Received: 12 April 2019 | Accepted: 19 May 2021

DOI: 10.1111/rssc.12511

ROYAL STATISTICAL SOCIETY
Series C Applied Statistics

ORIGINAL ARTICLE

RNN-based counterfactual prediction, with an application to homestead policy and public schooling

Jason Poulos[1] | **Shuxi Zeng**[2]

[1]Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts, USA

[2]Department of Statistical Science, Duke University, Durham, North Carolina, USA

Correspondence
Jason Poulos, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA.
Email: poulos@hcp.med.harvard.edu

Funding information
National Science Foundation Graduate Research Fellowship, Grant/Award Number: DGE-1106400; National Science Foundation, Grant/Award Number: DMS-1638521

Abstract

This paper proposes a method for estimating the effect of a policy intervention on an outcome over time. We train recurrent neural networks (RNNs) on the history of control unit outcomes to learn a useful representation for predicting future outcomes. The learned representation of control units is then applied to the treated units for predicting counterfactual outcomes. RNNs are specifically structured to exploit temporal dependencies in panel data and are able to learn negative and non-linear interactions between control unit outcomes. We apply the method to the problem of estimating the long-run impact of US homestead policy on public school spending.

KEYWORDS

counterfactual prediction, panel data, political economy, recurrent neural networks, synthetic controls

1 | INTRODUCTION

An important problem in the social sciences is estimating the effect of a binary treatment on a continuous outcome in a panel data setting. Two prevalent methods for causal inference with panel data are difference-in-differences (DID) and the synthetic control method (SCM). DID uses time-varying panel data to control for time-invariant unobserved confounding, and identifies causal effects by contrasting the change in average outcomes pre- and post-treatment, between treated and control units (e.g. Ashenfelter, 1978). DID assumes no time-varying unobserved confounding that affects both treatment and outcomes, which is highly restrictive and cannot be empirically tested. Moreover, the linear DID estimator assumes i.i.d. errors, which ignores the temporal aspect of the data and understates standard errors for estimated treatment effects when the regression errors are serially correlated,

which can arise when the time-series lengths are not sufficiently long to reliably estimate the data-generating process (Bertrand et al., 2004).

The SCM is a popular method that constructs a convex combination of control units that are similar to a single treated unit in terms of pre-treatment outcomes or covariates (Abadie & Gardeazabal, 2003; Abadie et al., 2010, 2015). The SCM estimator assumes there is a stable convex combination of the control units that absorbs all time-varying unobserved confounding and may be biased even if treatment is only correlated with time-invariant unobserved confounding, which is equivalent to the DID identification assumption (Ferman & Pinto, 2016). The SCM can be generalized to settings with staggered treatment adoption, where the time of initial treatment varies across multiple treated units (Ben-Michael et al., 2019; Dube & Zipperer, 2015) and to include features of DID estimation (Arkhangelsky et al., 2019; Ben-Michael et al., 2018) or Bayesian estimation (Brodersen et al., 2015; Pang et al., 2020).

We propose a method based on recurrent neural networks (RNNs), a class of neural networks that take advantage of the sequential nature of temporal data by sharing model parameters across multiple time periods (Graves, 2012; Hihi & Bengio, 1996). RNNs have been shown to outperform various linear models on time-series prediction tasks (Cinar et al., 2017). Unlike the SCM, RNNs are able to learn negative and non-linear interactions between control unit outcomes, and do not assume a specific activation function when learning representations of the control unit outcomes. RNNs are end-to-end trainable, whereas each component of the Bayesian structural time-series model proposed by Brodersen et al. (2015) must be assembled and estimated independently. RNNs are capable of sharing learned model weights for predicting multiple treated units, and can thus generate more precise predictions in settings where treated units share similar data-generating processes.

We train RNNs on the control unit outcomes data to learn a useful latent representation of outcomes in previous periods for predicting future outcomes. We weight the RNNs loss function by the propensity scores modelled in terms of pre-treatment covariates to ensure the weighted distribution of the observed confounders are balanced between treated and control units. The learned representation of control unit outcomes is then applied to the outcomes data of the treated units for predicting counterfactual outcomes. The causal effect of treatment on the treated units is estimated by contrasting the counterfactual predictions to the observed outcomes of the treated.

The RNN-based method is related to lagged regression models, which regress post-treatment outcomes on pre-treatment outcomes and covariates for control units and then use the model weights to predict the counterfactual outcome for treated units (e.g. Athey et al., 2018; Belloni et al., 2017; Carvalho et al., 2018). A closely related approach are linear factor models, which decompose the pre-treatment outcomes of control units into matrices of latent factors (i.e. time-varying coefficients) and factor loadings (i.e. unit-specific intercepts) and predict counterfactual treated unit outcomes based on the estimated factors and loadings (e.g. Amjad et al., 2018; Athey et al., 2017; Xu, 2017). These models typically use regularization or matrix factorization to reduce the dimensionality of the predictor set and thereby improve generalizability when applying the model fit on control units to treated units. These methods all assume unconfoundedness conditional on previous outcomes for control units.

The proposed method is also related to doubly robust estimators that combine both a propensity score model and an outcome model, which are consistent if either model is properly specified (Bang & Robins, 2005; Chernozhukov et al., 2018). Several studies independent of this work propose using neural networks for counterfactual prediction in non-panel observational data settings. For example, Farrell et al. (2021) provide inference results for semiparametric estimation of causal effects using multilayer perceptrons, while Hartford et al. (2017) and Bennett et al. (2019) integrate deep neural networks into an instrumental variables framework.

In the section immediately below, we state the problem of counterfactual prediction within the potential outcomes framework; Section 3 introduces the approach of using RNNs for counterfactual prediction; Section 4 presents the results of placebo tests; Section 5 applies the method to the problem of estimating the long-run impact of US homestead policy on state government investment in public schooling; Section 6 concludes and offers potential avenues for future research.

2 | POTENTIAL OUTCOMES FRAMEWORK

We explore a panel data setting where we observe a real-valued continuous outcome Y_{it} for each $i = 1, \ldots, N$ units and in each $t = 1, \ldots, T$ time periods, and where a subset of units is exposed to a binary treatment $W_{it} \in \{0, 1\}$ following an initial treatment period T_0. We also observe time-invariant pre-treatment covariates, V_{ip}, where p denotes the number of predictors.

