Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics and Operation Research

Assessment on predictors in Synthetic Control Method

Yu-Shan Cho (558516)

Abstract

The synthetic control method (SCM) is a widely used counterfactual estimation tool for studying the effects of policies or interventions. However, there are growing concerns regarding the lack of guidance on the selection of predictors in SCM. This paper aims to contribute to the ongoing discussion by investigating the use of different covariates across various simulation settings and an empirical setting.

The simulation study complements previous results shown by Ferman et al. (2020) and provides additional insights. One key finding is the preference for using time-invariant covariates as predictors, as opposed to the commonly employed practice of using the averaged value of outcomes in empirical studies. In the empirical analysis, an anti-tobacco program implemented in California is examined. By thoroughly examining the results obtained from different SCMs using different predictor sets, this study discovers that achieving approximate balance in certain covariates proves to be beneficial, and the relevance of these covariates in improving estimation is further verified by using an SCM that incorporates machine learning techniques. Furthermore, the study provides discussions on the linearity of covariate effects, supported by recent theoretical results.

Supervisor:	Dr. Wendun Wang
Second assessor:	Ph.D. Jens Klooster
Date final version:	2nd July 2023

Contents

1 Introduction

The Synthetic Control Method (SCM) has made remarkable advancements in the field of comparative case studies since its development by Abadie and Gardeazabal (2003), Abadie, Diamond and Hainmueller (2010), and Abadie, Diamond and Hainmueller (2015). This innovative technique for counterfactual estimation has garnered widespread adoption in estimating the effects of interventions or policies, primarily due to its transparent and interpretable nature. Precisely, it involves constructing a counterfactual scenario that enables researchers to estimate what would have been observed in the absence of the intervention.

The estimation procedure begins by constructing a pre-treatment trend that closely resembles the variable of interest, using data from untreated control units. This trend is then projected forward in the post-treatment period to estimate its potential trajectory. Then, by comparing the actual post-treatment outcome of the treated unit with the outcome estimated by this synthetic control (SC) unit, the effect of the intervention could be identified. Existing empirical studies of SCM cover a wide range of topics, including the impact of terrorist attacks on elections (Montalvo, 2011), the influence of the European monetary union on economic growth (Fernández & Garcia-Perea, 2015), the effects of right-to-carry laws on violent crime (Donohue et al., 2019), and the impact of lockdowns on air quality during the pandemic (Cole et al., 2020).

Despite that SCM being widely used in comparative analyses, there is an increasing number of critics regarding the lack of transparency surrounding the selection of outcome variables and other covariates used as predictors. This lack of transparency can tempt researchers to engage in specification searching, which ultimately leads to subjective statistical inferences and undermines the reliability of the method. Ferman et al. (2020) have extensively researched this issue within the context of the canonical SCM. Their theoretical and simulation findings suggest that using all pre-treatment outcome lags as matching variables is preferable, as this specification satisfies certain conditions that ensure convergence to some SC unit.

In contrast to the study conducted by Ferman et al. (2020) that primarily focuses on the number of pre-treatment outcomes, Kaul et al. (2022) investigate the influence of covariates on estimation. Their research demonstrates the existence of a trade-off between bias and variance when deciding whether to include covariates. In their simulation experiment, it is observed that using all pre-treatment outcome lags leads to the smallest variance, but also results in the largest bias compared to the scenario where covariates are included. They further confirm that the exclusion of covariates is the primary driver of this bias.

Given the ongoing debate regarding the selection of predictors and covariates in the SCM, this study aims to contribute by examining the effects of utilising different sets of predictors on various data structures and empirical data. The simulation study in this research makes a two-fold contribution. First, it identifies certain ambiguities in the study conducted by Ferman et al. (2020), offering a critical examination of their findings. Secondly, the study presents a series of analyses that successfully address and clarify these concerns, while also provides additional insights into the effectiveness of using different predictors.

Specifically, when the pre-treatment period is short, the specifications recommended by Ferman et al. (2020) are susceptible to over-fitting issues. In cases where the pre-treatment period is long, all the considered specifications generally perform well if the outcome variable is

stationary, whereas their effectiveness becomes more diverse when the outcome variable exhibits a trend. In such cases, the use of a time-invariant covariate results in smaller errors compared to using the averaged values of outcome variables, which is a common practice in empirical studies. Lastly, a newly developed SCM also provides additional insights into the impact of precisely balancing certain covariates with varying characteristics.

The empirical part of this study makes contributions to the existing literature by presenting an overview of results obtained from using different SCMs and diverse sets of predictors. By thoroughly examining the results, this research provides detailed discussions that are grounded in recent research on the role of covariates. These discussions encompass the relevance of covariates in estimation bias and the properties of their effects, particularly in terms of linearity.

The paper is structured as follows. Section 2 provides literature overview on the studies discussing covariates and the recent development in SCM. In Section 3, the standard SCM estimation procedure is introduced, followed by the two extended versions. Next, Section 4 demonstrates the simulation study, including discussions about the study design and the results. In Section 5, an empirical study is conducted. Finally, Section 6 concludes.

2 Literature

The growing number of promising applications of SCM has resulted in a surge of research in this field. An area of research examines the subjective nature of predictor and control unit selection, which can undermine the transparency of SCM. The lack of guidance in this area was first identified by Dube and Zipperer (2015a), although the study does not explore its implications about searching opportunities. Nevertheless, the authors propose a procedure based on cross-validation to identify the optimal set of predictors, using mean squared prediction error (MSPE) criteria. The issue of specification searching is explicitly expressed by Ferman et al. (2020), who conclude that using all pre-treatment outcomes as predictors is preferred over alternative options. The conclusion of this study is primarily grounded in the theoretical result that estimators are asymptotically equivalent across different specifications when the number of pre-treatment outcome lags used as predictors approaches infinity, given that the number of pre-treatment periods also tends to infinity. This theoretical finding is further supported by the results of their simulation study.

Regarding the role of predictors, several studies have provided insights into their importance and the implications for estimation. Kaul et al. (2015) delve into the inner workings of SCM under various specifications. Their key finding is that when all pre-treatment outcomes are used as predictors, the optimisation procedure does not take into account any covariates input as a predictor. This is because the algorithm essentially includes the same variables as both outcome and regressor. Furthermore, Botosaru and Ferman (2019) demonstrate that it is possible to derive a looser bound on the bias when only balance in pre-treatment outcomes is achieved. However, achieving balance in both covariates and pre-treatment outcomes can result in tighter bounds on the bias of the SC estimator compared to achieving balance in pre-treatment outcomes alone. The authors also show that when covariates have nonlinear effects on the potential outcomes or when their effects are multicollinear with the effects of other observed and unobserved covariates, achieving perfect balance on lagged outcomes does not imply approximate

balance for those covariates. This finding reinforces the idea of effectively utilising covariates as they may provide tighter bounds. Another study supporting the use of covariates with averaged outcomes is presented by Abadie (2021). The authors provide two additional reasons beyond the issue of biasedness. First, they argue that the co-movement of outcome variables across different units is still absorbed by the SC unit, as this variable across all units shares a similar movement over time. Additionally, using fewer predictors can enhance the interpretability of the results, as fewer control units receive positive weights.

Another area of research contributes towards the development of the techniques themselves. These advancements have greatly expanded the applicability and versatility of SCM. For instance, there are SCMs that can handle multiple pre-treatments across multiple units (Xu, 2017), or address the challenges posed by missing data to ensure reliable estimation and inference (Amjad et al., 2018). A recent innovation in SCM is the augmented SCM (ASCM) proposed by Ben-Michael et al. (2021). The ASCM is specifically designed to address situations where the estimated SC unit does not align well with the true values in the period prior to treatment. By correcting the estimation due to imbalance in outcomes and incorporating additional information about covariates, the ASCM improves the accuracy of SCM estimates. In addition, the sparse SCM (Sp-SCM) developed by Quistorff et al. (2021) automates the predictor selection procedure through regularisation, and reduces over-fitting by minimising the post-treatment errors to instead of the pre-treatment errors.

While there is an increasing amount of theoretical results available in the literature, each study tends to focus on different data assumptions. This study does not aim to generalize all of the assumptions or conduct extensive simulations encompassing all previously considered settings. However, by employing three distinct data generating processes and considering various predictor sets, this study establishes connections between theoretical findings and diverse assumptions presented in the literature.

3 Methodology - Synthetic Control Method

Assume that $J+1$ units indexed by j are arranged such that the treated unit is positioned first ($j = 1$), and that there are T time periods indexed by t. Besides, this treated unit is affected by the intervention from period $T_0 + 1$ to T.

The ultimate goal of SCM is to estimate the treatment effect $Y_{1t}^I - Y_{1t}^N$ for $t > T_0$, where Y_{1t}^N denotes the potential outcome that would be observed for unit 1 in period t in the absence of treatment and Y_{1t}^I is the observed outcome under treatment. Since Y_{1t}^I is observable, the goal of the algorithm is to estimate the counterfactual outcome resembling unit 1. This counterfactual outcome is constructed by taking a linear combination of the outcomes from the remaining J control units. Mathematically, it can be expressed as $\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}$, where $\hat{w}_j$ represents the weights assigned to the control units.

