Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis BSc2 Econometrics/Economics

Specification Searching in the Synthetic Control Method: Choice of Predictors and Splitting Period

Lala AlAsadi (537116)

Supervisor:	Wendun Wang
Second assessor:	Jens Klooster
Date final version:	2nd July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

This paper investigates whether the absence of clear guidelines regarding the selection of predictors used to create the Synthetic Control (SC) estimator gives rise to specification searching opportunities. We replicate the research design of Ferman, Pinto and Possebom (2019) by using extensive Monte Carlo (MC) simulations and analysis of real datasets to assess their theoretical findings which indicate that specification searching becomes asymptotically negligible when confined to a subset of SC specifications. We show that there exists considerable room for specification searching when the number of pre-treatment periods aligns with typical SC applications and when using alternative specifications that are commonly employed in SC applications. We extend this analysis by assessing an alternative choice of V matrix determination and assess the impact of this on the main conclusions. Our results indicate the increased vulnerability of SC estimations to specification searching when implementing the Cross-Validation technique. Notably, we add to the existing literature by finding substantial potential for cherry-picking the splitting periods of the training and validation period data used in the estimation of the V matrix. Overall, these findings highlight the significance of providing clear guidance in selecting predictors, and splitting periods in SC applications, as the lack of guidance allows for a notable level of discretion.

1 Introduction

The introduction of the Synthetic Control Method (SCM) transformed the field of quasi experimental methods used to establish Causal Inference in economics, social sciences, and even medical studies. Acclaimed as a transparent technique, it utilizes an explicit data-driven selection procedure to establish the weights needed to create a synthetic control unit from a pool of control units. This method is widely used in studies ranging across numerous fields to measure the causal effect of an implemented intervention or policy. For instance, in studies pertaining the effect of minimum wage laws (Reich, Allegretto & Godoey, 2017; Jardim et al., 2017; Neumark & Wascher, 2017; Allegretto, Dube, Reich & Zipperer, 2017), taxation (Kleven, Landais & Saez, 2013) and policies concerning public health such as the effects of anti-smoking legislation in California (Abadie, Diamond & Hainmueller, 2010). Peri and Yasenov (2017) apply SCM to immigration studies by studying the effect of the refugee inflow of Cubans in Miami on the labour market in 1980. These are a few instances of the myriad areas of research that have utilized this method highlighting its applicability to a wide range of research settings.

Of course, SCM isn't without its limitations. Firstly, the standard SCM proposed by Abadie and Gardeazabal (2003) imposes restrictive constraints. Namely that there is no intercept and weights must be non-negative and sum to one. However, these constraints are not likely to hold in practice and restricting to convex combinations of the weights is likely to bias the SCM estimator. Another precarious limitation of the standard synthetic control method that is overlooked in academia is the lack of explicit consensus on what variables and covariates should be included in the estimation of the Synthetic Control (SU) unit. In practice, research designs can widely vary in the combination of pre-treatment outcomes and covariates used as predictors for the synthetic control unit. Some research designs are exclusive to pre-treament outcomes used as predictors (Gobillon & Magnac, 2016; Hinrichs, 2012) while others also include other covariates (Abadie & Gardeazabal, 2003; Kleven et al., 2013; Dustmann, Schonberg & Stuhler, 2017). Apart from that, researchers employing SCM also face the choice of whether to include a fraction, linear combination or all available pre-treatment outcomes. This suggests that if the model specification plays a significant role in the estimation of the synthetic control weights such that the estimated treatment effects vary considerably, there is concern for specification searching. For instance, Billmeier and Nannicini (2013) study the effect of economic liberalization on real GDP per capita using all observations of the pre-treatment outcome as separate predictors alongside a set of covariates. In contrast to Billmeier and Nannicini (2013), Kaul Ashok and Manuel (2015) find drastically different results when re-estimating the effects of economic liberalization by restricting the number of pre-treatment outcomes used in the SC construction.

A recent study by Ferman et al. discusses how the standard SCM model is not as invariant to researcher's discretion as popularly acclaimed (2019). Rather, they show that in practice, the synthetic control method is prone to specification searching. While the researcher has little role in manipulating the weights allocated to each unit in the synthetic control group, there remains little to no consensus on the which predictor variables and covariates to include in the specification used to find the optimal weights. This flexibility in the choice of specification potentially provides room for specification searching. Namely, the ability to select specifications that yield significant estimates.

The popularity of the synthetic control method and its application in studies aiming to identify causal effects has been large and steadily growing. And as highlighted earlier, SCM has been implemented in many studies with considerable impact on society. For instance, Abadie, Diamond and Hainmueller (2015) evaluate the impact of the 1989 California Tobacco Control Program on health measures such as lung cancer and heart disease mortality, finding significant reductions. Such estimations are invaluable to public health advocates and policy makers worldwide in justifying the implementation of laws that limit tobacco use. Studies employing SCM often pertain aggregate level policy interventions which have profound impact on society. However, despite being employed in critical fields, the synthetic control method is relatively new in the realm of causal inference methods. Consequently, there has been limited exploration within academia regarding the potential consequences of modifying the inputs of the SC model, such as the number of pre-treatment periods and covariates used, on the estimated treatment effects. The social relevance of this study is that it sheds light on the potential selective use of data in research and policy making. This could undermine the integrity of studies employing SCM and policy decisions that are inspired by it, leading to biased conclusions. Understanding the specification searching possibilities using SCM is crucial for maintaining the credibility of research and policy. By highlighting the potential pitfalls and biases associated with cherry picking, such a study can raise awareness among researchers, policymakers, and the general public. Additionally, these findings potentially encourage researchers to increase transparency. As suggested by Ferman et al. (2019) it's important to perform the SC estimations using the numerous different specifications and present those results as well.

Ferman et al. provide theoretical conditions on the model specification that when satisfied eliminate the potential for specification searching. Using the placebo test suggested by Abadie et al., they test whether SCM is robust to specifications that satisfy the theoretical specifications and to those that don't, using Monte Carlo (MC) Simulations. Following their methodology, this paper will investigate the specification searching possibilities of the standard synthetic control method. Hence, we are interested in determining the probability of rejecting the null hypothesis of no effect, yielding a false positive, for at least one of the different specifications.

Additionally, the estimation of the synthetic control unit involves a nested optimization problem, consisting of an outer and inner optimization process (Kaul Ashok & Manuel, 2015). The outer optimization aims to closely match the pre-treatment outcome values of the synthetic counterfactual with those of the treated unit. On the other hand, the inner optimization ensures that the pre-intervention values of the predictors of the outcome variables are also well-matched. However, the inner optimization relies on the completion of the outer optimization to establish the optimal predictor weights. These weights play a crucial role in indicating the importance of each predictor in explaining the outcome variable and determining the significance of closely matching the pre-intervention outcomes.

In practical implementations, data-driven approaches are often employed for both optimization problems, such as the initial method proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010), which minimize the mean squared prediction error (MSPE) of the outcome variable during the pre-intervention periods. Subsequently, Abadie et al. (2015) followed with the introduction of a cross-validation method that selects predictor weights. Nevertheless, this approach requires the identification of a splitting period, which divides the available pre-treatment periods into a training and validation model. However, no formal instruction yet exists regarding what fraction of the pre-intervention data should be used for the training and validation periods. Therefore, considering the importance of predictor weights in estimating a SC, it becomes crucial to evaluate whether different splitting periods have a significant impact on the conclusions drawn from SCM. Hence, this paper aims to assess the opportunities for specification searching in another step of the SCM, specifically the approach taken in the outer optimization in predictor weight selection. More specifically, this paper will look into how the decision to split the pre-treatment periods into training and validation periods affects the conclusions of SCM.

More specifically, this paper will investigate specification searching opportunities by studying two dimensions of the SC implementation: the choice of pre-treatment outcome lags and choice of predictor weight selection method. To investigate the former, this paper will assess whether the theoretical conditions for asymptotically equivalent SCM estimators proposed by Ferman et al. (2019) do limit the potential for specification searching. Namely, we will test their theoretical results restricting to specifications with large pre-treatment periods and to ones where the number of pre-treatment periods used as predictors goes to infinity as the number of available pre-treatment periods goes to infinity. To do so, we employ Monte-Carlo (MC) simulations that generate simulated data without any treatment and employ the SCM using seven different specifications that are commonly used in SCM applications. To assess whether different specifications yield different conclusions, we will determine the probability of rejecting the null hypothesis of no treatment effect, particularly the rate at which we obtain false positives. The hypothesis of no effect is tested for each generated dataset for all seven specifications in line with the testing procedure proposed by Abadie et al. (2010) and employed by Ferman et al. (2019). We then calculate the probability that a researcher using these different specifications would find at least one specification that rejects the null hypothesis of insignificant effects. If the estimates produced by these specifications are similar, this probability is close to 5 (10) percent for the 5 (10) percent significance level tests.