We follow the Neyman–Rubin potential outcomes framework (Neyman, 1923; Rubin, 1974, 1990; Splawa-Neyman et al., 1990), where there exists a pair of potential outcomes, $Y_{it}(1)$ and $Y_{it}(0)$, corresponding to the response to treatment and control, respectively. The potential outcomes framework implicitly assumes treatment is well-defined to ensure that each unit has the same number of potential outcomes. 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. We only observe one of the two potential outcomes for each *it* value, while the other outcome is counterfactual. The observed outcomes are:

$$Y_{it} = \begin{cases} Y_{it}(0) \text{ if } W_i = 0 \text{ or } t < T_0 \\ Y_{it}(1) \text{ if } W_i = 1 \text{ and } t \geq T_0. \end{cases} \tag{1}$$

Define $\tau_{it} = Y_{it}(1) - Y_{it}(0)$ as the individual treatment effect. The causal estimand of interest is the average treatment effect on the treated (ATT):

$$\tau_t^{\text{ATT}} = \mathrm{E}\left[\tau_{it} | W_{it} = 1\right], \quad \text{for } t \in \left\{T_0, \ldots, T\right\}. \tag{2}$$

In the empirical application, we focus on estimating the ATT averaged over the post-treatment period:

$$\tau^{\text{ATT}} = \sum_{t=T_0}^{T} \tau_t^{\text{ATT}} / (T - T_0 + 1). \tag{3}$$

To estimate τ_t^{ATT}, we predict $Y_{it}(0)$ for all *it* values with $W_{it} = 1$; that is, the counterfactual outcome of the treated units had they not been exposed to treatment. The counterfactual predictions are subsequently plugged into the estimator:

$$\widehat{\tau}_t^{\text{ATT}} = \frac{\sum_{it} W_{it} \left(Y_{it}(1) - \widehat{Y}_{it}(0)\right)}{\sum_{it} W_{it}}. \tag{4}$$

The causal estimand τ_t^{ATT} is identified by assuming that treatment and potential outcomes under control are unconfounded conditional on the pre-treatment outcomes and covariates.

Assumption 1 Conditional unconfoundedness:

$$W_{it} \perp\!\!\!\perp Y_{it}(0) \,|\, Y_{i,1}, \ldots, Y_{i,T-1}, V_{ip}.$$

Assumption 1 ensures that treatment assignment affects potential outcomes under control only through covariates and the history of observed outcomes. The idea is that a hypothetical conditional randomization is taking place but, differently from observational studies under the usual strong ignorability assumption (Imbens & Rubin, 2015, Ch. 12), the conditioning set includes the outcome history.

Assumption 2 Overlap:

$$0 < e_{it} < 1, \quad \text{where} \quad e_{it} = \Pr(W_{it} = 1 \,|\, Y_{i,1}, \ldots, Y_{i,T_0-1}, V_{ip}).$$

Assumption 2 is needed to summarize the treatment assignment mechanism by the propensity score, e_{it}. As described in Section 3.1, we use the estimated propensity score to weight the RNNs loss function to correct for imbalances in the distributions of the conditioning set between the treated and control units.

3 | RNNS FOR COUNTERFACTUAL PREDICTION

RNNs operate on n inputs $X = \left(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\right)^\top = \left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\mathrm{T}_x}\right) \in \mathbb{R}^{n \times \mathrm{T}_x}$, where T_x is the input sequence length. The task is to predict outputs $Y = \left(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{\mathrm{T}_y}\right) \in \mathbb{R}^{n \times \mathrm{T}_y}$, where output length T_y can differ from T_x, and $\mathrm{T}_x + \mathrm{T}_y = \mathrm{T}$. Given X, we aim to learn a non-linear mapping $\mathscr{F}(\cdot)$ to predict the next values of the output sequence, $\widehat{y}_{t+1} = \mathscr{F}(X)$. RNNs capture non-linear correlations of the historical values of Y_{it} for $t = 1, \ldots, T-1$ to the future values of Y_{it}. The parameter sharing used in RNNs assumes the same learned model parameters are shared for all t; that is, the estimated non-linear correlations are stationary.

3.1 | Training process

At each t, the RNNs input $\boldsymbol{x}_t$ and pass it to a fixed-length vector $\boldsymbol{h}_t$ called the hidden state, which stores information from the history of inputs up to $\boldsymbol{x}_t$ and the previous hidden state, $\boldsymbol{h}_{t-1}$. Starting with an initial hidden state $\boldsymbol{h}_0$, $\boldsymbol{h}_t$ is updated from $t = 1, \ldots, \mathrm{T}_x$ according to the forward propagation equations:

$$\boldsymbol{a}_t = b + Q\boldsymbol{h}_{t-1} + R\boldsymbol{x}_t \tag{5}$$

$$\boldsymbol{h}_t = f\left(\boldsymbol{a}_t\right) \tag{6}$$

$$\widehat{\boldsymbol{y}}_t = d + U\boldsymbol{h}_t, \tag{7}$$

where Q, R, and U are weight matrices, and b and d are constants. The constants are initialized at zero and $\boldsymbol{h}_0$ is initialized by drawing values from a uniform distribution (Glorot & Bengio, 2010). The hidden state activation function in Equation (6) is a non-linear function such as the hyperbolic tangent (tanh), which is a shifted and scaled version of the logistic function that is commonly used with RNNs because the gradient computation is cheaper compared to the logistic function (Socher, 2016).

In Equations (5) and (6), the activation function computes the value of $\boldsymbol{h}_t$ using information from the previous hidden state $\boldsymbol{h}_{t-1}$ and the current input $\boldsymbol{x}_t$. The parameters used to compute $\boldsymbol{h}_t$ are shared

for each value of t. In Equation (7), the RNNs read information from $\boldsymbol{h}_t$ to output a sequence of predicted values $\hat{\boldsymbol{y}}^{(t)}$. The process of forward propagation culminates in producing a loss that compares $\hat{\boldsymbol{y}}^{(t)}$ and $\boldsymbol{y}^{(t)}$. Gradients for Equations (5) and (8) are computed by the back-propagation through time algorithm (Goodfellow et al., 2016, p. 384), which are subsequently used for gradient descent to estimate the network parameters.

We weight the RNNs objective function to minimize the MSE weighted by the propensity score, which we estimate by multiresponse lasso regression in order to share model parameters across multiple time periods and to shrink the coefficients of (all but one) correlated covariates towards zero (Simon et al., 2013; Tibshirani et al., 2012). In order to avoid extreme propensity weights, we employ overlap weighting so that observed values under treatment receive a weight of $1 - \hat{e}_{it}$ and observed values under control receive a weight equal to $\hat{e}_{it}$ (Li et al., 2018). The propensity-score weighted MSE is:

$$\mathrm{L} = \frac{1}{T_y} \sum_{t=1}^{T_y} W_{it}(1 - \hat{\mathbf{e}}_{it}) + (1 - W_{it})\hat{\mathbf{e}}_{it} \left(\hat{\boldsymbol{y}}_t - \boldsymbol{y}_t\right)^2 + \lambda \mathbf{u}_t^2, \tag{8}$$

where the right-hand side term is the ridge penalty on the learned weights, $U = \left(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{T_y}\right)$ from Equation (7), and $\lambda > 0$ controls the regularization strength.

