RNN-based counterfactual time-series prediction (Online Appendix)

May 10, 2018

Contents

1 Statistical significance

The following procedure constructs an exact distribution of *average* placebo effects under the null hypothesis:

1. Estimate the observed test static μ^* by estimating Eq. 2 (in the main text) for all J, which results in a matrix of dimension $(\tau - n) \times J$. Taking the row-wise mean results in a $\tau - n$-length array of observed average placebo treated effects.

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

3. Take a column-wise sum of the number of μ that are greater than or equal to μ^*.

Each element of the $(\tau - n) \times J$ matrix of counts obtained from the last step is divided by $\mathcal{Q}$ to estimate an array of exact two-sided p values, $\hat{p}$.

1.1 Randomization confidence intervals

I assume that treatment has a constant additive effect Δ and a 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 a significance level α. I find the endpoints of the confidence interval by randomly sampling 1,000 values of Δ.

2 RNN architecture and implementation details

The baseline LSTM take the form of a single unidirectional RNN. The encoder takes the form of a two-layer bidirectional LSTMs, each with 128 hidden units, and the decoder is a single-layer Gated Recurrent Unit (GRU) (Chung et al. 2014) with 128 hidden units (Fig. 1).

In the empirical applications, network weights are learned with stochastic gradient descent on $L^{(t)}$ using `Adam` stochastic optimization (Kingma and Ba 2014). As a regularization strategy, I apply dropout to the inputs and L2 regularization losses to the network weights.

1. $\mathcal{Q}$ can be computationally burdensome when there are many control units. I set $\mathcal{Q} = 10,000$ in applications in which $J > 16$ (e.g., California dataset).

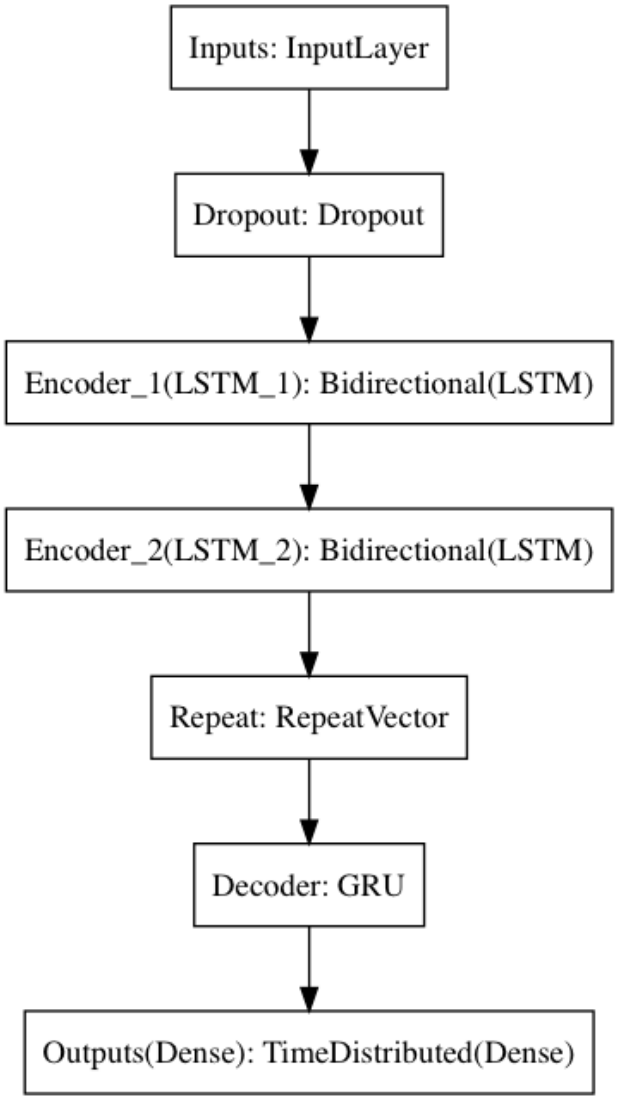
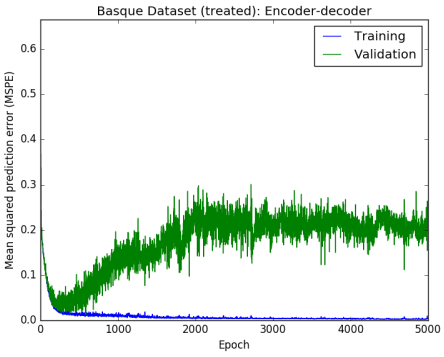
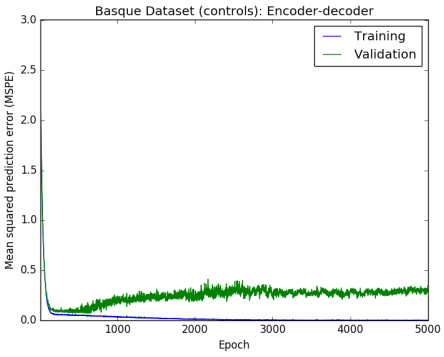


Figure 1: Encoder-decoder networks architecture. Dropout is applied to the visible input sequences, which are then fed to a two-layer bidirectional LSTM encoder. The encoder encodes the input sequences into a single vector that contains information about the entire sequence. The output of the encoder is repeated t times and fed to the single-layer GRU decoder, which translates the encoded sequence into the predicted sequence. Finally, a dense layer is applied to the decoder output to generate predictions.