Furthermore, as a preliminary step of assessing the effect of a different predictor weight selection method, we will perform the same procedure using the cross-validation technique (Abadie et al., 2010) and compare the rejection rates obtained for the different pre-treatment periods. Moreover, the principal assessment of specification searching opportunities provided by the choice of pre-treatment period split is performed by carrying-out MC simulations in a similar fashion and estimating the SCM method using the cross-validation technique (Abadie et al., 2015) technique with three different splitting periods. Correspondingly, we calculate the probability that a researcher using these different splitting periods will find at least one specification that yields an erroneous significant effect.

The main results of this study shed light on the specification searching opportunities within the Synthetic Control framework. Contrary to the theoretical results that suggest the possibility of equivalent estimators under specified conditions (Ferman et al., 2019), the results from the MC simulation suggests that there is still a probability of around 13 percent that at least one specification is significant at a 5 percent significance level. Even with pre-treatment periods as large as 400, specification searching is still not negligible. These results predominantly suggest that the different specifications commonly used in SC applications can lead to significantly different SC units, allowing for specification searching. However, the results exhibit that the possibilities for specification searching are driven by considering specifications that do not increase the number of pre-treatment lags used as predictors. Our findings confirm the concerns raised by Ferman et al. (2019) regarding the lack of guidance on the number and linear combination of pre-treatment outcome lags used as predictors.

The robustness of our conclusions is further supported by the adoption of a different optimization method for the V matrix, containing the predictor weights. Specifically the cross-validation technique proposed by Abadie et al. (2015) is utilized. Interestingly, estimating the V matrix using the cross validation method (Abadie et al., 2015), seems to worsen the specification searching possibilities. The probability of detecting a false significant effect in at least one specification for a 5 percent significance test can be as high as 16 percent with 12 pre-treatment periods compared to 14 percent (obtained under the standard nested optimization method). However, the results generally fall in line with those obtained under the standard optimization method using the full pre-treatment data. Particularly, that the rejection rates decrease for larger pre-treatment periods and that the specification searching possibilities are driven by specifications that do not satisfy the theoretical conditions.

This paper contributes to the existing literature in the following domains. Firstly, it reiterates that specification searching is a valid concern in Synthetic Control applications as investigated by Ferman et al. (2019). The study demonstrates that there is a significant probability of obtaining falsely significant results that indicate the presence of a treatment effect. This suggests that researchers have flexibility in selecting specifications where choices on the pre-treatment period data is made. This space for researcher discretion can yield specifications with significant estimates which potentially biases the results.

Furthermore, it builds on the integral detail of the Synthetic Control method's high sensitivity to the number of pre-treatment periods. As the probability of falsely concluding significant results decreases when considering specifications including higher number of pre-treatment periods. This paper also contributes to literature by looking into the impact of the V matrix approach. This study also finds that the utilization of the Cross-validation technique for selecting predictor weights exacerbates the issue of specification searching, amplifying the significance of determining the V matrix within the Synthetic Control Method (SCM). Lastly, this paper contributes to the literature by providing evidence for the possibility of cherry picking splitting periods of the training and validation data.

Overall this paper offers valuable insights into a relatively new field of causal inference, highlighting that it is not as independent from researchers' subjectivity as previously assumed.

The paper is structured as followed. Section 2 provides an explanation of the synthetic control model framework and incorporates a literature review of associated studies. Following that, Section 3 delves into the problem in detail and presents the Monte Carlo (MC) simulations conducted to investigate it. Furthermore, Section 4 outlines the results obtained from the MC simulations. Section 5 concludes the paper by discussing the results, implications, limitations, and recommendations for future research.

2 Synthetic Control Model Framework

2.1 Data and Notation

In this section, The Synthetic Control Method proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010) is introduced.

Considering the availability of panel data for which we observe data for $(J+1) \in \mathbb{N}$ units for $t = 1, \ldots, T$ time periods. This paper will focus on the case of a single treated unit with treatment taking place at $t = T_0$. Let j denote the unit for which we have data and, without loss of generality, set unit $j = 1$ as the treated unit. For each unit, we observe data on the outcome of interest, $Y_{j,t}$. Additionally, each unit has data on a set of k predictors, $X_{1,j}, \ldots X_{k,j}$. Let $\mathbf{X_j}$ be the vector containing the values of the predictors for each of the $j = 1, \ldots, J+1$ units. Moreover, we define a k x J matrix $\mathbf{X_0} = [\mathbf{X_2}, \ldots, \mathbf{X_{J+1}}]$ containing all values of predictors for the untreated units.

Following the potential outcome framework of Rubin's Casual Inference model (1974), we denote $Y^1_{j,t}$ as the potential outcome of unit j at time t under treatment and $Y^0_{j,t}$ in the absence of treatment. The treatment effect is defined as $\alpha_{j,t} = Y^1_{j,t} - Y^0_{j,t}$ for each time period $t = T_0 + 1, \ldots, T$. For this time period, $Y_{j,t} = Y^1_{j,t}$ is the observed outcome, hence the counterfactual outcome $Y^0_{j,t}$ is what needs to be estimated. The synthetic control estimator of the counterfactual uses a weighted average of the outcomes realized by the untreated units in the donor pool. Equation (1) denotes the SC estimator for the treated unit 1. Most estimators of the counterfactual share the following linear characterization of the counterfactual. However, deviations from this are also observed, such as non-linear generalizations (Athey & Imbens, 2006) and Bayesian methods that allow for time dependent coefficients in the regression (Brodersen et al., 2015) .

$$\hat{Y}^0_{1,t} = \sum_{j=2}^{J} w_j Y_{j,t} \tag{1}$$

The primary step in constructing the Synthetic Control for the treated unit begins with estimating the weights, w_j needed to construct the appropriate weighted average. Determining these weights requires a nested optimization problem which is characterized by an inner and outer optimization (Kaul Ashok & Manuel, 2015). As proposed by Abadie and Gardeazabal (2003); Abadie et al. (2010), the SC weights must be generated based on the criteria that the generated synthetic control closely resembles the pre-intervention predictor values of the treated unit ("inner optmization"). During the inner optimization process, a linear combination of the columns of $\mathbf{X_0}$ is sought to effectively represent $\mathbf{X_1}$. The objective is to find a combination of donor units that minimizes the difference between the predictor values of the treated unit and the counterfactual scenario. In other words, the goal is to identify the weights that result in the smallest possible discrepancy in predictor values between the treated unit and the synthetic control. Hence, for treated unit 1, $\mathbf{W} = [w_2 \ldots w_{j+1}]' := \hat{\mathbf{W}}(\mathbf{V}) \in \mathbb{R}^J$, are obtained by minimizing the following nested problem:

$$\hat{\mathbf{W}}(\mathbf{V}) := \arg\min_{\mathbf{W} \in \mathcal{W}} (\mathbf{X}_1 - \mathbf{X}_0\mathbf{W})'\mathbf{V}(\mathbf{X}_1 - \mathbf{X}_0\mathbf{W}) \tag{2}$$

where $\mathcal{W} := \Big\{ \mathbf{w} = [w_2, \ldots, w_{j+1}]' \in \mathbb{R}^J : w_j \geq 0 \ \forall j \in \{2, \ldots, J+1\} \text{ and } \sum_{j=2}^{J+1} w_j = 1 \Big\}$

The constraints on the SC weights as represented in $\mathcal{W}$ are subject to three constraints. Firstly, to avoid extrapolation non-negativity of weights is imposed. Furthermore, weights must sum to one, such that synthetic controls are weighted averages of the units in the donor pool and that the weights are sparse. This ensures that the synthetic controls are weighted averages of the outcomes of units in the donor pool. At the cost of extrapolation, these restrictions can be relaxed. Abadie et al. (2015) demonstrate how this can be performed through the regression estimator of the synthetic control with weights unrestricted besides the sum of the weights being equal to one. Doudchenko and Imbens (2016) propose a generalization where both constraints do not hold, namely that weights can be negative and the sum not restricted to one. They argue that the weights adding up to one is not ideal when the outcome value takes extreme values in the treatment unit. Furthermore, that it is implausible if the treated unit is an outlier relative to the other units. Another concern they raise is that when the number of controls is considerably larger than the number of pre-treatment periods, which is the case in most applications, the three constraints are unlikely to yield a unique set of weights and intercepts satisfying all of them.

Additionally, V is a diagonal positive semi-definite k x k matrix containing the weights $\mathbf{V} = (v_1, \ldots, v_k)$. These constants determine the relative significance of the generated SC in matching each of the k predictor values to that of the treated unit X_{11}, ..., X_{k1}. These weights are necessary as not all predictors have the same predictive power. Hence, $\mathbf{V}$ represents the relative importance of each predictor in predicting the outcome variable Y.