We train RNNs on control outcomes using a sliding window with step size of one. Each sliding window contains 10 time periods as input, and aims to predict following time period. Data are fed into the networks in batches of size 32 and the networks are trained for 500 epochs with mini-batch gradient descent on Equation (8). The last 20% of the training set is reserved for model validation. In order to prevent over-fitting, the training process stops early when there ceases to be improvement on the validation set loss within 25 epochs, and in the event of early stopping, the model weights associated with the lowest validation set loss are restored. In addition to applying a ridge penalty to Equation (8), we regularize the networks by applying dropout to both the hidden units and recurrent connections (Gal & Ghahramani, 2016).

3.2 | Identification

After training, we can impute the missing potential outcomes under control $\widehat{Y_{it}}(0)$ with the forward propagation equations in (5)–(7). We train on the control units to predict the missing outcome $\widehat{Y_{it}}(0)$ for those it values with W_{it}=1. We then estimate τ_t^{ATT} Equation (4) by contrasting $\widehat{Y_{it}}(0)$ with $Y_{it}(1)$. Under Assumptions 1 and 2, τ_t^{ATT} can be identified.

3.3 | Network architecture

We employ encoder–decoder networks, which are the standard for neural machine translation (Bahdanau et al., 2014; Cho et al., 2014; Vinyals et al., 2014) and are also widely used for predictive tasks, including speech recognition (e.g. Chorowski et al., 2015) and time-series forecasting (e.g. Zhu & Laptev, 2017). Encoder–decoder networks consist of an encoder and decoder RNN, both taking the form of long short-term memory (LSTM) networks. LSTMs are designed to resolve problems such as vanishing and exploding gradients, which prevent the networks from learning long-term dependencies in the data and tends to occur when the dimension of the hidden states is too small to summarize long input sequences (Bahdanau et al., 2014; Pascanu et al., 2013).

The encoder RNN reads in X sequentially and the hidden state of the network updates according to Equation (6). The final hidden state of the encoder is a fixed-size context vector $\boldsymbol{c}$ that summarizes the input sequence, which is copied over to the decoder RNN. Thus, the hidden state of the decoder is updated recursively by

$$\boldsymbol{h}_t = f\left(\boldsymbol{h}_{t-1}, \boldsymbol{y}_{t-1}, \boldsymbol{c}; \theta\right), \tag{9}$$

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

$$\Pr(\boldsymbol{y}_t \,|\, \boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{c}) = f\left(\boldsymbol{h}_{t-1}, \boldsymbol{y}_{t-1}, \boldsymbol{c}; \theta\right). \tag{10}$$

As in Equation (7), the decoder uses $\boldsymbol{h}_t$ to predict $\boldsymbol{y}_t$ at each t.

In our experiments and empirical application, the encoder–decoder networks consist of a two-layer LSTM encoder and single-layer gated recurrent unit (GRU) (Chung et al., 2014) decoder, each with 128 hidden units and tanh activation, stacked on a fully connected output layer with no activation function, that is, $f(x) = x$. We compare the encoder–decoder networks with a baseline LSTM consisting of a single-layer LSTM stacked on a fully connected output layer. The baseline LSTM has fewer network parameters than the encoder–decoder networks and is expected to be more suitable for smaller-dimensional data sets.

4 | PLACEBO TEST EXPERIMENTS

We conduct a series of placebo test experiments on data without real interventions in order to evaluate the ability of the RNN-based estimator to recover a null average treatment effect, that is, $\tau_t^{\text{ATT}} = 0$. The rationale for placebo test experiments is that we know the ground-truth and Assumption 1 is satisfied in expectation because placebo treatment is assigned randomly. For each trial run, we randomly select $N/2$ placebo treated units and predict their outcomes for periods following a given placebo initial time period under a staggered treatment adoption setting.

We benchmark the performance of the encoder–decoder networks described in Section 3.3 and a baseline single-layer LSTM against several estimators in terms of the root mean squared error (RMSE), comparing the actual and predicted values. A full description of the benchmark estimators we consider are provided in the supporting material (SM), Section SM-1.

1. DID Difference-in-differences regression of outcomes on treatment and unit and time fixed effects (Athey & Imbens, 2018);
2. MC-NNM Nuclear norm regularized matrix completion estimator (Athey et al., 2017);
3. SCM Generalized SCM with the restriction of non-negative weights and zero intercept of the original SCM, and weights estimated by gradient descent (Doudchenko & Imbens, 2016);
4. SCM-L1 Generalized SCM with an intercept and without weights restrictions, and weights estimated by lasso linear regression (Doudchenko & Imbens, 2016);
5. VAR Stationary vector autoregression, with weights estimated by lasso linear regression (Kock & Callot, 2015).

The comparison between the RNN-based estimators and the VAR is particularly important because the latter is capable of modelling a linear dependency component, which is present in most real-world data sets, while the former is suitable for modelling a non-linear dependency component that potentially spans across multiple time periods (Goel et al., 2017). However, RNNs typically require a large amount of training data to effectively capture non-linear and potentially long-term dependencies.

To facilitate the comparison, we run placebo test experiments on the data set underlying our empirical application along with three high-dimensional data sets. The education spending data set, described in Section 5.1, consists of historical data on the per-capita education spending of 36 US state governments over T = 203 years. We remove the treated units from the data set, leaving $N = 18$ control states. The Gaussian processes data are smooth signals generated using radial basis function to specify a Gaussian process with zero-valued mean function. The sine waves data are generated non-linear variations with frequencies in [1.0, 5.0], amplitudes in [0.1, 0.9] and random phases between $[-\pi, \pi]$. The Gaussian processes and sine waves data sets each consist of $N = 4956$ samples with sequence length of T=500. Lastly, the stock prices data set consists of stock market returns for N = 2453 stocks over T = 3082 days. The stock market is dynamic, non-stationary and complex in nature, and predicting stock market returns is a challenging task due to its unpredictable and non-linear nature.