To estimate the unit weights, SCM utilises the predictors of the pre-treatment observed outcomes, which can be linear combinations of Y_{1t} in the pre-treatment periods, or covariates that have explanatory power for Y_{1t}. Let a $(k \times 1)$ vector X_1 and a $(k \times J)$ matrix X_0 contain the predictors for the treated unit and for all control units, respectively, the optimisation process then searches for the linear combination of the columns of X_0 that represents X_1 as close as

possible. The estimation is sometimes referred to the inner optimisation in the literature, where the weights are chosen such that the distance metric is minimised:

$$\hat{w} = \underset{w \in W}{\operatorname{argmin}} \sqrt{(X_1 - X_0 w)' V (X_1 - X_0 w)} \tag{1}$$

$$= \underset{w \in W}{\operatorname{argmin}} \left(\sum_{k=1}^{K} v_k (X_{1k} - \sum_{i=2}^{N} X_{ik} w_i)^2 \right)^{1/2}, \tag{2}$$

where vector w contains the weights for each control unit to construct the synthetic control, while the relative importance of the predictors v are stored in a diagonal positive semi-definite matrix V. Besides, the typical SCM imposes a simplex constraint $W = \{\hat{w} \in \mathbb{R}^J | w_j \geq 0 \text{ and } \sum_{j \neq 1} w_j = 1\}$ on unit weights, requiring them to sum up to one while ensuring that each has a non-negative values. It is argued that this constraint prevents the method from extrapolating and reserve the interpretability (Abadie, 2021).

In the outer optimisation that concerns about the selection of V, one chooses from a set of positive semi-definite matrices to minimising the in-sample error:

$$V = \underset{V}{\operatorname{argmin}} (Y_1^* - Y_0^* w(V))' (Y_1^* - Y_0^* w(V)), \tag{3}$$

where Y_1^* and Y_0^* denote the subset of their corresponding outcome variables Y over the chosen pre-intervention periods.

3.1 Augmented synthetic control (ASCM)

When achieving a perfect pre-treatment fit in outcome variables is not feasible, particularly when the treated unit is outside the convex hull, the ASCM proposed by Ben-Michael et al. (2021) offers a solution to improve the estimates through extrapolation. Besides, with the proposed two-step approach, ASCM can even achieve perfect balance in covariates. In short, this procedure involves the following main steps: First, the pre- and post-treatment outcomes are residualised based on the covariates. Then, the ASCM weights are estimated using the residualised outcomes. The detailed estimation procedure are provided in Appendix C.

While Ferman et al. (2020) has demonstrated the existence of search opportunities regardless of whether a time-invariant covariate is included in the analysis, this study aims to delve deeper into the impact of balancing different covariates using ASCM. As discussed by Botosaru and Ferman (2019), obtaining good balance in both outcomes and covariates is beneficial as it can result in tighter bounds on estimation errors compared to solely focusing on the outcome variable. Therefore, the present study seeks to explore the importance of different covariates with distinct characteristics. To achieve this, several covariates are included in the data generating process, as introduced in Section 4. Additionally, in addition to investigating the covariates, the approach that focuses solely on de-biasing the original SCM estimates due to imbalance in pre-treatment outcomes is also conducted.

3.2 Sparse synthetic control (Sp-SCM)

Another extended version of SCM is the Sp-SCM proposed by Quistorff et al. (2021). The development of Sp-SCM is primarily motivated by two factors: the lack of guidance on selecting pre-treatment predictors and the issue of over-fitting that arises when the number of model parameters increases. Besides, it addresses the concern raised by Abadie (2021) regarding the non-uniqueness of unit weights, which can lead to arbitrary results and reduce reproducibility.

Sp-SCM tackles these issues by automating predictor selection through regularisation on both the unit and variable weights, resulting in a sparse set of predictors that capture the most informative characteristics. There are two important features of Sp-SCM. Firstly, Sp-SCM differs from other SCMs in that it optimises the variable weights using the outcome variables in the post-treatment period. This approach brings several advantages, the primary one being the elimination of over-fitting problems that can arise when pre-period outcomes are used as both targets and predictors. Furthermore, this adjustment in Sp-SCM helps mitigate the issue of degeneracy in covariates discussed by Kaul et al. (2015). The second feature is the relaxation of the simplex constraint on unit weights, similar to ASCM. The authors further demonstrate the advantage of this feature by arguing that it enables the SC to learn from donor units that exhibit counter-cyclical patterns.

Since the automated predictor selection is a unique feature of Sp-SCM that sets it apart from other existing SCMs, it is of interest to compare its results with those of other methods and explore the covariates selected by Sp-SCM. However, due to the limited compatibility with certain programming languages, this technique is only applied to the empirical study. Appendix D provides a review of this method, including two versions of Sp-SCM: the "Fast" version and the "Full-joint" version.

4 Simulation Study

To examine the presence of searching opportunities in the implementation of SCM, this study employs the Monte Carlo Simulations framework and builds upon the data generating process (DGP) utilised by Ferman et al. (2020). Moreover, it extends the framework by incorporating an additional DGP that places emphasis on covariates which exhibits a trending characteristic.

The DGP employed in the work by Ferman et al. is formulated as follows:

$$Y_{jt}^0 = \delta_t + \lambda_t^k + \varepsilon_{jt}, \qquad k = 1, ..., 10 \tag{4}$$

where a time-varying effect δ_t shared among units follows a standard normal distribution, a stationary trend denoted by λ_t^k follows an autoregressive process of order 1 with a serial correlation parameter of 0.5, and the transitory shock ε_{jt} is normally distributed with a standard deviation of 0.1. Moreover, the first two units adhere to the stationary trend λ_t^1, followed by the next two units adhering to the trend λ_t^2, and so forth. For simulations that involve a time-invariant covariate, the design also follows Ferman et al. (2020). In such DGP, an additional term $\theta_t Z_i$ is introduced in the equation, where θ_t is the parameter following a standard normal distribution, and binary variable Z_i takes the value of 1 for the first ten units while 0 for the last ten.

To evaluate the presence of trend in covariates as well as covariates with different levels of impact, this study utilises a DGP modified from the non-stationary model in the study of Ferman et al. (2020). While this previous study does not consider any covariate used as a matching variable, this study is particularly interested in the usage of covariates. Furthermore, this modification is also motivated by the need to capture real-world scenarios where covariates exhibit trends, and is common to see in empirical research. To allow investigation on covariates with different characteristics, three stationary covariates are introduced into the DGP, each with varying degrees of impact. The equation for this DGP is as follows:

$$Y_{jt}^0 = \delta_t + \lambda_t^k + \theta_1 Z_{i1} + \theta_2 Z_{i2} + \theta_3 Z_{i3} + \phi_t Z_{i4} + \varepsilon_{jt}, \qquad k = 1, ..., 10 \tag{5}$$

where Z_i is a binary variable, similar to before, and the first ten units receive strong, intermediate, and negligible effects denoted by $\theta_1 \sim N(2, 1)$, $\theta_2 \sim N(0, 1)$, and $\theta_3 \sim N(0, 0.1)$, respectively. Additionally, a non-stationary trend term ϕ_t follows a random walk process with a shock term drawn from a standard normal distribution. To imitate real-world situations where measurement error often occurs during data collection process, each value for the covariate term is subjected to a random error when being used as a matching variable in the SCM algorithm. More details related to this random error formulation are provided in Appendix B.

The simulation involves generating 100 rounds of datasets for each specified time period T_0 based on each equation discussed above, and no treatment is imposed on any unit. In each dataset, the twenty units are considered as a placebo treated unit iteratively, resulting in a total of 2000 SC units estimated for each T_0. It is worth noting that the original study generates 10000 SC units using 500 datasets, while this study conducts a smaller set of experiments.

The objective of this simulation is to examine the null hypothesis of no substantial treatment effect when employing various specifications. Specifically, Ferman et al. (2020) are interested in determining the probability of rejecting the null hypothesis for at least one specification. When searching opportunities are present, the probability of finding one specification that rejects the null can be significantly higher than an intended significance level, due to the widely different estimated SC units from different specifications. Their analysis method and the comments regarding their approach are provided in the subsequent subsections.

4.1 Specifications

Following Ferman et al. (2020), this study utilise seven specifications in the simulation study. The first specification includes all pre-treatment outcome values, the second one contains the first three-fourths, and the third one uses the first half of the pre-treatment outcome values. The fourth and fifth specifications consider the odd pre-treatment outcomes and their even counterparts, respectively. Lastly, specifications 6 and 7 use the mean value of pre-outcomes and three specific outcome values (the first one, the middle one, and the last one), respectively.

