Web-based supporting materials for "Homestead Acts and the Development of American State Capacity: An Application of RNN-Based Counterfactual Prediction" by Jason Poulos

August 14, 2018

Contents

1 Data

Table 1: Definitions and sources of variables.

Theme	Variable	Coverage	Definition	Source
Farms	Farm value	1860-1950 (decennial)	Log average value of farmland and buildings per acre ($)	Ibid.
	Land inequality	Ibid.	Gini coefficient based on distribution of farm sizes, adjusted for the share of propertyless farmers (see Vollrath (2013, pg. 273))	Ibid.
State Capacity	Revenues	1790-1982	Log per-capita state government total revenue (1982$)	Sylla, Legler, and Wallis (1993, 1995a, 1995b) and Haines (2010) (total free pop. data from Haines (2010))
Ibid.	Expenditures	Ibid.	Log per-capita state government total expenditure (1982$)	Ibid.
Ibid.	Education spending	Ibid.	Log per-capita state government education spending (1982$)	Ibid.
Land patents	Homesteads	1860-1950	Log per-capita cumulative number patents issued under the Homestead Act of 1862	U.S. BLM (`https://glorecords.blm.gov`) (total free pop. data from Haines (2010))

2 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. 8 (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, $\hat{p}$.

2.1 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.

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).

3 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.

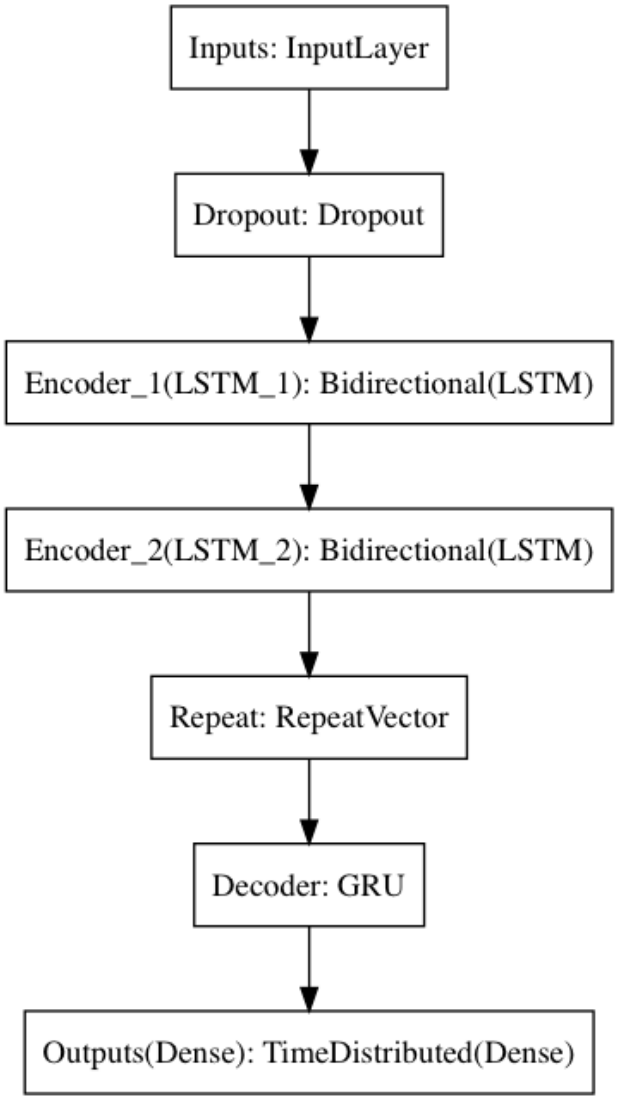
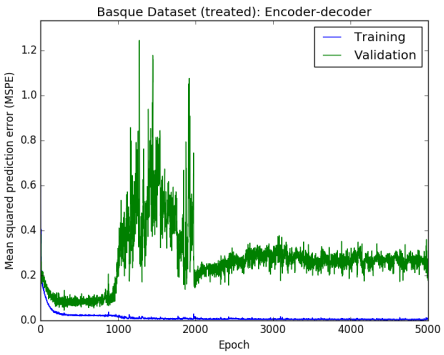
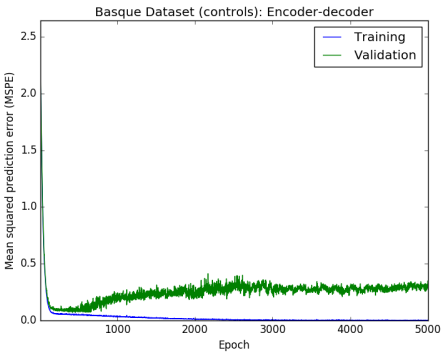