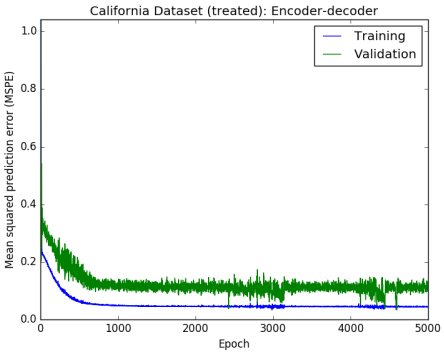
3 RNNs training history



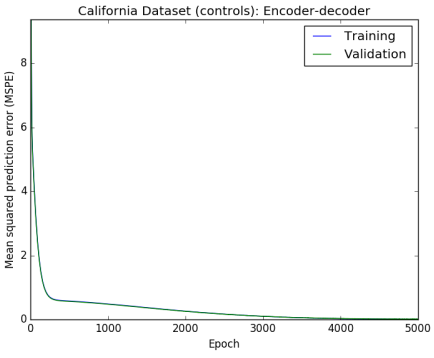
(a) Basque Country (treated)



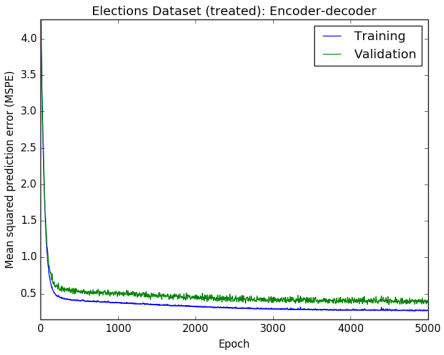
(b) Basque Country (controls)



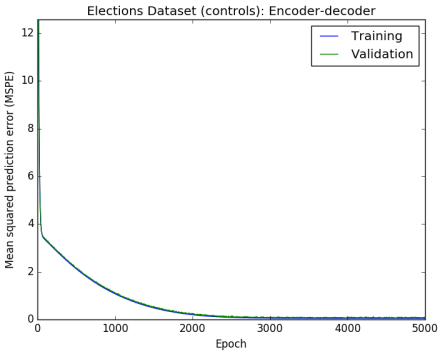
(c) California (treated)



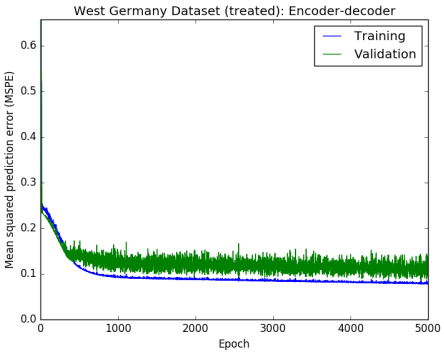
(d) California (controls)



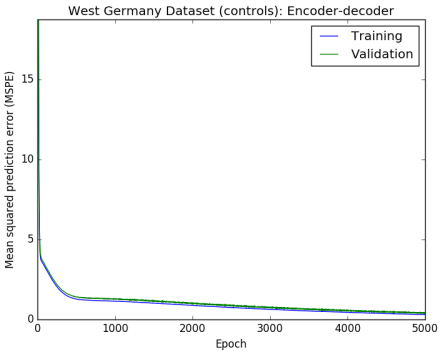
(e) Elections (treated)



(f) Elections (controls)

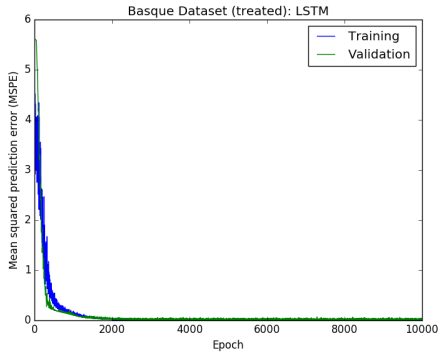


(g) West Germany (treated)

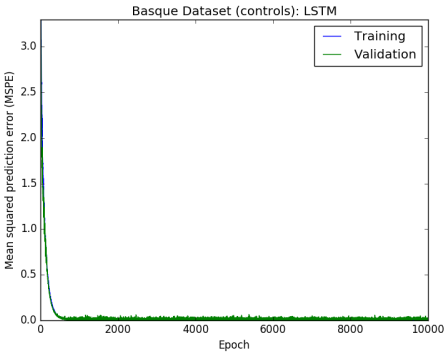


(h) West Germany (controls)

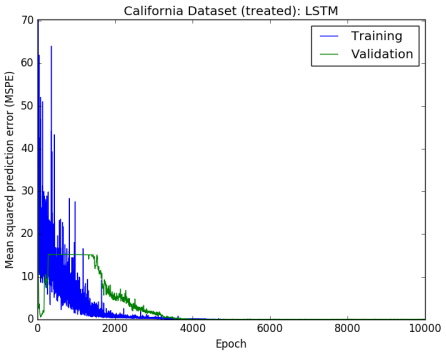
Figure 2: Evolution of encoder-decoder networks training and validation loss in terms of MSPE.