2.2 V Matrix Selection

Determining the optimal predictor weights, ultimately the V matrix, is characterized as the "outer optimization" process of SCM. Abadie and Gardeazabal (2003); Abadie et al. (2010) determine these constants using a data driven procedure that aims to minimize the Mean Squared Prediction Error (MSPE) of the synthetic control, formally shown below:

$$\hat{\mathbf{V}} = \arg\min_{\mathbf{V}} \Big((\mathbf{Y}_1 - \mathbf{Y}_0 \hat{\mathbf{W}}(\mathbf{V}))'(\mathbf{Y}_1 - \mathbf{Y}_0 \hat{\mathbf{W}}(\mathbf{V})) \Big) \tag{3}$$

Moreover, another data-driven way relies on a regression based method of selecting the predictor weights $\mathbf{V} = (v_1, \ldots, v_k)$. A regression is performed for each time period leading up to the intervention. The dependent variable being $Y_{j,t}$ which is regressed on all the available predictor variables $\mathbf{X_j}$. This regression yields regression coefficients $\beta_{t,k}$ for each predictor, which is used to determine the weight v_k for predictor k shown in equation 4. Predictors with larger squared regression coefficients are given more weight, indicating their higher significance in the synthetic control construction process. Kuosmanen, Zhou, Eskelinen and Malo (2021) propose a modification to thia existing approach by estimating a regression equation using panel data of predictors and then assigning weights based on the absolute values of the estimated coefficients. The authors note that this approach achieves a more equal balance between different predictors compared to using squared values.

$$v_k = \frac{\sum_t \beta_{t,k}^2}{\sum_{k=1}^K \sum_t \beta_{t,k}^2} \tag{4}$$

Another method used to determine $\mathbf{V}$ is an out-of-sample validation technique proposed by Abadie et al. (2015). By dividing the pre-treatment periods into a training period and validation period, we are able to assess the predictive power of each predictor in approximating the post-intervention outcome values. Initially, we divide the pre-intervention period into a training period and subsequently a validation period. We divide the training period such that it spans from t = 1 to t_0, while the validation period spans t = t_0 + 1 to T_0. While Abadie et al. (2015) assume that T_0 is even and select $t_0 = T_0/2$, the lengths of these periods may vary based on data availability and measurement timing of predictors. After splitting the data, the synthetic control weights $\hat{\mathbf{W}}$ are computed with the training period data and subsequently, $\hat{\mathbf{V}}$ is selected such that it minimizes the MSPE during the validation period.

As discussed by Albalate, Bel and Mazaira-Font (2021), the nested optimization problem in SCM, which they refer to as a bilevel optimization, has many flaws. Building on the findings of Albalate et al. (2021), they claim that finding an optimal solution is computationally challenging. Additionally, the hierarchical structure of bilevel programming can introduce difficulties such as non-convexity and disconnectedness, making the solution set disjointed and leading to highly unstable solutions and convergence to different local optima when dealing with simpler instances of bilevel optimization. Furthermore, this instability is further amplified by the strong reliance on the donor pool, consequently affecting the estimation of weights (Albalate, Bel & Mazaira-Font, 2020).

In the specific case of the SCM method, these flaws of bilevel optimization have implications. It means that the solution $\hat{\mathbf{V}}$ can be arbitrary and highly sensitive to small changes or perturbations. Consequently, the weights become unstable, and the resulting solution $\hat{\mathbf{V}}$ lacks reliability in terms of economic meaningfulness. It can be influenced by interpolation biases, which then renders the insights derived from the estimation less trustworthy or interpretable economically.

2.3 Cross-Validation Method in SCM

Broadly, cross validation can be adopted to assess the generalizability of a model and help evaluate how accurately it can predict outcomes on new data. When using Synthetic Controls to estimate a counterfactual, cross-validation can validate the accuracy of the estimation by comparing the predicted outcome for the treated unit with the actual observed outcomes.

Cross-validation techniques are applicable to Synthetic Control research designs for the following reasons. Firstly, the set of control units available to construct the SC unit are obligated to not be exposed to treatment. This condition allows for the identification of time-varying factors in the dataset, resulting in isolating the treatment effect by removing the remaining changes in the outcome variable of the treated unit. Moreover, Cross-validation is a commonly used method to estimate the generalization error when a large test dataset is not available (Rao & Fung, 2008) as is the case for common SC applications. Furthermore, the availability of information on the outcome variable of the treated unit before the treatment occurs naturally provides as a validation set that evaluates the SCM models.

Simultaneously, cross-validation used in Synthetic Control applications can also be used in the estimation of the counterfactual. As introduced by (Abadie et al., 2015) , a cross-validation procedure is used to determine the predictor weights used in the estimation procedure of the synthetic control weights. This method of cross-validation used in their paper falls under the category of a single training-validation period split, however various other forms of cross validation estimation exist in the literature. For instance, Kellogg, Mogstad, Pouliot and Torgovitsky (2021), in accordance with the approach described in the study by Abadie et al. (2015), maximize the accuracy of the treated unit's outcome series during the pre-treatment period. However, while Abadie et al. utilized a single training-validation split (2015), Kellogg et al. (2021) approach involves a sequence of one-step ahead forecasts. Each forecast is generated using exclusively past data preceding the forecasted time period. This proposed method falls under cross-validation with a rolling-window and advantageously preserves the temporal structure of the forecasting problem. Besides the one-step ahead rolling origin cross validation procedure, the authors also utilize a multi-step ahead criteria. However, they find that the RMSPE tends to be higher in the latter case.

Moreover, over the years there have been many adaptations of the SCM framework, such as regularized SCM which requires the identification of shrinkage parameters. Particularly, SCM is shown to be sensitive to model specification and is also plagued by the "small n large p" problem where the number of pre-treatment periods is limited compared to the wide array of potential explanatory variables to incorporate into the model. This feature exacerbates the model's sensitivity to specification. Hence, regularization methods can help by adding a penalty term to the estimation procedure, which encourages the model to shrink the coefficients of irrelevant or less important variables towards zero. In this case, a cross-validation procedure can also be used to identify the shrinkage parameters. This process helps prevent overfitting and ensures the inclusion of relevant variables, ultimately enhancing the model's accuracy and interpretability. Xu (2017) introduce a cross-validation scheme within a generalized Synthetic Control Method, which is based on an Interactive Fixed Effects (IFE) model which automatically selects the number of appropriate factors for estimation. This paper highlights a feature of data used in SCM designs, namely that it offers a natural validation dataset which uses the pre-treatment observations of the treated unit. The author introduces an algorithm based on a cross-validation procedure that selects the model with an appropriate number of factors to minimize the risk of overfitting and enhance accuracy of estimates. Doudchenko and Imbens (2016) also propose a cross-validation procedure to determine λ by choosing the value of the tuning parameters that minimize the cross-validation error.

Another paper by Abadie and L'Hour (2021) present two data-driven selectors for the penalty term that penalizes the pairwise discrepancies between the characteristics of the treated and synthetic control estimator. These procedures are aimed at using Cross-Validation to find the optimal penalty term that balances the prediction accuracy for the treated unit. Cross-validation can aid the identification of the optimal parameter values by identifying which lead to the best generalization and counterfactual estimation accuracy. The first method, a Leave-One-Out Cross-Validation of Post-Intervention Outcomes for the Untreated, a synthetic control is calculated for each control unit by minimizing the prediction error between the actual outcome and a weighted combination of control unit outcomes. The parameter λ is chosen by minimizing a loss measure, such as the sum of squared prediction errors for individual outcomes or the average squared prediction errors. The second method, Pre-Intervention Holdout Validation on the Outcomes of the Treated, divides the data in a training and validation period, the latter containing dates immediately preceding the intervention. The synthetic control is obtained by minimizing the mean squared prediction error between the actual outcome and the weighted combination of control unit outcomes for each treated unit and validation period using the training period data. This method, analogous to the single training-validation period split (Abadie et al., 2015), chooses the penalty term such that it minimizes a measure of error, such as the sum of prediction errors for the individual or aggregate outcomes within the validation period.

