RNN-Based Counterfactual Prediction

Jason Poulos

University of California, Berkeley

Abstract

This paper proposes an alternative to the synthetic control method (SCM) for estimating the effect of a policy intervention on an outcome over time. Recurrent neural networks (RNNs) are used to predict the counterfactual outcomes of treated units using only the pre-intervention outcomes of control units as predictors. The proposed approach is less susceptible to p-hacking because it not require the researcher to specify predictors or use pre-intervention covariates to construct the control group. RNNs do not assume a functional distribution, can learn nonconvex combinations of control units, and are specifically structured to exploit temporal dependencies in panel data. In placebo tests, the RNN-based approach outperforms the SCM and a matrix completion estimators in high-dimensional data settings where the number of time periods exceeds the size of the predictor set.

Keywords: Counterfactual Prediction; Encoder-Decoder Networks; Recurrent Neural Networks; Synthetic Controls; Variational Autoencoder

---

PhD Candidate, Department of Political Science, 210 Barrows Hall #1950, Berkeley, CA 94720-1950. *Email:* poulos@berkeley.edu. *Telephone:* +1-510-642-6323. I acknowledge support of the National Science Foundation Graduate Research Fellowship (DGE 1106400). This work used the computer resources of Stampede2 at the Texas Advanced Computing Center (TACC) under an Extreme Science and Engineering Discovery Environment (XSEDE) startup allocation (TG-SES180010). The Titan Xp GPU used for this research was donated by the NVIDIA Corporation.

1 Introduction

An important problem in the social sciences is estimating the effect of a discrete intervention on a continuous outcome over time. When interventions take place at an aggregate level (e.g., a state), researchers make causal inferences by comparing the post-intervention ("post-period") outcomes of affected ("treated") units against the outcomes of unaffected ("control") units. A common approach to the problem is the synthetic control method (SCM) (Abadie et al., 2010), which predicts the counterfactual outcomes of treated units by finding a convex combination of control units that match the treated units in term of lagged outcomes or pre-intervention ("pre-period") covariates.

The SCM has several limitations. First, the convexity restriction of the synthetic control estimator precludes dynamic, nonlinear interactions between multiple control units. Intuitively, one can expect that the treated unit may exhibit nonlinear or negative correlations with the control units. Ferman and Pinto (2016) demonstrate that the convexity restriction implies that the SCM estimator may be biased even if selection into treatment is only correlated with time-invariant unobserved covariates. Second, Ferman and Pinto (2018) demonstrate that the SCM is generally biased if treatment assignment is correlated with unobserved confounders, even when the number of pre-period periods grows. Moreover, the authors show that while the SCM minimizes imbalance in pre-period outcomes, the likelihood of finding exact balancing weights vanishes as the number of time periods increase, which results in bias.

While the strength of the SCM lies in its simplicity in setup and implementation, several problems arise from the lack of guidance on how to specify the SCM estimator. The specification of the estimator can produce very different results: Ferman et al. (2018) show, for example, how cherry-picking between common SCM specifications can facilitate p-hacking. Kaul et al. (2015) show that the common practice of including lagged outcomes as model inputs can render all other covariates irrelevant. Lastly, Klößner et al. (2017) demonstrates that the common practice of using cross-validation to select importance weights can yield

multiple values and consequently different results.

This paper proposes an alternative to the SCM that is capable of automatically selecting appropriate control units at each time-step, allows for nonconvex combinations of control units, and does not rely on pre-period covariates. The method uses recurrent neural networks (RNNs) to predict the counterfactual outcomes of treated units using only control unit outcomes as model inputs. RNNs are a class of neural networks that take advantage of the sequential nature of temporal data by sharing model parameters across multiple time-steps (El Hihi and Bengio, 1995). RNNs are nonparametric in that they do not assume a functional form when fitting the data. In addition, RNNs can learn the most useful non-convex combination of control unit outcomes at each time-step for generating counterfactual predictions. Relaxing the convexity restriction is useful when the data-generating process underlying the outcome of interest depends nonlinearly on the history of its inputs. RNNs have been shown to outperform various linear models on time-series prediction tasks (Cinar et al., 2017).

RNNs are end-to-end trainable and very flexible to a given sequential prediction problem. For example, they are capable of sharing learned parameters across time-steps and multiple treated units. while the SCM can be generalized to handle multiple treated units (e.g., Dube and Zipperer, 2015; Xu, 2017), the generalized the SCM is not capable of sharing model weights when predicting the outcomes of multiple treated units. Regularization methods such as dropout can easily be incorporated into RNN architectures to prevent overfitting during the training process, which is problematic when the networks learn an overreliance on a few model inputs.

The proposed method builds on a new literature that uses machine learning methods for data-driven counterfactual prediction, such as matrix completion (Athey et al., 2017; Poulos, 2019), or two-stage estimators that reduce data dimensionality via L1-regularized regression (Doudchenko and Imbens, 2016; Carvalho et al., 2018) or matrix factorization (Amjad et al., 2018) prior to regressing the outcomes on the reduced data. These methods

are data-driven in the sense that they are capable of finding an appropriate subset of control units for comparison in the absence of domain knowledge or pre-period covariates.

In the section immediately below, I describe the problem of counterfactual prediction and its relationship to matrix completion and the problem of covariate shift; Section 3 introduces the approach of using RNNs for counterfactual prediction; Section 4 presents the results of the placebo tests; Section 5 details the procedure for hypothesis testing and applies the RNN-based method and inferential procedure ot the problem of estimating the impact of homestead policy on long-run state government investment in public schooling; Section 6 concludes and offers potential avenues for future research.

2 Counterfactual prediction

The proposed method estimates the causal effect of a discrete intervention in observational panel data; i.e., settings in which treatment is not randomly assigned and there exists both pre- and post-period observations of the outcome of interest. Let $\boldsymbol{Y}$ denote a $\mathrm{N} \times \mathrm{T}$ matrix of outcomes for each unit $i = 1, \ldots, \mathrm{N}$, at time $t = 1, \ldots, \mathrm{T}$. $\boldsymbol{Y}$ is incomplete because we observe each element Y_{it} for only the control units and the treated units prior to time of initial treatment exposure, $\mathrm{T}_0 < \mathrm{T}$. Let $\mathcal{O}$ denote the set of (it) values that are observed and $\mathcal{M}$ the set of (it) missing values. Let the values of the $\mathrm{N} \times \mathrm{T}$ complete matrix $\boldsymbol{W}$ be $W_{it} = 1$ if $(it) \in \mathcal{M}$ and $W_{it} = 0$ if $(it) \in \mathcal{O}$.[1] The pattern of missing data is assumed throughout this paper to follow a simultaneous treatment adoption setting, where treated units are exposed to treatment at time T_0 and every subsequent period.

This setup is motivated by the Neyman (1923) potential outcomes framework, where for each it value there exists a pair of potential outcomes, $Y_{it}(1)$ and $Y_{it}(0)$, representing the response to treated and control regimes, respectively. The observed outcomes are

[1]Note that the process that generates W_{it} is referred to the treatment assignment mechanism in the causal inference literature (Imbens and Rubin, 2015) and the missing data mechanism in missing data analysis (Little and Rubin, 2014).

$$Y_{it} = \begin{cases} Y_{it}(0) & \text{if } W_{it} = 0 \text{ or } t < \mathrm{T}_0 \\ Y_{it}(1) & \text{if } W_{it} = 1 \text{ and } t \geq \mathrm{T}_0. \end{cases} \tag{1}$$

The problem of counterfactual prediction is that we cannot directly observe the missing potential outcomes and instead wish to impute the missing values in $\boldsymbol{Y}(0)$ for treated units with $W_{it} = 1$. The potential outcomes framework explicitly assumes unconfoundedness. In an observational setting, this assumption requires $\big(\boldsymbol{Y}(0), \boldsymbol{Y}(1)\big) \perp\!\!\!\perp \boldsymbol{W} | \boldsymbol{Y}(\mathcal{O})$, where $\boldsymbol{Y}(\mathcal{O})$ is the observed data.