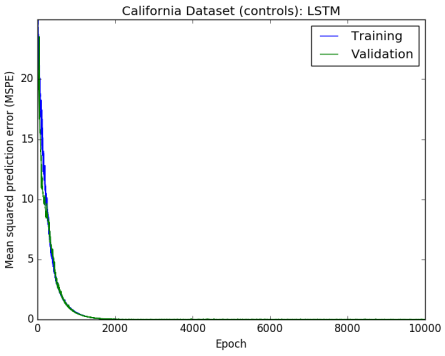
(a) Basque Country (treated)



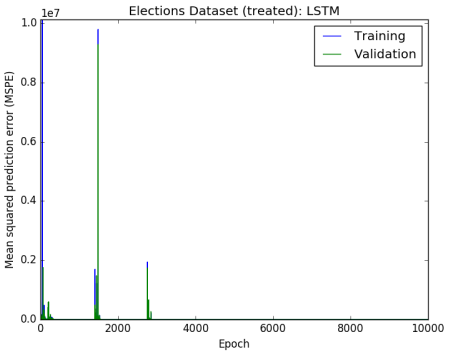
(b) Basque Country (controls)



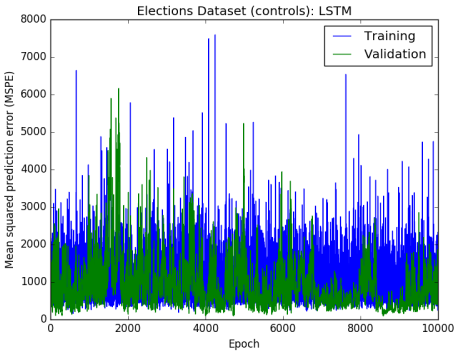
(c) California (treated)



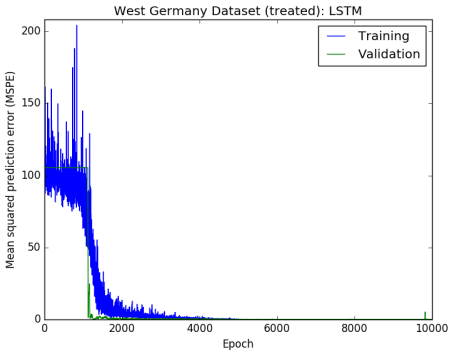
(d) California (controls)



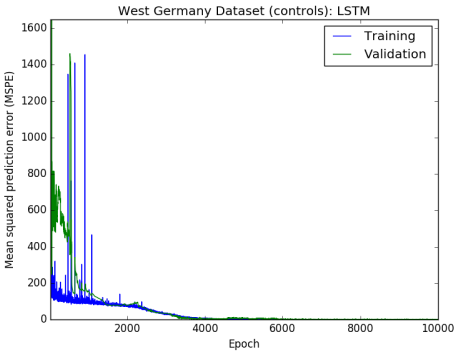
(e) Elections (treated)



(f) Elections (controls)



(g) West Germany (treated)

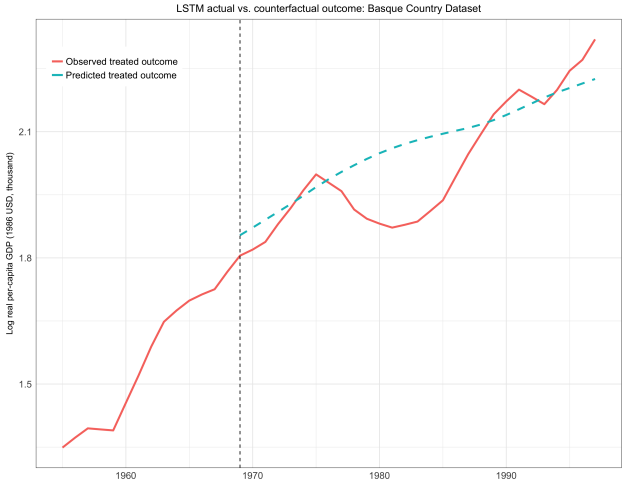


(h) West Germany (controls)

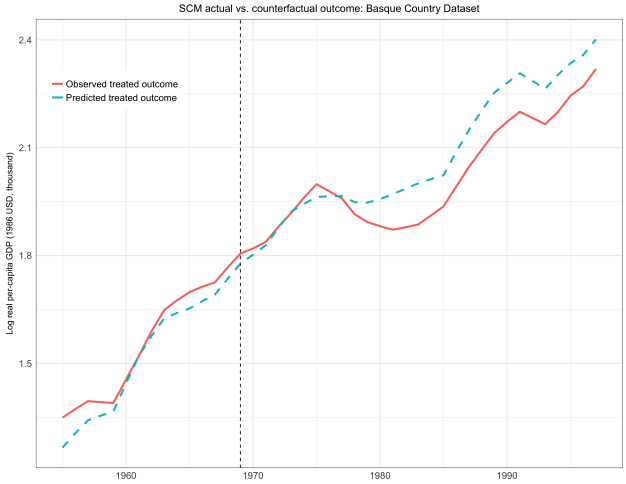
Figure 3: Evolution of LSTM training and validation loss in terms of MSPE.

