Forecasting Algorithms for Causal Inference with Panel Data

Jacob Goldin* Julian Nyarko† Justin Young‡

July 27, 2023

Abstract

Conducting causal inference with panel data is a core challenge in social science research. We adapt a deep neural architecture for time series forecasting (the N-BEATS algorithm) to more accurately predict the counterfactual evolution of a treated unit had treatment not occurred. The resulting estimator ("SyNBEATS") significantly outperforms two-way fixed effect, synthetic control, and matrix completion methods across a range of settings. In addition, it attains comparable or more accurate performance relative to a recently proposed method, synthetic difference-in-differences. Our results highlight how advances in the forecasting literature can be harnessed to improve causal inference in panel settings.

*University of Chicago and NBER, email: jsgoldin@uchicago.edu
†Stanford University. Corresponding Author, email: jnyarko@law.stanford.edu
‡Stanford University, email: justiny@stanford.edu

1 Introduction

Conducting causal inference with panel data is a core challenge in social science research. Consider a panel of states, one of which (the treated state) adopts a new policy. What is the effect of this policy on some outcome of interest? The question can be cast as a prediction problem: the (potential) untreated outcomes for the treated state can be observed in the time periods prior to new policy's adoption, but not afterwards. Meanwhile, the potential untreated outcomes for the control states can be observed both before and after the new policy's adoption. Identifying the causal effect of the policy entails estimating what the outcome would have been in the treated state during time periods after the policy's adoption, had the new policy not been adopted (Holland, 1986). Alternative tools for causal inference in panel data settings can be understood as different methods for predicting these counterfactual outcome on the basis of the data that is observed: namely, data from the treated state prior to the policy's adoption, and from the control states in time periods both before and after the policy's adoption.

In this paper, we draw on new advances in the time series forecasting literature to improve the accuracy of the predictions employed for causal inference in panel data settings. Over the past few years, the forecasting literature has developed a number of deep neural architectures that have significantly improved the the predictive capabilities of the models. However, these models are designed to be applied to data for single time series, i.e., predicting future values of a unit based on that same unit's past values. For causal inference with panel data, this is an important limitation because single-unit time series models do not incorporate information from the time series of control unit outcomes to help estimate missing values for the treated unit. We overcome this limitation by incorporating the time series of outcomes for control units into the forecasting model for the treated unit as additional features. That is, to predict the potential untreated outcome in the treated unit following the treatment, we feed into the model the outcome from the control states during the same time period – effectively casting contemporaneous and future outcomes of the control states as "leading indicators" for the potential untreated outcome of the treated state.

Among the different deep architectures for forecasting models, we focus on the neural basis expansion analysis for time series (N-BEATS) algorithm (Oreshkin et al., 2019). N-BEATS is a deep neural architecture designed to predict future values in a time series on the basis of past values. The algorithm has been shown to perform well in a range of forecasting tasks. The key innovation on which we rely is a recent adaptation of N-BEATS that incorporates time series other than the one being predicted as additional features (Olivares et al., 2021); in the panel data setting, this innovation allows us to use N-BEATS to predict

unobserved values of the treated unit on the basis of prior values of the treated unit as well as prior and contemporaneous values of the control units. Because our proposed approach essentially involves using the N-BEATS algorithm to estimate a synthetic (i.e., predicted) untreated outcome for the treated state during the post-treatment period, we refer to it as Synthetic N-BEATS ("SyNBEATS").

Although the N-BEATS algorithm has been shown to excel at a range of forecasting tasks, an important concern is whether its performance will be as strong when applied to the relatively small panel data sets typically employed in social science research. With limited data on which to train, simpler methods like synthetic controls (SC) or two-way fixed effects may yield more reliable causal estimates. To assess the suitability of SyNBEATS for panel data causal inference, we compare it to existing causal inference tools across two canonical panel data settings. Specifically, we contrast performance in data that has been used to estimate the effect of a cigarette sales tax in California (Abadie et al., 2010) and the German reunification on the West Germany economy (Abadie et al., 2015). In both of these settings, we find that SyNBEATS outperforms canonical methods such as SC and two-way fixed effect estimation. In addition, we compare the performance of these models on simulated events on abnormal returns in publicly traded firms (Baker and Gelbach, 2020). In this setting, where historical values would not be expected to provide much information about future outcomes, SyNBEATS still marginally improves performance relative to the other estimators. Finally, we compare SyNBEATS to two recent proposed causal inference methods for panel data settings: matrix completion (Athey et al., 2021) and synthetic difference-in-differences (Arkhangelsky et al., 2021). In the three settings we consider, we find that SyNBEATS generally achieves comparable (or slightly better) performance compared to SDID, and significantly outperforms matrix completion.

Our findings build on a growing literature applying machine learning tools to casual inference analysis. In addition to the new methods discussed above, two recent papers that are similar in spirit to ours include Mühlbach and Nielsen (2021), who adapt tree-based methods to synthetic control estimators, and Poulos and Zeng (2021), who apply a recurrent neural network model to predict treated units' outcomes based on pre-treatment observations. As discussed below, our approach differs from each of these in that it forms its counterfactual prediction directly using both pre-treatment values of the treated unit and post-treatment values of the untreated units.

The remainder of the paper proceeds as follows. Section 2 motivates our question of interest and provides background on the N-BEATS algorithm. Section 3 compares the performance of SyNBEATS to that of other causal inference methods in several panel data applications. Section 6 concludes.

2 Theoretical Framework

This section describes our notation and setup, introduces the N-BEATS estimator, and describes the other causal inference panel data estimators on which we focus.

2.1 Setup

We consider a panel data setting with N units across T time periods. One unit N adopts a new policy that takes effect for every period after T_0. We use $W_{it} \in \{0, 1\}$ to indicate whether a unit has adopted the policy, so that $W_{it} = 1$ for $i = N$ and $t > T_0$, and $W_{it} = 0$ otherwise. We are interested in the causal effect of the policy on some outcome of interest Y_{it}. Let $Y_{it}(1)$ and $Y_{it}(0)$ denote the (potential) outcomes of interest for unit i in period t corresponding to $W_{it} = 1$ and $W_{it} = 0$, respectively. For any unit and time period, the researcher observes only one potential outcome:

$$Y_{it} = W_{it}Y_{it}(1) + (1 - W_{it})Y_{it}(0)$$

The econometric challenge, illustrated in Figure 1(a), is to impute the missing counterfactual outcomes, $Y_{Nt}(0)$ for $t > T_0$, in order to estimate the average treatment effect on the treated:

$$\tau = E[Y_{it}(1) - Y_{it}(0)|W_{it} = 1]$$

2.2 Other Panel Data Casual Inference Estimators

A broad range of methods have been proposed to conduct causal inference in panel data settings. One popular approach is the synthetic control (SC) estimator (Abadie and Gardeazabal, 2003; Abadie et al., 2010), which constructs a synthetic counterfactual for the treated unit in the post-treatment periods by finding weights on relevant control unit values such that the synthetic unit fits the treated unit values well in the pre-treatment period. Doudchenko and Imbens (2016) generalize the SC literature and show these methods can be viewed as leveraging the structure of the data *vertically* across units (Figure 1(b)). In particular, the weights on the control units can be interpreted as regressing the treated unit outcomes on the control outcomes in the pre-treatment period:

$$\hat{\theta} = \arg\min_{\theta} \sum_{t \leq T_0} \left(Y_{Nt} - \theta_0 - \sum_{i=1}^{N-1} \theta_i Y_{it} \right)^2 \tag{1}$$

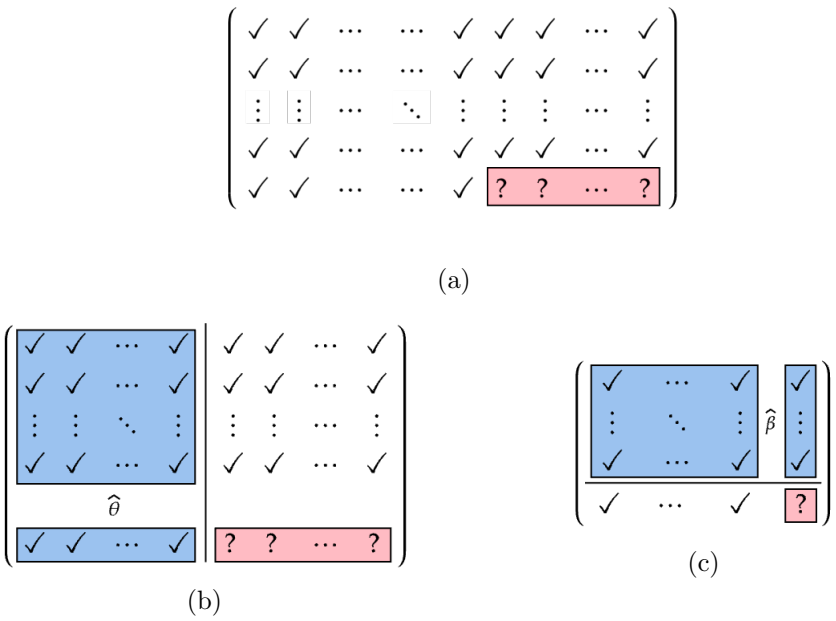


Figure 1: Problem statement and common estimators. Panel (a) illustrates the core challenge of imputing the counterfactual for the treated unit. Panels (b) and (c) respectively show the synthetic control and unconfoundedness approaches toward counterfactual estimation.