However, Rao and Fung (2008) demonstrate through controlled numerical experiments that when the sample size is small, dimensionality is increasing and the number of algorithms is high, cross validation becomes less effective as an estimate of generalization. Furthermore, they also investigate the use of cross validation, particularly leave-one-out cross-validation (LOOCV), in feature selection. This paper finds that as more features are selected, LOOCV significantly underestimates the true error and hence may select unnecessary features. This findings casts concern over the popular use of LOOCV in numerous SC applications (Xu, 2017; Abadie & L'Hour, 2021). Hence, this paper suggests that the performance of using a cross-validation technique to compute the predictor weights needed to construct the Synthetic Control unit is poor when the number of pre-treatment periods is small. Furthermore, Klößner, Kaul, Pfeifer and et al. (2018) examined the cross-validation technique introduced by Abadie et al. (2015) in the context of SCM and found that their results were not reproducible when using alternative software packages or changing the ordering of variables within the dataset. This failure was attributed to the cross-validation method's inability to uniquely define predictor weights. By performing MC simulations, the authors demonstrate the ambiguity in the estimation results that results from using this cross-validation technique. They also continued to show that the uncertainty in results depends on the difference between the number of predictors and the number of donor units with positive weights in the training period. In most cases, this difference was positive, rendering the cross-validation unreliable. To address this problem of ambiguity in predictor weights under the cross-validation method, Becker, Klößner and Pfeifer (2017) introduced a modified technique that ensures unique predictor weights by following two principles. Firstly, special predictors such as lagged values of the outcome variable are assigned certain minimum weights. Secondly, predictors generally are prevented from becoming irrelevant through accidentally obtaining small weights. In their work, Malo, Eskelinen, Zhou and Kuosmanen (2020) also address the design flaw of SCM by formulating it as a NP-hard bilevel optimization problem and proposing an iterative algorithm. They demonstrate that the original SCM problem has a unique optimum, which can be reached using their algorithm based on Tykhonov regularization, ensuring convergence to the true optimal solution.

2.4 Impact of Splitting Period in Cross-Validation

Changing the fraction of pre-intervention data that is split between training and validation models can have multiple implications. To start off, the less data is allocated to training, by providing a smaller fraction of the pre-intervention period to the training model, the less information can used to estimate the predictor weights . This may result in less precise or less accurate estimation of the treatment effect, as the model has access to fewer observations to learn from (Rao & Fung, 2008). Secondly, altering the fraction of pre-treatment data for training and validation models can affect the model's ability to capture the underlying dynamics and patterns in the data. It may cause for insufficient training as data provided for the V matrix estimation may be too limited to capture complex relationships which may result in overfitting or underfitting, where the estimated synthetic control fails to adequately reflect the treated unit's counterfactual outcome (Berrar, 2019). Hence, it is important to assess the impact of adopting a Cross-Validation method to determine the V matrix on the conclusions of the SCM.

2.5 Predictor Selection

Besides the different choices that need to be made in the estimation process of the **V** matrix, another crucial aspect of SCM is the selection of the variables to be used as predictors. Each specification s used in the case of T_0 pre-treatment variables can be characterized by the set of predictors used $\mathbf{X}_j(s, T_0)$ which encompasses the pre-treatment outcome lags, functions of the lags and additional covariates (Ferman et al., 2019). As there remains a lack of explicit consensus on which predictors and functions of the outcome values that should be included in $\mathbf{X_j}$, this detail provides room for specification searching. Ferman et al. (2019) describe this divergence in the choice of predictors used in the implementation of SC in academic papers. They draw attention to the wide differences in the number of pre-treatment periods, denoted by $L(s, T_0)$, used in different papers implementing SCM where some use all or half the available pre-treatment outcome values. Another variation that appears is also the inclusion of specifications that only include even-numbered pre-treatment lags. Numerous papers also only use the mean of the pre-treatment outcomes. Besides how many pre-treatment lags are included and which functions of these lags is used, there resides the matter of including additional covariates. More than a fourth of the papers discussed by Ferman et al. (2019) do not use additional covariates.

The importance of including covariates in the SCM specification is reasoned in the following section. Some of the few data requirements of the SCM as outlined by Abadie (2021) state that there must be sufficient post and pre-intervention observations. Abadie et al. (2010) show that under a good pre-treatment fit and a Data Generating Process that follows a linear factor model, bias is inversely proportional to the number of pre-treatment outcomes used. However, in practice the availability of sufficiently large data isn't available for pre-treatment period predictors. According to Abadie (2021), even with a perfect pre-treatment fit, a small number of pre-treatment periods may spuriously create a synthetic control appears to be a good fit. And in order to prevent that, strong predictors besides pre-treatment outcome values must be used. However, Kaul Ashok and Manuel (2015) show that when using all pre-treatment outcomes values in the specification, the role of the other covariates is rendered irrelevant. Predictors seem to have little impact on the synthetic control (Ben-Michael, Feller & Rothstein, 2018;

Doudchenko & Imbens, 2016). This presents a problem as the statistical properties of the SCM critically depend on its ability to reproduce outcome values of the features of the treated unit. Many SCM applications have resorted to using what is referred to as the "canonical SCM" (Ferman et al., 2019; Doudchenko & Imbens, 2016) which according to Kuosmanen et al. (2021) is due to the computational problems and design flaw in SCM optimization procedure. However, Abadie and L'Hour (2021) optimize the fit of the SC with respect to predictors disregarding the pre-treatment outcome lags. Hence, there resides considerable discourse over what the fitting procedure is and the trade-off problem of the fit with respect to the predictors versus the fit with respect to the pre-treatment outcomes remains unresolved.

2.6 Theoretical Results

Crucially, Ferman et al. (2019) provide a proof for a key Theoretical finding that the SC weights $\hat{\mathbf{W}}(s, T_0)$, converge in probability to the same $\bar{\mathbf{W}}$ if when $T_0 \to \infty$, $L(s, T_0) \to \infty$. Moreover, the assumption that the pre-treatment averages of the second moments of every subsequence of outcome values for all units converge to the same value is made. However, as T_0 is finite in practice, the applicability of these results to practical applications of the SCM will be tested in the following section.

2.7 Inference Method

Inference in the Synthetic Control framework is based on permutation methods suggested by Abadie et al. (2010). Their approach involves permuting the treated unit assumption and estimating the treatment effects for each unit and time period, by iteratively assigning treatment. Moreover, the permutation distribution is then constructed by pooling together the estimated effects of the treated unit with the placebo effects generated from the units in the donor pool. The treatment effect on the affected unit is considered significant when it is unusually large compared to the permutation distribution.

$$\frac{\frac{1}{T_0}\sum_{t=T_0+1}^{T}(Y_{j,t} - \hat{Y_{j,t}})^2}{\frac{1}{T-T_0}\sum_{t=1}^{T_0}(Y_{j,t} - \hat{Y_{j,t}})^2} \tag{5}$$

Furthermore, they compute a ratio of the mean squared prediction error as a test statistic, shown in equation 5, also used to calculate p-values. These p-values are interpreted as the probability of obtaining an estimate as extreme as the observed one if treatment assignment were random. However, the assumptions of randomization inference are restrictive in the SC framework since treatment assignment is typically not random. This makes the statistical interpretation of this placebo test unclear, nonetheless is it the most commonly used inference method in SC applications and hence will be adopted in this research.

Ferman et al. (2019) analyze the implications of their theoretical findings for the asymptotic equivalence between different specifications of the Synthetic Control (SC) method using this inference method. Furthermore, in the MC simulations they conduct, the probability of obtaining a test statistic in the top 5 and 10 percent of the distribution of placebo test statistics is also calculated to assess whether the estimated effects under SCM are significant.

In recent years, there have also appeared new methods of statistical inference for the SCM. For instance, Doudchenko and Imbens (2016) and Chen and Yan (2023) propose a mixed placebo test that combines the random assignment of treatment across units and time. Furthermore, Firpo and Possebom (2016) address the gap on formal inference theory in SCM and clarify necessary assumptions that ensure the validity of Fisher's Exact Hypothesis Testing Procedure for panel data. By establishing these sufficient hypotheses, they enable the testing of sharp null hypotheses and propose a novel approach for estimating Confidence Sets for the SCM. Furthermore, the confidence intervals obtained by inverting a combination of test statistics can be used to compute a confidence set that encompasses the different SC estimates obtained under different the specifications.

3 Specification Searching in Synthetic Control Method: Monte Carlo Simulation Study

To investigate the specification searching opportunities under the possibility of different specifications and test the reliability of the asymptotic results presented by Ferman et al. (2019), a Monte Carlo (MC) simulation is performed. This paper replicates their research design by generating 10,000 datasets using two different data-generating processes where we impose no treatment effects for all time periods. Then the null hypothesis of no treatment effect is tested using seven different specifications that are commonly used in SC applications. More explicitly, during each round of the MC simulation, the Synthetic Control Method is performed using seven different specifications. More explicitly, the Synthetic Control unit of the "treated" unit is constructed seven times using a different specification. Following the inference method discussed earlier, placebo tests were designed to have a rejection rate of α percent under the null hypothesis for a significance level of α percent. Subsequently, to investigate the opportunities for specification searching in the Synthetic Control Method, we calculated rejection rates, the probability of rejecting the null hypothesis at the α percent level for at least one specification. This provides insight into the likelihood that a researcher adopting the SCM to identify causal effects, would be able to yield a significant result in the absence of any. Hence, yielding rejection rates that are markedly higher than α would imply that the different specifications result in significantly different SC estimators which would indicate reasonable room for specification searching. However, if the different SC specifications produced similar SC weights, the rejection rate would be close to α percent, indicating a low risk of specification searching. If the seven specifications were to each yield remarkably different estimates of the Synthetic Control, then the the probability of finding a specification that rejects the null would be close to $1 - (1 - \alpha)^S$ (Ferman et al., 2019).