The potential outcomes framework also implicitly assumes treatment is well-defined to ensure that each unit has the same number of potential outcomes (Imbens and Rubin, 2015). It also excludes interference between units, which would undermine the framework by creating more than two potential outcomes per unit, depending on the treatment status of other units (Rubin, 1990).

2.1 Relationship to matrix completion and covariate shift

The intuition behind the proposed approach to counterfactual prediction is similar to that of the method of matrix completion via nuclear norm minimization (MC-NNM) proposed by Athey et al. (2017). Matrix completion methods attempt to impute missing entries in a low-rank matrix by solving a convex optimization problem via NNM, even when relatively few values are observed in $\boldsymbol{Y}$ (Candès and Recht, 2009; Candès and Plan, 2010). The estimator recovers a $\mathrm{N} \times \mathrm{T}$ low-rank matrix by minimizing the sum of squared errors via nuclear norm regularized least squares. The estimator reconstructs the matrix by iteratively replacing missing values with those recovered from a singular value decomposition (Mazumder et al., 2010).

Athey et al. (2017) note two drawbacks of MC-NNM. First, the errors may be autocorrelated because the estimator does not account for temporal dependencies in the observed data. The estimator estimate patterns row- and column-wise, but treat the data as perfectly synchronized (Yoon et al., 2018). In contrast, the SCM assumes that correlations across units are stable over time, while the RNN-based approach exploits the temporal component of the data and therefore does not have the problem of autocorrelated errors.

Second, the MC-NNM estimator penalizes the errors for each observed value equally without regard to the fact that the probability of missingness (i.e, the propensity score), increases with t. Athey et al. (2017) suggest weighting the loss function by the propensity score, which is similar to the importance weighting scheme proposed by Cortes et al. (2008) to address the problem of covariate shift, which is a special case of domain adaptation (Huang et al., 2007; Ben-David et al., 2007; Bickel et al., 2009; Cortes et al., 2010; Ganin et al., 2015).[2]

The covariate shift problem occurs when training and test data are drawn from different distributions. For notational ease, define the training set input-output pair as

$$\left(\boldsymbol{X}^{\text{train}}, \boldsymbol{Y}^{\text{train}}\right) = \left(\boldsymbol{Y}(\boldsymbol{W})^{(t<\text{T}_0)}, \boldsymbol{Y}(\boldsymbol{W})^{(t\geq\text{T}_0)}\right)$$

for units with $\boldsymbol{W} = 0$ and the test set pair $\left(\boldsymbol{X}^{\text{test}}, \boldsymbol{Y}^{\text{test}}\right)$ for units with $\boldsymbol{W} = 1$. In the proposed approach, the model weights learned on the training set is fit on $\boldsymbol{X}^{\text{test}}$ to predict $\boldsymbol{Y}^{\text{test}}$. The approach therefore assumes similarity between the distributions of $\boldsymbol{X}^{\text{train}}$ and $\boldsymbol{X}^{\text{test}}$. In order to minimize the discrepancy between the training and test set input distributions, I estimate the propensity score $\hat{e}_{it} = \Pr(W_{it} = 1|Z_{it})$, conditional on covariate matrix $\boldsymbol{Z}$ and then weight the training loss by the estimated propensity scores.

[2]Schnabel et al. (2016) first connected the matrix completion problem with causal inference in observational settings in the context of recommender systems under confounding. Johansson et al. (2016) formulates the general problem of counterfactual inference as a covariate shift problem.

2.2 Nonparametric regression

In its most basic form, counterfactual prediction can be represented as a nonparametric regression of the training set outputs on the inputs,

$$\hat{\boldsymbol{Y}}^{\text{train}} = \hat{f}_0\left(\boldsymbol{X}^{\text{train}}\right) + \epsilon^{(t)}, \tag{2}$$

where the noise variables $\epsilon^{(t)}$ are assumed to be i.i.d. standard normal and independent of the observed data. The nonlinear function $\hat{f}_0$ is estimated by minimizing the weighted mean squared error on the training set outputs,

$$\text{WMSE} = \sum \left(\boldsymbol{Y}^{\text{train}} - \hat{\boldsymbol{Y}}^{\text{train}}\right)^2 \cdot \frac{\hat{\boldsymbol{E}}^{\text{train}}}{|\boldsymbol{X}^{\text{train}}|}, \tag{3}$$

where $\hat{\boldsymbol{E}}^{\text{train}}$ is a matrix of estimated propensity scores.

At test time, the estimated function is used to predict $\hat{\boldsymbol{Y}}^{\text{test}} = \hat{f}_0\left(\boldsymbol{X}^{\text{test}}\right)$. The estimated causal effect of the intervention is then

$$\hat{\boldsymbol{\phi}} = \boldsymbol{Y}^{\text{test}} - \hat{\boldsymbol{Y}}^{\text{test}}. \tag{4}$$

The estimated average causal effect of the intervention on treated units is calculated by averaging over the time dimension, resulting in the vector $\bar{\boldsymbol{\phi}}$ of length $\mathrm{T}_\star = \mathrm{T} - \mathrm{T}_0$.

3 RNNs for counteractual prediction

RNNs (Graves, 2012; Goodfellow et al., 2016) consist of an input $\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n_x)}\right)$, an output $\boldsymbol{Y} = \left(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n_y)}\right)$, and a hidden state $\boldsymbol{h}^{(t)}$. In the plain vanilla RNN it is assumed $n_x = n_y = T$; in the encoder-decoder network architecture described below, n_x and n_y can vary in length.

At each t, RNNs input $\boldsymbol{x}^{(t)}$ and pass it to the $\boldsymbol{h}^{(t)}$, which is updated with a function $g^{(t)}$

using the entire history of the input, which is unfolded backwards in time:

$$\begin{aligned} \boldsymbol{h}^{(t)} &= g^{(t)}\left(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t-1)}, \ldots, \boldsymbol{x}^{(1)}\right) \\ &= f_1\left(\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)}; \theta\right). \end{aligned} \tag{5}$$

The activation function $f_1(\cdot)$, parameterized by θ, is shared for all t. Parameter sharing is particularly useful in the current application because it allows for better generalization when the dimension of the training data is relatively small. The updated hidden state (5) is used to generate a sequence of values $\boldsymbol{o}^{(t)}$ in the form of log probabilities corresponding to the output. The loss function computes $\hat{\boldsymbol{y}}^{(t)} = f_2\left(\boldsymbol{o}^{(t)}\right)$ and calculates the loss. The total loss for the input-output pair is the sum of the losses over all t.

The RNNs are trained to estimate the conditional distribution of $\boldsymbol{y}^{(t)}$ given the past inputs and also the previous output. This is accomplished by offsetting the input-output pairs by one time-step so that the networks receive $\boldsymbol{y}^{(1)}$ as input at $t+1$ to be conditioned on for predicting subsequent outputs. This popular training procedure is known as teacher forcing because it forces the networks to stay close to the ground-truth output $\boldsymbol{y}^{(t)}$ (Lamb et al., 2016). Specifically, the RNNs are trained to maximize the log-likelihood

$$\log \Pr\left(\boldsymbol{y}^{(t)} | \boldsymbol{x}^{(1)} \ldots \boldsymbol{x}^{(t)}, \boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(t-1)}\right). \tag{6}$$

3.1 Encoder-decoder networks

Encoder-decoder networks are the standard for neural machine translation (NMT) (Cho et al., 2014; Bahdanau et al., 2014; Vinyals et al., 2014) and are also widely used for predictive tasks, including speech recognition (Chorowski et al., 2015) and time-series forecasting (Zhu and Laptev, 2017).

The encoder RNN reads in $\boldsymbol{x}^{(t)}$ sequentially and the hidden state of the network updates according to (5). The hidden state of the encoder is a context vector $\boldsymbol{c}$ that summarizes the

input sequence, which is copied over to the decoder RNN. The decoder generates a variable-length output sequence by predicting $\boldsymbol{y}^{(t)}$ given the encoder hidden state and the previous element of the output sequence. Thus, the hidden state of the decoder is updated recursively by

$$\boldsymbol{h}^{(t)} = f_1\left(\boldsymbol{h}^{(t-1)}, \boldsymbol{y}^{(t-1)}, \boldsymbol{c}; \theta\right), \tag{7}$$

and the conditional probability of the next element of the sequence is

$$\Pr(\boldsymbol{y}^{(t)}|\boldsymbol{y}^{(t)}, \ldots, \boldsymbol{y}^{(t-1)}, \boldsymbol{c}) = f_1\left(\boldsymbol{h}^{(t)}, \boldsymbol{y}^{(t-1)}, \boldsymbol{c};\, \theta\right). \tag{8}$$

Effectively, the decoder learns to generate outputs $\boldsymbol{y}^{(t)}$ given the previous outputs, conditioned on the input sequence.