Although less common in panel data estimation, another method that can be employed stems from potential outcome imputation under unconfoundedness. With the assumption that temporal patterns are stable across units, this method uses the structure of the data *horizontally* (Figure 1(c)). In the panel data setting, this work can be interpreted as imputing missing values for a set of treated units in the post-intervention period based on outcomes during the pre-intervention period (Imbens and Wooldridge, 2009).

In particular, it extrapolates treated unit counterfactual values in $t > T_0$ by using weights obtained from regressing the control values in $t > T_0$ on those in periods $t = 1, \ldots, T_0 - 1$. The resulting estimates are a simple linear combination of the weights obtained from the controls and the associated $T_0 - 1$ pre-treatment values. Formally, with the number of treated units indexed by $\mathcal{I}$, we can view the weights as the solution to the following:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i \notin \mathcal{I}} \left(Y_{iT} - \beta_0 - \sum_{s=1}^{T-1} \beta_s Y_{is} \right)^2 \tag{2}$$

Whereas the methods discussed thus far rely primarily on either vertical or horizontal information, two-way fixed effects (TWFE) estimators simultaneously exploit stable patterns over time and across units (Angrist and Pischke, 2008). In the simplest case with no covariates, the TWFE estimate for the treatment effect is obtained from the following

regression:

$$Y_{st} = \alpha_s + \delta_t + \tau_{st}\mathbf{1}_{\{s=N\}}\mathbf{1}_{\{t>T_0\}} + \epsilon_{st}$$

Interactive fixed effects and more generally factor models extend this literature to allow for richer unobserved heterogeneity (Bai and Ng, 2002; Bai, 2003). These more general models assume the data generating process is given by a linear function of observed covariates and an unobserved low-rank matrix plus noise. Without covariates and with binary treatment, the outcomes can be described as

$$\boldsymbol{Y} = \boldsymbol{L} + \boldsymbol{W} \odot \boldsymbol{\tau} + \boldsymbol{\epsilon}$$

where $\boldsymbol{L}$ represents the target estimand we wish to recover. In the factor model literature, $\boldsymbol{L} := \boldsymbol{U}\boldsymbol{V}^T$ with loadings $\boldsymbol{U}$ and factors $\boldsymbol{V}$. Here, the underlying matrix $\boldsymbol{L}$ exhibits low-rank unobserved heterogeneity, allowing for arbitrary interactions beyond simple additive effects.

A recent estimator for $\boldsymbol{L}$ proposed by Athey et al. (2021) couples this literature with a parallel literature arising in statistics. Matrix completion (MC) aims to recover missing entries in a data matrix that are missing at random, in contrast to the "block" missingness commonly seen in economics. In our case, this structure takes the form of a counterfactual missing block on the lower right-hand corner of the matrix. Athey et al. (2021) combine both approaches by solving the low-rank matrix completion problem with fixed effects explicitly *not* regularized and show this resulting estimator performs well in a variety of settings.

Another recent method simultaneously incorporating both vertical and horizontal information is the *synthetic difference-in-difference* (SDID) estimator, which combines TWFE and SC methods to incorporate separate weighting across both units and time periods (Arkhangelsky et al., 2021). First, the authors employ a SC-type approach to reweight the unexposed control units to create a parallel trend. Unlike traditional SC methods, SDID allows for a level-shift from the actual outcomes as long as the synthetic outcomes and actual outcomes are parallel in the pre-treatment period for the unit of interest. Second, with parallel trends established, SDID applies a difference-in-differences type analysis on this reweighted panel. In contrast to standard DID however, the time periods also have weights placed on them, with the intuition that more recent periods will be more informative about the present. These weights $\hat{\lambda}_t$ are constructed in an analogous fashion to the unit weights but without regularization. With both sets of weights, the estimator recovers the treatment effect by solving:

$$(\hat{\tau}^{SDID}, \hat{\mu}, \hat{\alpha}, \hat{\delta}) = \arg\min_{\tau,\mu,\alpha,\delta} \left\{ \sum_{i=1}^{N}\sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \delta - W_{it}\tau)^2 \hat{\omega}_i \hat{\lambda}_t \right\}$$

2.3 SyNBEATS

We cast the challenge of causal inference in the panel data setting as a supervised learning problem, and propose drawing on both horizontal and vertical information to inform the prediction of the (missing) untreated potential outcomes for the treated unit in the post-treatment time period. Specifically, we propose a neural network architecture to determine prediction weights across time and units, depicted in Figure A.1.

Model training is based on the mean squared error loss over the preintervention period when Y_{Nt} is fully observed. To form predictions for these values at every time step, the model uses the contemporaneous control outcomes at t and the lagged treated unit's outcomes from $t - n_{lag}$ to $t - 1$, where n_{lag} denotes the number of lagged outcomes. Because n_{lag} is constant across all training examples, we can define $\underline{t} := t - n_{lag}$. Formally, at every $t = 1, \ldots, T_0$, our model predicts the observed Y_{Nt} with $\{Y_{it}\}_{i=1}^{N-1}$ and $\{Y_{Ns}\}_{s=\underline{t}}^{t-1}$. The model, $m : \mathbb{R}^{(N-1)+(n_{\text{lag}})} \to \mathbb{R}$, is trained via gradient descent methods to minimize the following loss:

$$J(\theta) = \frac{1}{T_0} \sum_{t \leq T_0} \left(Y_{Nt} - m\left(\{Y_{it}\}_{i=1}^{N-1}, \{Y_{Ns}\}_{s=\underline{t}}^{t-1} \; ; \; \theta \right) \right)^2 \tag{3}$$

In a standard feedforward neural network, each training input $\boldsymbol{x} \in \mathbb{R}^{d_0}$ is fed through multiple *layers*, with each layer ℓ generating an intermediate output $\boldsymbol{h}^{[\ell]} \in \mathbb{R}^{d_\ell}$.[1] This relationship is governed by

$$\boldsymbol{h}^{[\ell]} = f^{[\ell]}(\boldsymbol{W}^{[\ell]} \boldsymbol{h}^{[\ell-1]} + \boldsymbol{b}^{[\ell]})$$

Here, $\boldsymbol{W}^{[\ell]} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ and $\boldsymbol{b}^{[\ell]} \in \mathbb{R}^{d_\ell}$ represent weight and bias matrices akin to standard linear regression. In the first layer, $\boldsymbol{h}^{[0]} \equiv \boldsymbol{x}$. The function $f(\cdot)$ is taken here to be $f(x) = \max\{0, x\}$. In essence, f explicitly introduces non-linearity into the network to allow it to learn such relationships in the data. After an arbitrary number of layers L with intermediate outputs $\boldsymbol{h}^{[1]}, \ldots, \boldsymbol{h}^{[L]}$, the model yields a final output $\hat{\boldsymbol{y}} \in \mathbb{R}^{d_y}$. This in turn is compared to the observed outcome $\boldsymbol{y} \in \mathbb{R}^{d_y}$ and a standard loss function (e.g. MSE) is used to compute the overall loss. As each $\boldsymbol{h}^{[\ell]}$ depends on all components of $\boldsymbol{h}^{[\ell-1]}$ at each layer, we call such a network *fully connected*. As with many complex models, a closed-form solution does not exist generically, and the parameters must be estimated via gradient descent methods. In neural networks, this numerical method proceeds in a familiar manner: the gradient of the loss function propagates backward through the network via the chain rule to update all weights $\boldsymbol{W}^{[\ell]}$ and $\boldsymbol{b}^{[\ell]}$ until the weights converge.[2]

[1] For ease of notation, $\ell = 1, \ldots, L$ denotes which layer is of interest.

[2] Without any assumptions or additional structure on the problem, a global minimum is not guaranteed,

SyNBEATS largely follows Oreshkin et al. (2019) in its model architecture. For ease of exposition, we focus on the case with a single time series (i.e., data for one unit over time). However, we adapt the framework of Olivares et al. (2021) to allow for covariates, or other units. In SyNBEATS, the fully connected neural network architecture described above is compartmentalized into different *blocks*. Given a time series input $\boldsymbol{x} \in \mathbb{R}^{d_0}$, where d_0 is the length of the lookback period used to forecast, each block k adopts this standard neural network structure but generates not only a forecast $\hat{\boldsymbol{y}} \in \mathbb{R}^{d_y}$ with d_y as the length of the forecast horizon, but also a backcast $\hat{\boldsymbol{x}} \in \mathbb{R}^{d_0}$. As shown in Figure 2, the model does so in each block via a set of shared fully connected layers that branch into two separate forecast and backcast layers. For each block k, the shared fully connected layers yield intermediate outputs $\boldsymbol{h}_k^{[1]}, \ldots, \boldsymbol{h}_k^{[L]}$ as before. Next, it splits $\boldsymbol{h}_k^{[L]}$ into $\boldsymbol{h}_k^f$ and $\boldsymbol{h}_k^b$, which respectively serve as the final intermediate outputs for the forecast and backcast, respectively. The predictions in each block are finally given by:

$$\hat{\boldsymbol{y}}_k = g_k^f(\boldsymbol{W}_k^f \boldsymbol{h}_k^f + \boldsymbol{b}^f)$$
$$\hat{\boldsymbol{x}}_k = g_k^b(\boldsymbol{W}_k^b \boldsymbol{h}_k^b + \boldsymbol{b}^b)$$

Here, $g^f(\cdot)$ and $g^b(\cdot)$ are linear projection operators, although one can also use other basis functions as well (Oreshkin et al., 2019).