3.1 Data Generating Process (DGP)

To ensure robustness against the nature of the dataset, we consider both stationary and non-stationary data generating processes (DGP). In the first DGP, as shown in equation 7, we used a linear factor model where units were divided into groups following different stationary time trends.

$$Y_0 = \delta_t + \lambda_t^k + \varepsilon_{j,t} \tag{6}$$

Each unit's outcome variable $Y_{j,t}$ is determined by a trend component δ_t, a group-specific component λ_t^k, and an error term $\epsilon_{j,t}$. The simulations considered the case where the total number of units, $J + 1$, is set to 20 and split into $K = 10$ groups. Hence, this implies a split in groups of two where unit 1 and 2 follow a stationary time trend of λ_t^1 and so forth.

$$Y_0 = \delta_t + \lambda_t^k + \phi_t^r + \varepsilon_{jt} \tag{7}$$

The data generated from the non-stationary DGP is defined such that a subset of the common factors is non-stationary. This is realized by considering a DGP that includes a non-stationary trend ϕ_t^r that follows a random walk for some $k = 1, \ldots, K$ and $r = 1, \ldots, R$. Again, we consider $K = 10$ and set the subset of the common factors of $R = 2$ to include the non-stationary trend. This implies that units $j = 1, \ldots, 10$ follow the same non-stationary path $\phi_{1,t}$. However, only unit $j = 2$ follows the same stationary and non-stationary path λ_t^1 as the treated unit.

3.2 Specifications with Different Outcome Lags

By simulating datasets based on these DGPs, we aim to assess the impact of different specifications on the SC method and the potential for specification searching, taking into account the reliability of asymptotic results and the choices of T_0 commonly used in SC applications. Considering Non-stationary DGPs also introduce time-varying dynamics in the data allowing the relationship between the observed variables and common factors to change over time. The number of post-treatment periods is fixed such that $T - T_0 = 10$ and we vary the number of pre-intervention periods in the DGPs in the following manner: $T_0 \in \{12, 32, 100, 400\}$. Ferman et al. (2019) delineated that most papers adopting the SCM use a number of pre-treatment periods around between 8 and 16 which deems estimations using 12 pre-treatment periods has the most substance. Setting $T_0 = 400$ is primarily useful as it allows to test the reliability of the asymptotic approximations described in the theoretical results of Ferman et al. (2019). However, it is crucial to note that this is an extreme setting that does not hold in common SC applications where the largest pre-treatment period available in a SC context is 43 pre-treatment periods.

Following the common specifications used in SC applications, as outlined by Ferman et al. (2019), the SC estimator is calculated using all the following seven specifications for every generated dataset. These specifications vary in the linear combinations of pre-treatment outcome values used as predictors. Moreover, as justified earlier, time-invariant covariates are excluded form the specifications due to their inevitable triviality when extensively considering pre-treatment outcome lags in the specification (Kaul Ashok & Manuel, 2015).

1. All pre-treatment outcome values: $\mathbf{X_j} = [Y_{j,1} \ldots Y_{j,T_0}]'$
2. The first three-fourths of the pre-treatment outcome values: $\mathbf{X_j} = [Y_{j,1} \ldots Y_{j,3T_0/4}]'$
3. The first half of the pre-treatment outcome values: $\mathbf{X_j} = [Y_{j,1} \ldots Y_{j,T_0/2}]'$
4. Odd pre-treatment outcome values: $\mathbf{X_j} = [Y_{j,1} Y_{j,3} \ldots Y_{j,(T_0-3)} Y_{j,(T_0-1)}]'$
5. Even pre-treatment outcome values: $\mathbf{X_j} = [Y_{j,2} Y_{j,4} \ldots Y_{j,(T_0-2)} Y_{j,T_0}]'$
6. Pre-treatment outcome mean: $\mathbf{X_j} = \left[\frac{1}{T_0} \sum_{t=1}^{T_0} Y_{j,t}\right]$

7. Three outcome values (the first one, the middle one, and the last one): $\mathbf{X_j} = [Y_{j,1} Y_{j,T_0/2} Y_{j,T_0}]'$

Additionally, as an initial step of the outer-optimization analysis, this paper will assess the effects of adopting different techniques in the step of constructing the predictor weights. Hence, we will look into different ways of estimating the V matrix and what effect this has on the SC estimates. Specifically, this paper will extend the paper by Ferman et al. (2019) by adopting the cross-validation technique of the V matrix determination (Abadie et al., 2015). This is done by repeating the MC simulations outlined above by splitting the pre-treatment data into a training and validation period. Subsequently, the synthetic control weights $\hat{\mathbf{W}}$ are computed with the training period data after which a $\hat{\mathbf{V}}$ that minimizes the MSPE during the validation period is selected.

Following Abadie et al. (2015), the training period constitutes half the total pre-intervention period data. Hence, t_0 is selected such that it is equal to $T_0/2$. In each simulation, the Synthetic Control method is performed for all seven specifications by first estimating the SC using only half the available pre-treatment data under that specification. Subsequently, the V matrix estimated under that stage is adopted and the SC for the treatment period is estimated using the other half of the pre-intervention data. This provides insights into how adopting a different method could affect the possibilities of specification searching and whether the probabilities of finding an erroneous significant effect increases. The obtained rejection rates under this extended method of V matrix determination is compared to the rates obtained when replicating the Ferman et al. (2019) using the standard nested optimization procedure. After obtaining these rejection rates, we will compare them to those obtained following the estimation methods of Ferman et al. (2019) to assess whether the probability a researcher can report a significant result increases or decreases. Overall, performing the MC simulation using this alternative method of V matrix calculation allows us to assess the robustness of the results obtained from the simulation design of Ferman et al. (2019). Comparing the rejection rates obtained under this method to those estimated without using cross-validation allows us to assess if the specification searching conclusions are robust across different specifications which would enhance confidence in the findings. Additionally, conducting this analysis enables gaining insights into the extent to which the SCM is sensitive to the choice of V matrix determination.

3.3 Evaluating Different Splitting periods

However, while this method of estimation could be useful to adopt, it may potentially allow for specification searching in other dimensions such as the decision of how to split the pre-treatment periods into training and validation periods. When applying cross-validation to estimate the predictor weights, several different choices on the division of training and validation periods can be made. For instance, the division of pre-intervention period data into training and validation sets can be done equally, with half of the pre-treatment period used for training and the remaining half for validation. Alternatively, a different ratio could be used, such as allocating only 1/3 of the pre-treatment data for training purposes. The effects of changing the fraction of pre-treatment data allocated to training the model could be manifold as mentioned in earlier sections. When implementing cross-validation to estimate the predictor weights, various choices regarding the division of training and validation periods can be made. Thus, this paper extends on the literature of specification searching in the context of SCM by looking into different splitting periods.

Principally, to look into the effects of adopting different choices of splitting periods, MC simulations analogous to the one discussed earlier are performed. To select the appropriate V matrix in the SCM, we employed the cross-validation technique proposed by (Abadie et al., 2015). The outer optimization stage of the SCM involved performing cross-validation to determine the optimal V matrix. In each round of the simulation, we partitioned the data into splitting periods for the SCM. The training period was variable in the SC estimations, $t_0 \in \{T_0/3, T_0/2, (3/4)T_0\}$. More explicitly, during every round of the simulation three specifications, that differ only in the fraction of the pre-intervention periods used for training, are used to estimate the synthetic control for all units. Furthermore, these simulations were run for the specifications using three different pre-treatment periods $T_0 \in \{12, 32, 100\}$. This assessment allows for an examination of the asymptotic properties of SCM as the number of available pre-treatment periods increases. Furthermore, the Simulations were performed using two of the seven particular former mentioned specifications that were defined by the linear combination of pre-treatment lags used. Namely, Specification 1 (where all available pre-treatment outcome lags are used) and Specification 6 (where only the mean of the outcome lags is used). These specifications provide insights into the impact of different linear combinations of pre-treatment lags on the synthetic control estimation. Both of these specifications represent different extremes within the SCM specifications. The first specification satisfies the asymptotic conditions outlined by Ferman et al. (2019), while the second specification fails to meet them. By comparing the simulation results obtained from each specification, we can gain valuable insights into how different linear combinations of the pre-treatment lags can impact the effectiveness of specification searching in a cross-validation estimation setting. Consequently, an objective will be to assess whether the rejection rates are higher when using the 6th specification.