Figure 1: Encoder-decoder networks architecture. Dropout is applied to the visible input sequences, which are then fed to a two-layer bidirectional LSTM encoder. The encoder encodes the input sequences into a single vector that contains information about the entire sequence. The output of the encoder 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.

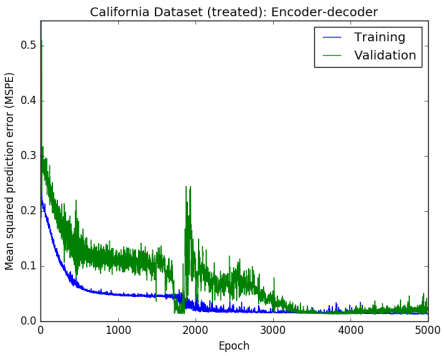
4 RNNs training history: SCM datasets



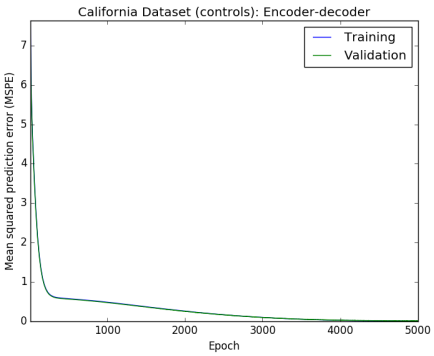
(a) Basque Country (treated)



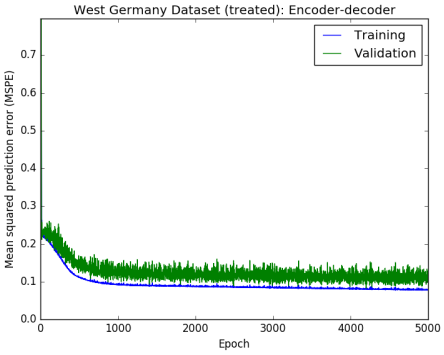
(b) Basque Country (controls)



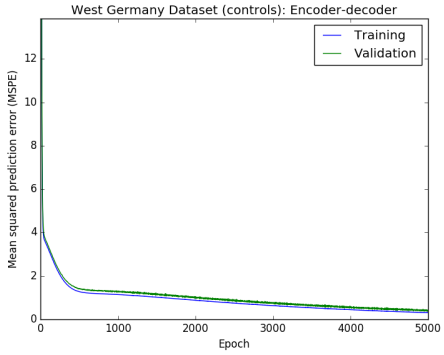
(c) California (treated)



(d) California (controls)

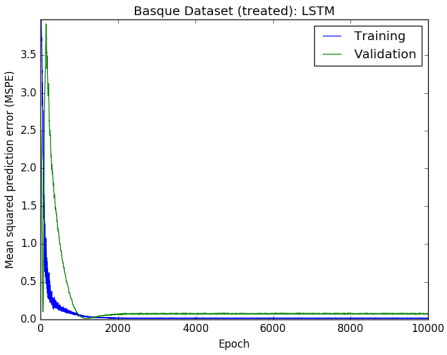


(e) West Germany (treated)