The backcast at block k, $\hat{\boldsymbol{x}}_k$, is fed to the next block $k+1$ as the new input:

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - \hat{\boldsymbol{x}}_{k-1}$$

The forecast at block k $\hat{\boldsymbol{y}}_k$ is added iteratively to the forecasts at all other previous blocks:

$$\hat{\boldsymbol{y}}_{\text{block}} = \sum_k \hat{\boldsymbol{y}}_k$$

In other words, this block structure breaks down the signal from input $\boldsymbol{x}$ step-by-step, with each $\hat{\boldsymbol{x}}_k$ representing the signal leftover that could not be explained by the previous neural networks in blocks $1, \ldots, k$. Although with infinite data, an arbitrarily large neural network should be able to flexibly mine all the signal in the input and achieve arbitrarily low error, in practice these methods cannot do so within reasonable computational limits, which helps explain the success of this architecture.

but in practice even local minima may perform well in RMSE compared to standard econometric models.

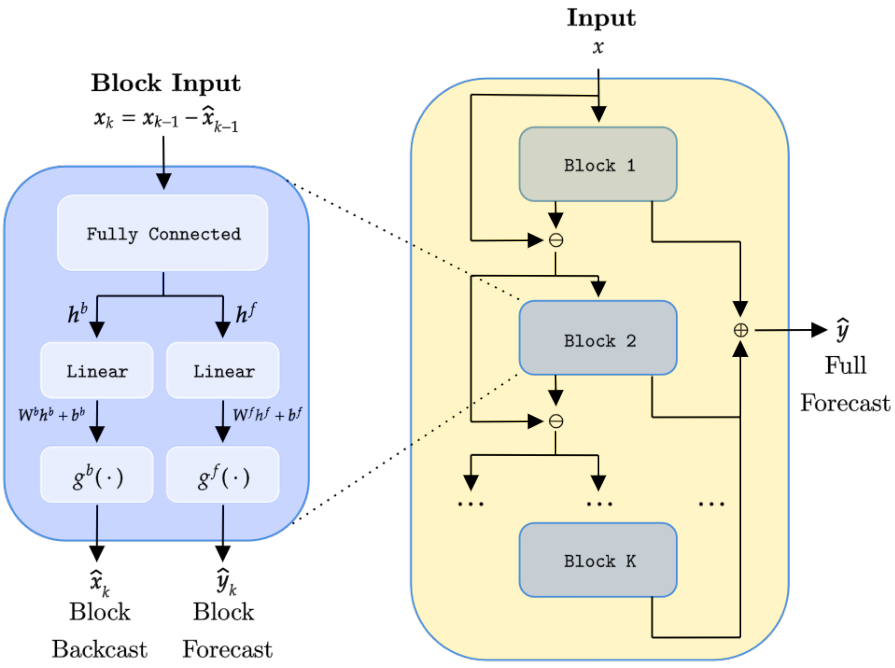


Figure 2: Model Architecture. *Left*: A single block k. *Right:* The whole model.

Statistical Inference Formally deriving analytic expressions for SyNBEATS' asymptotic distribution is outside the scope of this paper. However, estimators of design-based uncertainty that have been proposed for SC-type estimators can be readily adapted to SyNBEATS, such as by applying SyNBEATS to placebo treatments obtained by permuting the treatment year or treatment unit (Abadie et al., 2010; Shaikh and Toulis, 2021). Alternatively, instead of modeling the assignment mechanism directly, one can quantify uncertainty in model prediction using conformal inference methods to obtain finite sample coverage guarantees (Chernozhukov et al., 2021).

3 Comparison of SyNBEATS to Popular Estimators

In this section, we compare the performance of SyNBEATS to two popular methods for causal inference in panel settings: synthetic controls and two-way fixed effects estimation. We analyze these methods in various data settings using two placebo exercises.

3.1 Framework

In each setting, we collect outcome data Y_{it} for a balanced panel of $i \in \{1, 2, ..., N\}$ untreated units and $t \in \{1, 2, .., T_0, T_0 + 1, ..., T\}$ periods. We next assume that a single unit (N) is affected by a pseudo treatment W, beginning in year T_0, so that $W_{it} = 1$ if and only if $i = N$ and $t > T_0$. Our goal is to estimate the average treatment effect on the treated of the pseudo-treatment on the pseudo-treated unit N in post-treatment years $t > T_0$:

$$\tau_t^{ATT} = \mathbb{E}[Y_t(1) - Y_t(0)|W_t = 1] \tag{4}$$

As $Y_{Nt}(1)$ for $t > T_0$ is observed, and because a simple sample average (the mean of Y for unit N in the post-intervention period) is consistent and unbiased for $\mathbb{E}[Y_t(1)|W_t = 1]$, to obtain an overall unbiased and consistent estimator of τ^{ATT} requires estimating $\mathbb{E}[Y_t(0)|W_t = 1]$. We compare alternative estimators for this term by analyzing the prediction error on a range of datasets and simulations. To that end, we conduct placebo exercises assuming $Y_{Nt}(0)$ is unobserved for $t > T_0$ and we predict $Y_{Nt}(0)$ (out-of-sample) using other data from the panel that would be available to the researcher if the unit really had been treated (i.e., if $Y_{Nt}(0)$ really was unobserved). By comparing estimates of $Y_{Nt}(0)$ derived from alternative methods to the true $Y_{Nt}(0)$, we can evaluate the performance of the various methods in estimating τ^{ATT}. We consider two main validation exercises: (1) the prediction error obtained from untreated control units in the post-treatment period, and (2) the prediction error for the treated unit in time periods during the pre-treatment period.

3.2 Implementation of Estimators

This subsection briefly discusses our implementation of SyNBEATS, Synthetic Controls, and OLS with two-way fixed effects.

Synthetic Controls We implement the SC method as proposed by Abadie et al. (2010). As recommended in the subsequent literature, we use pre-treatment outcomes to assess a control state's suitability for inclusion in the SC (Cavallo et al., 2013; Doudchenko and Imbens, 2016).

Two-Way Fixed Effects We consider a standard two-way fixed effects (TWFE) regression model. For consistency with the other methods we consider, we allow the treatment

effect to vary by post-treatment year, and estimate the following model:

$$Y_{st} = \alpha_s + \delta_t + \sum_{i=T}^{T+K} \tau_i \mathbf{1}_{\{s=N\}} \mathbf{1}_{\{t=i\}} + \varepsilon_{st}$$

Unlike the other approaches considered here, we estimate the TWFE model on both the pre-treatment values of $Y_{st}(0)$ as well as the post-treatment values of $Y_{st}(1)$. The model's implied prediction of the potential untreated outcome in a given post-treatment year can be derived from the estimated treatment effect in the corresponding year:

$$\widehat{Y}_{st}(0) = Y_{st}(1) - \widehat{\tau}_t$$

SyNBEATS We implement our SyNBEATS algorithm as described in Section 2. We use the original hyperparameters as described by Oreshkin et al. (2019), as they have shown to yield high performance in a large number of time series settings and our applications have limited data available for hyperparameter tuning.[3] For the same reason, we set $n_{lag} = 1$. Although in principle n_{lag} is a hyperparameter that can be tuned to improve performance, for convenience we treat it as fixed. To predict beyond a single post-treatment period, we iteratively use the predicted value of the preceding period as an input to the algorithm; for example, in order to predict $\widehat{Y}(0)_{s,t+1}$, we use $\widehat{Y}(0)_{st}$ as an input.

3.3 Proposition 99 in California

In 1988, California implemented Proposition 99, raising the state tax on cigarettes from 10 cents to 35 cents per pack. Abadie et al. (2010) apply the SC method to this setting to estimate the effect of the tax on cigarette sales. Comparing the real California to its synthetic counterpart, they find that the tax reduced cigarette sales in the years following its adoption.

Below, we perform two exercises to assess the performance of SyNBEATS in this setting. The period of observation in the data set ranges from 1970 to 2000. The post-treatment period begins in 1989. Our outcome of interest is per-capita cigarette sales (in packs), which we observe at the state-year level. Following Abadie et al. (2010), we take as our control group the 38 states that did not adopt tax increases during this period.

[3]Specifically, our architecture includes 30 stacks, a single block with 4 layers and a layer width of 256. We use the Adam optimizer with default settings, including a learning rate of 0.001.

Exercise 1: Pseudo-Treated States

In our first exercise, we exclude California (the true treated state). Instead, we assume that one other state from the control group (the "pseudo-treated" state) has been treated with a cigarette sales tax in 1989. We mask the post-1988 cigarette sales of the pseudo-treated state and apply each of the alternative causal inference methods. For each method, we assess the prediction error for the (actually untreated) outcome in the pseudo-treated state in the post-treatment period. This exercise will yield a valid assessment of the prediction errors for California in 1989 to the extent that, for each method we consider, the 1989 prediction error for California is drawn from the same distribution as the 1989 prediction error for the control states.