Following the SCM estimation, we performed hypothesis testing to assess the specification choices on the probability of rejecting the null hypothesis. We conducted a placebo test for each specification using the root mean squared prediction error (RMSPE) test statistic, as suggested by Abadie et al. (2010). In each round of the simulation, the synthetic control is estimated for all units, besides the unit defined as treated. Then the null hypothesis of no treatment effect is rejected at a 5 percent significance level if the treated unit has the highest RMSPE among the 20 units considered. Correspondingly, our main interest lies in determining the probability of rejecting the null hypothesis at the 5 percent significance level in at least one specification. This probability represents the likelihood that a researcher would report a significant result, even in the absence of any true effect, if they engage in specification searching. If the three specifications lead to the same synthetic control unit, we would expect the probability of rejecting the null hypothesis in at least one specification to be around 5 percent for the 5 percent significance level. However, this probability may be higher if the choice of specification affects the synthetic estimator, particularly in finite samples (Abadie et al., 2015).

4 Results

The results obtained from the MC simulation study outlined above showed that there is substantial evidence of room for specification searching both in terms of the linear combination of pre-treatment outcomes used in the specification, and when considering a V matrix obtained with cross validation estimation using different splitting periods.

To assess whether specifications including different linear combinations of pre-treatment outcomes can yield significant treatment effects in the absence of any, we will examine Table 1. The results we obtain when replicating the simulation design of Ferman et al. (2019) are equivalent to those obtained by the authors. Columns 1 and 2 outline the probabilities of rejecting the null hypothesis at 5 and 10 percent significance levels when estimated for the stationary model. Likewise columns 3and 4 represent the rejection rates for the non-stationary model. Panel A represents the probability of obtaining at least one significant synthetic control estimate of the treatment effect when considering all specifications. These also include specifications that do not satisfy the theoretical conditions that are outlined by Ferman et al. (2019). Panel B in turn, excludes these specifications, namely specifications 6 and 7. When first assessing the case where $T_0 = 12$, it is observed that the probability of finding a significant specification is at around 14.3 percent (25 percent) for the 5 and 10 percent significance levels for the stationary models. The rates obtained under the non-stationary model for $T_0 = 12$ are almost the same. These rejection rate estimates are markedly higher than the corresponding α percent significance level. These results display that under 12 pre-treatment periods, which is relatively small yet what is seen in average SC applications, researchers have ample opportunity to select statistically significant specifications even when the null hypothesis is true.

Now to assess whether the specification searching opportunities diminish as the number of pre-treatment periods tends to infinity, we take a look at the rejection rates obtained for $T_0 = 32$ and $T_0 = 100$. Ferman et al. (2019) hypothesize that if when the number of pre-treatment periods approaches infinity and the variation in SC weights across different specifications disappears, we would anticipate the rejection rate to approach 5 percent as the number of pre-treatment periods increases. In such a scenario, all specifications would yield similar SC units and, consequently, produce comparable treatment effect estimates. However, this is not observed from the results. As outlined in Table 1 for T_0 set at 32 and 100 pre-treatment periods, the probability of rejecting the null hypothesis essentially remains the same around 14 and 25 percent for the 5 and 10 percent significance levels respectively. The probabilities of rejecting the null are evidently still significantly higher than the test size which allows for the conclusion that when considering all seven specifications which includes both ones that satisfy the conditions of Ferman et al. (2019) and those that don't, specification searching is a valid concern, even when the number of pre-intervention periods is large. This is further confirmed when considering $T_0 = 400$. Such extensive data on pre-treatment periods does not exist in real applications, however, assessing the behavior of the SC estimations under this scenario allows us to study the asymptotic properties. However, again when considering all specifications, the rejection rates still remain significantly higher than the test size under both stationary and non-stationary models. Furthermore, when comparing the rejection rate estimates across those obtained under stationary and non-stationary models, no remarkable differences are observed. Hence, we can conclude that the estimation is robust to the nature of the dataset and any differences between the two are not substantial enough to significantly affect the rejection rate estimates.

Table 1: Rejection rates for different specifications

	Stationary Model		Non-Stationary Model	
	5%	10%	5%	10%
PANEL A: All seven specifications				
T0=12	0.143	0.250	0.142	0.254
	(0.003)	(0.004)	(0.004)	(0.004)
T0=32	0.146	0.255	0.158	0.275
	(0.003)	(0.004)	(0.004)	(0.005)
T0=100	0.143	0.254	0.152	0.264
	(0.003)	(0.004)	(0.004)	(0.004)
T0=400	0.134	0.241	0.145	0.255
	(0.003)	(0.004)	(0.004)	(0.005)
PANEL B: Specifications meeting conditions				
T0=12	0.106	0.190	0.110	0.198
	(0.003)	(0.004)	(0.003)	(0.004)
T0=32	0.100	0.179	0.109	0.191
	(0.003)	(0.004)	(0.004)	(0.005)
T0=100	0.090	0.157	0.094	0.162
	(0.003)	(0.004)	(0.003)	(0.004)
T0=400	0.077	0.138	0.081	0.142
	(0.003)	(0.004)	(0.004)	(0.005)

Furthermore, the probability of rejecting the null hypothesis is also generated when only considering 5 specifications that whose number of pre-treatment outcomes used as predictors increases with T_0. Hence, specifications 6 and 7 are excluded when estimating the rejection rates. These probabilities are outlined in Panel B of Table 1. We observe that the rejection rates greatly decrease for all pre-treatment periods. When considering the SCM estimation with $T_0 = 12$ for the stationary model, we see that when excluding specifications 6 and 7 the rejection rates drops from 14.3 percent to 10.6 percent. Thus, these results show that when we exclude specifications that violate the theoretical conditions for asymptotic equivalence (Ferman et al., 2019), the specification searching possibilities are weakened. However, the problem still remains as the rejection probabilities still remain markedly higher than the test size. Furthermore, the theoretical conditions (Ferman et al., 2019) when held suggest that the specification searching possibilities should be very small asymptotically. However, while the probability of rejecting the null hypothesis comes very close to the test size when $T_0 = 400$ (7.7 percent for a test size of 5 percent) it is still not negligible. Furthermore, this result not mitigate the problem in reality as such large number of pre-treatment periods does not exist.

The mechanism proposed by Ferman et al. (2019) to explain the observed differences in rejection rates when considering specifications that violate the theoretical conditions is as follows: For specifications 1 through 5, the weights converge to the same set of values as T_0 tends to infinity. However, they claim that the weights of specifications 6 and 7 may converge to different values. In order to investigate this, they assess in their simulation design the proportion of misallocated weights for the specifications. Misallocated weights refer to the proportion of weights assigned to control units that do not exhibit the same trends as the treated unit. Their findings indicate that the proportion of misallocated weights is much larger for specifications 6 and 7 and does not decrease when T_0 increases. However, for the specifications satisfying the conditions, the misallocated weights are smaller and decrease with the number of pre-treatment periods. The results also suggest that specifications 6 and 7 cannot capture the time-series dynamics of the units.

4.1 Results using Alternative V Matrix

This paper extends the study of specification searching by (Ferman et al., 2019) performing the above simulation exercises using an alternative method of V matrix calculation as known as the cross validation method proposed by Abadie et al. (2015). The choice of splitting period in this assessment is $t_0 = T_0/2$, alternatively, half the pre-intervention data is used to train the model and calculate the V matrix. The results of these simulations are presented in Table 2. We will now assess the robustness of the results discusses earlier to the choice of V matrix determination. Compared to 1, the estimated rejection rates are predominantly higher under this choice of V matrix calculation as shown in Table 2. Panel A represents the probability of obtaining at least one significant synthetic control estimate of the treatment effect when considering all specifications, including ones that don't satisfy the theoretical conditions (Ferman et al., 2019). Furthermore, in columns 1 and 2 the rejection rates obtained for the stationary model are presented. Under data generated from the stationary DGP, we estimate that the probability of finding a significant treatment effect for $T_0 = 12$ is 16.3 percent. When adopting this method, it appears that the specification searching possibilities increase around 2 percent. However, the properties observed in the rejection rates when removing specifications that don't adhere to the theoretical conditions do remain valid under this method. When looking at Panel B, we clearly see that the probability of rejecting the null hypothesis decreases. However, comparable to the original MC simulations, we see that the rejection rates are still significantly higher than the test size. However, even at $T_0 = 100$ where the theoretical results of Ferman et al. (2019) suggest that the rejection rates will be negligible, it remains at 10.3 percent for the 5 percent significance test.

