# The Synthetic Control Method

Matthew Chen

## 1 Introduction

The synthetic control method seeks to provide estimates for causal effects after some intervention or treatment on time series data (i.e., a comparative case study). Naive methods for estimating the treatment effects, i.e., the effect of the intervention, rely roughly on comparing the average change in the response or outcome variable of interest for the treatment group to the average change in the response for the control group. However, the fundamental problem with this approach is the units in the control group may be fundamentally different than those in the treated group, and thus are not comparable to each other. The synthetic control method provides a solution to this problem by taking a weighted average of the control group units so that they are more comparable to the treated unit.

## 2 Methodology

In this section, I provide a brief overview of the synthetic control method, as provided in (Abadie et al., 2010). Suppose there are J+1 units observed for T timesteps, and assume (without loss of generality) that only the first one is exposed to the intervention at some time $T_0$. This leaves J potential control units, which we refer to as the "donor pool." Additionally, we assume that the intervention has no effect before time $t=T_0$. Let $Y_{it}^{I}$ and $Y_{it}^{N}$ be the potential outcomes, corresponding to the outcome observed with and without treatment, respectively, for the $i^{th}$ unit at the $t^{th}$ timestep.

We define the effect of intervention (i.e., the causal effects) at time t as the difference between the potential outcomes:

$$\alpha_{it} = Y_{it}^{I} - Y_{it}^{N}$$

The goal is to estimate the casual effects of the intervention on the treated unit (j=1) for all timesteps $t > T_0$, i.e., we want to estimate the sequence $\alpha_{1T_0+1}, \dots, \alpha_{1T}$. These can be written as:

$$\alpha_{1t} = Y_{1t}^{I} - Y_{1t}^{N} = Y_{1t} - Y_{1t}^{N}, \qquad t > T_0 \tag{1}$$

The universal missing data problem of casual inference becomes apparent: for the treated unit we do not observe its outcomes as if it never had received the treatment and vice versa for the untreated units. Thus, the overall goal of the synthetic control method is to use a weighted average of the control units in the donor pool to estimate the control/untreated effects for the treated unit.

If we assume the following model structure:

$$Y_{it}^{N} = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \epsilon_{it} \tag{2}$$

Here, $\delta$ is an unknown common factor, $\theta$ is a vector of unknown parameters, Z is a vector of observed covariates, $\lambda$ responds to unobserved common factors with its corresponding unknown factor loadings $\mu$, and $\epsilon$ is a zero-mean error term.

Suppose there exists a vector of weights $\left(\mathrm{w}_2^*, \ldots., \mathrm{w}_{\mathrm{J}+1}^*\right)$ such that the following are true simultaneously:

$$\sum_{j=2}^{J+1} w_j^* Y_{jt} = Y_{1t}, \qquad t \in 1, \ldots, T_0 \tag{3}$$

$$\sum_{j=2}^{J+1} w_j^* Z_j = Z_1 \tag{4}$$

Then consider the following quantity:

$$Y_{1t}^N - \sum_{j=2}^{J+1} w_j^* Y_j t \tag{5}$$

Under mild regularity conditions, (Abadie et al., 2010) showed that under expectation, (5) is upper-bounded by a quantity that goes to zero, and were also able to obtain similar results for a more complicated autocorrelation model as opposed to the structure stated in (2). This is their underlying motivation for proposing an approximately unbiased estimator for $\mathrm{Y}_{1\mathrm{t}}^{\mathrm{N}}$ as a weighted average or convex combination of units in the donor pool, granted that they match the covariates and outcomes for the treated unit in the pre-intervention period such as in (4) and (5). We use the estimator:

$$\widehat{\alpha_{1t}} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}, \qquad t \in \{T_0 + 1, \ldots, T\} \tag{6}$$

### 2.1 Implementation

Let $W \in \mathbb{R}^{J+1}$ be a vector of positive weights that sum to 1, which when estimated, will be used to compute Equation 6. For the treated unit, let $X_1 = \left(Z_1, Y_{pre,1}\right)$ where $Z_1$ is a vector of observed covariates and $Y_{pre,1}$ are the pre-intervention outcomes. Similarly, define $X_0$ as a (k x J) matrix containing the observed covariates and pre-intervention outcomes for the controls.

