establish pre- and post-treatment trends. Additionally, relevant institutional details suggest that the commission may have disrupted the political status quo. Results suggest that the AIRC did indeed reduce partisan gerrymandering, but fall short of full statistical significance. At a minimum, the AIRC does not seem to have made gerrymandering outcomes worse.

I use the SCUL method to undertake a first step towards rigorous casual analysis in this field. Beyond gerrymandering, the SCUL method is potentially useful in any study that investigates the effects of region-specific policies. In addition to contributing to our understanding of the role redistricting commissions play in gerrymandering outcomes, this paper demonstrates an empirical method that is useful to political scientists and legal scholars at large.

References

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: 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 effect 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 effect 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 Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences 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 effect 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 efficiency 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 Griffith. *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 Issacharoff. Gerrymandering and political cartels. *Harv. L. Rev.*, 116:593, 2002.

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

Loren Dean Kruschke. Measuring partisan efficiency in redistricting. forthcoming.

Jeffrey 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 effect 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 affect 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. Our electoral exceptionalism. *U. Chi. L. Rev.*, 80:769, 2013b.

Nicholas O Stephanopoulos and Eric M McGhee. Partisan gerrymandering and the efficiency 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 effective by examining the extent to which synthetic and observed outcomes diverge, after it goes into effect.

Abadie et al. (2015) provide a compelling example of how the standard synthetic control should work, in practice. In this paper, they evaluate the causal impact of the 1990 German reunification on West Germany's GDP. Predictive variables include trade openness, inflation rate, industry share, schooling, and investment rate. The synthetic control group is comprised of a weighted combination of subset of OECD nations: Austria, Japan, Netherlands, Switzerland, and the United Kingdom. It is shown that the synthetic control's predictive variable values closely match West Germany's, and vastly outperform a naive average of OECD nations. In turn, the synthetic control's predicted GDP also closely matches West Germany far better than the OECD average prior to reunification. After reunification, synthetic West Germany's predicted GDP remains substantially larger than what was actually observed. All told, this provides strong credibility that the synthetic control is correctly acting as a counterfactual, and enables researchers to interpret the divergence in economic outcomes as the causal effect of reunification.

The natural question, then, is how the synthetic control is constructed. This boils down to choosing how to weight both characteristics and units in the synthetic control, and requires some mathematical exposition. Typically, one may observe a sample of $N + 1$ units, where the first unit (denoted unit 0) receives treatment and units $1, ..., N$ do not. For each of these units, the researcher observes K characteristics of interest. Let X_0 represent the $K \times 1$ vector of statistics of interest for the treated

group, and X_D represent the $K \times N$ matrix of statistics of interest for each unit in the donor pool. In Abadie et al. (2015), there were five 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 define two sets of weights. First, one defines weights for each donor characteristic. Then, one must define weights for each donor unit. For this purpose, let V be the $K \times K$ positive semi-definite matrix of characteristic weights.[34] Furthermore, let W be the $N \times 1$ 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:

$$y_t^* = Y_{Dt}^{'} W \tag{1}$$

where Y_{Dt} represents the $N \times 1$ vector of outcomes for donor pool units in period t. The vector of unit weights, W, are chosen to minimize the weighted difference between observed and synthetic control group characteristics:

$$\sqrt{(X_0 - X_D W)' V (X_0 - X_D W)} \tag{2}$$

The matrix of characteristic weights, V, is typically chosen so that the synthetic control best matches the treated group over the pre-treatment period. Specifically, for $W^*(V)$ that minimizes equation (2), V^* is chosen as follows:

$$V^* = \arg\min_{V \in \mathcal{V}} \left(\sum_{t=1}^{T_{pre}} (y_{0t} - Y_{Dt}^{'} W)^2 \right) \tag{3}$$

where $\mathcal{V}$ represents the set of all non-negative $K \times K$ matrices, y_{0t} represents the outcome variable for the treatment group in period t, and T_{pre} represents the length of the pre-treatment period.[35] Synthetic control weights are then $W_{synth} = W^*(V^*)$.

Of course, this is not the only way that weights could be chosen. There are an infinite number of potential alternatives, and so the pros and cons of any particular method should be considered. For example, the restriction that weights must be non-negative and sum to one is meant to guard against extrapolation. This can certainly be a desirable characteristic; a synthetic control group based on extreme extrapolation is unlikely to be a convincing counterfactual. However, this method is not

[34] This allows researchers to emphasize the relative importance of different characteristics of interest.

[35] T_{pre} may be adjusted to a subset of the pre-treatment period, so that the synthetic control predicts both pre-treatment and post-treatment outcomes. This modification further supports the predictive value of included characteristics, and lends credence to the synthetic control's use as a counterfactual.

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:

$$W_{SCUL} = \arg\min_W \left(\sum_{t=1}^{T_{pre}} (y_{0t} - Y_{Dt}^{'} W)^2 + \lambda |W|_1 \right) \tag{4}$$

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 to include in the model; it is an objective way of retaining only variables which most heavily influence the outcome variable in the model. Third, coefficient weights may be negative numbers, which allows variables negatively correlated with the outcome variable to be used to generate synthetic estimates.