3.2 Recurrent variational autoencoder

While the encoder-decoder architecture is effective for many sequential prediction tasks, the model does not learn a vector representation of the entire input. The variational autoencoder (VAE) (Kingma and Welling, 2013) is a generative model that learns a latent variable model for $\boldsymbol{x}^{(t)}$ such that new sequences $\boldsymbol{x'}^{(t)}$ can be generated by sampling from the latent space q. Similar to encoder-decoder networks, the VAE has an encoder that learns a latent representation of the input sequence and a decoder that maps the representation back to the inputs. The VAE architecture differs from encoder-decoder networks in that the VAE doesn't have a final dense layer that compares the decoder outputs to $\boldsymbol{x'}^{(t)}$; i.e., it is a "self-supervised" technique. Another difference is that the VAE learns parameter weights by mapping the inputs to a distribution over parameters of q.

The recurrent VAE (RVAE) (Fabius and van Amersfoort, 2014; Chung et al., 2015; Bowman et al., 2015) consists of an encoder RNN that maps $\boldsymbol{x}^{(t)}$ to a distribution over parameters of q. The model then randomly samples $\boldsymbol{z}$ from the latent distribution, $q(\boldsymbol{z}|\boldsymbol{x}^{(t)}) =$

$q(\boldsymbol{z}; f_3(\boldsymbol{x}^{(t)}; \theta))$. Finally, a decoder RNN takes the form of a conditional probability model $\Pr(\boldsymbol{x}^{(t)}|\boldsymbol{z})$. The parameters of the model are learned by maximizing the loss function, which takes the difference between the log-likelihood between the decoder outputs $\boldsymbol{x'}^{(t)}$ and $\boldsymbol{x}^{(t)}$ and the relative entropy between $q(\boldsymbol{z}|\boldsymbol{x}^{(t)})$ and the model prior $\Pr(\boldsymbol{z})$. The latter component of the loss function acts as regularizer by forcing the learned latent distribution to be similar to the model prior.

4 Placebo tests

I conduct placebo tests on actual datasets in order to benchmark the accuracy of RNN-based estimators. There are no actual treated units in the placebo tests, so the estimators are evaluated on their ability to recover a null effect.

For each trial run, I randomly select half of the units in the dataset to be treated and predict their counterfactual outcomes for periods following a selected T_0. I compare the predicted values to the observed values by calculating the root-mean squared error (RMSE). I benchmark the encoder-decoder networks and RVAE against the following estimators:[3]

(a) DID Regression of $\mathbf{Y}$ on $\mathbf{W}$ and unit and time fixed effects

(b) MC-NNM Matrix completion via nuclear norm minimization, with the regularization term on the nuclear norm selected by cross-validation (Athey et al., 2017)

(c) SCM Approached via exponentiated gradient descent (Abadie et al., 2010)

(d) VT-EN Vertical regression with elastic-net regularization, with the regularization and mixing parameters selected by cross-validation (Zou and Hastie, 2005; Athey et al., 2017).

[3]Implementation details for the encoder-decoder networks and RVAE are provided in Section A. In the placebo tests, the networks are trained using an unweighted MSE loss function for 500 epochs on a 12GB NVIDIA Titan Xp GPU.

4.1 Synthetic control datasets

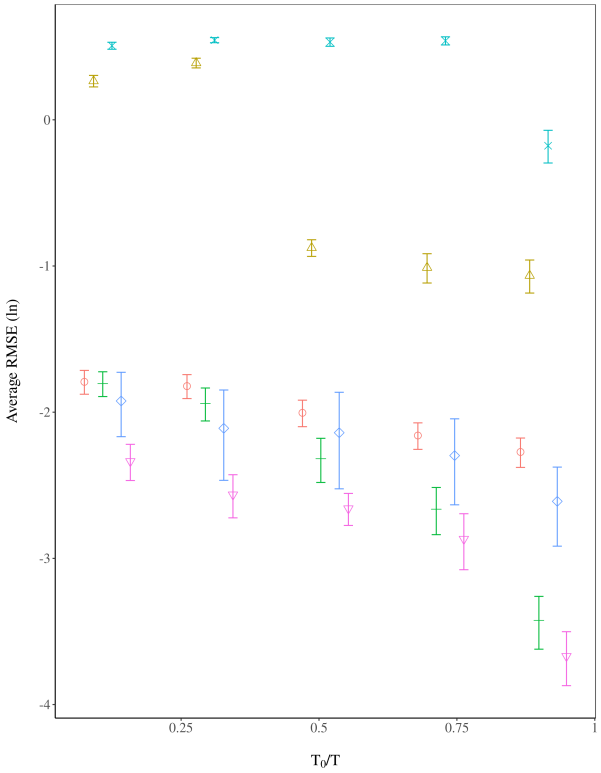
I first conduct placebo tests on three datasets common to the synthetic control literature, with the actual treated unit removed from each dataset: Abadie and Gardeazabal's (2003) study of the economic impact of terrorism in the Basque Country during the late 1960s (N = 16, T = 43); Abadie et al.'s (2010) study of the effects of a large-scale tobacco control program implemented in California in 1988 (N = 38, T = 31); and Abadie et al.'s (2015) study of the economic impact of the 1990 German reunification on West Germany (N = 16, T = 44). Each dataset is log-transformed to alleviate exponential effects.

Figure 1 reports the estimated average prediction error with the estimates jittered horizontally to reduce overlap. Error bars are calculated using the standard deviation of the error distribution generated by multiple runs. The RNN-based estimators yield comparable error rates vis-à-vis the alternatives only for high ratios of T_0/T, which reflect the need for sizeable training sets for the RNN-based approach. The RVAE performs the worse on comparatively small training data since it is learning from less information than the encoder-decoder networks; i.e., without the post-period observations of the control units. The MC-NNM estimator does comparatively well in the simulations due to the fact that it is capable of using additional information in the form of pre-period observations of the treated units, whereas the other estimators train only on the control observations.

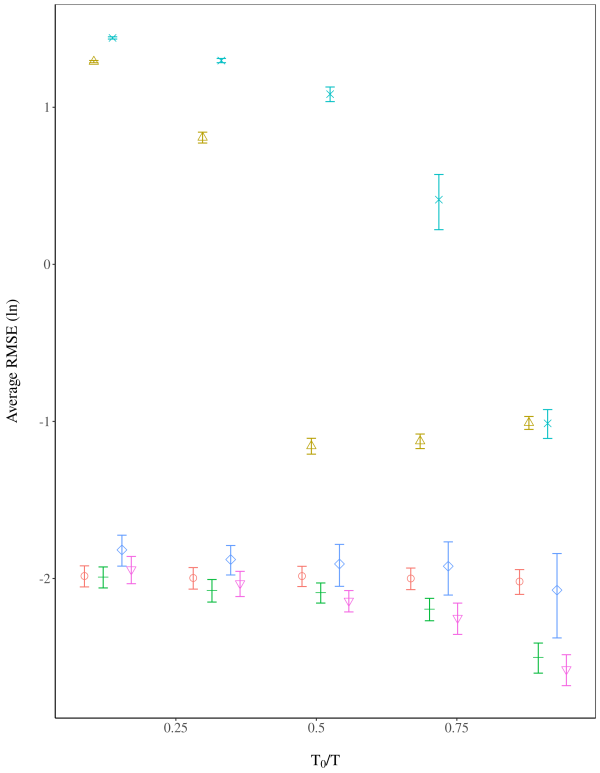
4.2 Stock market data

The second battery of placebo tests draws on a dataset of stock market returns compiled by Athey et al. (2017). The dataset consists of daily returns for 2,453 stocks over 3,082 days. In order to track how the error rates vary according to the dimensionality of the data, I create six sub-samples of the first T daily returns of N randomly selected stocks for the pairs (N, T) = (10, 490), (20, 245), (50, 98), (70, 70), (100, 49), and (140, 35). In each sub-sample, half of the units are randomly selected as treated, and $T_0 = T/2$.