The vector of weights is obtained by solving the following optimization problem.

$$W^* = argmin_W^* \left|\left|X_1 - X_0 W^*\right|\right|_V \tag{7}$$

The matrix V is a (k x k) positive semi-definite matrix such that:

$$\left|\left|X_1 - X_0 W\right|\right|_V = \sqrt{\left((X_1 - X_0 W)^t V (X_1 - X_0 W)\right)} \tag{8}$$

Note that the matrix V weights the importance of the different covariates and pre-intervention outcomes in X. Abadie et al., 2010 recommended jointly choosing V (diagonal and positive definite) and W such that the mean square error of predicting the control units using the synthetic control is minimized, though other choices are possible. This includes the choice of V using domain expertise, or naively, simply as the inverse variance of the included variables (Abadie & Gardeazabal, 2003).

### 2.2 Inference

We are interested in testing the null hypothesis of no intervention effect, i.e., we want to know if the estimated intervention effect is significant and not likely to be a product of only random chance. Inference for synthetic control is similar to other randomization inference methods where treatment assignments are randomly permutated to empirically derive the null distribution. To derive placebo effects, we permutate the assignment of units to the treatment and control groups, and re-conduct the synthetic control method on the placebo units to estimate the causal effects.

## 3 Data Analysis

In this section I demonstrate the synthetic control method using data from California's Proposition 99. This is the same data as in Abadie et al., 2010, though my goal is not to replicate their analysis but to understand the synthetic control method and how it is implemented.

### 3.1 Background

Proposition 99 was the first largescale statewide tobacco control program in the United States, which was passed in 1988 and went into effect in 1989. The legislature increased the cigarette excise tax by 25 cents per pack and funded a broad range of tobacco prevention programs across the state (Siegel, 2002). In this data analysis, we answer the question if causal effects can be found between decreasing cigarette sales and the passing of Proposition 99 in California.

Assuming no other unmeasured confounding variables, the state level variables that we control in this study are the following: log – GDP per capita, per capita beer consumption, the proportion of the population between ages 15-24, and the cigarette price. The data spans between 1970 and 2000, for a pre-intervention period of 19 years. Additionally note that Abaldie found that the results were robust to the inclusion of many other state-level demographic variables.

Further, following the example of California, the states of Massachusetts, Arizona, Oregon, and Florida also introduced large tobacco control programs, and as a result were not included in this analysis. Other states also introduced increasing in the cigarette excise tax, and these were also excluded. In total, we obtain a donor pool of 38 states.

### 3.2 Results

To solve for the optimal choice of the weight matrix W and the importance matrix V, I utilize the Python package SyntheticControlMethods. After solving, we can evaluate the fit of the synthetic control in terms of the pre-intervention outcomes and observed covariates (Table 1). We find that the covariates for the synthetic control are reasonably close to the actual values, and the per-capita cigarette sales (in the pre-intervention period) are much closer compared to the average of the control units. This gives confidence toward the use of the synthetic control.

*Table 1. Covariates (averaged in the pre-intervention period) for the synthetic California compared to the real California and the average of the control units. Importance scores are shown, and correspond to the diagonals of the estimated V matrix.*

| | California | Synthetic California | Average Control Units | Importance |
|---|---|---|---|---|
| per capita cigarette sales | 116.21 | 116.21 | 130.57 | 0.152 |
| Ln(GDP per capita) | 10.03 | 9.72 | 9.79 | 0.050 |
| per capita beer consumption | 24.28 | 22.9 | 23.7 | 0.303 |
| proportion age 15-24 | 0.18 | 0.19 | 0.18 | 0.405 |
| cigarette retail price | 66.64 | 64.21 | 64.50 | 0.090 |

We also can plot the synthetic control compared to the observed treated unit (Figure 1). Note that in the pre-intervention period, the synthetic control matches the trend of the observed outcome reasonably well. This provides confidence in its use as an estimate for California's cigarette sales if Proposition 99 never took place. Plotting the estimated causal effects (Figure 2), or the difference between the observed treated data and the synthetic control, we see cigarette sales declined sharply.