4.2 Rejection rate

Following Ferman et al. (2020), the ratio of the MSPE (RMSPE) is used to infer the implications of the resulting SC units regarding specification searching. This metric is specified as:

$$RMSPE_j := \frac{\frac{\Sigma_{t=T_0+1}^{T}\left(Y_{jt}-\widehat{Y}_{jt}^{N}\right)^2}{T-T_0}}{\frac{\Sigma_{t=1}^{T_0}\left(Y_{jt}-\widehat{Y}_{jt}^{N}\right)^2}{T_0}}. \tag{6}$$

Based on the concept of a placebo test, the authors reject the null hypothesis at a 5 percent significance level if the treated unit exhibits the highest RMSPE among the twenty units. This procedure is commonly used in SC empirical studies to assess significance. Additionally, it is typically accompanied by the calculation of a p-value, as proposed by Abadie et al. (2010): $p := \frac{\sum_{j=1}^{J+1} 1[RMSPE_j \geq RMSPE_1]}{J+1}$, where the treated unit is assumed to be indexed as 1. In this simulation design comprising twenty units, this approach suggests that the placebo test would have a rejection rate of 5 percent by design, when solely considering a single specification.

In the context of specification searching, the objective is to determine the probability of rejecting the null hypothesis at the 5 percent significance level in at least one specification. To achieve this, the RMSPE rankings for all twenty units are recorded in each simulation round for every specification. In an ideal scenario, it would be expected that the probability of rejecting the null hypothesis in at least one specification closely aligns with the 5 percent threshold. However, this ideal outcome is only realised when all specifications consistently identify the same highest-ranking RMSPE among the twenty units. If different specifications produce conflicting results, the probability of rejecting the null hypothesis in at least one specification will be higher. This demonstrates the potential chances of specification searching when applying SCM.

Nevertheless, potential concerns arise in this procedure. Firstly, the use of a ratio metric to indicate searching opportunities presents challenges in accurately determining whether certain specifications truly result in worse outcomes. This is because a ratio alone does not provide sufficient information to ascertain whether a higher RMSPE is due to small errors in the pre-period or if the post-period error is truly relatively large. This limitation can introduce ambiguity and uncertainty in the interpretation of the results. Moreover, the authors focus on the ranking of RMSPEs across units for each specification rather than using the true numeric values. While this approach may provide insights into the relative performance of each specification on the twenty placebo units, it can lead to ambiguity and misunderstandings when comparing between the specifications. For example, it is possible that a particular specification has smaller errors overall but exhibits a different ranking of RMSPEs for each placebo unit compared to other specifications with higher errors. Lastly, it would be more appropriate to acknowledge that the variations in the ranking of RMSPE across all specifications contribute to the final probability found. The theoretical evidence derived by Ferman et al. (2020) suggests that the unit weights calculated based on specifications under a certain condition will converge to the same values as the number of pre-treatment periods increases. Consequently, the resulting synthetic unit and RMSPE rankings tend to be similar among these specifications. However, it is important not to immediately interpret this as evidence that alternative specifications specifically contribute to the issues of specification searching, as the variations in RMSPE rankings are a collective result of all specifications. These concerns still exists even under the condition of good pre-period fit.

To address these concerns, it is crucial to investigate the actual errors and differences between specifications rather than relying solely on a derived ratio metric. This study aims to add

more context on the simulation analysis, providing a more comprehensive understanding of the strengths and weaknesses of each specification.

4.3 Methods for analyses

This section introduces the measure to analyse the simulation results, with a specific focus on the pre-treatment periods $T_0 = 12$ and 100.

The primary objective is to assess the performance of various specifications under different DGPs via post-treatment errors. The section first introduces a measure to determine the quality of pre-treatment fit, and conducts a supplementary analysis on unit weights to further investigates the results shown by Ferman et al. (2020). Lastly, the errors are analysed.

Normalised mean squared error for outcome variables assessment

As emphasised by Abadie et al. (2010) and Abadie et al. (2015), achieving a strong pre-treatment fit for both outcome variables and covariates is crucial when employing SCM. To assess the fit of pre-treatment outcome variables, the normalised mean squared error ($\widetilde{R}^2$) is utilised, following the approach adopted by Ferman et al. (2020). It is defined as follows:

$$\widetilde{R}^2 = 1 - \frac{\Sigma_{t=1}^{T_0}(Y_{1t} - \widehat{Y}_{1t}^N)^2}{\Sigma_{t=1}^{T_0}(Y_{1t} - \overline{Y}_{1t}^N)^2}, \quad \text{with} \quad \bar{Y}_1 = \frac{\Sigma_{t=1}^{T_0} Y_{1t}}{T_0}. \tag{7}$$

A value of one indicates a perfect fit. This study follows Ferman et al. (2020) in adopting a more lenient restriction, where the the analysis is restricted to condition of "at least one specification has good fit." As noted by the authors, using a more stringent restriction would result in a substantially higher probability of rejecting the null in at least one specification. This is because the test statistics RMSPE for placebo units are not conditional on a good pre-treatment fit. Therefore, the use of a more stringent would result in over-rejection Ferman and Pinto (2017).

Convergence in unit weights

In addition to the rejection rate discussed in Section 4.2, which calculates the probability of at least one specification indicating significance, this study delves deeper into the scenario where a subset of specifications concur on significance.

By examining the ratio of the counts of significant effects based on at least one of the first five specifications to the counts based on a subset of specifications, this analysis offers additional insights into the convergence of unit weights across those specifications. Specifically, this analysis focuses solely on the treatment units where the associated five synthetic units all exhibit good fit. This definition of good-fit is stricter compared to the one used when calculating the rejection rate, where the treatment unit is included in the analysis if any of the five specifications indicates a significant effect. This restriction provides a good chance for those specifications to converge in the same weights, and allows for a more intuitive understanding.

Post MSPE

Finally, the quality of the estimate is evaluated using the associated post-treatment errors. This evaluation follows a similar approach employed by Dube and Zipperer (2015b).

The authors of the study acknowledge the lack of guidance regarding the selection of predictors, and aim to determine the optimal choice of predictors using a placebo framework to assess the prediction error associated with a given set of predictors. In their empirical analysis, they treat each donor unit as a placebo since no treatment is assigned to any of them. Then, the square root of the average of post-treatment MSPE is calculated over all placebo units. The optimal specification is determined based on the smallest average MSPE in the post-period.

In contrast to Dube and Zipperer (2015b), this study calculates the MSPE without taking the square root. This decision is made to provide a clearer understanding and interpretation of the results, as it aligns with the formulation of the RMSPE in Equation 6. The post-MSPE for a certain set of predictors and specification is calculated as:

$$MSPE_j = \frac{\Sigma_{t=T_0+1}^{T} \left(Y_{jt} - \widehat{Y}_{jt}^{N}\right)^2}{T - T_0}. \tag{8}$$

Furthermore, as the focus in this analysis is the performance of each specification individually, only the SC units that exhibit a good pre-treatment outcome fit for the specific specification being analysed are considered. As different specifications may result in different amount of SC units with good-fit, the MSPE values over all those SC units are averaged separately for each specification and compared between them.

Even though Ferman et al. (2020) analyse the quality of specification by calculating the proportion of unit weights that are misallocated for each specification, one should note that the analysis is conducted without any condition on the pre-treatment fit, which is an essential part in empirical applications. The results of specifications 6 and 7 failing to provide accurate weights may be attributed to their higher occurrence of poor-fit, while results may be differ when the SC units being analysed are constrained to those with good-fit condition. The quality examination in this study, although using a different approach that is based on errors, it aims to compensate the evaluation analysis in the previous study by conditioning on good pre-treatment fit for each specification separately.

4.4 Simulation Result

Rejection Probability

Table 1 provides the probabilities of rejecting the null hypothesis at the 5% significance level for at least one specification, which partially replicates the findings of Ferman et al. (2020). However, it is important to note that the analysis under the good-fit condition sets the criteria for $\widetilde{R}^2$ at 0.9, which is around the middle of the range used by Ferman et al. (2020) (0.8 to 0.95). This adjustment is made because this study constructs 2000 SC placebo units for each T_0, which is smaller than the 10,000 units used in their analysis. Consequently, using a higher constraint such as 0.95 would result in too few observations, while using a lower constraint such as 0.8 would yield too many observations.

Table 1: The rejection rate

	noCOV	inCOV			trCOV		
		invar	no	both	invar	no	trMean
No Restriction							
Panel A: 1-7							
$T0 = 12$	0.149	0.134	0.134	0.134	0.141	0148	0.139
$T0 = 32$	0.150	0.140	0.138	0.141	0.147	0.151	0.149
$T0 = 100$	0.147	0.156	0.148	0.136	0.149	0.150	0.143
$T0 = 400$	0.139	0.132					
Panel B: 1-5							
$T0 = 12$	0.108	0.101	0.101	0.105	0.111	0.110	0.104
$T0 = 32$	0.101	0.098	0.094	0.103	0.101	0.110	0.104
$T0 = 100$	0.091	0.108	0.091	0.095	0.097	0.090	0.093
$T0 = 400$	0.079	0.079					
$\widetilde{R}^2 \geq 0.9$							
Panel A: 1-7							
$T0 = 12$	0.191	0.154	0.154	0.154	0.161	0.161	0.156
$T0 = 32$	0.168	0.149	0.151	0.149	0.146	0.155	0.156
$T0 = 100$	0.155	0.161	0.154	0.137	0.158	0.155	0.147
$T0 = 400$	0.140	0.137					
Panel B: 1-5							
$T0 = 12$	0.143	0.119	0.120	0.123	0.123	0.121	0.119
$T0 = 32$	0.122	0.107	0.102	0.111	0.110	0.114	0.114
$T0 = 100$	0.102	0.103	0.096	0.099	0.104	0.100	0.098
$T0 = 400$	0.080	0.082					

Note: Panel A provides the results with all seven specifications being considered, while Panel B presents the results when specifications 6 and 7 are excluded.