4 Estimates on Basque Country data

(a) Encoder-decoder

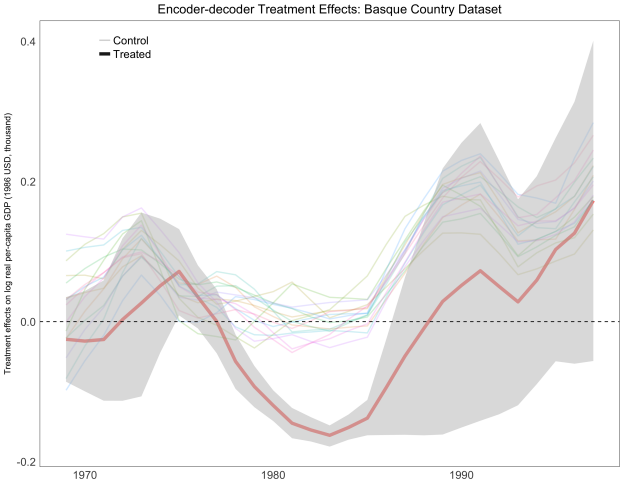


(b) LSTM

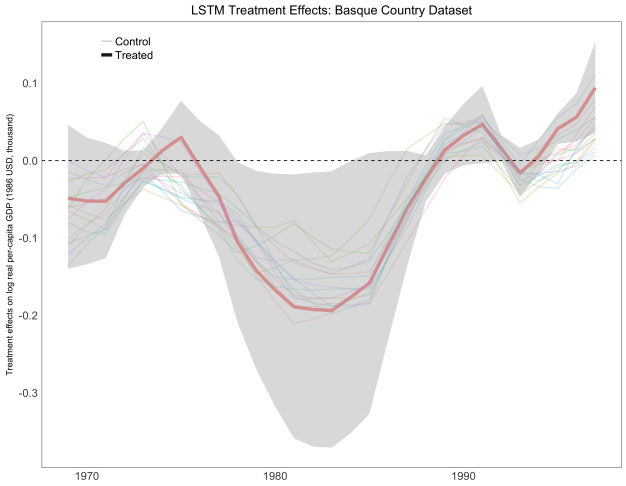


(c) SCM

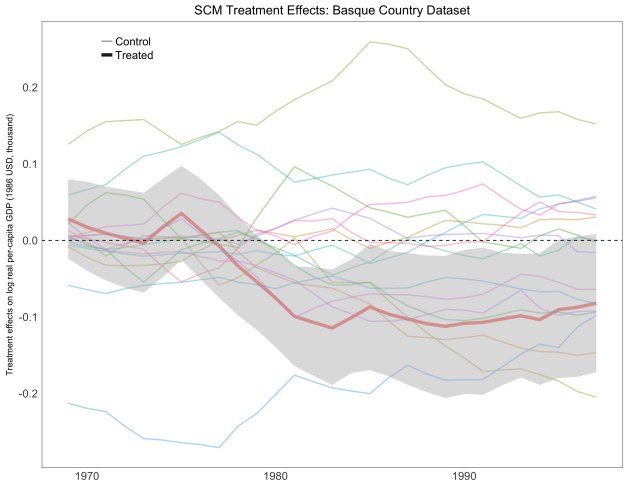
Figure 4: Observed and counterfactual predicted outcomes for treated unit in Basque Country dataset.



(a) Encoder-decoder

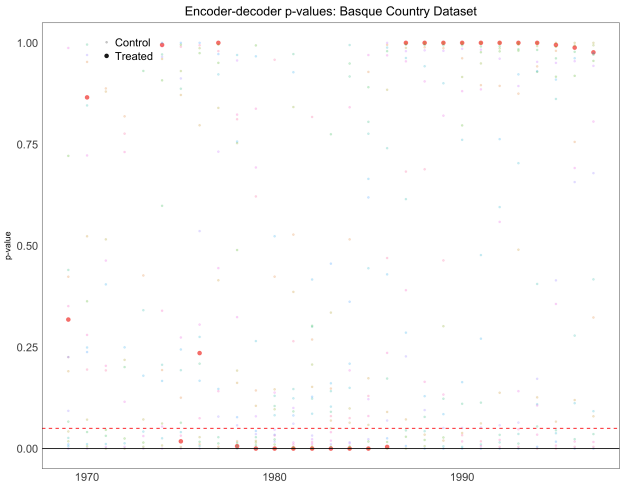


(b) LSTM

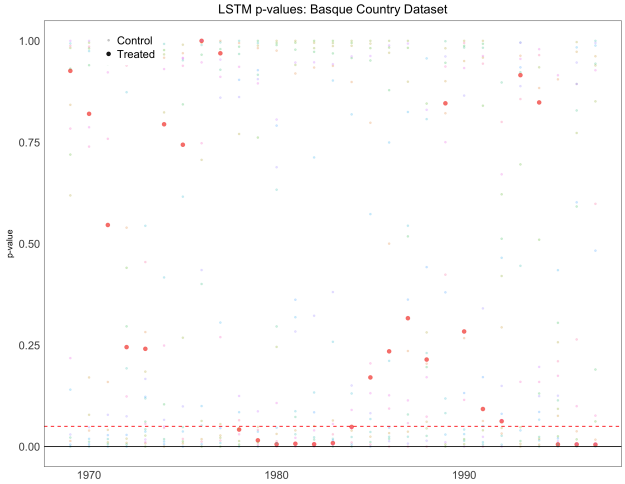


(c) SCM

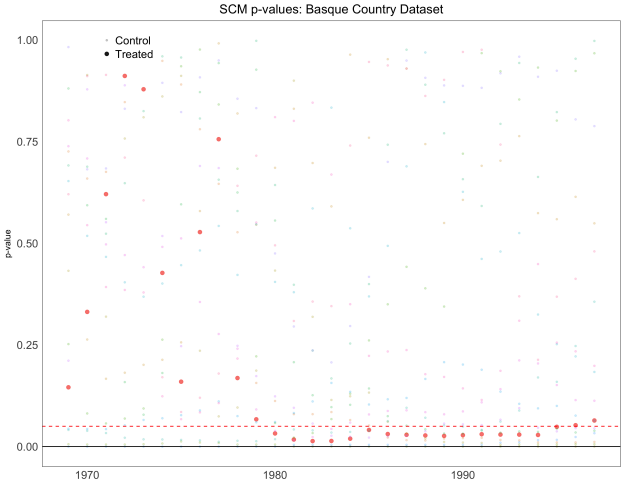
Figure 5: Time-series of post-period treatment effects in Basque Country dataset. Darker line represents the effect on the actual treated unit and each lighter line represents the effects on control units. Shaded regions represent 95% randomization confidence intervals.