A.3 Choosing the Penalty Parameter, λ

Utilizing lasso regression necessitates careful consideration of how the penalty parameter, λ, is chosen. Small values of λ improve in-sample model fit, but to an extreme; when λ is very close to zero, the model is almost identical to OLS. This will almost always result in fitting on noise, which removes the model's ability to make out of sample predictions. In turn, this prevents one from using the synthetic control as a counterfactual, which directly contradicts the method's objective. Instead, SCUL selects the optimal λ by using a procedure known as rollin-origin cross-validation.

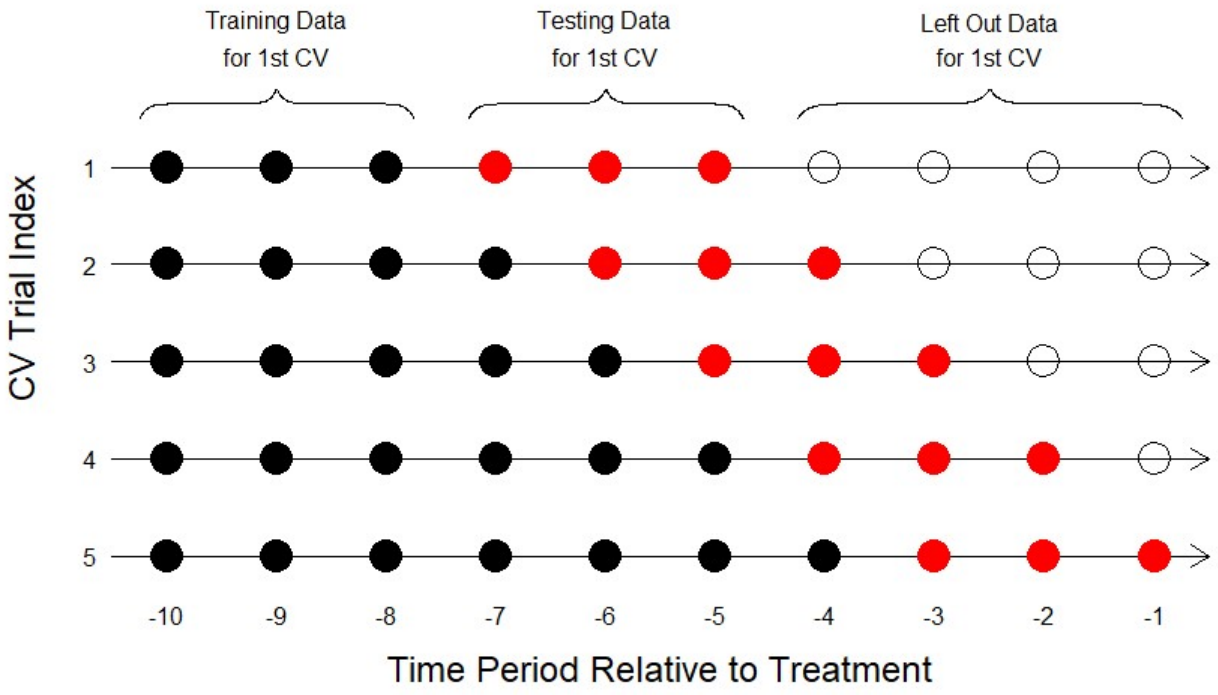
Rolling-origin cross-validation works by partitioning the pre-treatment portion of the data into multiple subsets that include training and testing segments. Lasso regression is performed on the training segment of the data, and the optimal λ value is determined using the testing data. This is done for every potential subset of training and testing data, which generates many potential λ values to choose from. Of these, the median λ value is selected. Given this λ, the corresponding lasso regression is fitted to the entire pre-treatment period, and used to construct the synthetic counterfactual.

There are many potential ways to partition the data into training and testing periods. In this setting, the goal is to predict a sequence of future outcomes, and so the training and testing data are constructed so that they are uninterrupted time periods where the testing period immediately follows the training period. Hollingsworth and Wing (2020) suggest including as many time periods in the testing data as there are in the post-treatment period, and to include at least as many time periods in the training data as there are in the testing data.[37] Each iteration of the cross-validation procedure adds one additional time period to the training data, with the final iteration resulting in a training period that spans 1982 - 1994. This is last training period that allows for three time periods of data in the testing data set (1996 - 2000). Figure A1 provides a visual representation of this procedure.[38]

[37] Because there are 10 time periods of pre-treatment data and 9 time periods of post-treatment data, this is not possible in the current context. Still, there are reasons to be optimistic about the SCUL method's ability to predict post-treatment outcomes in this study. This is discussed further in section 4.1.

[38] Figure A1 is constructed to mimic Figure 1 in Hollingsworth and Wing (2020).

Figure A1: Rolling-Origin Cross-Validation Visualization



A.4 Synthetic Control Weights Using SCUL

Equation (1) makes it clear that the synthetic control is constructed as the summed product of donor series outcomes and weights. This means that donor series weights cannot be interpreted as the share of the synthetic prediction in a given time period. This is true for both the standard synthetic control method and SCUL method.