The second column in Table 1 displays the results for the stationary data generating DGP without additional covariate (noCOV), while the next three columns represent the stationary DGP with an additional stationary covariate (inCOV). These DGPs are consistent with the ones used by Ferman et al. (2020). Subsequently, the remaining columns are associated with a DGP that incorporates a trend component (trCOV).

For the inCOV case, there are three scenarios considered. The first scenario involves matching one time-invariant covariate, which aligns with the approach used by Ferman et al. (2020). The second scenario assumes there is no covariate, meaning the SCM does not consider any covariate to be balanced. The third scenario involves including all the covariates in parallel to the pre-treatment outcomes as separate predictors, rather than using a value from linear combination. For the trCOV case, the analysis is conducted in two ways. First, the time-invariant covariate is included, similar to the approach used in the inCOV case. The alternative approach is to match the averaged values of the trend variables. Besides, for the extension scenarios, only three cases of T_0 are simulated. Panel A provides the results when all seven specifications are taken into account, while Panel B presents the results when specifications 6 and 7 are excluded.

In the inCOV DGP with a pre-treatment period $T_0 = 12$, the analysis reveals a 14.9 percent probability of obtaining a significant result in at least one specification out of the seven considered. This finding underscores the issue of specification searching when applying SCM. Moreover, this issue persists even when a larger amount of pre-treatment data is available, although the probability of reporting a significant specification generally slightly decreases with a

longer pre-treatment period. Furthermore, restricting the analysis to samples that meet good-fit criteria does not eliminate the chance for specification searching. However, by excluding specifications 6 and 7 from the analysis, the probability of rejecting the null hypothesis aligns more closely with the desired 5 percent significance level, particularly when the pre-treatment period T_0 is large. In general, these findings hold true across different DGPs with different sets of predictors used in the SCM. Therefore, with a smaller number of simulations, this study arrives at a conclusion that is generally consistent with the findings of Ferman et al. (2020), even though the probability values differ to some degree.

Despite the discussions above, it is noted that all the probabilities reported in Table 1 are higher than the expected test size of 5% in the context of specification searching. According to the conclusions drawn by Ferman et al. (2020), this phenomena may suggest the asymptotic results might not provide reliable approximations in most SC applications, as theoretical results indicate that the possibility of specification searching within a certain class of predictor sets should asymptotically become very small.

Except for this inference, these rejection rates with values higher than 5% may also be linked to the discussions by Abadie (2021) and Quistorff et al. (2021), which suggest that there may exist multiple ways to select unit weights to precisely match a given set of matching variables. The estimation procedure of SCM relies on optimisation methods and their implementations. Consequently, the wide range of potential solutions in the estimation process can contribute to variations in the final estimated results. The similar issue is also discussed by Kaul et al. (2022) that the optimisation process also depends on the software used. Lastly, the good-fit condition specifying "at least one specification provides good fit" may lead to a large amount of placebo treatment samples being included to calculate the rejection rates, which can be potentially increase the rejection rate.

Convergence in unit weights

To assess the convergence behavior in unit weights across the five specifications, Table 2 presents the percentage of the count of significant results when considering all five specifications compared to the count of significant results when considering any of the five specifications. Additionally, the table also displays the results corresponding to considering any one to any four of the specifications. In an expected scenario where the unit weights converge across the considered specifications, the resulting percentage associated with all five specifications would be close to one. This implies a high level of agreement among the specifications in terms of indicating significance. Conversely, the percentage associated with considering only one specification would be close to zero, indicating minimal variation or disagreement among the specifications.

Note that, as discussed in Section 4.3, this analysis are restricted to those placebo samples for which all five specifications produce good-fit SC units, which is different from the previous table that considers all the samples associated with at least one good-fit SC unit.

Table 2 presents evidence that the expected outcome of convergence in unit weights across various time periods, DGPs, and sets of predictors is not consistently met. The percentages shown in the table indicate that there is limited convergence among the different specifications, even when considering the fit for each specification individually. Notably, the results corres-

Table 2: The percentage of the counts of significance indicated by a subset of specifications

DGP		T0=12						T0=100					
		Either	5 (%)	4 (%)	3 (%)	2 (%)	1 (%)	Either	5 (%)	4 (%)	3 (%)	2 (%)	1 (%)
noCOV		99	0.23	0.09	0.15	0.19	0.33	122	0.24	0.17	0.10	0.14	0.35
inCOV	invar	106	0.20	0.21	0.19	0.20	0.63	145	0.17	0.18	0.17	0.20	0.37
	no	106	0.20	0.18	0.18	0.16	0.28	152	0.22	0.18	0.11	0.13	0.38
	both	123	0.16	0.14	0.17	0.21	0.32	156	0.21	0.17	0.12	0.17	0.33
trendCOV	invar	155	0.16	0.14	0.13	0.14	0.43	163	0.20	0.16	0.15	0.18	0.33
	no	147	0.18	0.14	0.13	0.17	0.39	153	0.21	0.13	0.16	0.17	0.33
	trMean	142	0.17	0.12	0.15	0.18	0.39	147	0.21	0.15	0.16	0.18	0.29

Note. The columns named "Either" present the count of the significance results for either of the five specifications. The subsequent columns, labelled with numbers, represent the percentage of that count attributed to a specific number of specifications being met simultaneously. For example, number 4 represents that any of the four specifications agree on the significanct results.

ponding to only one specification range from 30% to 60%, which is significantly higher than expected. These findings raise concerns about the reliability and consistency of the estimated unit weights across specifications 1 to 5, irrespective of the asymptotic scenario with a large amount of data or the small data setting. This lack of convergence observed highlights the potential uncertainties regarding the estimated unit weights, even under the certain specification condition proposed by Ferman et al. (2020).

However, note that this analysis does not provide information on the number of cases that satisfy the good-fit condition. Further investigation or derivation are necessary to gain a comprehensive understanding on this aspect.

MSPE

Based on the previous two tables, it is evident that specification searching occurs across the three DGPs and across different data inputs. Moreover, given the relatively low level of agreement by all the first five specifications, it might be premature to attribute the searching opportunity solely to the alternative specifications. These analyses lead to the need of more details to thoroughly assess the advantages and drawbacks associated with using different specifications.

Table 3 presents the results for the average MSPE for each specification with $T_0 = 12$ and 100, which serves as a metric to evaluate the quality of the SC units under each specification. Moreover, since the objective of this analysis is to assess the performance of each specification, it includes the cases where the specification being analysed produces a synthetic control with good fit ($\widetilde{R}^2 \geq 0.9$), rather than the more lenient condition that is used for Table 1. The results without this restriction are also provided for comparative purposes in Table 4.

For a shorter pre-treatment period, there are some notable observations in Table 3. Firstly, contrary to the conclusions drawn by Ferman et al. (2020), specifications 6 and 7 occasionally yield good synthetic control units and sometimes even outperform the specifications that meet the condition proposed in their study. This is particularly evident in the case of stationary DGPs. These findings challenge the notion that specifications 6 and 7 perform worse than the others and are responsible for the observed specification searching opportunities. Although One potential concern is that the two specifications have lower probability to achieve good fit, as shown by Ferman et al. (2020), this study still emphasises that more details are needed to determine the performance of each specification accurately. Specifically, under the more lenient

Table 3: MSPE results conditioning on good fit

Model		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
		$T0 = 12$							$T0 = 100$						
noCOV		0.26	0.27	0.28	0.27	0.27	0.21	0.24	0.20	0.20	0.20	0.20	0.20	0.20	0.20
inCOV	invar	0.26	0.28	0.29	0.27	0.26	0.23	0.26	0.20	0.20	0.21	0.20	0.21	0.20	0.21
	no	0.26	0.28	0.29	0.27	0.26	0.23	0.26	0.20	0.20	0.20	0.20	0.20	0.20	0.22
	both	0.26	0.27	0.27	0.26	0.26	0.26	0.25	0.20	0.20	0.21	0.20	0.20	0.21	0.21
trCOV	invar	0.29	0.32	0.31	0.29	0.29	0.39	0.30	0.20	0.21	0.21	0.21	0.21	0.61	0.29
	no	0.29	0.33	0.34	0.32	0.31	0.38	0.35	0.20	0.20	0.21	0.20	0.20	1.28	0.42
	mean	0.29	0.32	0.33	0.30	0.31	0.39	0.32	0.20	0.21	0.21	0.20	0.20	0.80	0.31

Note: The results are obtained by the average values of MSPE produced by those SC units with good fit ($\widetilde{R} > 0.9$).