(a) Encoder-decoder



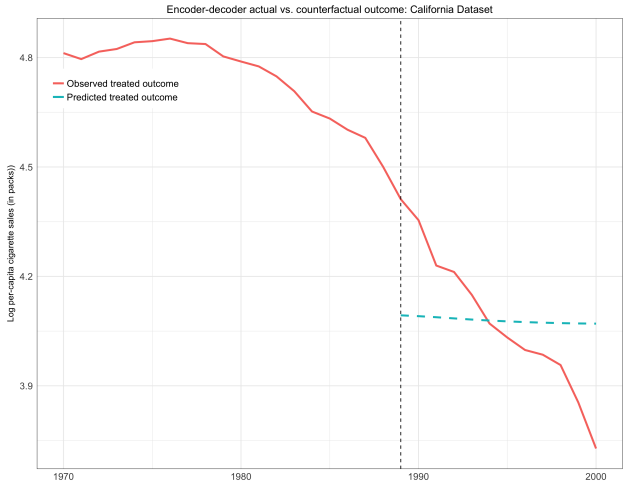
(b) LSTM



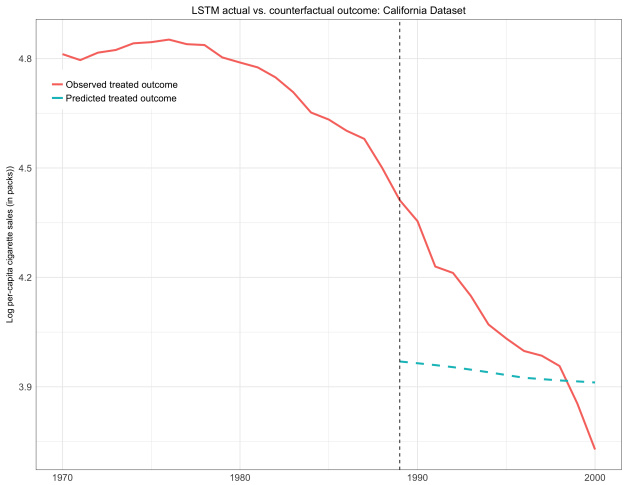
(c) SCM

Figure 6: Per-period randomization p-values corresponding to treatment effects on treated and control units in Basque Country dataset.