In Figure 1, we calculate the autocorrelation function $\rho(k) = \text{Cov}(Y_t, Y_{t+k})/\text{var}(Y_t)$ for the observations in different placebo test data sets and a given lag k. The correlation decays quickly in the stock prices and Gaussian processes data sets, which suggests that the outcomes depend mostly on the observations in the short run. On the other hand, the correlation in the education spending and sine waves data sets remain prominent for the order lagged before 55 and 54, respectively, indicating that the dependency along the time dimension is longer term and of a more complicated structure. The RNN-based estimators and VAR are expected to hold an advantage over the other estimators in capturing the temporal information or dependency in the long-run.

Figure 2 reports the average RMSE for each estimator on the education spending and Gaussian processes data sets, varying the placebo initial treatment time under randomly assigned treatment in a staggered treatment adoption setting. The horizontal axis is the ratio of the initial placebo treatment time to the number of periods in the placebo data, so higher values represent more training data, and estimates are jittered across the horizontal axis to avoid overlap. When trained on the education spending data, the average RMSE for the RNN-based estimators generally decreases as the amount of training data increases, reflecting the need for sufficient time periods. The LSTM and encoder–decoder networks outperform DID and matrix completion on the education spending data, and perform comparatively to the SCM estimators and VAR when trained on the higher-dimensional Gaussian processes data.

In Figure 3, we create 10 different sub-samples by selecting the first T daily returns of N randomly selected stocks, keeping the overall data dimension fixed at $N \times T = 400{,}000$, focusing on encoder–decoder networks, SCM estimators and VAR. The sub-sampled matrices range from very thin, $N \times T = (200 \times 2000)$, to very fat, $N \times T = (2000 \times 200)$. In each case, $N/2$ units are randomly selected for placebo treatment starting at an initial placebo time period of $T/2$. The average RMSE is the highest for the encoder–decoder networks when the data are very thin, which reflects the benefit of training on high-dimensional data. All estimators perform comparatively across the remaining sub-samples.

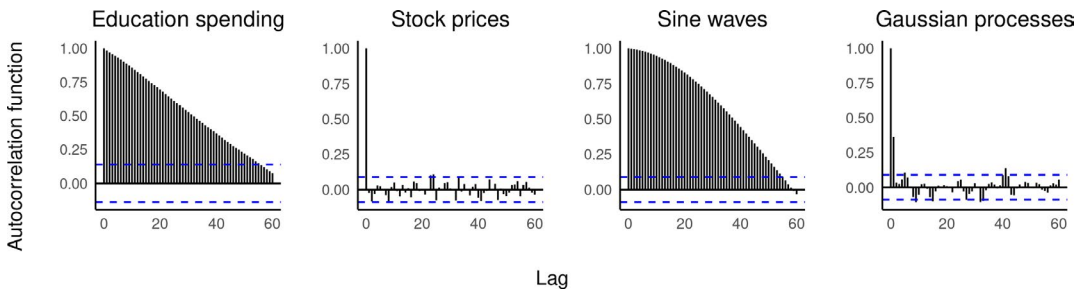
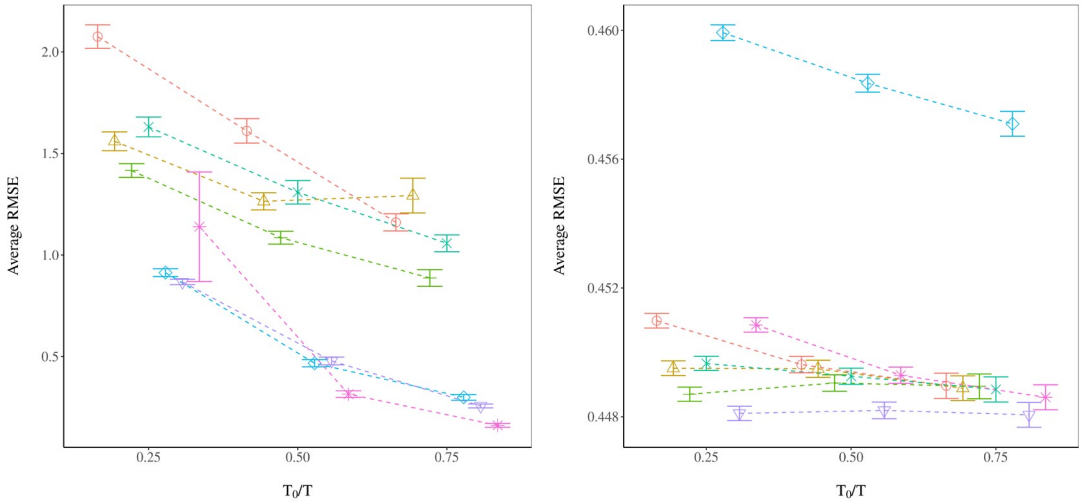


FIGURE 1 Autocorrelation function of placebo test data sets. Dashed horizontal lines represent 95% confidence bands



(a) Education spending: $(N \times T) = (18 \times 203)$†. (b) Gaussian processes: $(N \times T) = (1,000 \times 500)$‡.

FIGURE 2 Placebo tests under staggered treatment adoption. Vertical lines represent ± 1.96 times the standard error of the average RMSE across 100 runs. †: Subset from $(N \times T) = (38 \times 203)$; ‡: sub-sampled from $(N \times T) = (4956 \times 500)$. Method: ⊖, DID; △, Encoder–decoder (ours); +, LSTM (ours); ×, MC-NNM; ◇, SCM; ▽, SCM-L1; ✳, VAR [Colour figure can be viewed at wileyonlinelibrary.com]

Table 1 reports the average RMSE for the estimators with the placebo initial treatment period set to $T/2$. For the stock prices data, the first $T = 500$ daily returns of $N = 1000$ stocks are randomly sub-sampled from the larger data set. The baseline LSTM outperforms the encoder–decoder networks in the education spending and sine waves data sets, underscoring how the higher complexity of the encoder–decoder networks can hinder performance on data sets with smaller dimensions or those with simple signals, such as with the sine waves data. The VAR is capable of capturing the linear interdependencies among multiple time series and achieves the lowest average RMSE on the education spending data set, whereas the SCM estimators perform better on the other data sets. The RNN-based estimators perform comparably to the SCM and VAR estimators in the Gaussian processes, sine waves and stock prices data sets, and outperform both DID and matrix completion estimators in the education spending and sine waves data sets.

5 | HOMESTEAD POLICY AND PUBLIC SCHOOLING IN THE US

In the empirical application, we are interested in estimating the impact of mid-19th century homestead policy on the development of state government public education spending in the US. Social scientists have long viewed the rapid development of public schooling in the US as a nation-building policy (e.g. Alesina et al., 2013; Bandiera et al., 2018; Meyer et al., 1979). According to this view, states across the US adopted compulsory primary education to homogenize the population during the Age of Mass Migration, when tens of millions of foreign migrants arrived to the country between 1850 and 1914.