Table 4: MSPE results without conditioning on good-fit

Model		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
		$T0 = 12$							$T0 = 100$						
noCOV		0.27	0.34	0.40	0.32	0.32	1.14	0.61	0.20	0.20	0.21	0.20	0.20	1.11	0.62
inCOV	invar	0.26	0.36	0.44	0.33	0.33	1.38	0.73	0.20	0.20	0.24	0.24	0.23	1.15	0.50
	no	0.26	0.35	0.44	0.33	0.33	1.38	0.73	0.20	0.20	0.21	0.20	0.20	1.43	0.65
	both	0.26	0.31	0.38	0.28	0.28	0.92	0.43	0.20	0.20	0.21	0.20	0.20	0.88	0.39
trCOV	invar	0.29	0.37	0.39	0.31	0.32	0.90	0.45	0.20	0.21	0.21	0.21	0.21	0.82	0.36
	no	0.30	0.39	0.48	0.37	0.36	2.34	0.89	0.20	0.20	0.21	0.20	0.20	2.13	0.78
	mean	0.29	0.38	0.45	0.34	0.36	1.38	0.58	0.20	0.20	0.21	0.20	0.20	1.18	0.39

Note: The results are obtained by the average values of MSPE produced by all SC units.

condition, the rejection rate analysis may not provide a fair assessment for specifications 6 and 7, as it includes cases where they have poorer fits.

Upon closer inspection of the performance, it is observed that specification 6, which utilises the mean values of the outcome variables, tends to exhibits smaller errors when dealing with stationary DGPs if the period is short. The reason that other specifications demonstrate larger errors could be explained by the issue of over-fitting created by the small T_0. As discussed by Abadie and Vives-i Bastida (2022), good pre-treatment fit may be attained solely through the variation in the individual transitory shocks, ε_{jt}, and eventually lead to large post-treatment estimation errors.

Conversely, when confronted with the trCOV DGP, specifications 6 and 7 face greater challenges in delivering reliable SC units. These findings imply that the limited number of predictors utilised in these specifications may be especially inadequate for accurately capturing the underlying trend in the DGP. Moreover, by comparing different covariate inputs used in each specification, it is observed that incorporating a time-invariant covariate slightly enhances the performance in the case of trCOV, in comparison to the commonly employed approach of using mean values in empirical analyses. This consistently holds true for all specifications, except specification 1, which uses all lagged outcomes. This exception is likely related to the study by Kaul et al. (2015), which suggests that including all lagged outcomes renders the other covariates meaningless in the estimation procedure.

When analysing a longer pre-treatment period, most specifications are generally considered

effective, except for specifications 6 and 7. Specification 6, in particular, performs significantly worse, as the limitation of using fewer match variables is amplified by the longer pre-treatment period. On the other hand, specification 7 exhibits only slightly inferior results when handling with trCOV, especially when it incorporates a covariate. Similar to the case of shorter period, matching on a time-invariant covariate is found to produce fewer errors compared to matching on a mean value. Nevertheless, this suggests that specification 7, when combined with a certain covariate, may still be acceptable for estimating the treatment effect even when the variable of interest is non-stationary.

When the analysis is not conducted under the good-fit condition, the magnitude errors for specifications 6 and 7 are greatly affected, as illustrated in Table 4. These errors can be two to three times larger compared to the errors obtained by the other specifications, when considering the shorter period, and even larger when analysing the longer period. This indicates that a significant portion of SC units constructed using specifications 6 and 7 likely have a bad fit. Consequently, including these units in the analysis substantially raises errors. This finding also implicitly supports the previous discussion that the lenient good-fit requirement used when calculating the rejection rate may introduce bias in the results for these two specifications.

MSPE when considering covariates in ASCM

Based on the previous subsection, it can be concluded that for the trCOV DGP, the choice of covariates has a substantial impact on the quality of SC units. Consequently, it becomes important to further investigate the specific influence of balancing covariates in trCOV. To accomplish this, the ASCM method is employed, as it possesses the capability to achieve perfect balance on the specified covariates.

Besides, the results obtained from ASCM that focus on adjusting unit weights due to pre-treatment outcome imbalance are also presented. This allows for a comprehensive analysis of the effects of both covariate balancing and pre-treatment outcome adjustments on the results.

Table 5 provides the results for the MSPE with pre-treatment periods of 12 and 100. These results can be directly compared to Table 3 and 4 since they utilise the same DGP.

Table 5: Post-period MSPE results when using ASCM

DGP	T0=12							T0=100						
	SCM	**x**	$\boldsymbol{Z_1}$	$\boldsymbol{Z_2}$	$\boldsymbol{Z_3}$	$\boldsymbol{Z_4}$	$\boldsymbol{Z_{1\text{-}4}}$	**SCM**	**x**	$\boldsymbol{Z_1}$	$\boldsymbol{Z_2}$	$\boldsymbol{Z_3}$	$\boldsymbol{Z_4}$	$\boldsymbol{Z_{1\text{-}4}}$
trCOV ($\widetilde{R}^2$)	0.29	0.33	0.58	0.47	0.51	0.50	0.87	0.20	0.20	0.47	0.41	0.53	0.25	0.58
trCOV (All)	0.29	0.32	0.81	0.62	0.69	0.58	1.08	0.20	0.20	0.59	0.55	0.81	0.27	0.62

Note: The light gray row describes the data input. The first element, "SCM," denotes the original SCM method using all lagged outcomes. The "x" element represents the ASCM method without using covariates as input, but purely adjusting for the imbalance in outcomes. The subsequent elements indicate the specific covariates used in the ASCM.

When $T_0 = 12$, improving the fit for pre-treatment outcomes can hinder the estimation of the counterfactual in the post-period. In contrast, the impact of correcting the outcome bias becomes negligible when T_0 is increased to 100. This observation may again be attributed to the risk of over-fitting to noise, which is more pronounced under a shorter pre-treatment period. More precisely, despite the proposed cross-validation procedure by (Ben-Michael et al., 2021) to select a regularisation parameter and control the level of extrapolation, this procedure does not

effectively alleviate the issue of over-fitting in a small finite sample.

When analysing on covariates, it is found that adjusting unit weights with the purpose of achieving perfect fit in some covariates actually leads to increased estimation errors. These findings are contrary to the results reported in the simulation study conducted by Ben-Michael et al. (2021). However, this phenomenon can be attributed to the simulation design used in this analysis. The simulation design incorporates additional error terms in the covariates to mimic real-world scenarios, resulting in noisy characteristics. As a result, exact matching on these pre-treatment covariates is more likely to introduce errors in the out-of-sample period. To address this concern, the next table presents the results obtained using the same methodology but with smaller variations assigned to each error term in the covariates. The specific parameters used in the simulation can be found in Appendix B.

Table 6: Post-period MSPE results when using aSCM with smaller noise in covariates

DGP	T0=12							T0=100						
	SCM	**x**	Z_1	Z_2	Z_3	Z_4	$Z_{1\text{-}4}$	**SCM**	**x**	Z_1	Z_2	Z_3	Z_4	$Z_{1\text{-}4}$
trCOV ($\widetilde{R}^2$)	0.29	0.33	0.30	0.50	0.61	0.41	0.63	0.20	0.20	0.20	0.50	0.54	0.21	0.51
trCOV (All)	0.29	0.33	0.30	0.61	0.84	0.45	0.73	0.20	0.20	0.20	0.65	0.66	0.23	0.55

Note: The light gray row describes the data input. The first element, "SCM," denotes the original SCM method using all lagged outcomes. The "x" element represents the ASCM method without using covariates as input, but purely adjusting for the imbalance in outcomes. The subsequent elements indicate the specific covariates used in the ASCM.

The errors generally decrease when covariates are assigned with smaller noises, indicating that ASCM is sensitive to the level of noise present in the covariate inputs. This finding has a direct implication for empirical applications, highlighting the need to carefully assess the stability of covariates prior to their usage. This implication also extends to the standard SCM. While the standard SCM allocates variable weights to predictors based on optimisation procedures and does not specifically focus on exact balancing, if there are relatively large noises in the matching covariates, the optimisation process can be significantly affected and distorted.

However, it is also evident that using the original SCM with all lagged outcomes still yields the smallest errors in the shorter period. This suggests that even with smaller variations in the covariates, the over-fitting issues and potentially excessive extrapolation involved in ASCM due to the limited sample size persist. In contrast, for the longer period, where these concerns are less prominent, the impact of matching different covariates with varying characteristics becomes more apparent. Specifically, the covariates Z_2 and Z_3, by design, have smaller effects on the outcomes compared to Z_1 and Z_4. Exact balancing on Z_2 and Z_3, as well as balancing on all covariates, proves to be detrimental to the estimation. On the other hand, the cases of adjusting weights due to imbalance in pre-outcomes and matching on Z_1 and Z_4 produce similar errors as the considered SCM. While the specific differences may not be discernible when observing errors with only two decimal places, it is theoretically demonstrated by Ben-Michael et al. (2021) that adjusting weights due to pre-outcome imbalance and matching on relevant covariates are expected to result in smaller errors. However, the simulation in this study shows slightly larger errors when matching on the trend covariate Z_4, contrary to the theoretical expectation. The following discussions provide some explanations for this discrepancy.