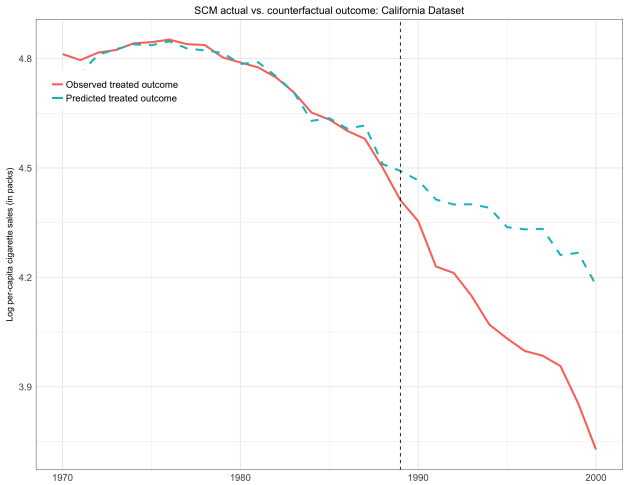
5 Estimates on California data



(a) Encoder-decoder

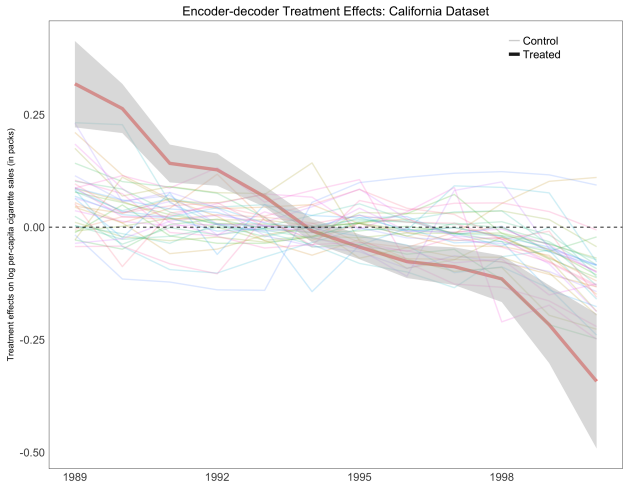


(b) LSTM

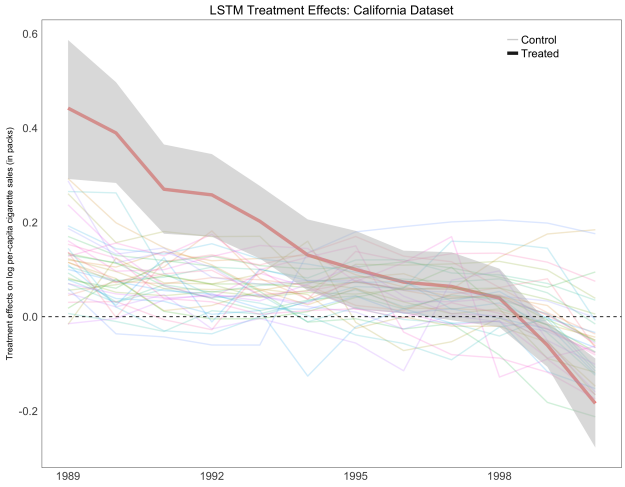


(c) SCM

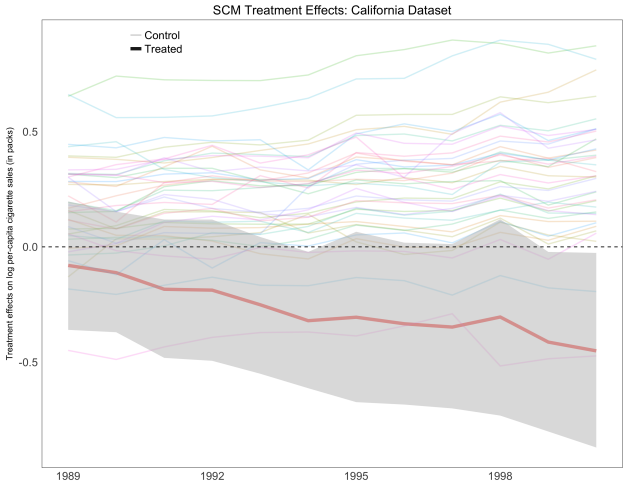
Figure 7: Observed and counterfactual predicted outcomes for treated unit in California dataset.



(a) Encoder-decoder

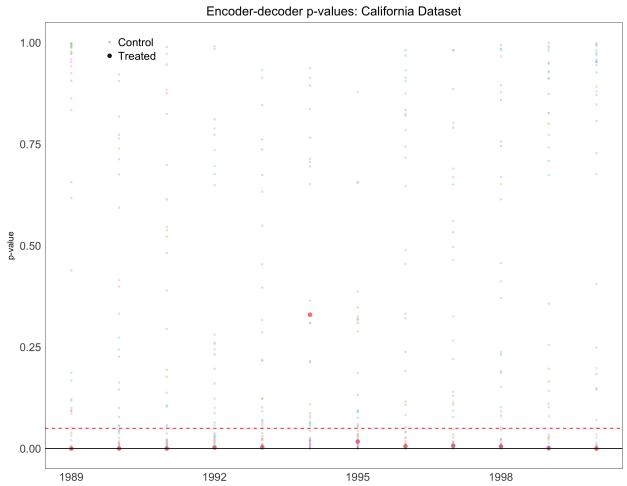


(b) LSTM

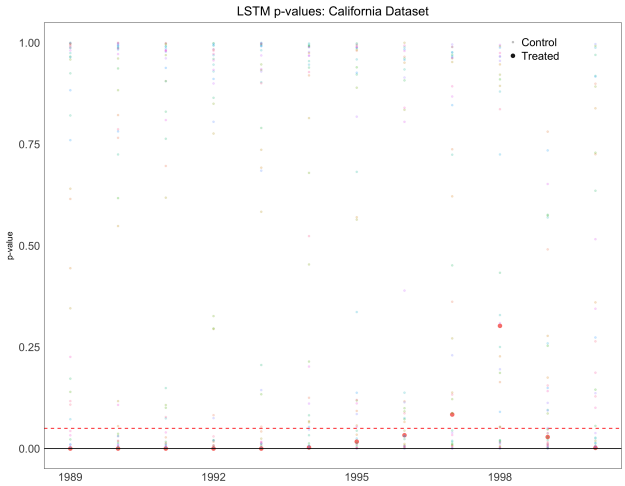


(c) SCM

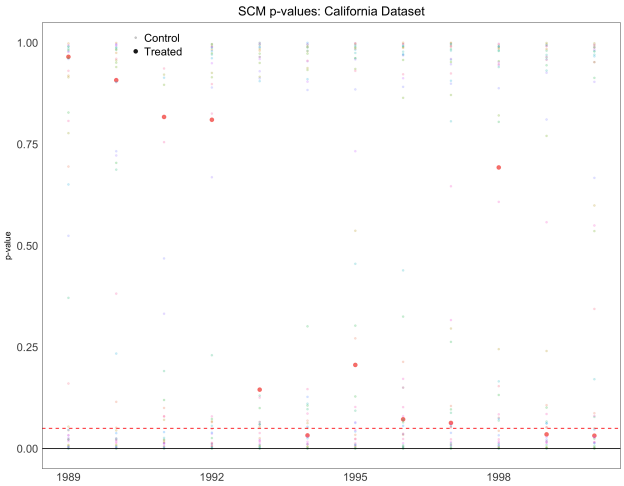
Figure 8: Time-series of post-period treatment effects in California dataset. See notes to Fig. 5.