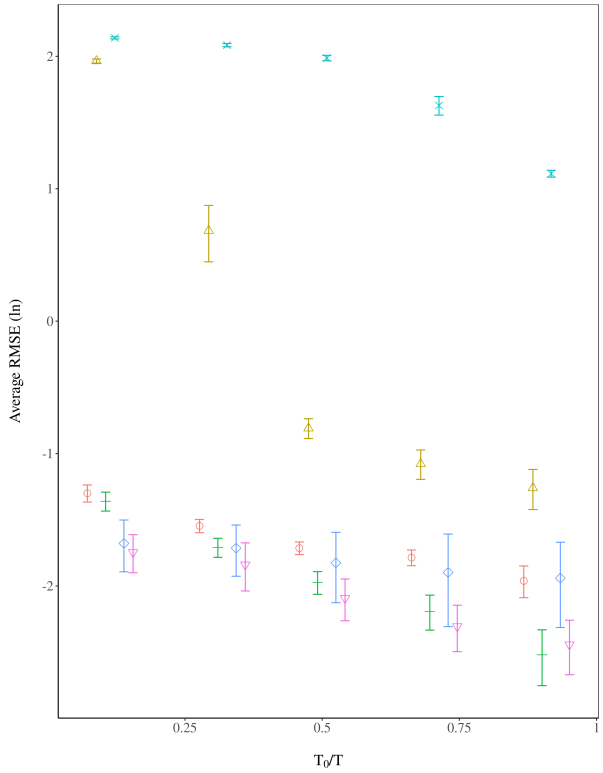
Figure 2 reports the average RMSE for each pair with standard errors informed by the



(A) Basque Country terrorism data



(B) California smoking ban data



(C) West German reunification data

Figure 1: Placebo tests on synthetic control datasets: ○, DID; △, ED; +, MC-NNM; ×, RVAE; ◇, SCM; ▽, VT-EN.

error distribution generated by five trial runs. The average RMSE is the lowest for all estimators in the sub-sample $(\mathrm{N}, \mathrm{T}) = (10, 490)$, which reflects the benefit of training on a large number of time periods. Within this sub-sample, encoder-decoder networks and RVAE achieve the lowest average RMSE, followed by MC-NNM, SCM, DID, and lastly, vertical regression. The RNN-based estimators do comparatively less well when $N \gg T$ since there is not an adequate number of training set pre-periods to learn a concise representation of the inputs.

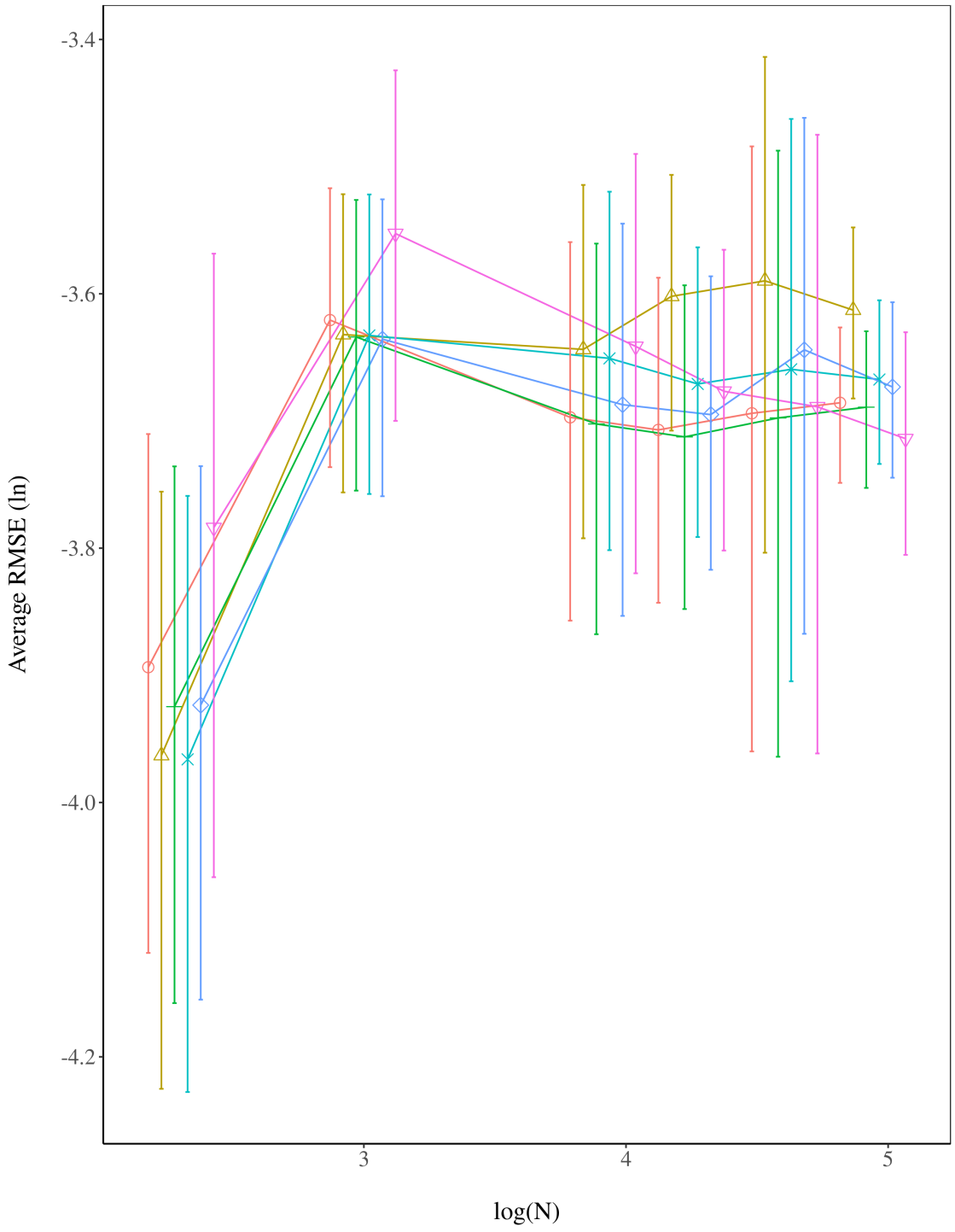


Figure 2: Placebo tests on stock market data: ⊖, DID; △, ED; +, MC-NNM; ×, RVAE; ◇, SCM; ▽, VT-EN.

5 Application: Homestead policy and public schooling

Sociologists and political economists (e.g, Meyer et al., 1979; Alesina et al., 2013; Bandiera et al., 2018) have viewed the rapid development of public schooling in the U.S. during the 19th century as a nation-building policy. It is argued that states across the U.S. adopted compulsory primary education means to homogenize the population during the 'Age of Mass Migration', when of tens of millions of foreign migrants arrived to the country between 1850 and 1914.

An alternative explanation for the rise of public schooling is the view of Engerman and Sokoloff (2005) that frontier state governments sought to increase public investments in order to attract eastern migrants following the passage of the Homestead Act (HSA) of 1862, which opened for settlement hundreds of millions of acres of frontier land. Any adult citizen could apply for a homestead grant of 160 acres of land, provided that they live and make improvements on the land for five years. According to the authors, the sparse population on the frontier meant that state and local governments competed with each other to attract migrants in order to lower local labor costs and to increase land values and tax revenues. Frontier governments offered migrants broad access to cheap land and property rights, unrestricted voting rights, and a more generous provision of schooling and other public goods.

The HSA may have also increased state schooling expenditures by reducing the degree of land inequality on the frontier. Policies that led to the decentralization of public land are expected to lower land inequality by fixing land grants to 160 acres, thereby encouraging farm sizes to approach their ideal scale. Political economy frameworks (e.g., Acemoglu and Robinson, 2008; Besley and Persson, 2009) emphasize that greater economic power of the ruling class reduces public investments. In the model of Galor et al. (2009), wealthy landowners block education reforms because public schooling favors industrial labor productivity and decreases the value in farm rents. Inequality in this context can be thought of as a proxy for the amount of *de facto* political influence elites have to block reforms.