Engerman and Sokoloff (2005) propose an alternative explanation for the rise of public schooling: state governments on the western frontier expanded investments in public education to attract eastern migrants following the passage of the Homestead Act (HSA) of 1862. The HSA opened for settlement hundreds of millions of acres of frontier land, and any adult citizen could apply for a homestead grant of

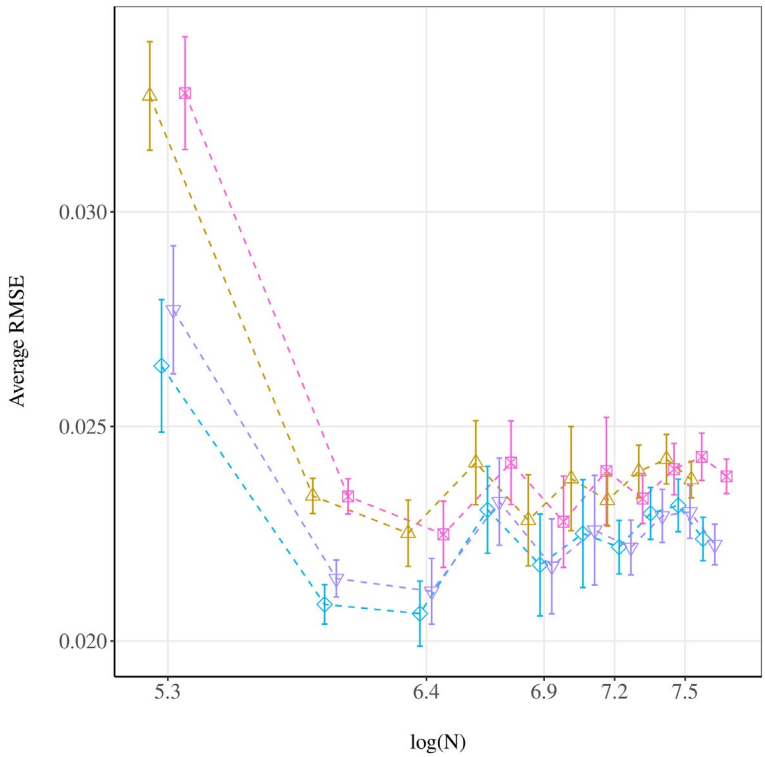


FIGURE 3 Stock prices data: placebo tests under staggered treatment adoption, with varying dimensions and keeping $N \times T = 400{,}000$. Vertical lines represent ± 1.96 times the standard error of the average RMSE across 10 runs. *Method:* △, Encoder–decoder (ours); ◇, SCM; ▽, SCM-L1 ✳, VAR [Colour figure can be viewed at wileyonlinelibrary.com]

160 acres of land, provided that they live and make improvements on the land for five years. According to this view, the sparse population on the frontier meant that state governments competed with each other to attract migrants in order to lower local labour costs and to increase land values and tax revenues. State governments in public land states—that is, states crafted from the public domain and open to homesteading—offered migrants broad access to cheap land and property rights, unrestricted voting rights, and access to public schooling. State land states, which include the original 13 states, Kentucky, Maine, Tennessee, Texas, Vermont and West Virginia, were not open to homesteading because the state government had primary authority to distribute public land (Murtazashvili, 2013).

Another alternative view is that the HSA led to larger investments in public schooling by reducing the degree of land inequality on the frontier as a consequence of fixing land grants to 160 acres. Political economy frameworks (e.g. Acemoglu & Robinson, 2008; Besley & 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 favours industrial labour 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 education reforms.

5.1 | Data

We draw data on state government education spending from the records of 48 state governments during the period of 1789–1932 (Sylla et al., 1993), 16 state governments during the period of 1933–1937

TABLE 1 Average RMSE on test set under staggered treatment adoption

	Education spending (18×203)[a]	**Gaussian processes (1000×500)[b]**	**Sine waves (1000×500)[b]**	**Stock prices (1000×500)[c]**
DID	1.611 ± 0.030	0.449 ± 10^{-4}	0.467 ± 0.001	$0.023 \pm 2 \times 10^{-4}$
Encoder–decoder (ours)	1.264 ± 0.021	0.449 ± 10^{-4}	0.405 ± 0.001	0.023 ± 10^{-4}
LSTM (ours)	1.085 ± 0.016	0.449 ± 10^{-4}	0.398 ± 0.001	0.023 ± 10^{-4}
MC-NNM	1.308 ± 0.029	0.449 ± 10^{-4}	0.457 ± 0.001	$0.023 \pm 2 \times 10^{-4}$
SCM	0.467 ± 0.009	0.458 ± 10^{-4}	0.349 ± 0.001	$\mathbf{0.022 \pm 2 \times 10^{-4}}$
SCM-L1	0.478 ± 0.009	$\mathbf{0.448 \pm 10^{-4}}$	$\mathbf{0.285 \pm 0.001}$	$0.023 \pm 2 \times 10^{-4}$
VAR	$\mathbf{0.314 \pm 0.008}$	0.449 ± 10^{-4}	0.326 ± 0.002	0.024 ± 10^{-4}

Notes: Average RMSE ± 1.96 times the standard error across 100 runs, with $N/2$ treated units and $T/2$ treated periods. Bold indicates lowest average RMSE.

[a]Subset from $(N \times T) = (38 \times 203)$.

[b]Sub-sampled from $(N \times T) = (4956 \times 500)$.

[c]Sub-sampled from $(N \times T) = (2453 \times 3082)$.

(Sylla et al., 1995a,b), and US Census special reports for the years 1902, 1913, 1932, 1942, 1944, 1946, 1948 and 1950–2008, covering 48 states (Haines, 2010; U.S. Census Bureau, 2010). Removing states and years with zero or near-zero variance results in a data set consisting of $T = 203$ observations for $N = 38$ US states, half of which are treated. We inflation-adjust the education spending data according to the US Consumer Price Index and scale by the total free population in the decennial census. We impute the 34.1% of values in the data set that are missing by Last Observation Carried Forward (LOCF), which replaces each missing value with the most recent non-missing value prior to it, with remaining missing values carried backward. We visualize the extent of the missing data in Figure SM-1 and evaluate the sensitivity of the causal estimates to alternative imputation methods in Section 5.3. Lastly, we log-transform the data to alleviate exponential effects.