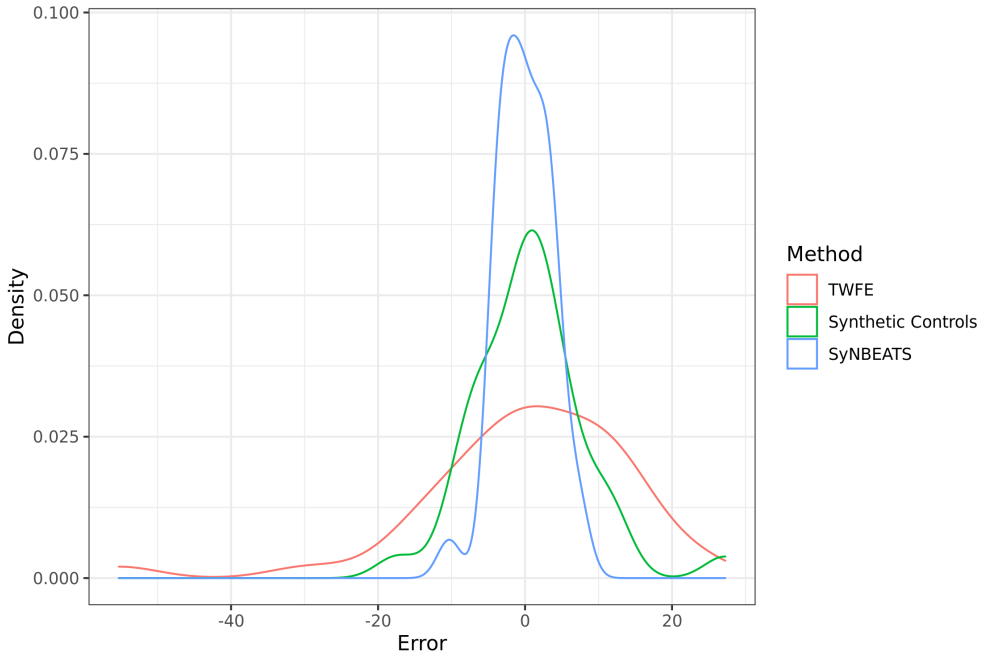
We repeat this process for the 38 control states, iteratively defining each control state as the pseudo-treated state. We first focus on the prediction error for the first post-treatment year (1989). This yields 38 prediction errors for each estimation method – one for each pseudo-treated state. To summarize the performance of a method, we compare the root mean squared error (RMSE) of the predictions across control states.

We also evaluate each model based on a longer-term, 5-year prediction window (1985–1989). In this case, each state will have five prediction errors, one for each post-treatment period. For the longer-term predictions, we calculate mean squared error based on the prediction errors in each pseudo-treated state over each post-treatment year.

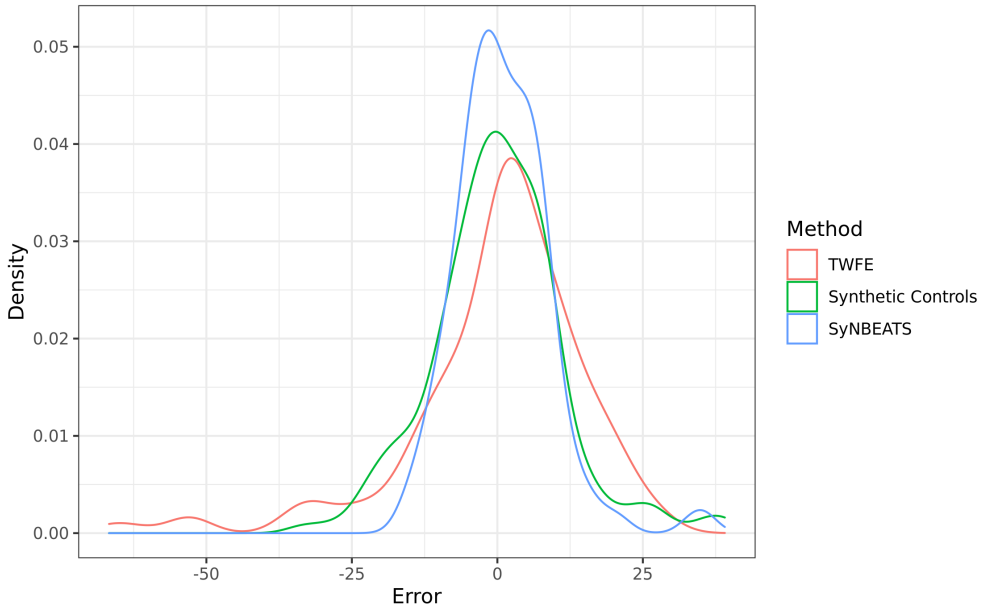
Because the out-of-sample prediction error determines the accuracy of the estimated treatment effect (Equation (4)), we compare the various estimation methods along this dimension. The model predictions are visualized in Figure 3, and the resulting distribution of prediction errors is summarized in Table 1. Among the methods we consider, SyNBEATS yields the most accurate prediction, with an RMSE of 3.59, 54% improvement over the second-best alternative of SC. Similarly, for longer-term predictions, SyNBEATS yields the best performance, with an RMSE of 8.17, which is a 27% improvement over SC as the second-best alternative.

Exercise 2: Pseudo-Treated Years

In this second Exercise, we maintain California as the treated state, but we (counterfactually) assume that Proposition 99 took effect during some year prior to 1989 (the "pseudo-treated year"). We iteratively define as our pseudo-treated year each year between 1975 and 1988. To assess longer-term predictions, we also consider 5-year pseudo-post-treatment periods in the same calendar year window (i.e., 1975–1979 through 1984–1988). In each iteration, we use data from years prior to the pseudo-treated year to estimate the predictive



(a) 1-Year Predictions



(b) 5-Year Predictions

Figure 3: Prop 99 Prediction Errors from Placebo Treated States

The figure shows the distribution of prediction errors for pseudo-treated states, by prediction method, for the Proposition 99 data set. In Panel A, the outcome being predicted is the pseudo-treated state's smoking rate in 1989 ($N = 38$); the model is trained on data from the pseudo-treated state from 1970-1988 and from the other control states for 1989. In Panel B, the outcome being predicted is the pseudo-treated state's smoking rate in years 1985–1989 ($N = 190$); the model is trained on data from the pseudo-treated state from 1970–1984 and from the other control states for the year being predicted.

models.[4] This exercise will yield a valid assessment of the prediction errors for California in 1989 to the extent that, for each method we consider, California's prediction error in 1989 is drawn from the same distribution as its prediction error for earlier years.

As shown in Table 1, SyNBEATS outperforms other traditional estimators in their short-term predictions, improving the RMSE by 31% compared to the second-best alternative (SC). To further facilitate a direct comparison of short- to long-term predictions for each estimator, in Figure 4, we contrast predictions in the first year after treatment to those obtained in the fifth year after treatment. Not surprisingly, all estimators perform worse for longer-term predictions, with the performance difference between SyNBEATS and SC closing as well. Focusing on the full five-year window, SyNBEATS slightly outperformed SC.

Finally, to assess the source of SyNBEATS' performance gains relative to other algorithms, we consider a pure forecasting approach based on the N-BEATS algorithm. Unlike SyNBEATS, this pure forecasting algorithm does not take as inputs the outcomes for the control states in years at or after the treatment. Instead, it uses information only on pre-treatment outcomes of the treated unit. This approach is consistent with the implementation of N-BEATS as originally proposed by Oreshkin et al. (2019). We thus refer to it simply as "N-BEATS".

Appendix Table A.4 compares the performance of N-BEATS with SyNBEATS. Not surprisingly, N-BEATS attains worse performance than SyNBEATS. In addition, depending on the exercise, N-BEATS attains similar or slightly better performance as the other methods for one-year predictions. This suggests that much of the performance gains in SyNBEATS might be traced back to its ability to efficiently learn the time series structure of the treated unit's outcomes. In contrast, N-BEATS performs substantially worse than competing estimators for the longer-term predictions, suggesting that incorporating the post-treatment information from the control units becomes increasingly important over the prediction horizon.

Application

In this subsection, we apply SyNBEATS and the other estimators we consider to estimate the effect of Proposition 99 on California. Specifically, we train the model to predict smoking rates in California during years prior to 1989, and apply the model out-of-sample to predict California smoking rates in 1989 onward. Our estimated treatment effects correspond to the difference between these predictions and the actual, observed smoking rates in California during these years. The results of this exercise are displayed in Figure A.2.

[4]The one exception is the TWFE model, which, as described above, we estimate using data that includes pseudo-treated years.