(a) Encoder-decoder



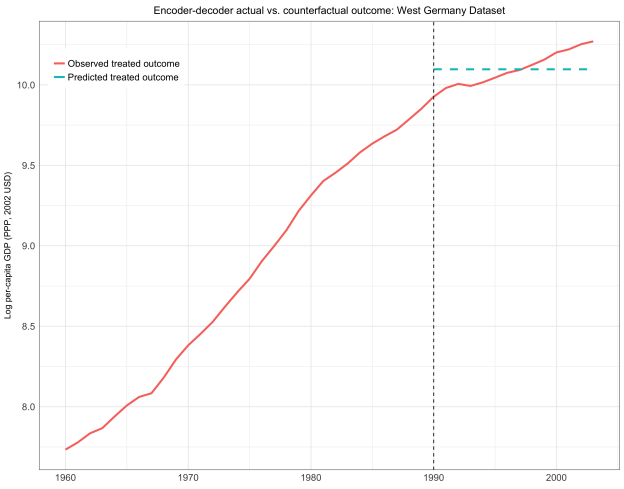
(b) LSTM



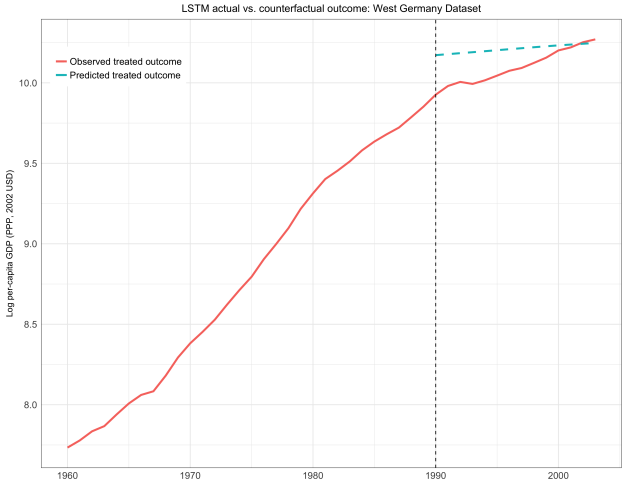
(c) SCM

Figure 9: Per-period randomization p-values corresponding to treatment effects on treated and control units in California dataset.