5.1 Data and assumptions

I create a state-level measure of state government education spending from the records of 48 state governments during the period of 1783 to 1932 (Sylla et al., 1993) and the records of 16 state governments during the period of 1933 to 1937 (Sylla et al., 1995a,b). Comparable measures for 48 states are drawn from U.S. Census special reports for the years 1902, 1913, 1932, 1942, 1962, 1972, and 1982 (Haines, 2010).

The data pre-processing steps are as follows. The measure is inflation-adjusted according to the U.S. Consumer Price Index (Williamson, 2017) and scaled by the total free population in the decennial census (Haines, 2010). Missing values are imputed separately in the pre- and -post-periods by carrying the last observation forward and remaining missing values are imputed by carrying the next observation backward. The data are log-transformed to alleviate exponential effects. Lastly, I remove states with no variance in the pre-period outcomes, resulting in a complete matrix of size $(\mathrm{N} \times \mathrm{T}) = (32 \times 156)$.

In this application, public land states — i.e., states crafted from the public domain — serve as treated units (i.e., the test set). State land states, which include states of the original 13 colonies, Maine, Tennessee, Texas, Vermont, and West Virginia, were not directly affected by homestead policies and therefore serve as control units (i.e., the training set). The RNN-based approach assumes the distribution of $\boldsymbol{X}^{\text{train}}$ and $\boldsymbol{X}^{\text{test}}$ are similar.

I weight the training loss by propensity scores in order to minimize distributional discrepancy between the training and test set inputs. The propensity scores are estimated via logistic regression with unit-specific, pre-period covariates including state-level average farm sizes measured in the 1860 and average farm values measured in the 1850 and 1860 censuses (Haines, 2010) to control for homesteaders migrating to more productive land. To control for selection bias arising from differences in access to frontier lands, I create a measure of total miles of operational track per square mile aggregated to the state-level using digitized railroad maps provided by Atack (2013). Fig. A2 shows that the training and test set input

distributions weighted by the propensity scores are visually similar.[4]

Aggregating to the state level approximately 1.46 million individual land patent records authorized under the HSA, I determine that the earliest homestead entries occurred in 1869 in about half of the frontier states, about seven years following the enactment of the 1862 Homestead Act.[5] Using this information, I set $\mathrm{T}_0 = 87$, which leaves $\mathrm{T} - \mathrm{T}_0 = 69$ time periods when half of the states are exposed to treatment. While the approach assumes that treatment adoption is simultaneous across states, the date of initial treatment exposure varied as new frontier land opened between the period of 1869 to 1902. Also note that while the no interference assumption cannot directly be tested, it is likely that state land states were indirectly affected by the out-migration of homesteaders from frontier states.

5.2 Estimates

Prior to analyzing the data, I conduct placebo tests on the education spending data similar to those described in Section 4.1. Figure A1 presents the average RMSE calculated on the control unit outcomes with standard errors originating from 10 runs. In line with the previous placebo tests, the RNN-based estimators yield error rates comparable to the alternative estimators only when there are sufficient pre-period observations to train on; in this case, when $\mathrm{T}_0/\mathrm{T} \geq 0.5$. We can be reasonably confident that the RNN-based estimators will be at least as accurate as the other estimators since $\mathrm{T}_0/\mathrm{T} = 0.55$ in this application.

Next, I train a encoder-decoder network on the training set of state land states and use the learned weights to predict the counterfactual outcomes of public land states. The top panel of Figure 3 compares the average outcomes of treated units and control units along with the average predicted outcomes of treated units. The dashed vertical line represents the first year of treatment exposure in 1869. We are primarily interested in the difference

[4]However, a weighted two-sided t-test rejects the null of equivalence for the difference-in-means between the two distributions ($t = \bar{\boldsymbol{X}}^{\text{train}} - \bar{\boldsymbol{X}}^{\text{test}} = -0.86$; $\sigma_t = 0.07$; $p < 0.01$).

[5]Land patent records provide information on the initial transfer of land titles from the federal government and are made accessible online by the U.S. General Land Office (`https://glorecords.blm.gov`).

in the observed and predicted treated unit outcomes, which is the quantity $\bar{\boldsymbol{\phi}}$. These per-period average causal impacts are plotted in the bottom panel and are bounded by 95% randomization confidence intervals, which are estimated following the procedure described in Section B.

Counterfactual predictions of state government education spending in the absence of the HSA generally tracks the observed control time-series until the turn of the 19$^{\text{th}}$ century, at which the counterfactual flattens and diverges from the increasing observed control time-series. This delay can potentially be explained by the fact that homestead entries did not substantially accumulate until after Congress prohibited the sale of public land in 1889 in all states except Missouri (Gates, 1941, 1979).

Taking the mean of post-period impacts, I estimate that the impact of the HSA on the state government spending of states exposed to homesteads is 0.69 [-0.19, 2.01]. The confidence intervals surrounding this estimate contains zero, which implies that the estimated impact is not significantly more extreme than the exact distribution of average placebo effects under the null hypothesis. Examining the time-specific causal estimates reveals that fifty years after the first homestead entry, the estimated impact of the HSA on state government education spending in 1919 is 0.68 log points [0.13, 1.24]. The confidence intervals surrounding this time-specific estimate does not contain zero, which implies that the estimated impact is significantly more extreme than the average placebo effects. To put the magnitude of the point estimate in perspective, it represents about 3% of the total school expenditures per-capita in 1929.[6]

[6]Data on total school expenditures per-capita of total population are from Snyder and Dillow (2010).