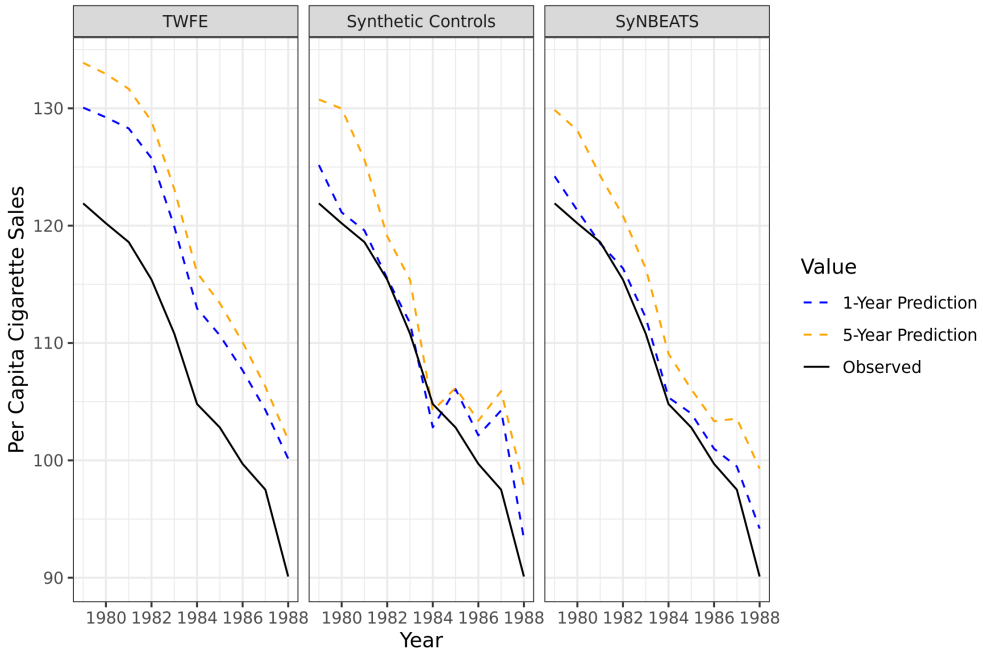


Figure 4: Prop 99 Prediction Errors from Placebo Treated Years

The figure compares short-term (year 1 post-treatment) and long-term (year 5 post-treatment) prediction errors obtained from predicting the smoking rate in California in pseudo-treated years using various estimators. The blue, dashed line depicts the prediction error in the first pseudo-treated year. The orange, dashed line depicts the predictions in the fifth pseudo-treated year. The predictions are formed using data from California from 1970 through the year prior to the first pseudo-treated year, and from the control states using data from the pseudo-treated year.

Using SyNBEATS, we estimate that Proposition 99 reduced cumulative cigarette sales per capita by a total of 192 packs in the ten years following its passage. Over the same time horizon, the other estimators imply reductions of 207 packs (SC) and 292 packs (TWFE). In relative terms, the SyNBEATS estimate corresponds to a reduction in cigarette sales of 22% (p=0.03).[5].

3.4 German Reunification

The Berlin Wall fell on November 9, 1989. A year later, on October 3, 1990, West Germany and East Germany officially reunited, effectively marking the end of the Cold War. In a canonical study, Abadie et al. (2015) apply the SC method to examine the causal effect of the reunification on the economy of West Germany, using twenty other countries as control units. They find that the reunification reduced West Germany's per capita GDP.

Again, we perform two exercises to assess the accuracy of alternative causal inference

[5]This p-value is derived from a placebo test over pseudo-treated units. We estimate a percentage reduction in smoking that is as large as California's in only one of the 38 placebo states we consider.

methods in this setting. First, we iteratively predict the 1990 GDP of each country in the control group. Second, we iteratively predict West Germany's GDP from 1963 to 1989. In this setting as well, SyNBEATS consistently outperforms the other methods we consider, with respect to both short-term and long-term predictions, and with respect to both predicting pseudo-treatment country outcomes in the true treatment year (Figure A.3) and in predicting true treatment country outcomes in pseudo-treatment years (Figure A.4). As in the other experiment, a pure forecasting algorithm (N-BEATS) performs reasonably well for short-term predictions, but incorporating post-treatment control unit information appears important for maintaining prediction accuracy over longer time horizons (Appendix Table A.4).

As above, we apply SyNBEATS and the other estimators we consider to estimate the effect of the German reunification on West Germany's GDP. Specifically, we train the model to predict the GDP in West Germany during years prior to 1990, and apply the model out-of-sample to predict West Germany GDP in 1990 onward. Our estimated treatment effects correspond to the difference between these predictions and the actual, observed GDP in West Germany during the post-treatment years. The results of this exercise are displayed in Figure A.5. Using SyNBEATS, we estimate that reunification reduced GDP per capita in West Germany by a total of \$17,766 per capita in the 13 years after it occurred. In percentage terms, this corresponds to a decrease of 5% (p=0.05).[6]. In contrast, over the same time horizon, SC implies a GDP reduction of \$15,116. The two-way fixed effects estimator suggests an average *increase* in GDP per capita of \$9,577.

3.5 Predicting Abnormal Returns for Simulated Stock Market Events

In this exercise, we rely on data used in Baker and Gelbach (2020) to predict abnormal stock returns for simulated events. This exercise is well suited for assessing the performance of the various estimators for use in financial event study analyses – i.e., estimates of the causal effect of a shock (such as securities litigation or a merger announcement) on stock prices. The data set includes returns for firms with a share price above \$5 between 2009 and 2019. It also contains 10,000 randomly selected, unique firm-level pseudo-events (i.e., the events do not correspond to anything that would be expected to affect the firm's stock price). For each firm-level event, we use returns data for the 250 trading days prior to the event to predict the returns on the event date. As in Baker and Gelbach (2020), for each firm, our

[6]As above, this p-value is derived from a placebo test over untreated units; only one of the 16 placebo countries we consider has an estimated percentage GDP reduction as large as West Germany's.

pool of control units contains all firms with the same four-digit SIC industry code; if there are fewer than eight such firms, we include peers with the same three-digit SIC industry code.

Using 100 randomly selected pseudo-events from this data, we compare the predictions of each estimator for the stock price on the day after the simulated event (Table 1). As in the previous settings, SyNBEATS tends to generate the most accurate predictions, although here the performance gains are more modest. One potential explanation for why SyNBEATS does not improve performance as dramatically in this setting is that there is limited information in prior stock performance that can be used to predict future stock performance.

4 Comparison to Recent Methods

So far, we have provided evidence that SyNBEATS outperforms SC and TWFE, the two most commonly employed methods by social scientists for causal inference in panel data. In this section, we compare SyNBEATS to two recently proposed estimators, matrix completion (MC) and synthetic difference-in-differences (SDID), described in Section 2.

4.1 Evaluation Using Empirical Applications

We begin by comparing the performance of SyNBEATS to MC and SDID using the three empirical applications described in Section 3. To implement MC, we follow Candes and Tao (2010) and the panel data modified version of MC introduced in Athey et al. (2021).

To implement SDID, we follow Arkhangelsky et al. (2021) and their implementation in the associated code release. In particular, we first construct the level-shifted synthetic control with the unit weights described in Section 2. Next, a weighted DID analysis is performed, where the pre-treatment average value of the treated unit and its level-shifted synthetic control are both given as weighted averages determined by the SDID time weights. The post-treatment average values are just simple averages, and a standard DID analysis is then applied. We note that SDID yields an average treatment effect on the treated over the entire post-treatment period of interest. To obtain predictions over varying treatment horizons, we recover the per-period treatment effects by conducting the last DID analysis step on *each* value in the post-treatment period rather than the simple average over all periods.

To compare SyNBEATS with MC and SDID, we replicate the analyses in the previous section with these estimators instead of SC and TWFE. The results are presented in Table A.3. SyNBEATS dramatically outperforms MC in each analysis we consider with the data sets corresponding to Prop 99 and German Reunification. With respect to SDID, the story

is more nuanced: SyNBEATS performs better in 6 out of the 8 analyses we consider using this data, but SDID out-performs SyNBEATS in the remaining two. Notably, the performance gains of SyNBEATS compared to SDID are smaller than with the other estimators we consider. Finally, with respect to the stock analysis, SyNBEATS yields the lowest RMSE, but all three of the estimators yield comparable performance – again, consistent with the hypothesis that time series forecasts are unlikely to greatly improve predictive power in this context.

Table 1: Comparison of Panel Data Methods

Method	Prop 99				German Reunification				Stocks
	Units		Years		Units		Years		
	Short	Long	Short	Long	Short	Long	Short	Long	
TWFE	14.420	15.386	8.369	10.193	2,254.893	2,103.349	797.574	824.017	0.0284
Synthetic Controls	7.705	11.152	2.716	4.264	878.390	1,122.949	115.919	305.509	0.0281
SyNBEATS	**3.591**	**8.176**	**1.882**	**4.088**	**325.792**	**876.349**	**70.747**	**214.964**	**0.0275**

Notes: The table compares the performance of the different estimators based on the RMSE. Columns labeled "Prop 99", "German Reunification" and "Stocks" describe our different evaluation data sets. Columns labeled "Units" refer to analyses that consider pseudo-treated units and columns labeled "Years" refer to analyses that consider pseudo-trated years. Columns labeled "Short" refer to one-year predictions and columns labeled "Long" refer to five-year predictions.

4.2 Evaluation Using Simulations

Our findings thus far suggest that SyNBEATS consistently outperforms TWFE, SC, and MC. In contrast, the results were more mixed with respect to SDID: SyNBEATS achieved lower RMSE than SDID for 7 out of the 9 evaluations we considered, whereas SDID yielded lower prediction error for 2 of the 9 evaluations.