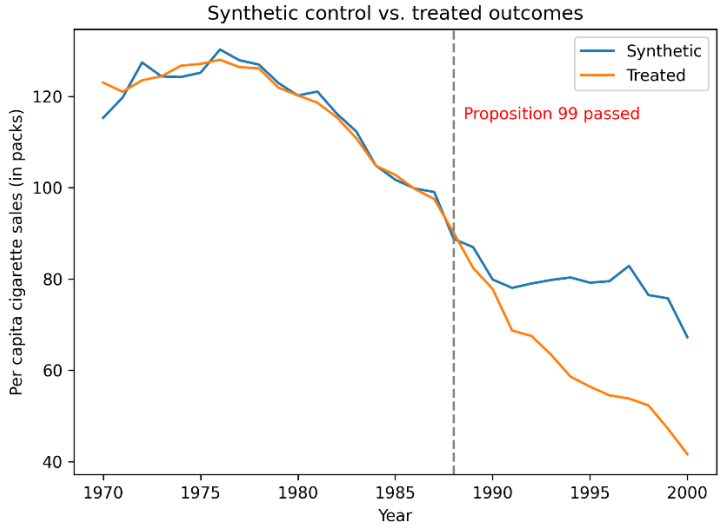


*Figure 1. Synthetic control vs. treated outcome.*

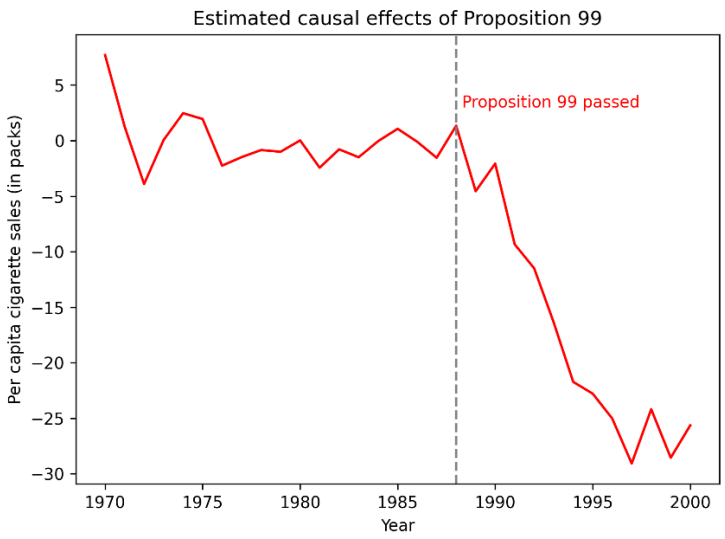


*Figure 2. Estimated causal effects of proposition 99 on cigarette sales*

### 3.3 Inference

Figure 3 plots the estimated effects for all of the placebos after conducting randomization inference. Immediately, we notice several important results. First, California has a very good pre-treatment fit which again confirms that we found a reasonable synthetic control. If California did not fit well in preintervention, the researcher cannot judge the validity of the estimated effects – it may only be a product of poor fit. Second, the effects for California in the post-intervention period seem significant (more dramatic) compared to most of the placebo effects. And finally, we note that several states did not have a good fit in the pre-intervention period. It does not make sense to compare placebos with good fits with placebos without good fits. Deleting placebos that have poor pre-intervention fits based on the pre-intervention root mean squared prediction error (RMSPE) is an option, but the choice of the RMSPE threshold can be subjective.