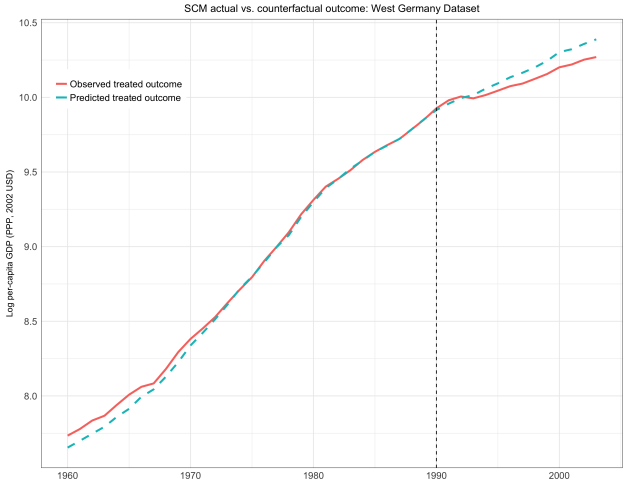
6 Estimates on West Germany data



(a) Encoder-decoder

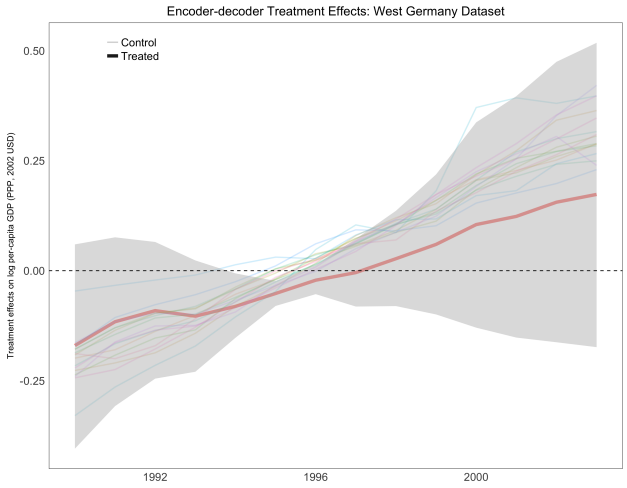


(b) LSTM

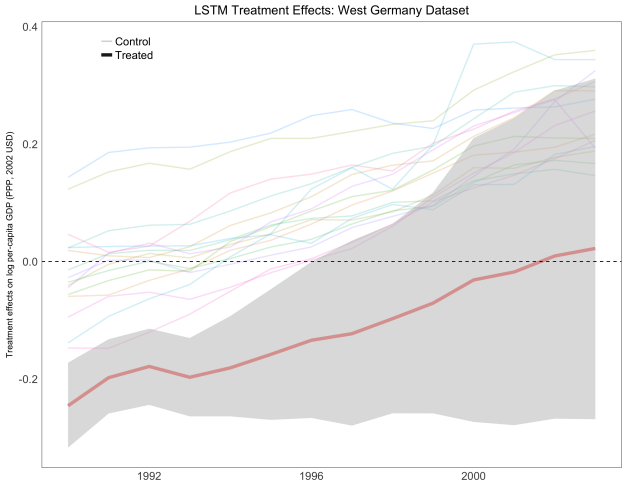


(c) SCM

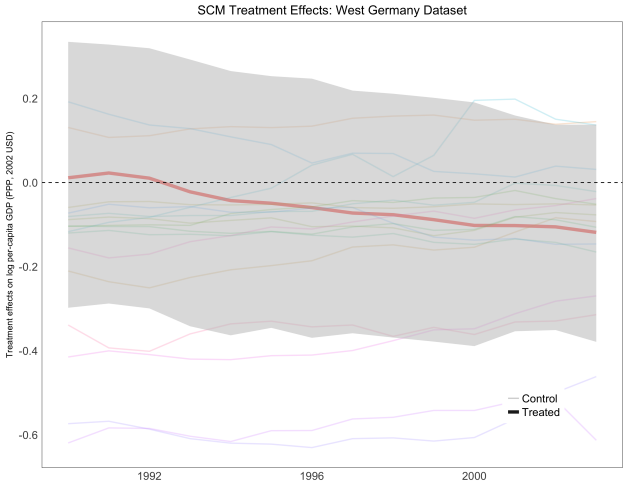
Figure 10: Observed and counterfactual predicted outcomes for treated unit in West Germany dataset.



(a) Encoder-decoder

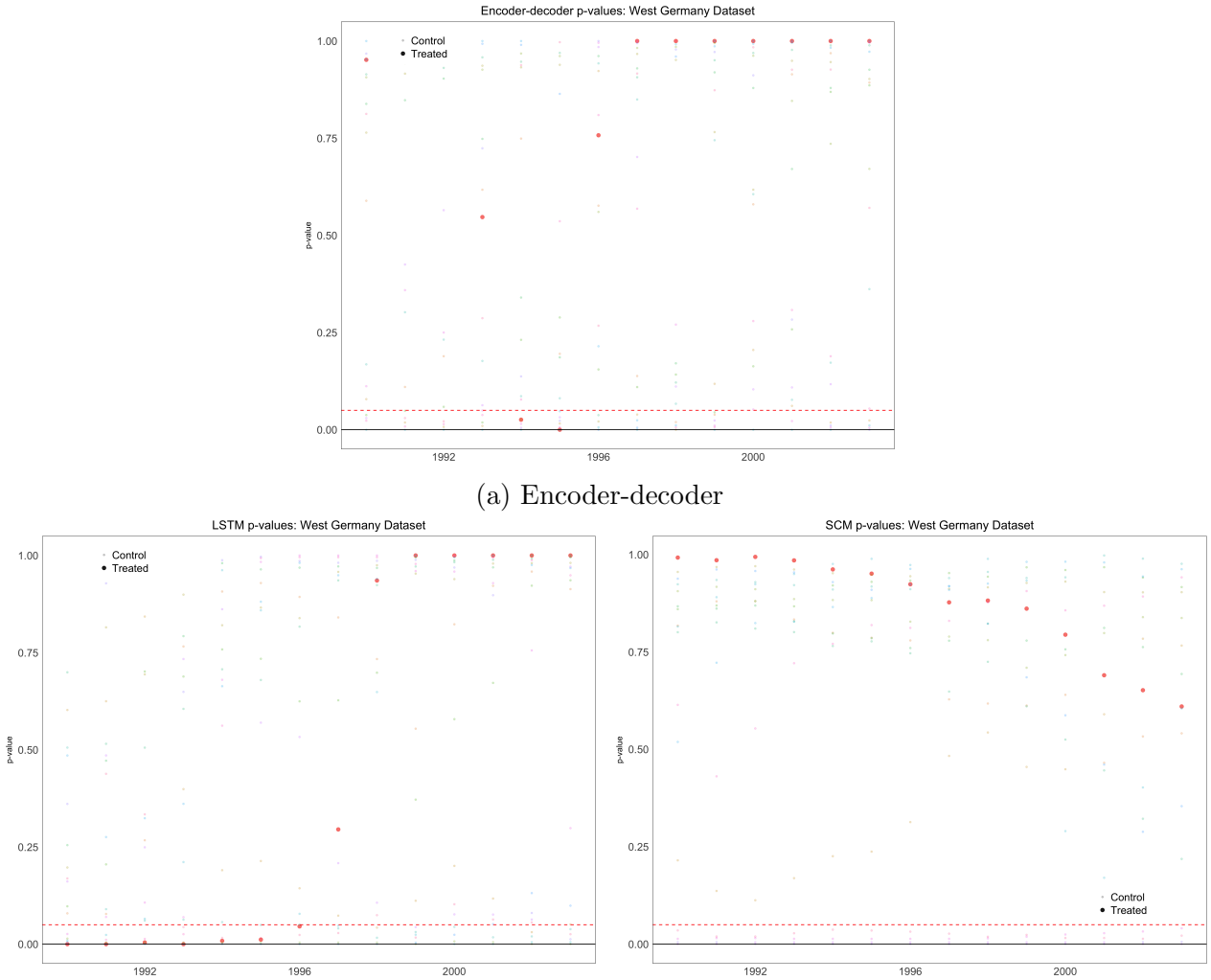


(b) LSTM



(c) SCM

Figure 11: Time-series of post-period treatment effects in West Germany dataset. See notes to Fig. 5.



(a) Encoder-decoder

(b) LSTM

(c) SCM

Figure 12: Per-period randomization p-values corresponding to treatment effects on treated and control units in West Germany dataset.

References

Chung, Junyoung, Caglar Gulcehre, Kyung Hyun Cho, and Yoshua Bengio. 2014. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling." *arXiv preprint arXiv:1412.3555.*

Kingma, Diederik, and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization." *arXiv preprint arXiv:1412.6980.*