To shed additional light on the factors shaping the relative performance of SyNBEATS and SDID, we next conduct a simulation exercise. We consider a data generating process via a linear factor model (Xu, 2017; Bai, 2009). We adapt the model in Xu (2017) to allow for variation in the relative importance of fixed and interactive effects in generating the outcomes. Specifically, our DGP is given as:

$$Y_{it} = c * (\alpha_i + \xi_t) + (1 - c) * (x_{it,1} \cdot 1 + x_{it,2} \cdot 3 + \lambda_i' f_t + 5) + \epsilon_{it} \tag{5}$$

where $f_t = (f_{1t}, f_{2t})'$ and $\lambda_i = (\lambda_{i1}, \lambda_{i2})'$ are two-dimensional time-varying factors and unit-specific factor loadings, respectively. Here, $\epsilon_{it}, \eta_{it,1}, \eta_{it,2} \sim N(0, 1)$. Factors and time fixed effects are similarly drawn: $f_{1t}, f_{2t}, \xi_t \sim N(0, 1)$. Factor loadings and unit fixed effects are drawn uniformly with zero mean and unit variance: $\lambda_{i1}, \lambda_{i2}, \alpha_i \sim U[-\sqrt{3}, \sqrt{3}]$. The key in this model specification is that the regressors are positively correlated with factors, loadings, and their product. For $k = 1, 2$, regressors are given by

$$x_{it,k} = 1 + \lambda_i' f_t + \lambda_{i1} + \lambda_{i2} + f_{1t} + f_{2t} + \eta_{it,k}$$

In our simulations, we iteratively vary three parameters, each time comparing the performance of SyNBEATS to SDID. The parameters we vary were selected to capture a range of panel data settings that empirical researchers may confront in practice.

- We vary the number of control units, N, from 10 to 100 in increments of 10.
- We vary the number of pre-treatment periods, T, from 10 to 100 in increments of 10.
- We vary the relative importance of fixed effects, c vis-a-vis the other components contributing to the outcome Y.

We simulate data for each parameter choice 100 times and estimate outcomes over 4 periods. Model performance is evaluated as the average RMSE over all simulations and post-treatment periods. Results are depicted in Figure 5.

As shown in the figure, SyNBEATS consistently outperforms SDID under a variety of parameter choices. However, the relative advantage of SyNBEATS decreases (and even

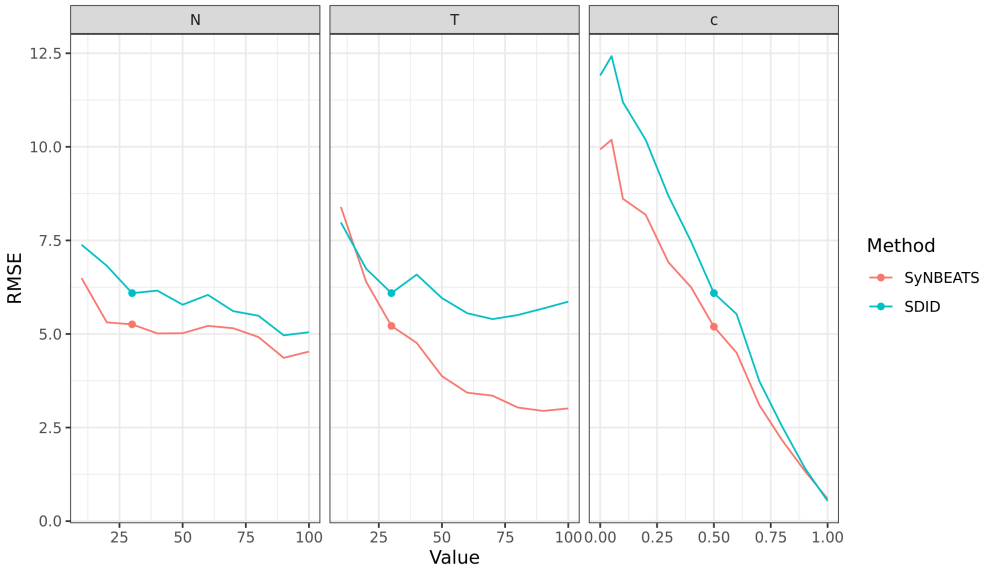


Figure 5: Simulation Comparison of SDID and SyNBEATS

The figure compares root mean squared error (RMSE) of predictions generated by SyNBEATS and Synthetic Difference-in-Differences under a range of simulated panel data settings. The baseline simulation assumes 30 control units, 30 pre-treatment time periods, and equal importance of fixed effects and interactive components (i.e., $c = 0.5$ in Equation 5). For ease of reference, the points depict the performance under the baseline specification. The panel labeled 'N' varies the number of control units (from 0 to 100); the panel labeled 'T' varies the number of pre-treatment time periods (from 0 to 100); and the panel labeled 'c' varies the relative importance of fixed effects to interactive effects, with higher values indicating greater importance of the fixed effects component of the data generating process. The plotted RMSE corresponds to the average per-period RMSE over 4 post-treatment time periods from 100 draws of a data generating process with the specified parameters.

reverses) for shorter pre-treatment periods and fewer controls. This is consistent with the assumption that the greater flexibility in the functional form of SyNBEATS requires more training data to generate reliable estimates. Further, the relative advantage of SyNBEATS decreases as the relative importance of fixed effects increases, consistent with the functional form assumptions of SDID being well-suited for that scenario.

Overall, our results suggest that SyNBEATS appears more accurate if researchers have sufficient training data. However, if training data is sparse, or if the data generating process is known to be dominated by unit fixed effects, SDID may yield lower prediction errors.

5 Towards a Simpler Model

In this section we aim to identify whether the gains to SyNBEATS arise from the model architecture itself or the usage of the data. If we use the same data pattern of lagged outcomes and contemporaneous control units but instead predict using a simple linear model, how well can we do?

- OLS results
  - Comment that it's surprising as its overparametrized, and show biggest beta norm
    - Maybe some discussion about overparametrization and neural nets and how for neural nets its usually fine?
  - Could just be noisy
  - Connect it to SC and the constraints, and connect that a bit to Doudchenko and Imbens (2016) to set the context of relaxing constraints
  - Connect to Doudchenko and Imbens (2016) more by running tests with constraints/regularization.
- Compare to off-the-shelf NN/RF models
- Compare also to XGboost

6 Conclusion

This paper introduced the SyNBEATS algorithm as a tool for causal inference in panel data settings. In the applications we considered, SyNBEATS consistently yielded lower prediction errors compared to commonly employed estimators (SC and TWFE), as well as

compared to the more recently developed MC method. SyNBEATS also yielded comparable or stronger performance than SDID. Simulations suggest that SDID may outperform SyNBEATS when training data is sparse or when the data generating process is well described by a relatively simple (i.e., non-interactive) fixed effects model. Overall, compared to alternative estimators, SyNBEATS may yield more accurate estimates of the causal effect of policies of interest in a range of realistic panel-data settings.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 2010, *105* (490), 493–505.

_ , _ , and _ , "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 2015, *59* (2), 495–510.

_ and Javier Gardeazabal, "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, March 2003, *93* (1), 113–132.

Angrist, Joshua D. and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, December 2008.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager, "Synthetic Difference-in-Differences," *American Economic Review*, December 2021, *111* (12), 4088–4118.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi, "Matrix Completion Methods for Causal Panel Data Models," *Journal of the American Statistical Association*, 2021, *116* (536), 1716–1730.

Bai, Jushan, "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 2003, *71* (1), 135–171.

_ , "Panel Data Models With Interactive Fixed Effects," *Econometrica*, 2009, *77* (4), 1229–1279.

_ and Serena Ng, "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 2002, *70* (1), 191–221.

Baker, A. and J.B. Gelbach, *Machine Learning and Predicted Returns for Event Studies in Securities Litigation: Preliminary and Incomplete* Rock Center for Corporate Governance at Stanford University working paper series, Rock Center for Corporate Governance, Stanford University, 2020.

Candes, Emmanuel J. and Terence Tao, "The Power of Convex Relaxation: Near-Optimal Matrix Completion," *IEEE Transactions on Information Theory*, 2010, *56* (5), 2053–2080.

Cavallo, Eduardo, Sebastian Galiani, Ilan Noy, and Juan Pantano, "Catastrophic Natural Disasters and Economic Growth," *The Review of Economics and Statistics*, 12 2013, *95* (5), 1549–1561.

Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu, "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," *Journal of the American Statistical Association*, 2021, *116* (536), 1849–1864.

Doudchenko, Nikolay and Guido W Imbens, "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis," Working Paper 22791, National Bureau of Economic Research October 2016.

Holland, Paul W., "Statistics and Causal Inference," *Journal of the American Statistical Association*, 1986, *81* (396), 945–960.

Imbens, Guido W and Jeffrey M Wooldridge, "Recent developments in the econometrics of program evaluation," *Journal of economic literature*, 2009, *47* (1), 5–86.

Mühlbach, Nicolaj and Mikkel Slot Nielsen, "Tree-based synthetic control methods: Consequences of relocating the US embassy," 2021.

Olivares, Kin G., Cristian Challu, Grzegorz Marcjasz, Rafal Weron, and Artur Dubrawski, "Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx," *CoRR*, 2021, *abs/2104.05522.*

Oreshkin, Boris N., Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," *CoRR*, 2019, *abs/1905.10437.*

Poulos, Jason and Shuxi Zeng, "RNN-based counterfactual prediction, with an application to homestead policy and public schooling," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2021, *70* (4), 1124–1139.