Table 2: Rejection rates for different specifications using Cross Validation

	Stationary Model		Non-Stationary Model	
	5%	10%	5%	10%
PANEL A: All seven specifications				
T0=12	0.163	0.273	0.163	0.279
	(0.004)	(0.004)	(0.004)	(0.005)
T0=32	0.164	0.285	0.179	0.309
	(0.004)	(0.004)	(0.004)	(0.005)
T0=100	0.150	0.271	0.170	0.294
	(0.004)	(0.004)	(0.004)	(0.004)
PANEL B: Specifications meeting conditions				
T0=12	0.136	0.238	0.137	0.241
	(0.003)	(0.004)	(0.003)	(0.004)
T0=32	0.128	0.229	0.139	0.245
	(0.003)	(0.004)	(0.003)	(0.004)
T0=100	0.103	0.191	0.117	0.204
	(0.003)	(0.004)	(0.003)	(0.004)

Ultimately, the analysis of specification searching possibilities using cross-validation to calculate the V matrix yields similar results. This robustness in the choice of V matrix optimization supports the conclusion that there is room for specification in the SCM. Moreover, these findings also suggest that the method used to determine predictor weights in synthetic control estimations plays a significant role. Specifically, when employing the cross-validation method for V matrix calculation, there is a higher probability of rejecting the null hypothesis of no effect in at least one specification, even when there is no true effect present. This observation emphasizes the potential impact of the choice of V matrix determination on the statistical significance of the estimated effects.

4.2 Results using Alternative Splitting Periods

Finally, this paper set out to estimate the specification searching possibilities in another dimension of SCM: the choice of splitting the pre-treatment data into training and validation sets. To assess this, a Monte Carlo simulation was conducted using three distinct specifications that only differed in the choice of the splitting period. The rejection rates, presented in Table 3, indicate the probability of a researcher using the cross-validation method to estimate the V matrix successfully finding a splitting period that leads to rejecting the null hypothesis of no treatment effect. The first two columns of the table represent the estimation done for a stationary data generating process while columns 3 and 4 represent the analysis on the non-stationary DGP. Panel A conducts the synthetic control analysis using Specification 1, which uses all available pre-treatment lags in the estimation. We see that for $T_0 = 12$ the rejection rate for a 5 percent significance level test is around 8.7 percent. However when the number of pre-treatment lags increases, $T_0 = 32$, the probability of rejecting the null hypothesis decreases to 6.2 percent.

Table 3: Rejection rates using different splitting periods

	Stationary Model		Non-Stationary Model	
	5%	10%	5%	10%
PANEL A: S1				
T0=12	0.087	0.166	0.097	0.181
	(0.003)	(0.003)	(0.003)	(0.004)
T0=32	0.068	0.131	0.100	0.181
	(0.003)	(0.003)	(0.003)	(0.004)
T0=100	0.062	0.117	0.085	0.154
	(0.003)	(0.003)	(0.003)	(0.004)
PANEL B: S6				
T0=12	0.088	0.160	0.082	0.156
	(0.003)	(0.004)	(0.003)	(0.004)
T0=32	0.086	0.160	0.091	0.169
	(0.003)	(0.004)	(0.003)	(0.004)
T0=100	0.089	0.166	0.093	0.176
	(0.003)	(0.004)	(0.003)	(0.004)

Furthermore, Panel B outlines the rejection rates obtained when the SCM is performed using Specification 6, the mean of pre-treatment outcome lags as predictors. For the stationary model, we see that the rejection rates for pre-treatment periods 12, 32, and 100 all fall around 8.8 percent. The results when estimating the data from the non-stationary model lie very close to those obtained for the stationary model data. Thus, it suggests that the results are robust to the nature of the dataset increasing the confidence in the estimation method and its ability to provide reliable results. In the absence of specification searching we would expect the probability of rejecting at least one specification to be below the α percent significance level for a test size of α percent. Hence, the rejection rates being over 8 percent hint at a potential for specification searching when considering different splitting periods. However,it is interesting to note that the rejection rates seem to remain around the same rate and do not appear to follow any asymptotic patterns when the pre-treatment periods increases.

Interestingly, Table 3 shows that the rejection rates do not differ very much between those estimated under Specification 1 and Specification 6. Hence, these results suggest that the choice of the linear combination of pre-treatment lags used as predictors does not have an effect on the probability of yielding a significant result that rejects the null hypothesis. According to the theoretical conditions by Ferman et al. (2019), ideal we would expect that the rejection rates would be higher in the estimations using Specification 6. As the number of pre-treatment periods increases, the weights under Specification 1 converge to the same value while those under Specification 6 do not (Ferman et al., 2019). This would imply that conclusions obtained under estimating the SCM with the latter specification would yield disparate conclusions which could potentially imply higher odds of yielding a spurious significant result. However, the comparing between the results in Panel A and B in Table 3, we do note the presence of this phenomena.

5 Conclusion

This paper assesses the specification searching opportunities under the Synthetic Control framework in two dimensions. Firstly, it looks into the lack of guidance on the number and linear combination of pre-treatment outcome lags that can be used as predictors. This paper replicates the research design of Ferman et al. (2019), and tests their theoretical results that claim the possibility of specification searching becomes asymptotically irrelevant if the number of pre-treatment outcome lags used as predictors goes to infinity when the pre-treatment periods tends to infinity. In this study, a Monte Carlo simulation was conducted to investigate specification searching opportunities and test the reliability of asymptotic results where 10,000 datasets were generated using different data-generating processes. To assess the impact of different linear combinations of pre-treatment outcomes used as predictors, seven different specifications are defined for which we perform the SCM using them every round. The results provide insights into the likelihood of obtaining at least one spurious significant result which would provide a researcher implementing SCM with the opportunity to specification search. Congruent to Ferman et al. (2019) we find that specification searching remains a problem in SC applications as many SC applications don't have a large number of pre-treatment periods which is needed for the asymptotic results to hold. Furthermore, a large proportion of the papers implementing SCM adopt specifications that do not satisfy the conditions satisfied by the theoretical conditions. We extend this analysis by considering a different choice of optimization method of the V matrix to assess the robustness of the conclusions made on the specification searching possibilities. More explicitly this paper adopts the cross-validation technique proposed by Abadie et al. (2015). Under this optimization procedure, the conclusion of room for specification searching in the SCM is substantiated. The rate of incorrectly rejecting the null hypothesis is even higher under this estimation which further enhances confidence in specification searching findings.

Furthermore, we assess the potential implications of adopting different choices for splitting the pre-treatment periods into training and validation periods in the context of synthetic control methods. While this method of estimation is useful, it introduces the possibility of specification searching in other dimensions, such as the division of training and validation periods. The research extends the existing literature by investigating the effects of different splitting periods through Monte Carlo simulations. The simulations involve partitioning the data into splitting periods for SCM and selecting the appropriate V matrix using cross-validation. The training period varies in each simulation round, with three specifications using different fractions of the pre-intervention periods for training. Additionally, the simulations are performed for different numbers of pre-treatment periods to examine the asymptotic properties of SCM as the number of available periods increases. This paper finds evidence for the possibility of specification searching through adjusting the splitting period used in the determination of predictor weights. However contrary to what would be expected from literature, the probability of incorrectly rejecting the null hypothesis appears to be the same when considering specifications that satisfy the theoretical condistions of Ferman et al. (2019) and those that do not. Moreover, contrary to expectations, the possibility of specification searching does not appear to diminish as the number of available pre-intervention data increases.

Of course, this research is not without limitations. Firstly, this paper excludes an analysis with time-invariant covariates. As discussed in earlier sections, Kaul Ashok and Manuel (2015) point out that using all pre-treatment outcomes as predictors, while important to ensure a synthetic control with good fit, rends all other covariates irrelevant in the optimization steps. While Abadie and Gardeazabal (2003) and Abadie et al. (2015) stress the importance of closely matching the values of other covariates between the treated and other non-treated units, the optimization procedure will not attempt to match on the observed covariates when including all pre-treatment outcome lags . However, according to Botosaru and Ferman (2019) the bias of the SC estimator can potentially remain bounded when only balance for pre-treatment outcomes is achieved (Botosaru & Ferman, 2019). Kaul Ashok and Manuel (2015) claim however that the effects of observed covariates can provide a better control when the number of pre-treatment periods is small. Ferman et al. (2019) also consider an alternative simulation that includes time-invariant covariates and show that the same patterns in specification searching are observed across different pre-treatment periods and settings excluding specifications that do not satisfy the theoretical conditions. Hence, the conclusions do not seem to change when taking other covariates into account. Still it is important to extend the analysis of different splitting periods used for cross-validation estimation of V using specifications that include other covariates as well.