The SCUL method allows researchers to evaluate both the donor series weights and the corresponding shares of the synthetic control that each variable constitutes. [39] Because reporting weights for each time period is, in most cases, too overwhelming to be useful, the SCUL method restricts these reports to the first and last time period. This provides a general sense of which donor series are most impactful for synthetic control estimation. Moreover, it provides a general indication of whether donor contributions change in importance over time.

[39] Recall that donor series weights in the SCUL method are the lasso regression coefficient for each corresponding donor variable.

A.5 Evaluating Synthetic Control Fit

There is little consensus on how synthetic control fit should be determined. That is, there is not an agreed-upon benchmark (or even metric) for whether a synthetic control group matches the observed group "well enough." The closest most studies come to this is to measure the synthetic control's mean squared prediction error (MSPE) for the treated group, and to rely on naive "eye-tests" of model fit. This is generally used as a baseline to determine whether placebo groups are reproduced well enough by the synthetic control method to warrant use for statistical inference. While this can be useful, concerns exist about how these practices should be best implemented, in general.

In particular, measuring MSPE is unit dependent, and therefore forms a poor basis of comparison across different dependent variables. Although this study does not utilize a broad array of dependent variables, using MSPE to measure model fit is clearly sub-optimal in general. Furthermore, using MSPE for the treated group as a baseline for acceptable placebo fit is a questionable practice. If the MSPE is too large, placebos with poor fit will be included. If the MSPE is too small, too many placebos will be eliminated, making statistical inference impossible.

Hollingsworth and Wing (2020) propose to measure model fit with a modified version of Cohen's D. This is the average difference between the observed outcome and its synthetic counterpart, measured in standard deviations during the pre-treatment period. Specifically, for a measured outcome, s, they let $\sigma_s = \sqrt{\frac{1}{T_{pre}} \sum_{t=1}^{T_{pre}} (y_{st} - y_s)^2}$ represent its average standard deviation over the pre-treatment period. In this context,Cohen's D is $D_s = \sum_{t=1}^{T_{pre}} \left| \frac{y_{st} - y^*_{st}}{\sigma_s} \right|$. This provides a useful metric for evaluating model fit that is standardized across outcome variables. The exact threshold for which Cohen's D indicates "good fit" is arbitrary, but, in general, model fit improves as Cohen's D decreases. Hollingsworth and Wing (2020) suggest a threshold of 0.25; if the synthetic prediction is more than a quarter of a standard deviation removed from its observed counterpart, it is considered to have poor fit, and discarded. This study adopts the same benchmark, though it is relaxed in some unreported robustness checks.

A.6 Estimating Treatment Effects

Estimating treatment effects is straightforward using the SCUL method. Assuming the synthetic control is accepted as a valid counterfactual, any divergence between the observed outcome and its synthetic counterpart during the post-treatment period constitutes a treatment effect. Typically, this is averaged over the entire post-treatment period. That is, the average effect of treatment on the

treated group is $ATT = \frac{1}{(T-T_{pre}-1)} \sum_{t=T_{pre}+1}^{T} (y_{st} - y_{st}^{*})$, 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 effect 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 fit observed outcomes sufficiently well for use in the placebo distribution. To this end, placebos should be evaluated using the same Cohen's D threshold for model fit as the treatment group. Only those placebo variables that meet this threshold for model fit should be used to construct the null distribution. This will almost always eliminate some placebo variables from the null distribution. For example, in this study, 31 states are used to construct Arizona's synthetic counterfactual. Of these, 11 have a have a pre-treatment fit within the 0.25 Cohen's D threshold, and are used to construct the null distribution in baseline results.

Lastly, identification of the ATT dictates that researchers think carefully about which variables are included in the donor pool. Here, there are two primary issues that must be guarded against. First, it is imperative to rule out simultaneity issues between the outcome variable of interest and donor variables in its synthetic counterpart. This means ensuring that there are no donor variables included in the model that are causally affected by the outcome variable being predicted. Typically, this is done by eliminating donor variables from the same group as the outcome variable. In this study's context, the outcome variable of interest is Arizona's gerrymandering metric. To protect against simultaneity concerns, donor variables from Arizona are not used to construct its synthetic counterfactual.[41]

Second, one must rule out the possibility that donor pool variables are themselves the result of treatment. Failure to do so leaves analysis vulnerable to criticism that the estimated ATT is not the effect of treatment itself, but the result of differences between various treated groups. To illustrate this point, this study's goal is to compare Arizona's observed election outcomes to what would have happened had the AIRC not been created. Using states that *did* implement redistricting commissions to

[40] On the other hand, if donor and outcome elements are highly related, there may be a more substantial overlap between donor and placebo variables. For example, if one were to study the effect of a sugar tax on soft drink sales, sport drink sales could be used as both a placebo and donor variable.

[41] This process is also followed when constructing synthetic counterfactuals from other states.

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.