Shaikh, Azeem and Panos Toulis, "Randomization Tests in Observational Studies with Staggered Adoption of Treatment," 2021.

Xu, Yiqing, "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, 2017, *25* (1), 57–76.

Appendix

Traditional estimators exploit the structure of the data either via a "vertical" (Abadie and Gardeazabal, 2003; Abadie et al., 2010) or "horizontal" (Imbens and Wooldridge, 2009)regression, as shown in panels B and C of Figure 1. The blue regions in each image indicate which subsets of the data are being used to find the respective weights for each method. Our method, SyNBEATS, aims to exploit both patterns simultaneously by casting this imputation problem as a supervised learning task, shown in Figure A.1. Note that the number of lags is a hyperparameter.

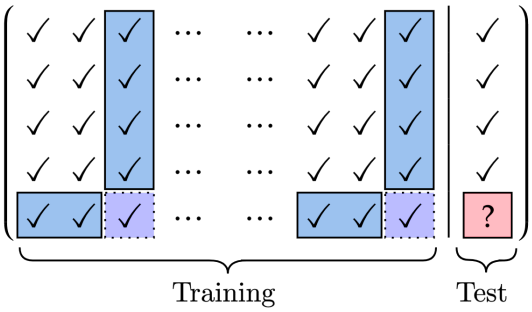


Figure A.1: SyNBEATS on $\boldsymbol{Y(0)}$

Our model casts the dataset imputation as a supervised learning problem. In particular we mask historical outcomes (purple) and learn to predict them with lagged outcomes and contemporaneous controls (blue). We then apply this model out of sample to the true missing data (red).

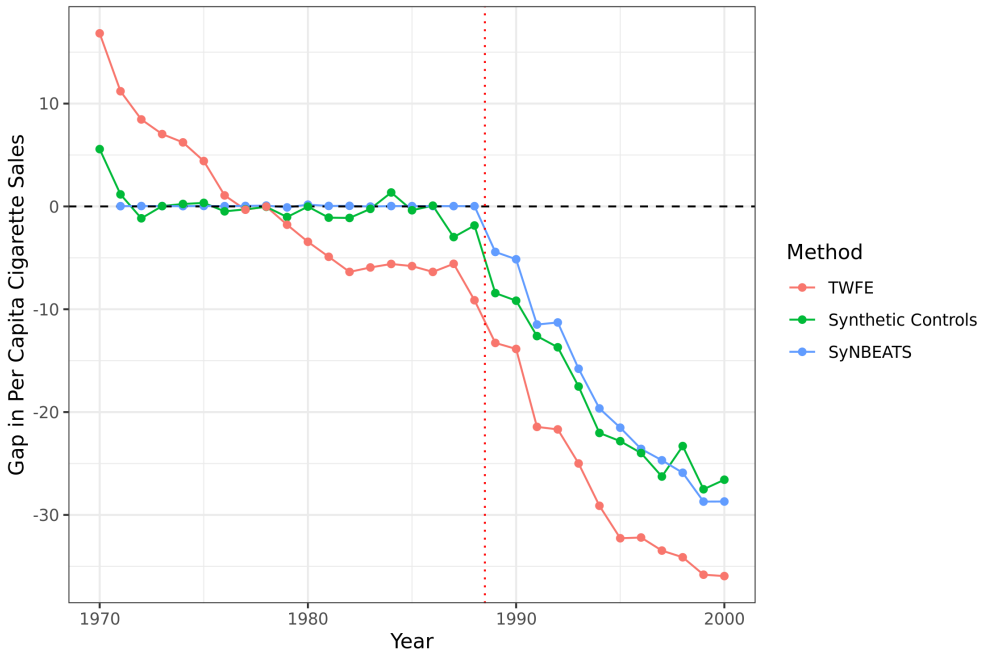
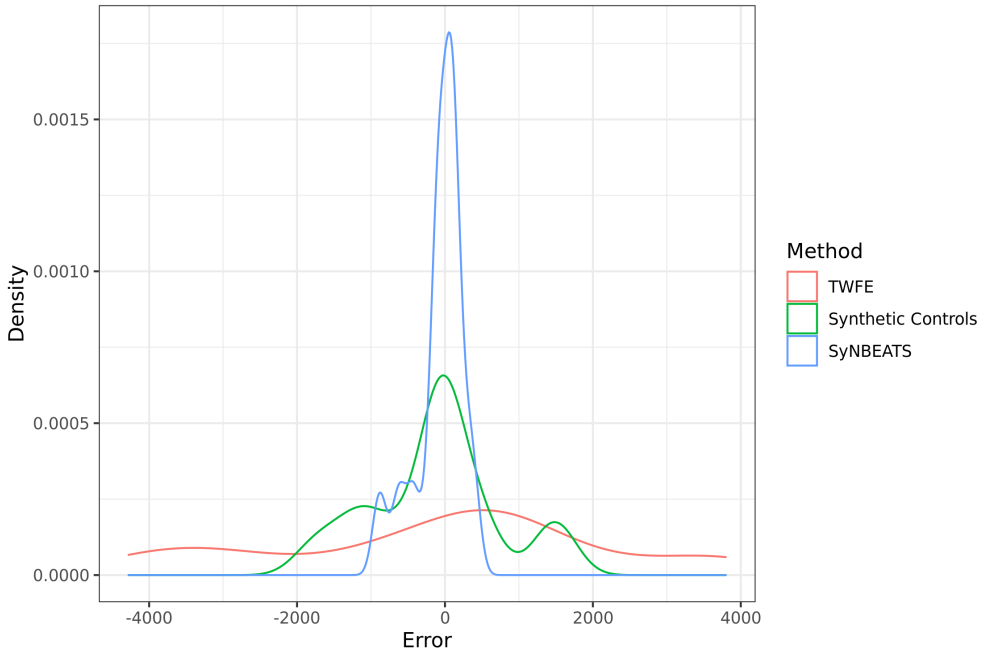
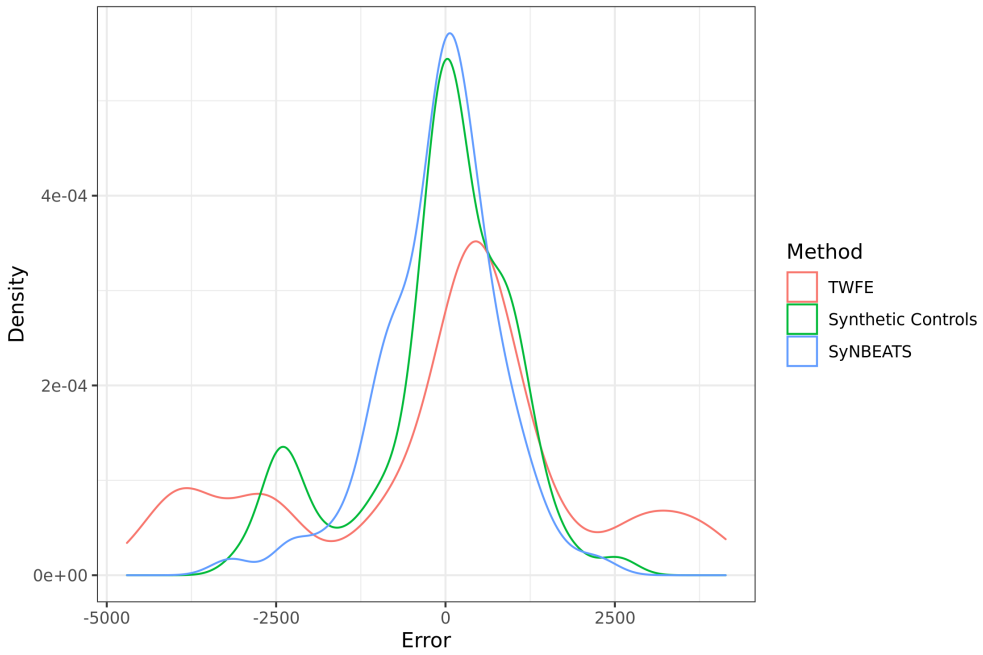


Figure A.2: Effect of Prop 99: Predicted Versus Observed Cigarette Sales

This figure compares the estimated effect of Proposition 99 on cigarette sales from 1970 to 2000 across the different estimators. The predictions are formed using data from California in 1970-1988 and for the control states from 1970-2000. The red dashed line represent the treatment year.



(a) 1-Year Predictions



(b) 5-Year Predictions

Figure A.3: German Reunification Prediction Errors from Placebo Treated Countries

The figure shows the distribution of prediction errors for pseudo-treated states, by prediction method, for the German reunification data set. In Panel A, the outcome being predicted is the pseudo-treated country's GDP in 1990 ($N = 16$); the model is trained on data from the pseudo-treated state from 1960-1989 and from the other control states for 1990. In Panel B, the outcome being predicted is the pseudo-treated country's GDP in years 1986-1990 ($N = 80$); the model is trained on data from the pseudo-treated country from 1960-1985 and from the other control countries for the year being predicted.