The staggered treatment adoption setting is appropriate for this application because T_0 varies across states that were exposed to homesteads following the passage of the HSA. We aggregate approximately 1.46 million records of individual land patents authorized under the HSA to the state level in order to determine how the initial treatment time varies across states (General Land Office, 2017). Using these records, we determine that the earliest homestead patents were filed in 1869 in about half of the public land states, while the remaining public land states had patents filed in subsequent years.

To minimize the discrepancy between the covariate distributions of public land states and state land states, we weight the training loss by propensity scores given per-capita education spending during pre-treatment years. We also include in the conditioning set state-level averages of farm sizes and farm values measured in the 1850 and 1860 censuses (Haines, 2010) and the state-level share of total miles of operational railroad track per square mile (Atack, 2013). These pre-treatment covariates control for homesteaders migrating to more productive land and for selection bias arising from differences in access to frontier lands.

While Assumption 1 cannot be directly tested, the placebo tests on pre-treatment data reported in Section 5.4 provide indirect evidence that unconfoundedness is not violated. The no interference assumption also cannot directly be tested; however, it is likely that state land states were indirectly affected by the out-migration of homesteaders from public land states. Interference in this case would likely cause the estimated treatment effect to be understated.

5.2 | Main estimates

We train RNNs on the state land states (i.e. control units) and use the learned weights to predict the counterfactual outcomes of public land states (i.e. treated units). The top panel of Figure 4 plots the counterfactual predictions of encoder–decoder networks along with the observed outcomes of treated and control units. Prior to first homestead patent in 1869, the predicted outcomes of the public land states closely track their observed outcomes, which indicates that the networks perform well in the prediction task. The bottom panel plots the differences in the observed and predicted outcomes of the public land states, which are bounded by 95% randomization confidence intervals. We estimate the confidence intervals by constructing a distribution of average placebo effects under the null hypothesis (Cavallo et al., 2013; Firpo & Possebom, 2018; Hahn & Shi, 2017), and describe the estimation procedure in Section SM-2. The confidence intervals generally include zero for time periods prior to 1869, when no treatment effect is expected.

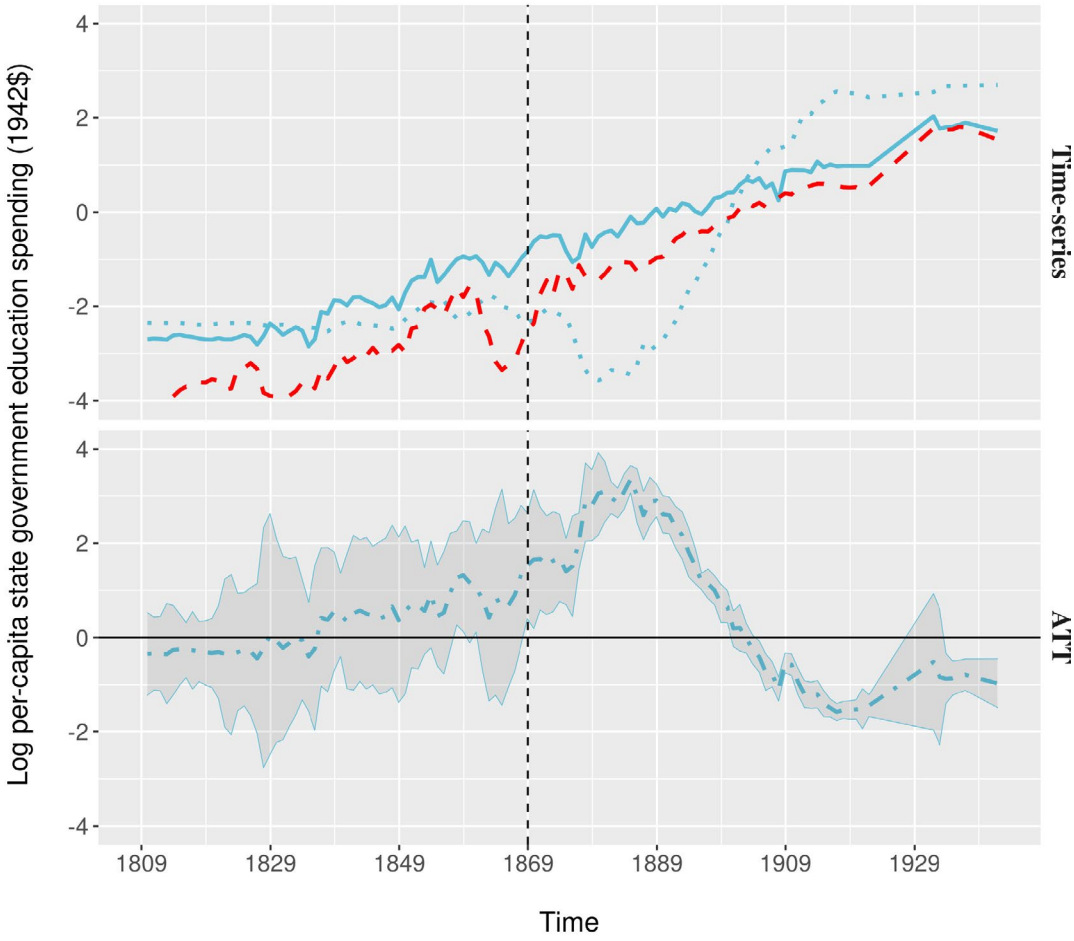


FIGURE 4 Encoder–decoder estimates of the impact of the Homestead Act on state government education spending, 1809 to 1942, with Last Observation Carried Forward imputed missing values. The dashed vertical line represents the first treatment time in 1869. *Key:* ——, observed treated; - - - -, observed control; ·····, counterfactual treated; — · —, $\hat{\tau}_t^{\text{ATT}}$ [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 ATT estimates by estimator

	$\widehat{\tau}^{\text{ATT}}$	$\widehat{\tau}^{\text{ATT}}_{\text{placebo}}$
DID	−0.597 [−2.352, 1.351]	−0.030 [−1.212, 1.396]
Encoder–decoder (ours)	**0.681** [**0.175, 1.186**]	0.213 [−1.662, 2.019]
LSTM (ours)	0.266 [−0.342, 0.876]	−0.083 [−2.146, 1.343]
MC-NNM	−0.603 [−2.429, 0.994]	0.132 [−0.661, 1.136]
SCM	0.404 [−0.075, 1.089]	0.878 [−0.187, 2.463]
SCM-L1	0.348 [−0.316, 1.029]	0.517 [−0.799, 1.977]
VAR	**0.407** [**0.231, 1.062**]	0.868 [−1.381, 5.110]

Notes: First column is ATT estimates of the impact of the HSA on log per-capita state government education spending. Second column is placebo ATT estimates, with the placebo initial treatment time set to $T_0/2$. Bracketed values are 95% randomization confidence intervals. Bold indicates statistical significance at the $\alpha = 0.05$ level.