According to Botosaru and Ferman (2019), when covariates have linearly independent effects on outcomes, achieving a perfect balance on pre-treatment outcomes also results in approximate

balance on those covariates. In the case of the original SCM using all lagged outcomes as predictors, there is a high probability of achieving a perfect fit in the outcomes, which in turn contributes to good balance in the covariates. This approximate balance across all observed predictors then is expected to help minimise errors in the SC estimates. However, when using ASCM to precisely balance a specific covariate through extrapolation, some distortions may be unintentionally introduced and disrupt the approximate balance achieved by the SCM. This may explain why the ASCM produces larger errors compared to the SCM with all lagged outcomes, even when the relevant covariates are matched.

In addition to extrapolation properties of ASCM, the features of the standard SCM can also explain its better performance in some cases. Kaul et al. (2022) shed light on the trade-off between using covariates or excluding them in the standard SCM. They consider a model where the outcome variable Y_{jt} can be expressed as $Y_{jt} = \widetilde{Y}_{jt} + \theta_t Z_j$, where $\widetilde{Y}_{jt} = \delta_t + \lambda_t^k + \varepsilon_{jt}$ represents the variable of interest if it was not affected by any covariate. Then, the estimation error can be written as $Y_{1t} - \sum_{j=2}^{J+1} w_j Y_{jt} = \widetilde{Y}_{1t} - \sum_{j=2}^{J+1} w_j \widetilde{Y}_{jt} + \theta_t \left(Z_1 - \sum_{j=2}^{J+1} w_j Z_j \right)$. The latter term involving covariates could be arbitrarily large if one does not emphasise balance in the covariates, which leads to small-sample bias. On the other hand, by ignoring covariates, the SCM would capture their effects as if they were unobserved components (λ_{jt}). Kaul et al. (2022) provide detailed discussions and a simulation study that further explore this context.

5 Empirical Application

To examine the relevance of covariates in the real-world setting, the discussed SCMs each with different advantages are employed to estimate the effects of California's tobacco control program. This policy was first studied by Abadie et al. (2010) and had been extensively considered in the filed of SC research. This study aims to present the comprehensive analysis over different SCMs and particularly assess the impact of balancing certain covariates.

The main evaluation for model quality is based on the average errors in post-treatment outcome variables under the placebo framework, as usually done in synthetic control literature. The average fit in terms of r^2 in this period are also provided. Based on the conclusion on model quality, the relevance of covariates in constructing synthetic unit for California can be assessed via pre-treatment fit in each covariates.

5.1 Background and data

Proposition 99 is an anti-tobacco legislation implemented by California's government in January 1989. It was a significant step in the modern era as a large-scale anti-smoking law, which involved increasing the cigarette excise tax by 25 cents per pack. The motivation behind this legislation was the growing awareness of health issues associated with smoking. In addition to raising taxes, the revenue generated from Proposition 99 was directed towards funding health programs and anti-smoking education initiatives. The funds were also allocated for conducting anti-smoking media campaigns to further promote awareness and discourage smoking.

The dataset compiled by Abadie et al. (2010) includes various variables related to tobacco consumption and socio-economic factors. The primary focus is on per-capita cigarette consump-

tion, which was recorded for both the pre-treatment period (1970-1988) and the post-treatment period (1989-2000). In addition to cigarette consumption, other variables such as the average retail price of cigarettes (1980-1988), the percentage of the population aged 15-24 (1980-1988), the logarithm of per-capita personal income (1980-1988), and per-capita beer consumption (1984-1988) were collected. Besides, the dataset consists of data from 38 American states that did not implement large-scale tobacco control programs during the specified time period.

5.2 SCM & Data input

Several SC methods each with different conditions are employed to estimate the policy effect. For the canonical SCM, this study follows the suggestion of Ferman et al. (2020) to utilise all pre-treatment outcomes as predictors. The specification proposed by Abadie et al. (2010) is also re-estimated, incorporating cigarette sales data from the years 1975, 1980, and 1988. In both specifications, the averaged values of all four covariates are included. Two approaches are used to determine predictor weights: an automatic procedure and directly assigning equal weights. For ASCM, three sets of predictors are considered. The first set includes only the retail price of cigarettes, which exhibits non-stationary patterns. The second set includes only the age-related covariate, which has more stable patterns. The third set includes all four available covariates. For the sparse SCM, both the fast version and the full-joint versions are implemented. The data input for both methods consists of two cases. The first case involves estimating the policy effect without any covariate, while the second case includes the averaged values of the four covariates.

5.3 Empirical results

Table 7 depicts the average errors and coefficients of determination over the 37 placebo SC units generated by each model. By regressing the true value of cigarette consumption on the estimated value, the r^2 value represents the proportion of the variance in the true values that is explained by the estimated values.

Table 7: Evaluation on the average post-period outcome fit under the placebo framework

	All		Abadie		ASCM			Sp-SCM (Fast)		Sp-SCM	
		eqV		eqV	Price	Age	all 4		all 4		all 4
r^2	0.892	0.864	0.871	0.820	0.908	0.903	0.888	0.511	0.511	0.560	0.545
RMSE	8.251	9.190	8.625	11.523	6.427	6.374	6.989	9.038	9.389	9.083	9.766
MAE	6.063	7.232	6.600	9.709	4.355	4.347	4.937	7.759	8.259	8.012	8.701

Note: The rows shaded in a lighter gray specifies the conditions for each estimation method. "eqV" indicates that the variable weights are equally weighted. "Price" and "Age" indicates using only the retail price or age as a covariate, while "all 4" signifies the use of all four covariates.

The darker gray shading represents the type of SCM utilised. The first two entries correspond to the standard SCM with all pre-treatment outcomes and the specification proposed by Abadie et al. (2010), respectively. The following entries pertain to the ASCM and Sp-SCM.

The analysis reveals that the ASCM outperforms both the standard SCM and the Sp-SCM across all evaluation metrics, indicating its overall better performance. This finding is in contrast to the simulation results presented in Table 5 and 6, where the errors of ASCM with relevant covariates approximate those of the standard SCM. This difference in ASCM performance could

be attributed to the nature of covariate effects. While the simulation study constructs the outcome values based on the covariates linearly and independently, empirical settings are more likely to involve non-linear and multicollinear effects. In such cases, there is an specific advantage of balancing the covariates as it can lead to tighter bounds on the bias of the SC unit, as derived by Botosaru and Ferman (2019).

On the other hand, it remains unclear why the Sp-SCM performs worse than the standard SCM with Albadie's specification, considering that Quistorff et al. (2021) find the lower mean square errors (MSE) is yielded by Sp-SCM Fast. One potential explanation for this disparity could be that the authors of the study mention the occurrence of errors when estimating the SC unit for some placebo units, which are subsequently eliminated from their analysis. Nevertheless, the differences between the performance of the models are not particularly large.

Tuning back to ASCM, it is also found that the errors depend on the choice of covariates to some extent. Achieving a perfect balance in all four covariates is not the most favorable option as there is a larger error and smaller r^2. This may lead us to carefully assess the impact of different covariates being balanced. The following analysis aims to provide a detailed assessment on this aspect, based on the results of MSPE as an indicator of estimator performance. While the primary focus is on the covariate values during the pre-treatment period, Table 8 also includes the corresponding content over the post-period, for completeness.

Table 8: Mean values of covariates

		Pre-treatment				Post-treatment			
		Price	Age	Income	Beer	Price	Age	Income	Beer
California		**66.637**	**0.179**	**10.032**	**24.280**	**204.292**	**0.152**	**10.135**	**21.144**
All		66.029	0.181	9.817	23.506	187.780	0.149	9.978	23.045
	eqV	65.965	0.180	9.945	22.968	186.328	0.144	10.146	21.681
Abadie		65.126	0.181	9.832	24.124	186.294	0.147	9.988	23.721
	eqV	65.865	0.180	9.964	23.714	184.843	0.142	10.167	22.278
ASCM	Price	66.637	0.178	9.861	25.352	189.255	0.143	10.003	24.586
	Age	66.421	0.179	9.875	23.248	192.330	0.147	10.027	22.385
	all 4	66.637	0.179	10.032	24.280	193.460	0.145	10.190	22.304
Sp-SCM		74.817	0.202	11.378	32.360	210.879	0.165	11.592	31.081
Fast	all 4	75.007	0.202	11.407	32.458	211.424	0.165	11.621	31.180
Sp-SCM		64.511	0.179	9.850	23.998	177.252	0.149	10.039	23.638
	all 4	69.999	0.194	10.801	29.026	196.576	0.160	11.008	28.093

Note: "eqV" indicates that the variable weights are equally weighted. "Price" and "Age" indicates using only the retail price or age as a covariate, while "all 4" signifies the use of all four covariates. Since the periods in which the data is available differ between covariates, the values are averaged over their corresponding pre- and post-periods.

SCM

When combining the results from Table 7 and 8, it becomes evident that although attaining precise balance on all four covariates during the pre-treatment period may not be the most optimal approach, there are still benefits to approximately matching these covariates. This inference is supported by the subpar fit results of the Sp-SCM and its higher errors, in comparison to both the standard SCM and ASCM, which exhibit similarly good fit across all covariates.