Abadie et al. (2010) and (Abadie et al., 2015) emphasize the importance of achieving a good pre-treatment fit before using the SCM to establish causal inference. Hence, it is important to assess whether the results generated from the simulation were due to the estimation of Synthetic Control units that did not ensure a good pre-treatment fit. If it were the case that the Synthetic Controls estimated for the simulated data had poor treatment fits and induced the spurious treatment effects, then the specification searching problem would not be as critical since such specifications would not be selected by researchers (Ferman et al., 2019; Abadie, 2021). While not performed in this paper, Ferman et al. (2019) show that the probability of rejecting the null hypothesis in at least one of the seven specifications remains significantly higher than the test size when restricting to synthetic controls with a good pre-treatment fit. However, it would be relevant to extend this analysis on synthetic controls with good pre-treatment fit when performing the MC simulation on different splitting periods.

Furthermore, some recommended areas of future research in the context of cherry picking with synthetic controls would be to perform the analysis for different training periods using the other specifications among the 7 discussed and assess if the cross validation method performs worse for certain linear combinations of pre-treatment outcomes used as predictors.

We will conclude with a few recommendations based on existing literature that could help abate the lack of guidance on the predictors used in the model, choice of V matrix and splitting period when adopting cross validation. The lack of consensus on which predictors to include in the estimation of the synthetic control unit clearly poses a threat of specification searching problems. Ferman et al. (2019) suggest using all pre-treatment outcomes in the specification as it minimizes the MSPE and does not result in ambiguity in the number of pre-treatment outcomes to be used. This specification, including all pre-treatment outcome lags, is also recommended by (Doudchenko & Imbens, 2016). Furthermore, if these covariates are deemed as unimportant for predicting the outcome, lack theoretical foundation for their use, or subject to data availability restrictions, ignoring them may result in variance reduction (Kaestner, Garrett, Chen, Gangopadhyaya & Fleming, 2017).

However, if a researcher aims to use other covariates besides the pre-treatment outcome lags they must not use all pre-treatment outcome lags as they risk giving these covariates no weight in the optimization. However, estimating such specifications is still possible and as long as a data driven procedure is used to estimate the V matrix, the weight given to covariates that should not be matched on is low. Furthermore, following the results of this paper which is in line with Ferman et al. (2019), if the specification using all pre-treatment outcome lags is not used, then one of the other specifications that satisfies the theroetical conditions must be used instead. This would at least ensure that with a large number pre-treatment periods, the specification searching problem is attenuated.

This paper shows however, that using an objective criteria that minimizes the MSPE as an objective criteria (Donohue, Aneja & Weber, 2018), is not a reliable method of specification selection. This is because we provide evidence for the potential of specification searching when it comes to the splitting period which requires subjective decision making from the researcher. Furthermore, as guided by the results which depict higher incorrect rejection rates of the null hypothesis when researchers employ the corss-validation technique of V matrix estimation (Abadie et al., 2015), conclusions using this method of optimization needs to be proceeded with caution. Hence, a critical step if using the cross-validation technique to estimate the V matrix would be to supplement the analysis by also performing the principal methods proposed by Abadie and Gardeazabal (2003) and (Abadie et al., 2010) that use the full sample to estimate V.

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. & L'Hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, *116*(536), 1817–1834.

Albalate, D., Bel, G. & Mazaira-Font, F. (2020, April). Ensuring stability, accuracy and meaningfulness in synthetic control methods: the regularized shap-distance method. *IREA-B, Working Paper*(2020-05). (Posted on 26th April 2020)

Albalate, D., Bel, G. & Mazaira-Font, F. A. (2021). Decoupling synthetic control methods to ensure stability, accuracy and meaningfulness. *SERIEs*, *12*, 549–584. doi: 10.1007/s13209-021-00242-8

Allegretto, S., Dube, A., Reich, M. & Zipperer, B. (2017). Credible research designs for minimum wage studies: A response to neumark, salas, and wascher. *ILR Review*, *70*(3), 559–592.

Athey, S. & Imbens, G. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), 431–497.

Becker, M., Klößner, S. & Pfeifer, G. (2017). Cross-validating synthetic controls. *Economics Bulletin*, *38*, 603-609.

Ben-Michael, E., Feller, A. & Rothstein, J. (2018). The augmented synthetic control method. *arXiv preprint arXiv:1811.04170*.

Berrar, D. (2019). Cross-validation. *Data Science Laboratory*. doi: 10.1016/B978-0-12-809633-8.20409-6

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

Botosaru, I. & Ferman, B. (2019, 5). On the role of covariates in the synthetic control method. *The Econometrics Journal*, *22*(2), 117–130. Retrieved from `https://doi.org/10.1093/ectj/utz001` doi: 10.1093/ectj/utz001

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., Scott, S. L. & et al. (2015). Inferring causal impact using bayesian structural time-series models. *The*, *9*(1), 247–274.

Chen, Q. & Yan, G. (2023). A mixed placebo test for synthetic control method. *Economics Letters*, *224*, 111004. Retrieved from `https://doi.org/10.1016/j.econlet.2023.111004` doi: 10.1016/j.econlet.2023.111004

Donohue, J. J., Aneja, A. & Weber, K. D. (2018). *Right-to-carry laws and violent crime: A comprehensive assessment using panel data, the lasso, and a state-level synthetic controls analysis* (Tech. Rep. No. 23510). Cambridge, MA: National Bureau of Economic Research. Retrieved from `https://www.nber.org/papers/w23510`

Doudchenko, N. & Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis.
(Advocate higher predictive power of outcomes of control units)

Dustmann, C., Schonberg, U. & Stuhler, J. (2017). Labor supply shocks, native wages, and the adjustment of local employment. *The Quarterly Journal of Economics*, *132*, 435–483. doi: 10.1093/qje/qjw052

Ferman, B., Pinto, C. & Possebom, V. (2019). Cherry picking with synthetic controls. *Journal of Business & Economic Statistics*, *37*(4), 532–545.

Firpo, S. & Possebom, V. (2016). Synthetic control estimation method: A generalized inference procedure and confidence sets.

Gobillon, L. & Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, *98*, 535–551. doi: 10.1162/REST$_{a0}$0608

Hinrichs, P. (2012). The effects of affirmative action bans on college enrollment, educational attainment, and the demographic composition of universities. *Review of Economics and Statistics*, *94*, 712–722. doi: 10.1162/REST$_{a0}$0232

Jardim, E., Long, M. C., Plotnick, R., van Inwegen, E., Vigdor, J. & Wething, H. (2017). Minimum wage increases, wages, and low-wage employment: Evidence from seattle. *NBER Working Paper*, *23532*.

Kaestner, R., Garrett, B., Chen, J., Gangopadhyaya, A. & Fleming, C. (2017). Effects of ACA medicaid expansions on health insurance coverage and labor supply. *Journal of Policy Analysis and Management*, *36*(3), 608–642. doi: 10.1002/pam.21993

Kaul Ashok, P. G., Klößner Stefan & Manuel, S. (2015). Synthetic control methods: Never use all pre-intervention outcomes together with covariates. *Journal Name*.

Kellogg, M., Mogstad, M., Pouliot, G. A. & Torgovitsky, A. (2021). Combining matching and synthetic control to tradeoff biases from extrapolation and interpolation. *Journal of the American Statistical Association*, *116*(536), 1804–1816. doi: 10.1080/01621459.2021.1979562

Kleven, H. J., Landais, C. & Saez, E. (2013). Taxation and international migration of superstars: Evidence from the european football market. *The American Economic Review*, *103*(5), 1892–1924. doi: 10.1257/aer.103.5.1892

Klößner, S., Kaul, A., Pfeifer, G. & et al. (2018). Comparative politics and the synthetic control method revisited: a note on abadie et al.(2015). *Swiss Journal of Economics and Statistics*, *154*(1), 11. doi: 10.1186/s41937-017-0004-9

Kuosmanen, T., Zhou, X., Eskelinen, J. & Malo, P. (2021, February). Design flaw of the synthetic control method. Retrieved 2021-03-05, from `https://mpra.ub.uni-muenchen.de/106390/` (MPRA Paper No. 106390)

Malo, P., Eskelinen, J., Zhou, X. & Kuosmanen, T. (2020). Computing synthetic controls using bilevel optimization.
(MPRA Paper No. 104085)

Neumark, D. & Wascher, W. (2017). Reply to 'credible research designs for minimum wage studies'. *ILR Review*, *70*(3), 593–609.

Peri, G. & Yasenov, V. (2017). The labor market effects of a refugee wave: Synthetic control method meets the mariel boatlift. *Social Science Research Network*. doi: 10.2139/ssrn.2940595

Rao, R. & Fung, G. (2008, 04). On the dangers of cross-validation. an experimental evaluation. In (p. 588-596). doi: 10.1137/1.9781611972788.54

Reich, M., Allegretto, S. A. & Godoey, A. (2017). *Seattle's minimum wage experience 2015–16* (Tech. Rep.). CWED Policy Brief, University of California, Berkeley.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. Retrieved from `https://doi.org/10.1037/h0037350` doi: 10.1037/h0037350

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, *25*(1), 57–76. doi: 10.1017/pan.2016.2