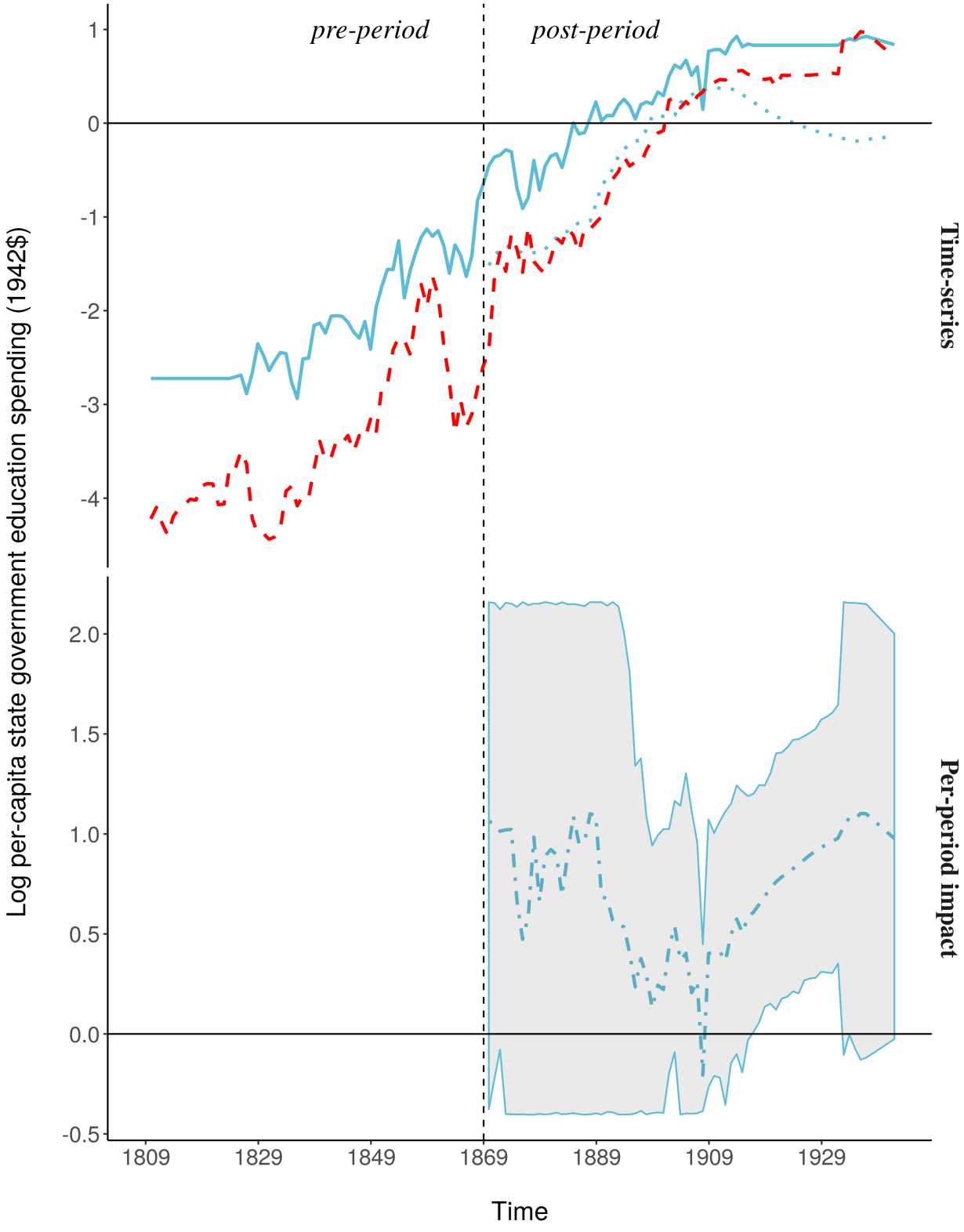


Figure 3: Encoder-decoder estimates of the impact of the HSA on state government education spending, 1809 to 1982: ——, observed treated; - - - -, observed control; ·····, counterfactual treated; —·—, $\bar{\mathbf{\Phi}}$.

6 Conclusion

This paper makes a methodological contribution in proposing a novel alternative to the SCM for estimating the effect of a policy intervention on an outcome over time in settings where appropriate control units are unavailable. The SCM is growing in popularity in the social sciences despite its limitations — the most obvious being that the choice of specification can lead to different results, and thus facilitate p-hacking. By inputting only control unit outcomes and not relying on pre-period covariates, the proposed method offers a more principled approach than the SCM.

The RNN-based approach joins a new generation of data-driven machine learning techniques for generating counterfactual predictions. Machine learning techniques in general have an advantage over the SCM in that they automatically choose appropriate predictors without relying on pretreatment covariates; this capability limits "researcher degrees of freedom" that arises from choices on how to specify the model. RNNs do not assume a specific functional distribution, can learn nonconvex combinations of control units, and are specifically structured to exploit temporal dependencies in the data. RNNs are also capable of handling multiple treated units, which is useful because the model can share parameters across treated units, and thus generate more precise predictions in settings in which treated units share similar data-generating processes.

In placebo tests, RNN-based estimators perform comparatively worse than the alternatives on small dimensional datasets such as those featured in the original synthetic control papers. Both RNN-based estimators require sufficient pre-period observations in order to learn an informative representation of the control units. The RVAE in particular requires a a large amount of training data since it is a self-supervised method that learns without outputs. In higher dimensional datasets such as the stock market data, the RNN-based methods generally outperform the alternatives when $N \ll T$. The estimators underperform when $N \gg T$, which again reflects the need for sufficient pre-period observations.

The matrix completion method performs well in either case, despite of its disadvantage

of treating the data as static and thus ignoring the temporal component of the data. A built-in advantage of the matrix completion approach is that it does not assume a specific structure to the treatment assignment mechanism and thus can accommodate settings in which the time of initial treatment exposure varies across treated units. One potential avenue for future research is to integrate RNNs into the matrix completion approach by training multidirectional RNNs (e.g., Yoon et al., 2018) to both impute missing values across the unit dimension and interpolate missing values within the time dimension.

A second area of future research would explore ways to relax the assumption of equivalence between the distributions of pre-period outcomes between control and treated units. An alternative approach to the one currently proposed is to treat the problem of counterfactual prediction like a NMT problem by training the networks on the pre-period outcomes of control units to predict those of treated units. The learned model weights would then be fit on the post-period outcomes of control units at test time. This setup would instead assume equivalence between the distributions of pre-and post-period outcomes of control units, which is more likely to be satisfied in the absence of interference between treated and control units.

References

Abadie, A., Diamond, A. and 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**, 493–505.

— (2015) Comparative politics and the synthetic control method. *American Journal of Political Science*, **59**, 495–510.

Abadie, A. and Gardeazabal, J. (2003) The economic costs of conflict: A case study of the Basque Country. *The American Economic Review*, **93**, 113–132.

Acemoglu, D. and Robinson, J. A. (2008) Persistence of power, elites, and institutions. *American Economic Review*, **98**, 267–293.

Alesina, A., Giuliano, P. and Reich, B. (2013) Nation-building and education. *Working Paper 18839*, National Bureau of Economic Research. Available at: `http://www.nber.org/papers/w18839`.

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

Atack, J. (2013) On the use of geographic information systems in economic history: The American transportation revolution revisited. *The Journal of Economic History*, **73**, 313–338.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G. and Khosravi, K. (2017) Matrix Completion Methods for Causal Panel Data Models. *ArXiv e-prints*.

Bahdanau, D., Cho, K. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv e-prints*.

Bandiera, O., Mohnen, M., Rasul, I. and Viarengo, M. (2018) Nation-building through compulsory schooling during the age of mass migration. *The Economic Journal*, **129**, 62–109.

Ben-David, S., Blitzer, J., Crammer, K. and Pereira, F. (2007) Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, 137–144.

Besley, T. and Persson, T. (2009) The origins of state capacity: Property rights, taxation and politics. *American Economic Review*, **99**, 1218–1244.

Bickel, S., Brückner, M. and Scheffer, T. (2009) Discriminative learning under covariate shift. *Journal of Machine Learning Research*, **10**, 2137–2155.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R. and Bengio, S. (2015) Generating sentences from a continuous space. *arXiv:1511.06349*.

Candès, E. J. and Plan, Y. (2010) Matrix completion with noise. *Proceedings of the IEEE*, **98**, 925–936.

Candès, E. J. and Recht, B. (2009) Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, **9**, 717.

Carvalho, C., Masini, R. and Medeiros, M. C. (2018) ArCo: An artificial counterfactual approach for high-dimensional panel time-series data. *Journal of Econometrics*, **207**, 352–380.

Cavallo, E., Galiani, S., Noy, I. and Pantano, J. (2013) Catastrophic natural disasters and economic growth. *Review of Economics and Statistics*, **95**, 1549–1561.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014) Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *ArXiv e-prints*.

Chollet, F. et al. (2015) Keras. `https://keras.io`.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y. (2015) Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, 577–585.

Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2014) Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *ArXiv e-prints*.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C. and Bengio, Y. (2015) A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, 2980–2988.

Cinar, Y. G., Mirisaee, H., Goswami, P., Gaussier, E., Aït-Bachir, A. and Strijov, V. (2017) Position-based content attention for time series forecasting with sequence-to-sequence RNNs. In *International Conference on Neural Information Processing*, 533–544. Springer.

Cortes, C., Mansour, Y. and Mohri, M. (2010) Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, 442–450.

Cortes, C., Mohri, M., Riley, M. and Rostamizadeh, A. (2008) Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory*, 38–53. Springer.

Doudchenko, N. and Imbens, G. W. (2016) Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. *ArXiv e-prints*.

Dube, A. and Zipperer, B. (2015) Pooling multiple case studies using synthetic controls: An application to minimum wage policies. IZA Discussion Paper No. 8944. Available at: `http://ftp.iza.org/dp8944.pdf`.

El Hihi, S. and Bengio, Y. (1995) Hierarchical recurrent neural networks for long-term dependencies. In *Neural Information Processing Systems*, vol. 400, 409.

Engerman, S. L. and Sokoloff, K. L. (2005) The evolution of suffrage institutions in the new world. *The Journal of Economic History*, **65**, 891–921.

Fabius, O. and van Amersfoort, J. R. (2014) Variational recurrent auto-encoders. *arXiv:1412.6581*.

Ferman, B. and Pinto, C. (2016) Revisiting the synthetic control estimator. Available at `https://mpra.ub.uni-muenchen.de/81941/`.