When analysing the results of standard SCM, it is observed that using all lagged outcomes achieves approximate balance across covariates. Since covariates would be rendered meaningless under this setting (Kaul et al., 2015), their good balance implies that some of them are relevant and their effects on the potential outcome is linear, as the optimisation procedure does not explicitly aim to achieve such balance. The theory behind it is discussed by Botosaru and Ferman (2019) and a similar conclusion is drawn in their empirical study.

Comparing the SCM with all lagged outcomes to the SCM with Albadie's specification, it is noted that Albadie's specification performs better in fitting the covariates Income and Beer. This finding partially aligns with a simulation study conducted by Kaul et al. (2022), where they find that when covariates have linear and independent effects, using all outcome lags can lead to poorer fitting of covariates compared to estimators that effectively utilise some outcome-related predictors. On the other hand, the weaker fit for Price can be attributed to the findings discussed in the study by Botosaru and Ferman (2019). They explain that the data-driven procedure used to determine variable weights tends to assign relatively small weights to covariates that should not be matched on, irrespective of whether their effects are linear or non-linear. Overall, these analyses suggest that the effect of Price on the outcome may be non-linear, while Income and Beer have a suspected linear effect on cigarette consumption. Besides, achieving balance for Price is deemed less important in the optimisation procedure compared to Income and Beer, under Albadie's specification.

However, this conclusion is based on limited theoretical considerations and may oversimplify the complex dynamics among the variables. While there are justifications for achieving different levels of balance in each covariate, they cannot directly explain the inferior performance of Albadie's specification compared to using all lagged outcomes. One plausible explanation for this result could be the difference in the number of predictors utilised by the two specifications.

ASCM

In the context of achieving balance in covariates, the use of ASCM can offer additional insights, as it allows for precise control to achieve perfect balance in covariates. Since ASCM is modified from SCM with all lagged outcomes, the results of ASCM are compared to this specification. First, it is found that ASCM generally produces a better fit for all covariates compared to SCM, except for the Beer covariate, where the results are less clear. This suggests that achieving balance on the three covariates considered in the study could be the reason for ASCM's superior performance. Besides, it is also found that exactly matching on Beer, in addition to the other three covariates, leads to increased errors. This may be attributed to the limited representativeness of the Beer covariate, as it only covers a short period of five years. As a result, aiming for balance specifically on the Beer covariate is not recommended due to its inherent lack of reliability.

Sp-SCM

The last part of analysis focuses on Sp-SCM. Although Sp-SCM produces larger errors compared to the other approaches, its automated selection of predictors and its different estimation procedure focusing on post-period can offer insights into the importance of covariates. Since the results from the Full-joint version of Sp-SCM does not give information on the match variables,

Table 9 only depicts the results for the Fast version.

Table 9: The predictors selected by Sp-SCM (Fast)

Sp-SCM (Fast)	all 4	1970	1974	1977	1982	1984	1986	1987	1988
		1985	1987	1988	Income	Age	Price		

Note. The number denotes the year of which cigarette consumption is selected.

When employing Sp-SCM (Fast) with all lagged outcomes, it identifies eight years, with the majority of them being close to the intervention year of 1988. This indicates that the lagged outcomes from these years are deemed more influential in estimating the outcome of interest in the post-period. On the other hand, when additional information on covariates is incorporated, it selects three covariates as well as three years that are close to 1988. This finding highlights the relevance of the covariates Price, Age, and Income, and reinforces the idea that achieving better balance in these three covariates may be the reason that the ASCM performs better. Moreover, this finding confirms predictor selection considered by Abadie et al. (2010), and the inclusion of the Income covariate as a relevant predictor also aligns with the results obtained in the study conducted by i Bastida (2022), where the analysis involves forty covariates.

6 Conclusion

The simulation design expands upon the framework employed by Ferman et al. (2020), incorporating diverse data-generating processes and covariate inputs. While the authors suggest the use of certain specifications, this study raises potential concerns and conducts analyses to provide clarification. The findings challenge previous research conclusions and introduce new insights. Particularly, the results emphasise the characteristics of outcome variables. While most results can be explained plausibly based on existing theories, further investigation is required to determine the cause of improved performance when employing time-invariant covariates compared to using the averaged value of outcome variables under certain conditions.

In the empirical study evaluating the impact of the anti-tobacco policy, better balance in covariates related to retail price, age, and income corresponds to smaller estimation bias. Additionally, the study includes discussions grounded in concrete theories to evaluate whether the effects of these covariates on outcome variables are linear or non-linear. However, further analysis is necessary to accurately determine the specific properties of their effects and to quantify the extent of improvement in estimation when achieving balance.

The analysis of covariate fit in empirical study is also limited as it simplifies to consider only the case of California. It would be beneficial to follow the approach of Billmeier and Nannicini (2013), who provide covariate fit analysis for each placebo country and discuss their economic background. Moreover, this study lacks a discussion on the potential presence of multicollinearity among covariates, which is likely to exist in real-world data and can have relevant implications for estimation bias (Botosaru & Ferman, 2019). This aspect should be taken into consideration in future studies. Additionally, the discussions on ASCM in empirical study is limited to the three scenarios. Comparing results from exactly balancing different combinations of covariates with ASCM may offer further understanding on their importance of fit.

References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, *59*(2), 391–425.

Abadie, A., Diamond, A. & Hainmueller, 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.

Abadie, A., Diamond, A. & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495–510.

Abadie, A. & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American economic review*, *93*(1), 113–132.

Abadie, A. & Vives-i Bastida, J. (2022). Synthetic controls in action. *arXiv preprint arXiv:2203.06279*.

Amjad, M., Shah, D. & Shen, D. (2018). Robust synthetic control. *The Journal of Machine Learning Research*, *19*(1), 802–852.

Ben-Michael, E., Feller, A. & Rothstein, J. (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, *116*(536), 1789–1803.

Billmeier, A. & Nannicini, T. (2013). Assessing economic liberalization episodes: A synthetic control approach. *Review of Economics and Statistics*, *95*(3), 983–1001.

Botosaru, I. & Ferman, B. (2019). On the role of covariates in the synthetic control method. *The Econometrics Journal*, *22*(2), 117–130.

Cole, M. A., Elliott, R. J. & Liu, B. (2020). The impact of the wuhan covid-19 lockdown on air pollution and health: a machine learning and augmented synthetic control approach. *Environmental and Resource Economics*, *76*(4), 553–580.

Donohue, J. J., Aneja, A. & Weber, K. D. (2019). Right-to-carry laws and violent crime: A comprehensive assessment using panel data and a state-level synthetic control analysis. *Journal of Empirical Legal Studies*, *16*(2), 198–247.

Dube, A. & Zipperer, B. (2015a). Pooling multiple case studies using synthetic controls: An application to minimum wage policies.

Dube, A. & Zipperer, B. (2015b). Pooling multiple case studies using synthetic controls: An application to minimum wage policies.

Ferman, B. & Pinto, C. (2017). Placebo tests for synthetic controls.

Ferman, B., Pinto, C. & Possebom, V. (2020). Cherry picking with synthetic controls. *Journal of Policy Analysis and Management*, *39*(2), 510–532.

Fernández, C. & Garcia-Perea, P. (2015). The impact of the euro on euro area gdp per capita.

i Bastida, J. V. (2022). Predictor selection for synthetic controls..

Kaul, A., Klößner, S., Pfeifer, G. & Schieler, M. (2015). Synthetic control methods: Never use all pre-intervention outcomes together with covariates.

Kaul, A., Klößner, S., Pfeifer, G. & Schieler, M. (2022). Standard synthetic control methods: The case of using all preintervention outcomes together with covariates. *Journal of Business & Economic Statistics*, *40*(3), 1362–1376.

Montalvo, J. G. (2011). Voting after the bombings: A natural experiment on the effect of terrorist attacks on democratic elections. *Review of Economics and Statistics*, *93*(4),

1146–1154.

Quistorff, B., Goldman, M. & Thorpe, J. (2021). *Sparse synthetic controls: Unit-level counterfactuals from high-dimensional data* (Vol. 5) (No. 1).

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, *25*(1), 57–76.

A Programming code

The standard SCM and the augmented SCM are implemented with packages `Synth` and `augsynth` in R programming language, respectively. The sparse SCM are provided by Quistorff et al. (2021) with Python code, while this study convert the environment to R interface, allowing for a direct comparison between each type of SCM.

For the simulation study, the Monte Carlo results are summarised using the STATA file provided by Ferman et al. (2020). The resulted files are then again imported into R interface, in which the further analyses are performed. As for the empirical study regarding California's Proposition 99, all procedures are conducted in R programming.

B Appendix - The noise specification in covariates

Regarding the DGP with three stationary covaraites and a random walk covariate formulated as equation 5 (trCOV), the random errors associated with Z_1 to Z_4 are drawn from normal distributions $N(1,2)$, $N(0,2)$, $N(0,0.5)$ and $N(0,0.5)$. These choices are made to strengthen the their corresponding strong, intermediate, weak, and trending properties. Besides, to alleviate the impact of noises when analysing MSPE in Section 4.4, the distributions of their random errors become $N(0,0.1)$, $N(0,0.1)$, $N(0,0.1)$ and $N(0,0.2)$.