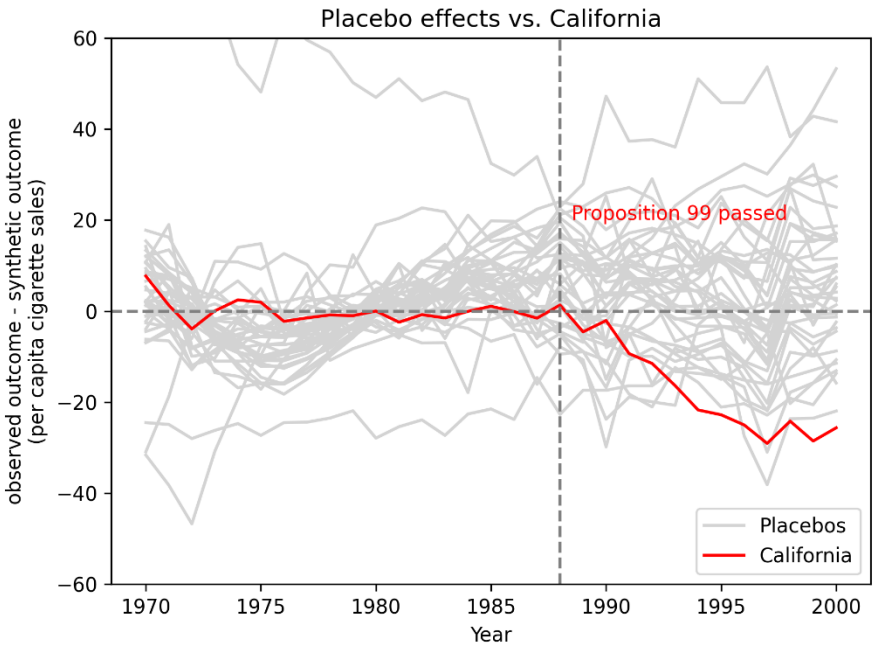


*Figure 3. Results from randomization inference*

Instead, Abadie et al., 2010 proposes analyzing the ratio between the RMSPE in the post-intervention period and the pre-intervention period. This way, no threshold for removing placebos is needed (i.e., if the fit is poor in both pre and post intervention, the ratio is still small). We can plot the empirical null distribution (on the null hypothesis of no treatment effect) which is shown in Figure 4.

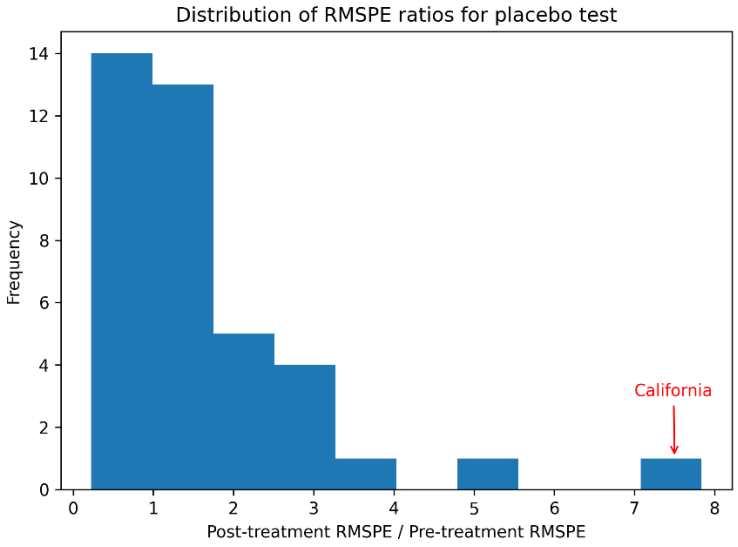


*Figure 4. Empirical null distribution of the RMSPE ratio*

We find that California's RMSPE ratio is the largest among the placebos, by a large margin. Calculating the p-value for California (using the percentile on the empirical null distribution), we have a probability of 1/39 or 0.026 of observing results as extreme as California if the treatment assignment was chosen at random.

### 3.4 Other Considerations

In causal inference, the assumption of no interference between units is commonly made. This is a strong assumption for this study since we know that Proposition 99 in California inspired similar programs in other states. It is possible that anti-tobacco sentiment in California would spread to other states. There was also increased advertising backlash from tobacco companies, which significantly increased their spending budget in California after the passing of Proposition 99. This may affect the funding available for other states, potentially lowering smoking rates outside of California. However, these effects would only lower smoking rates in the donor pool and lower the estimated effects of Proposition 99, and yet we still find significant results in California. Thus, these concerns are unlikely to undermine the results.

## 4 Conclusion

In this report, I introduced and explained the use of the synthetic control method, as proposed by (Abadie et al., 2010; Abadie & Gardeazabal, 2003). In the synthetic control method, the overall idea is to take weighted combinations of observations in the donor pool to produce a suitable synthetic control to compare the treated unit with, where the outcome before the intervention and other covariates are reasonably matched. I also applied the method to data from Proposition 99, as found in (Abadie et al., 2010). Here, we make the same conclusions as Abadie, and find that Proposition 99 significantly lowered per-capita cigarette sales in California.

## 5 References


Abadie, A., Diamond, A., & Hainmueller, A. J. (2010). 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. https://doi.org/10.1198/jasa.2009.ap08746

Abadie, A., & Gardeazabal, J. (2003). *The Economic Costs of Conflict: A Case Study of the Basque Country*.

Siegel, M. (2002). The effectiveness of state-level tobacco control interventions: A review of program implementation and behavioral outcomes. In *Annual Review of Public Health* (Vol. 23, pp. 45–71). https://doi.org/10.1146/annurev.publhealth.23.092601.095916