— (2018) Synthetic controls with imperfect pretreatment fit. Available at: `https://sites.google.com/site/brunoferman/research`.

Ferman, B., Pinto, C. and Possebom, V. (2018) Cherry picking with synthetic controls. Available at: `https://mpra.ub.uni-muenchen.de/85138/1/MPRA_paper_85138.pdf`.

Firpo, S. and Possebom, V. (2018) Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, **6**.

Galor, O., Moav, O. and Vollrath, D. (2009) Inequality in landownership, the emergence of human-capital promoting institutions, and the great divergence. *The Review of Economic Studies*, **76**, 143–179.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. and Lempitsky, V. (2015) Domain-Adversarial Training of Neural Networks. *arXiv e-prints*, arXiv:1505.07818.

Gates, P. W. (1941) Land policy and tenancy in the prairie states. *The Journal of Economic History*, **1**, 60–82.

— (1979) Federal land policies in the southern public land states. *Agricultural History*, **53**, 206–227.

Glorot, X. and Bengio, Y. (2010) Understanding the difficulty of training deep feedforward neural networks. In *Artificial Intelligence and Statistics*, vol. 9, 249–256.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. Cambridge, MA: MIT press.

Graves, A. (2012) Neural networks. In *Supervised Sequence Labelling with Recurrent Neural Networks*, 15–35. Springer.

Hahn, J. and Shi, R. (2017) Synthetic control and inference. *Econometrics*, **5**, 52.

Haines, M. R. (2010) Historical, Demographic, Economic, and Social Data: The United States, 1790-2002. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-05-21. `doi.org/10.3886/ICPSR02896.v3`.

Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B. and Smola, A. J. (2007) Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, 601–608.

Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Johansson, F., Shalit, U. and Sontag, D. (2016) Learning representations for counterfactual inference. In *International Conference on Machine Learning*, 3020–3029.

Kaul, A., Klößner, S., Pfeifer, G. and Schieler, M. (2015) Synthetic control methods: Never use all pre-intervention outcomes together with covariates. Available at: `http://www.oekonometrie.uni-saarland.de/papers/SCM_Predictors.pdf`.

Kingma, D. P. and Ba, J. (2014) Adam: A Method for Stochastic Optimization. *arXiv e-prints*, arXiv:1412.6980.

Kingma, D. P. and Welling, M. (2013) Auto-encoding variational bayes. *arXiv:1312.6114*.

Klößner, S., Kaul, A., Pfeifer, G. and Schieler, M. (2017) Comparative politics and the synthetic control method revisited: A note on Abadie et al. (2015). *Swiss Journal of Economics and Statistics*.

Lamb, A. M., Goyal, A. G. A. P., Zhang, Y., Zhang, S., Courville, A. C. and Bengio, Y. (2016) Professor forcing: A new algorithm for training recurrent networks. In *Advances In Neural Information Processing Systems*, 4601–4609.

Little, R. J. and Rubin, D. B. (2014) *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley & Sons.

Mazumder, R., Hastie, T. and Tibshirani, R. (2010) Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, **11**, 2287–2322.

Meyer, J. W., Tyack, D., Nagel, J. and Gordon, A. (1979) Public education as nation-building in America: Enrollments and bureaucratization in the American states, 1870-1930. *American Journal of Sociology*, **85**, 591–613.

Neyman, J. (1923) On the application of probability theory to agricultural experiments. *Annals of Agricultural Sciences*, **51**. Reprinted in Splawa-Neyman et al. (1990).

Poulos, J. (2019) State-Building through Public Land Disposal? An Application of Matrix Completion for Counterfactual Prediction. *arXiv e-prints*, arXiv:1903.08028.

Rubin, D. B. (1990) Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, **5**, 472–480.

Schmidhuber, J. and Hochreiter, S. (1997) Long short-term memory. *Neural Computation*, **9**, 1735–1780.

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N. and Joachims, T. (2016) Recommendations as treatments: Debiasing learning and evaluation. *arXiv:1602.05352*.

Snyder, T. D. and Dillow, S. A. (2010) Digest of education statistics, 2009. National Center for Education Statistics. Available at: `https://nces.ed.gov/programs/digest/index.asp`.

Splawa-Neyman, J., Dabrowska, D., Speed, T. et al. (1990) On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, **5**, 465–472.

Sylla, R. E., Legler, J. B. and Wallis, J. (1993) Sources and Uses of Funds in State and Local Governments, 1790-1915: [United States]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2017-05-21. `doi.org/10.3886/ICPSR06304.v1`.

— (1995a) State and Local Government [United States]: Sources and Uses of Funds, Census Statistics, Twentieth Century [Through 1982]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2017-05-21. `doi.org/10.3886/ICPSR06304.v1`.

— (1995b) State and Local Government [United States]: Sources and Uses of Funds, State Financial Statistics, 1933-1937. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2017-05-21. http://doi.org/10.3886/ICPSR06306.v1.

Vinyals, O., Kaiser, L., Koo, T., Petrov, S., Sutskever, I. and Hinton, G. (2014) Grammar as a Foreign Language. *ArXiv e-prints*.

Williamson, S. H. (2017) Seven ways to compute the relative value of a us dollar amount, 1774 to present. `MeasuringWorth.com`. [Online; accessed 01-October-2017].

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

Yoon, J., Zame, W. R. and van der Schaar, M. (2018) Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*.

Zhu, L. and Laptev, N. (2017) Deep and Confident Prediction for Time Series at Uber. *ArXiv e-prints*.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.

Appendix

A Implementation details

The networks are implemented with the `Keras` neural network library (Chollet et al., 2015) in Python on top of a TensorFlow backend. When implementing encoder-decoder networks, the encoder takes the form of a two-layer Long Short-Term Memory (LSTM) network (Schmidhuber and Hochreiter, 1997), each with 128 hidden units, and the decoder is a single-layer Gated Recurrent Unit (GRU) (Chung et al., 2014) also with 128 hidden units. Each recurrent layer uses a linear activation function (f_1) with weights initialized using Xavier initialization (Glorot and Bengio, 2010). The loss function internally computes the predicted outputs as a linear function (f_2) of the log probabilities.

RNN weights are learned with mini-batch gradient descent on the WMSE using `Adam` stochastic optimization with the learning rate set to $5 \cdot 10^{-4}$ (Kingma and Ba, 2014). As a regularization strategy, I apply dropout to the inputs and L2 regularization losses to the network weights. The networks are trained for 1,000 epochs, which takes 10 minutes to run on a laptop CPU. The model is validated on the last 20% of the training set input-out pairs.

The RVAE is implemented similarly, but with the following differences: the encoder takes the form of a single-layer LSTM with 32 hidden units and the decoder is a two-layer LSTM with the number of hidden units equal to 32 and the number of predictors, respectively. The latent space $\boldsymbol{z}$ is implemented as a densely-connected layer with a dimension of 200 units and $f_3(\cdot)$ takes the form of a log-normal distribution. The RVAE is trained with stochastic gradient descent for 5,000 epochs, which takes seven minutes to run on the same CPU.

B Hypothesis testing

Abadie et al. (2010) propose a randomization inference approach for calculating the exact distribution of placebo effects under the sharp null hypothesis of no effect. Cavallo et al. (2013) extends the placebo-based testing approach to the case of multiple (placebo) treated units by constructing a distribution of *average* placebo effects under the null hypothesis.[7]

Randomization p-values are obtained following these steps:

1. Estimate the observed test static $\hat{\boldsymbol{\phi}}$ from (4). Averaging over the time dimension results in a $\mathrm{T}_\star$-length array of observed average treatment effects.

2. Calculate every possible average placebo treated effect μ by randomly sampling without replacement which $\mathrm{J}-1$ control units are assumed to be treated. There are $\mathcal{Q} = \sum_{\mathrm{g}=1}^{\mathrm{J}-1} \binom{\mathrm{J}}{\mathrm{g}}$ possible average placebo effects.[8] The result is a matrix of dimension $\mathcal{Q} \times \mathrm{T}_\star$

3. Sum over the time dimension the number of μ that are greater than or equal to $\hat{\boldsymbol{\phi}}$.

Each element of the vector obtained from Step 3 is divided by $\mathcal{Q}$ to estimate a $\mathrm{T}_\star$-length vector of exact two-sided p values, $\hat{p}$.