Counterfactual predictions of state government education spending in the absence of the HSA generally tracks the observed treated outcomes until the first treatment time, at which the counterfactual diverges from the increasing observed treated time-series. Taking the mean of the post-treatment impacts, $\widehat{\tau}_t^{\text{ATT}}$, the encoder–decoder estimate of the impact of the HSA on the education spending of public land states is 0.681 log points [0.175, 1.186], as reported in the first column of Table 2. The confidence intervals surrounding this estimate do not contain zero, which indicates that the estimated effect is significantly more extreme than the exact distribution of average placebo effects under the null hypothesis. To put the magnitude of this point estimate in perspective, it represents about 2.5% of the per-capita total expenditures for public schools in 1929 (Snyder & Dillow, 2010). Table 2 reports the causal estimates recovered by each of the benchmark estimators used in the placebo tests. The VAR estimated effect is slightly smaller and also statistically significant, whereas the other estimators yield wider confidence intervals that contain zero.

5.3 | Sensitivity

In the main analyses, we impute values in the education spending data that are missing due to lack of coverage by LOCF. Table SM-1 presents ATT estimates on differently imputed data sets using four alternative imputation methods: k-nearest neighbour (k-NN), linear interpolation, multivariate imputation by chained equations (MICE) and random forests. The encoder–decoder estimates remain significant when missing values are replaced by linear interpolation, and lose their significance when missing values are replaced by the other imputation methods.

We also evaluate the sensitivity of the causal estimates to different configurations of RNNs hyperparameters by varying the hidden activation function (tanh or sigmoid); the number of hidden units per hidden layer (128 or 256 units); early stopping patience (25 or 50 epochs); and the dropout probability ($p = 0.2$ or $p = 0.5$). We report the ATT estimates by hyperparameter configuration in Table SM-2. The encoder–decoder causal estimates are positive and significant for half of the 16 different hyperparameter configurations.

5.4 | Placebo tests

We assess the accuracy of the estimators by conducting placebo tests on the pre-treatment data, when no treatment effect is expected. Among the actual treated units, we assign treatment times that are

equally spaced between the placebo T_0, which is half of the actual T_0, and $T_0 - 1$. We then construct randomization confidence intervals for the placebo counterfactual trajectories, which we report in the second column of Table 2. For each estimator, the confidence intervals contain zero, providing indirect evidence that Assumption 1 is not violated.

6 | CONCLUSION

This paper makes a methodological contribution in proposing an RNN-based estimator for estimating the effect of a binary treatment in panel data settings. RNNs are specifically structured to exploit temporal dependencies in the data and can learn non-linear combinations of control units; the latter is useful when the data-generating process underlying the outcome depends non-linearly on the history of its inputs. Most real-world time series data have a linear dependency component, for which VARs are suitable, and a non-linear dependency component that potentially spans across multiple time periods, for which RNNs are suitable. RNNs are unable to handle both linear and non-linear patterns, which are often both present in real-world time-series, and typically require a large amount of training data to effectively capture non-linear and potentially long-term dependencies. In placebo tests, we find that RNN-based estimators perform well in terms of minimizing out-of-sample error compared to VAR and other linear estimators on both small- and high-dimensional data sets with varying degrees of temporal dependency. An area of further research is extending the method to combine linear time-series models, such as VAR, with RNNs in order to more accurately model complex autocorrelation structures in the data (e.g. Goel et al., 2017).

In the empirical application, we estimate the impact of mid-19th century homestead policy on the development of state government public education spending in the United States. We train RNNs on states unaffected by homestead policy and use the learned weights to predict the counterfactual outcomes of public land states, which were open to homesteading. The encoder–decoder estimate of the impact of homestead policy on the education spending of public land states is 0.681 log points [0.175, 1.186], which represents about 2.5% of the per-capita total expenditures for public schools in 1929. This estimate is generally robust to the configuration of RNN hyperparameters and the missing data imputation method. The result is consistent with both the historical view that state governments in public land states expanded public education investments to attract eastern migrants, and the political economy view that homesteading deterministically lowered land inequality on the frontier and consequently prevented wealthy landowners from blocking public schooling reforms.

ACKNOWLEDGEMENTS

Poulos acknowledges the support of the National Science Foundation Graduate Research Fellowship (DGE-1106400) and the National Science Foundation under Grant DMS-1638521 to the Statistical and Applied Mathematical Sciences Institute. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Comet GPU at the San Diego Supercomputer Center through allocation SES180010. Code to reproduce the results of the paper is available at https://github.com/jvpoulos/rnns-causal.

REFERENCES

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

Abadie, A., Diamond, A. & Hainmueller, J. (2010) Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105, 493–505.

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

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

Alesina, A., Giuliano, P. & 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. & Shen, D. (2018) Robust synthetic control. *The Journal of Machine Learning Research*, 19, 802–852.

Arkhangelsky, D., Athey, S., Hirshberg, D.A., Imbens, G.W. & Wager, S. (2019) Synthetic difference in differences. *Working Paper 25532*, National Bureau of Economic Research. Available at: http://www.nber.org/papers/w25532

Ashenfelter, O. (1978) Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, 60, 47–57.

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. & Imbens, G. (2018) Design-based analysis in difference-in-differences settings with staggered adoption. *arXiv e-prints*, arXiv:1808.05293.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G. & Khosravi, K. (2017) Matrix completion methods for causal panel data models. *arXiv e-prints*, arXiv:1710.10251.

Athey, S., Imbens, G. & Wager, S. (2018) Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B*, 80, 597–623.

Bahdanau, D., Cho, K. & Bengio, Y. (2014) Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, arXiv:1409.0473.

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

Bang, H. & Robins, J.M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.

Belloni, A., Chernozhukov, V., Fernández-Val, I. & Hansen, C. (2017) Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85, 233–298.

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

Ben-Michael, E., Feller, A. & Rothstein, J. (2019) Synthetic controls with staggered adoption. *arXiv e-prints*, arXiv:1912.03290.

Bennett, A., Kallus, N. & Schnabel, T. (2019) Deep generalized method of moments for instrumental variable analysis. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. & Garnett, R. (Eds.) *Advances in neural information processing systems*, vol. 32. Red Hook, NY: Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper/2019/file/15d185eaa7c954e77f5343d941e25fbd-Paper.pdf

Bertrand, M., Duflo, E. & Mullainathan, S. (2004) How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119, 249–275.