C Appendix - Review of the augmented SCM

This section provides a review of the augmented SCM proposed by Ben-Michael et al. (2021). Since their main proposal focuses on incorporating ridge regression technique in SCM and is implemented in this study, this section only covers the parts related to Ridge ASCM.

When the Ridge ASCM only involves the outcome variables without any covariate, the estimator for the post-treatment outcome is specified as $\hat{\eta}_0^{ridge} + \mathbf{X_i'}\hat{\boldsymbol{\eta}}^{ridge}$, where the two coefficients are obtained from ridge regression of the post-treatment outcome Y_{0t} on the pre-treatment outcome X_0. The estimation of these coefficients, along with the estimated synthetic value, are obtained through the following procedures:

$$\{\hat{\eta}_0^{ridge}, \hat{\boldsymbol{\eta}}^{ridge}\} \quad = \quad \underset{\eta_0, \boldsymbol{\eta}}{\operatorname{argmin}} \quad \frac{1}{2}\sum_{j=2}^{J+1}(Y_i - (\eta_0 + X_i'\boldsymbol{\eta}))^2 \quad + \quad \lambda^{ridge}||\boldsymbol{\eta}||_2^2, \tag{9}$$

$$\widehat{Y}_{1t} \quad = \quad \sum_{j=2}^{J+1} w_i^{SCM} Y_{it} \quad + \quad \left(X_1 - \sum_{j=2}^{J+1} \hat{w}_i^{SCM} X_i\right) \cdot \hat{\boldsymbol{\eta}}^{ridge}, \quad \forall t > T_0. \tag{10}$$

where the penalty hyper-parameter λ^{ridge} (level of extrapolation...) is determined through a cross-validation approach. As can be seen from equation 10, this approach directly corrects the SCM estimates by the imbalance in pre-treatment outcomes.

Furthermore, the authors introduce two ways to incorporate auxiliary covariates in ASCM, aiming to achieve good balance not only in outcome variables but also in covariates. Since ASCM uses lagged outcomes and covariates differently during the estimation, the notions are separated as X and Z in the equations below, respectively. Similar to before, X_0 and Z_0 are notations for donor units, while X_1 and Z_1 are for the treated unit. The approach used in this study augments the SCM weights with an outcome model $\hat{\eta}_0 + \mathbf{X_i'}\hat{\boldsymbol{\eta}}_{\boldsymbol{x}} + \mathbf{Z'}\hat{\boldsymbol{\eta}}_{\boldsymbol{z}}$. Therefore, the estimation of these coefficients are obtained by modifying equation 9:

$$\underset{\eta_0,\boldsymbol{\eta}_x,\boldsymbol{\eta}_z}{\operatorname{argmin}} \quad \frac{1}{2}\sum_{j=2}^{J+1}(Y_i - (\eta_0 + X_i'\boldsymbol{\eta}_x + Z_i'\boldsymbol{\eta}_z))^2 \quad + \quad \lambda_x||\boldsymbol{\eta}_x||_2^2 \quad + \quad \lambda_z||\boldsymbol{\eta}_z||_2^2. \tag{11}$$

Moreover, since the number of covariates is relatively small relative to the number of units in the simulation setting, this study follows the suggestion by Ben-Michael et al. (2021) to fit a model that only regularises the lagged outcome coefficients $\boldsymbol{\eta}_x$ by setting λ_z to zero, resulting in perfect balance in covariates. Subsequently, it implies the estimation comprising two folds: First, residualise the pre- and post-treatment outcomes on the covariates. Then, estimate ASCM weights on the residulised outcomes.

D Appendix - Review of the sparse SCM

This section provides a review of the sparse SCM developed by Quistorff et al. (2021). In Sp-SCM, the unit weight and the variable weights are estimated by the following formulas, given the regularisation parameters λ_v and λ_w:

$$w(v,\lambda_w) \quad = \quad \underset{w}{\operatorname{argmin}} \left|\left|X_1 - X_0 w\right|\right|_V^2 \quad + \quad \lambda_w \left|\left| w - \frac{1}{J} \right|\right|_2^2, \tag{12}$$

$$v(\lambda_v,\lambda_w) \quad = \quad \underset{v\geq 0}{\operatorname{argmin}} \sum_{j=2}^{J+1} \left|\left|Y_1^{post} - Y_0^{post} w(v,\lambda_w)\right|\right|_2^2, \quad + \quad \lambda_v||v||_1, \tag{13}$$

where λ_v and λ_w are determined by a cross-validation procedure. To enhance the prediction performance on new data for the treated units and encourage a sparse set of matching variables with non-zero weight, variable weights are applied with L1-regularization. This regularization technique ensures a unique solution for the variable weights and facilitates automatic feature selection from the full set of predictor inputs. In contrast, L2-regularization is applied to the unit weights. This regularization method helps reduce computational complexity by providing a closed-form solution for the unit weights. Quistorff et al. (2021) discuss detailed motivations on these choices.

The authors provide two types of Sp-SCM, including a Fast version and a Full-joint version. For the Full-joint version, the optimisation procedure perform a full nested estimation to find v and w jointly, such that the resulting SC for controls have smallest squared prediction error on

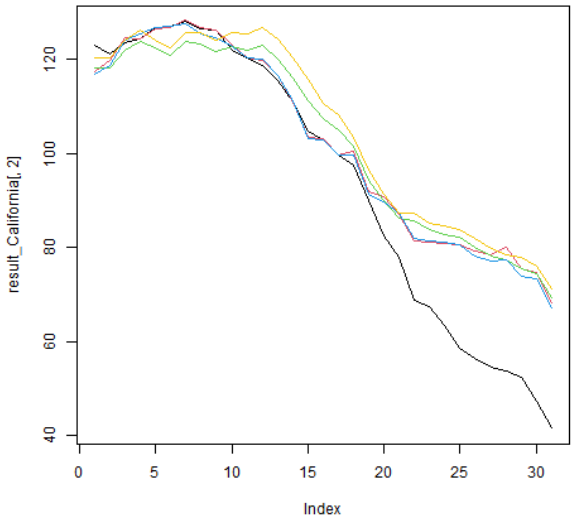
the post-treatment period. For another case, the Fast version, it constructs an approximate v by estimating a linear model of pre-outcomes on predictors. This procedure significantly reduces the amount of computation time.

D.1 California graph

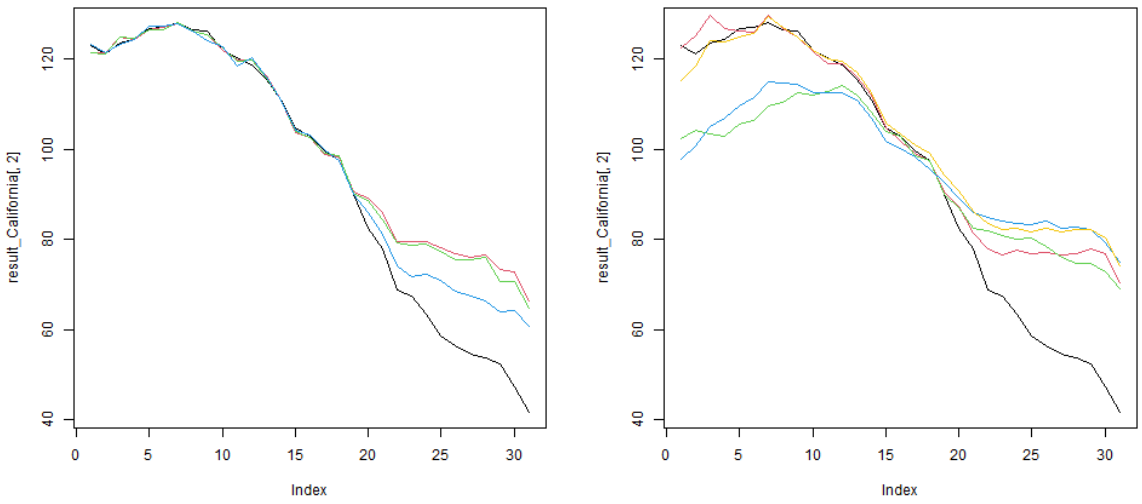
The figures visually illustrate the time series of the outcome variable, the cigarette sales. The true value of cigarette consumption in California is represented by a black line, while synthetic control units are represented by coloured lines.

Figure 1 illustrates the four cases of standard SCMs. In Figure 2a, three cases of ASCMs are plotted, and Figure 2b showcases four cases of Sp-SCMs.

Figure 1: The plot for standard SCMs



The red line depicts the scenario where all pre-treatment outcomes are utilised, while the green line represents the counterpart based on equal variable weights. Similarly, the blue and yellow lines display the results obtained using Albadie's specifications.



(a) ASCMs (b) SP-SCMs

Figure 2: The plot for ASCMs and SP-SCMs

In Figure 2a, the red line corresponds to the case where the covariate related to price is precisely balanced, the green line corresponds to the covariate related to age being balanced, and the blue line represents the results achieved when all four covariates are balanced.

In Figure 2b, the red and green lines represent the results obtained using the Fast version, with the green line incorporating covariates. The blue and yellow lines depict the counterparts for the Full-joint versions.