B.1 Randomization confidence intervals

Under the assumption that treatment has a constant additive effect Δ, I construct an interval estimate for Δ by inverting the randomization test. Let δ_Δ be the test statistic calculated by subtracting all possible μ by Δ. I derive a two-sided randomization confidence interval by collecting all values of δ_Δ that yield $\hat{p}$ values greater than or equal to significance level $\alpha = 0.05$. I find the endpoints of the confidence interval by randomly sampling 500 values of Δ.

[7] Firpo and Possebom (2018) derive the conditions under which the randomization inference approach is valid from a finite sample perspective and Hahn and Shi (2017) analyze the approach from a repeated sampling perspective.

[8] Since calculating $\mathcal{Q}$ can be computationally burdensome for relatively high values of J, I artificially set $\mathcal{Q} = 10,000$ in cases when $\mathrm{J} > 16$.

C Appendix Figures

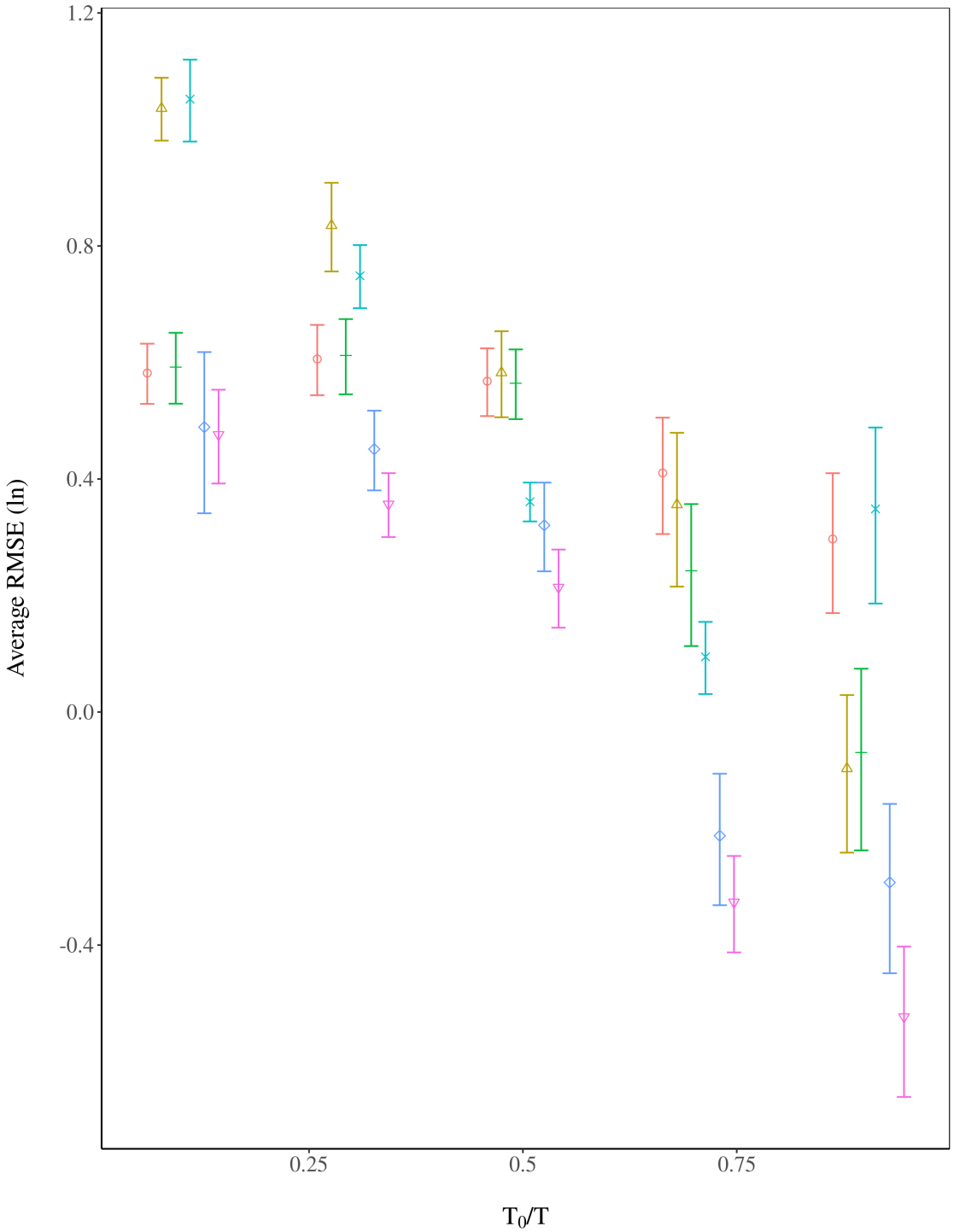
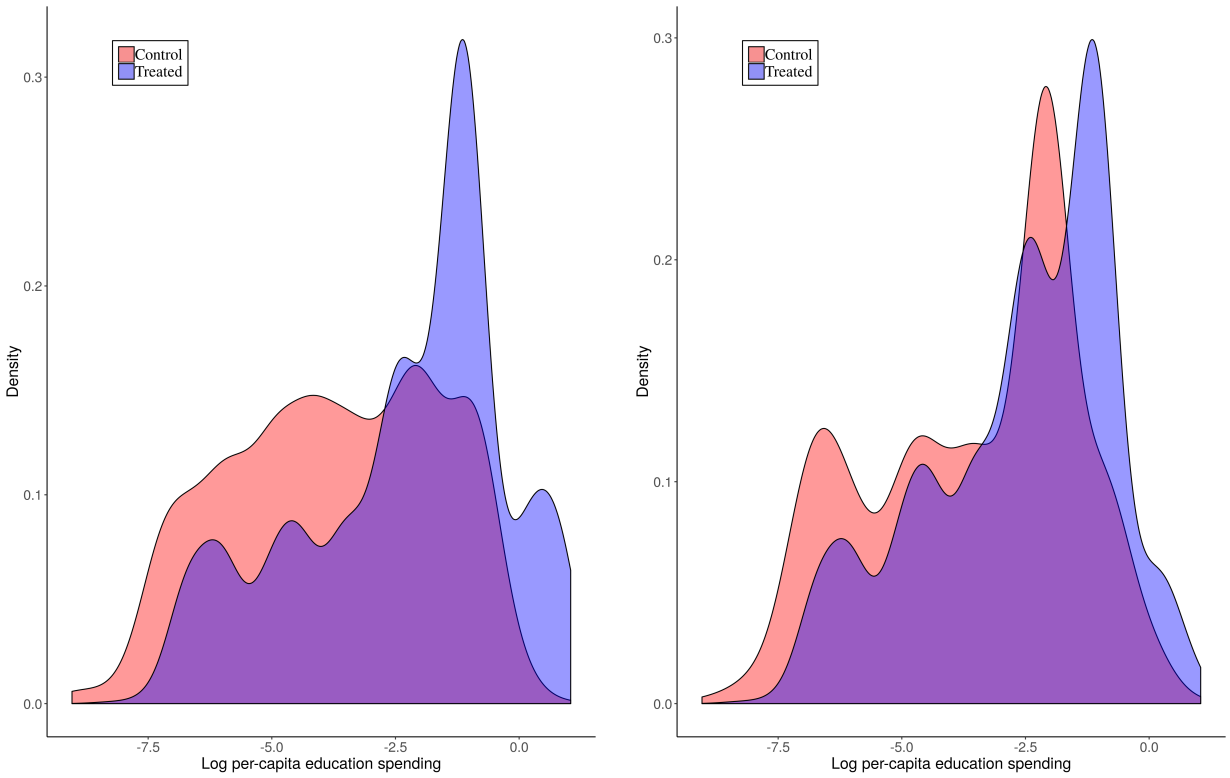


Figure A1: Placebo tests on education spending data: ⊖, DID; △, ED; +, MC-NNM; ×, RVAE; ◇, SCM; ▽, VT-EN.



(A) Unweighted (B) Weighted

Figure A2: Pre-period densities of log per-capita state government education spending by treatment status. Density in Figure A2b weighted by propensity score.