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

Brodersen, K.H., Gallusser, F., Koehler, J., Remy, N. & Scott, S.L. (2015) Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, 9, 247–274.

Carvalho, C., Masini, R. & 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. & Pantano, J. (2013) Catastrophic natural disasters and economic growth. *Review of Economics and Statistics*, 95, 1549–1561.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. et al. (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21 C1–C68.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. et al. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv e-prints*, arXiv:1406.1078.

Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K. & Bengio, Y. (2015) Attention-based models for speech recognition. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M. & Garnett, R. (Eds.) *Advances in neural information*

processing systems, vol. 28. Red Hook, NY: Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper/2015/file/1068c6e4c8051cfd4e9ea8072e3189e2-Paper.pdf

Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv e-prints*, arXiv:1412.3555.

Cinar, Y.G., Mirisaee, H., Goswami, P., Gaussier, E., Aït-Bachir, A. & Strijov, V. (2017) Position-based content attention for time series forecasting with sequence-to-sequence RNNs. In: Liu D., Xie S., Li Y., Zhao D., El-Alfy ES. (Eds.)*International conference on neural information processing*, vol. 10638. Cham: Springer, pp. 533–544.

Doudchenko, N. & Imbens, G.W. (2016) Balancing, regression, difference-in-differences and synthetic control methods: a synthesis. *arXiv e-prints*, arXiv:1610.07748.

Dube, A. & 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

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

Farrell, M.H., Liang, T. & Misra, S. (2021) Deep neural networks for estimation and inference. *Econometrica*, 89, 181–213.

Ferman, B. & Pinto, C. (2016) Revisiting the synthetic control estimator. Available at https://mpra.ub.uni-muenchen.de/81941/1/MPRA_paper_81941.pdf

Firpo, S. & Possebom, V. (2018) Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, 6, 20160026.

Gal, Y. & Ghahramani, Z. (2016) A theoretically grounded application of dropout in recurrent neural networks. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. & Garnett, R. (Eds.) *Advances in neural information processing systems*, vol. 29. Red Hook, NY: Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper/2016/file/076a0c97d09cf1a0ec3e19c7f2529f2b-Paper.pdf

Galor, O., Moav, O. & 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.

General Land Office (2017) General Land Office (GLO) Records Automation. Bureau of Land Management, Washington, DC. Available at: https://glorecords.blm.gov/

Glorot, X. & Bengio, Y. (2010) Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y.W. & Titterington, M. (Eds.) *Proceedings of machine learning research*, vol. 9. PMLR, pp. 249–256. Available at: http://proceedings.mlr.press/v9/glorot10a.html

Goel, H., Melnyk, I. & Banerjee, A. (2017) R2N2: residual recurrent neural networks for multivariate time series forecasting. *arXiv e-prints*, arXiv:1709.03159.

Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep learning*. Cambridge, MA: MIT Press.

Graves, A. (2012) Neural networks. In: Kacprzyk, J. (Ed.) *Supervised sequence labelling with recurrent neural networks*. Berlin, Germany: Springer-Verlag, pp. 5–13.

Hahn, J. & 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. https://doi.org/10.3886/ICPSR02896.v3

Hartford, J., Lewis, G., Leyton-Brown, K. & Taddy, M. (2017) Deep IV: a flexible approach for counterfactual prediction. In: Precup, D. & Teh, Y.W. (Eds.) *Proceedings of machine learning research*, vol. 70. PMLR, pp. 1414–1423. Available at: http://proceedings.mlr.press/v70/hartford17a.html

Hihi, S. & Bengio, Y. (1996) Hierarchical recurrent neural networks for long-term dependencies. In: Touretzky, D., Mozer, M.C. & Hasselmo, M. (Eds.) *Advances in neural information processing systems*, vol. 8. Cambridge, MA: MIT Press, pp. 493–499. Available at: https://proceedings.neurips.cc/paper/1995/file/c667d53acd899a97a85de0c201ba99be-Paper.pdf

Imbens, G.W. & Rubin, D.B. (2015) *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press.

Kock, A.B. & Callot, L. (2015) Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186 325–344.

Li, F., Morgan, K.L. & Zaslavsky, A.M. (2018) Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113 390–400.

Meyer, J.W., Tyack, D., Nagel, J. & 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.

Murtazashvili, I. (2013) *The political economy of the American frontier*. New York, NY: Cambridge University Press.

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

Pang, X., Liu, L. & Xu, Y. (2020) A Bayesian alternative to synthetic control for comparative case studies. Available at: https://ssrn.com/abstract=3649226

Pascanu, R., Mikolov, T. & Bengio, Y. (2013) On the difficulty of training recurrent neural networks. In: Dasgupta, S. & McAllester, D. (Eds.) *Proceedings of machine learning research*, vol. 28. PMLR, pp. 1310–1318. Available at: http://proceedings.mlr.press/v28/pascanu13.html

Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

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

Simon, N., Friedman, J. & Hastie, T. (2013) A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv e-prints*, arXiv:1311.6529.

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

Socher, R. (2016) Deep learning for natural language processing lecture 6: neural tips and tricks and recurrent neural networks. Available at: https://cs224d.stanford.edu/lectures/CS224d-Lecture6.pdf.

Splawa-Neyman, J., Dabrowska, D.M. & Speed, T.P. (1990) On the application of probability theory to agricultural experiments. *Statistical Science*, 5, 465–472.

Sylla, R.E., Legler, J.B. & 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. http://doi.org/10.3886/ICPSR06304.v1

Sylla, R.E., Legler, J.B. and Wallis, J. (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. http://doi.org/10.3886/ICPSR06304.v1

Sylla, R.E., Legler, J.B. & Wallis, J. (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

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J. et al. (2012) Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74 245–266.

U.S. Census Bureau (2010) Data base on historical finances of federal, state and local governments. Available at: https://www.census.gov/programs-surveys/gov-finances/data/historical-data.html

Vinyals, O., Kaiser, L., Koo, T., Petrov, S., Sutskever, I. & Hinton, G. (2014) Grammar as a foreign language. *arXiv e-prints*, arXiv:1412.7449.

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

Zhu, L. & Laptev, N. (2017) Deep and confident prediction for time series at Uber. *arXiv e-prints*, arXiv:1709.01907.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Poulos J, Zeng S. RNN-based counterfactual prediction, with an application to homestead policy and public schooling. *J R Stat Soc Series C*. 2021;70:1124–1139. https://doi.org/10.1111/rssc.12511