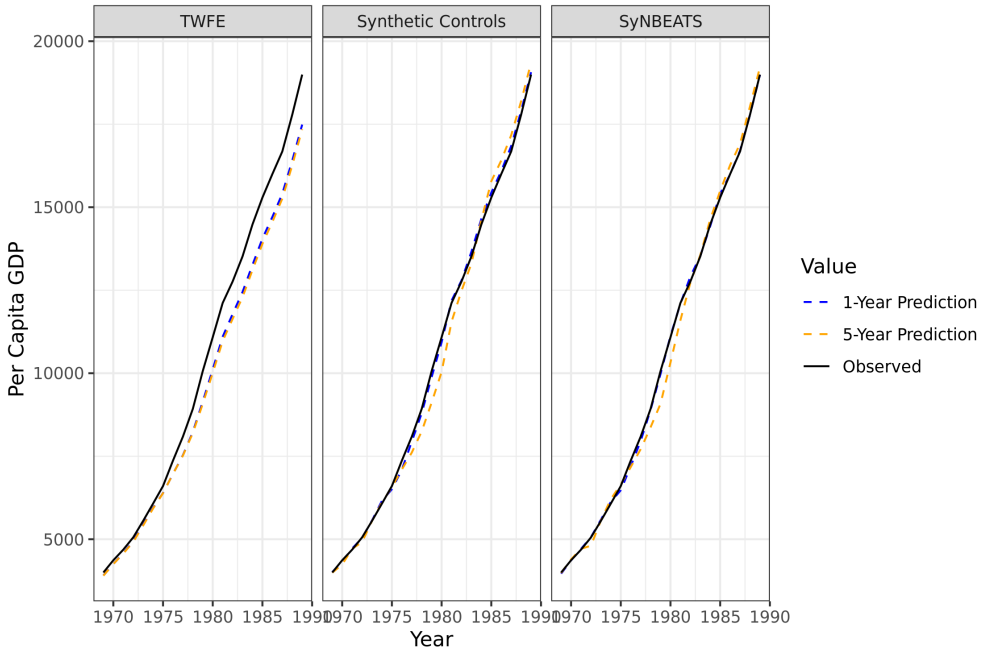


Figure A.4: German Reunification Prediction Errors from Placebo Treated Years

The figure compares short-term (year 1 post-treatment) and long-term (year 5 post-treatment) prediction errors obtained from predicting the GDP in West Germany in pseudo-treated years using various estimators. The blue, dashed line depicts the prediction error in the first pseudo-treated year. The orange, dashed line depicts the predictions in the fifth pseudo-treated year. The predictions are formed using data from West Germany from 1960 through the year prior to the first pseudo-treated year, and from the control countries using data from the pseudo-treated year.

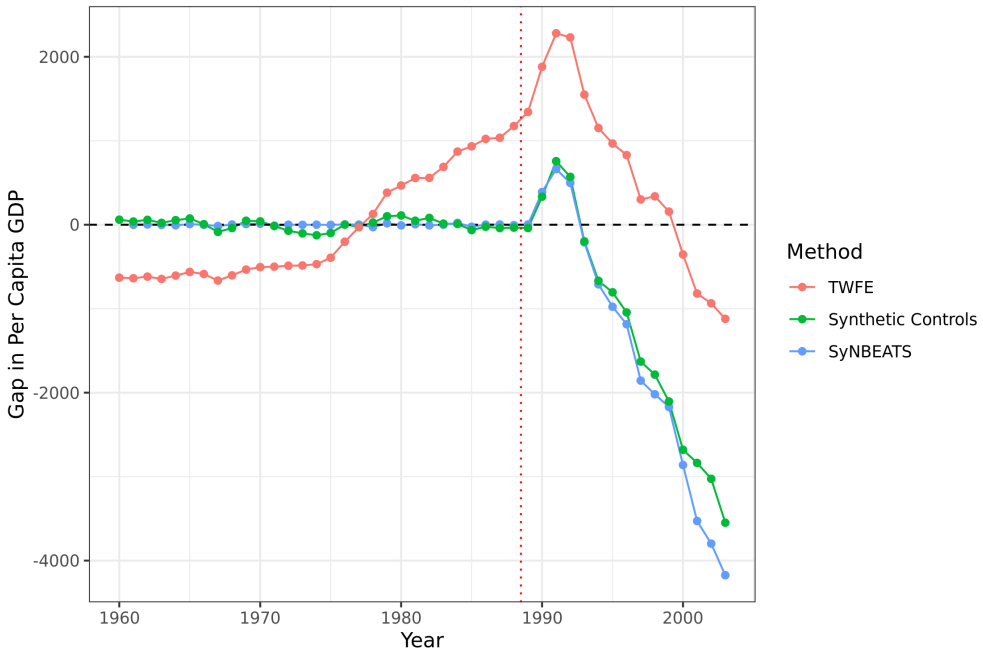


Figure A.5: Effect of German Reunification: Predicted Versus Observed GDP

This figure compares the estimated effect of the German reunification on the GDP of West Germany from 1990 to 2003 across the different estimators. The predictions are formed using data from West Germany in 1962–1989 and for the control states from 1962–2002. The red dashed line represent the treatment year.

Table A.1: German Reunification Analysis with Placebo Treated Countries

Method	RMSE	MAPE	Best
	1 Year Predictions		
TWFE	$2,254.893$	11.498	0
Synthetic Controls	878.390	3.884	0.312
SyNBEATS	**325.792**	**1.423**	**0.688**
	5 Year Predictions		
TWFE	$2,103.349$	11.713	0.188
Synthetic Controls	$1,122.949$	6.049	**0.438**
SyNBEATS	**876.349**	**4.330**	0.375

Notes: The table summarizes performance of the estimators we consider at predicting the 1990 GDP for each pseudo-treated country, using the German reunification data set. Each model is trained on data from the pseudo-treated country from 1960–1989 and from the other control states for the year(s) being predicted. In Panel A, root mean-squared error (RMSE) and mean absolute percentage error (MAPE) are calculated from the distribution of prediction errors across pseudo-treated countries for the year following treatment (1990). In Panel B, RMSE and MAPE are calculated from the distribution of average annual prediction errors across pseudo-treated countries for the five-year period following treatment (1986-1990). The column "Best" reports the share of pseudo-treated countries for which the specified estimation method yields the lowest prediction error over the specified time horizon (i.e., either one year or average over the five-year period).

.

Table A.2: German Reunification Analysis with Placebo Treated Years

Method	RMSE	MAPE	Best
	1 Year Predictions		
TWFE	797.574	5.159	0.120
Synthetic Controls	115.919	1.180	0.160
SyNBEATS	**70.747**	**0.840**	**0.720**
	5 Year Predictions		
TWFE	824.017	5.679	0.095
Synthetic Controls	305.509	2.258	0.190
SyNBEATS	**214.964**	**1.686**	**0.714**

Notes: The table summarizes performance of the estimators we consider at predicting the GDP in West Germany for each pseudo-treated year, using the German reunification data set. Each model is trained on data from West Germany from 1960 through the year prior to the first pseudo-treated year, and from the control countries using data from the pseudo-treated year. In Panel A, root mean-squared error (RMSE) and mean absolute percentage error (MAPE) are calculated from the distribution of prediction errors across pseudo-treated years (1963–1989). In Panel B, RMSE and MAPE are calculated from the distribution of average annual prediction errors across for each five-year pseudo-treatment period (i.e., 1963–1967 through 1985–1989). The column "Best" reports the share of pseudo-treated years for which the specified estimation method yields the lowest prediction error over the specified time horizon (i.e., either one year or average over the five-year period).

Table A.3: Comparison of Modern Panel Data Estimators

Method	Prop 99				German Reunification				Stocks
	Units		Years		Units		Years		
	Short	Long	Short	Long	Short	Long	Short	Long	
Matrix Completion	7.173	13.446	4.422	9.083	905.532	1,305.247	207.155	670.672	0.0279
SDID	3.742	8.745	**1.769**	4.747	382.673	**656.548**	71.623	219.838	0.0283
N-BEATS	5.272	12.788	3.098	9.585	367.328	1,467.527	237.523	944.738	0.0308
SyNBEATS	**3.591**	**8.176**	1.882	**4.088**	**325.792**	876.349	**70.747**	**214.964**	**0.0275**

Notes: The table compares the performance of different modern estimators based on the RMSE. The N-BEATS estimator differs from the baseline SyNBEATS algorithm in that its predictions for the treated unit are based entirely on prior observations from the treated unit, and do not include values of the control units.

Table A.4: Comparison of SyNBEATS to N-BEATS Without Covariates

Method	Prop 99				German Reunification				Stocks
	Units		Years		Units		Years		
	Short	Long	Short	Long	Short	Long	Short	Long	
N-BEATS (No Covariates)	5.272	12.788	3.098	9.585	367.328	1,467.527	237.523	944.738	0.0308
SyNBEATS	**3.591**	**8.176**	**1.882**	**4.088**	**325.792**	**876.349**	**70.747**	**214.964**	**0.0275**

Notes: The table replicates the analyses reported in Tables 1-3 and Appendix Tables 1 and 2 for the comparison of SyNBEATS with the N-BEATS estimator. The N-BEATS estimator differs from the baseline SyNBEATS algorithm in that its predictions for the treated unit are based entirely on prior observations from the treated unit, and do not include values of the control units.