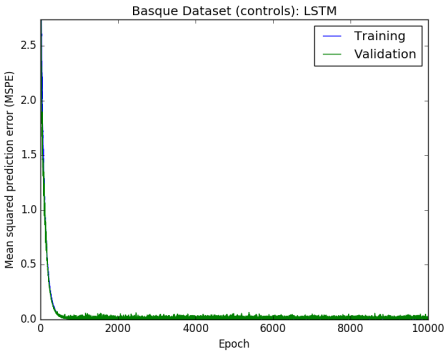


(f) West Germany (controls)

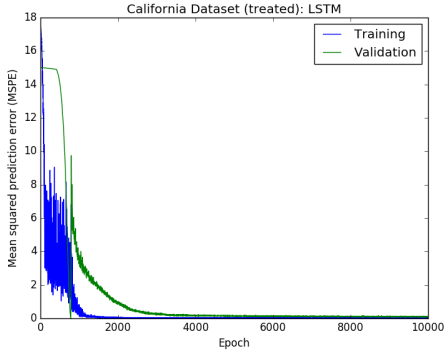
Figure 2: Evolution of encoder-decoder networks training and validation loss in terms of MSPE.



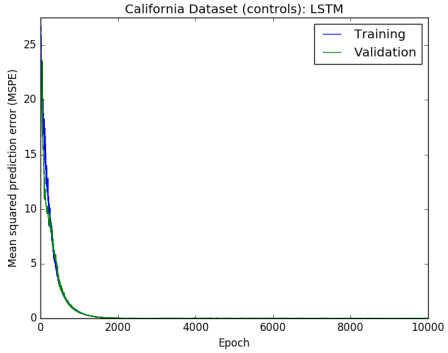
(a) Basque Country (treated)



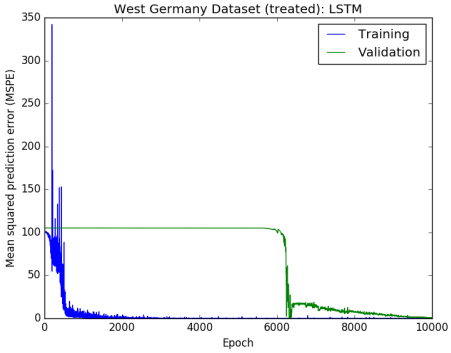
(b) Basque Country (controls)



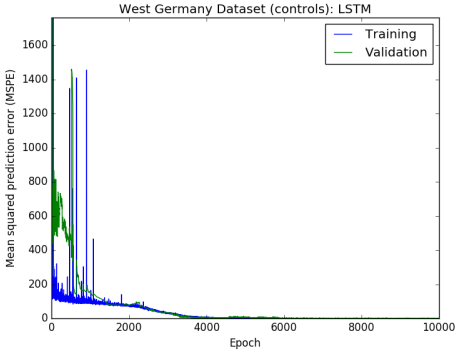
(c) California (treated)



(d) California (controls)



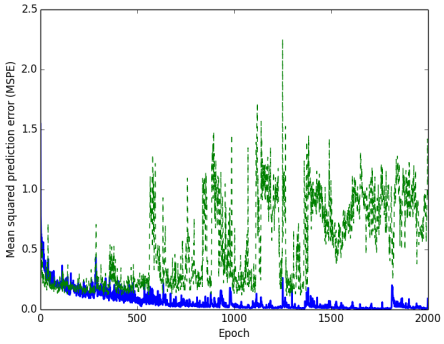
(e) West Germany (treated)



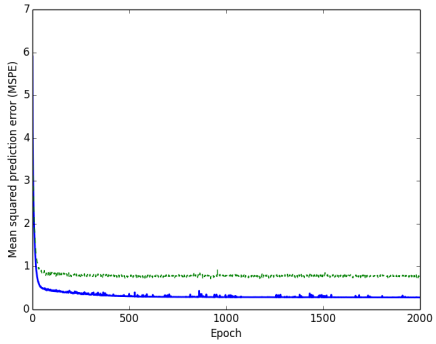
(f) West Germany (controls)

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

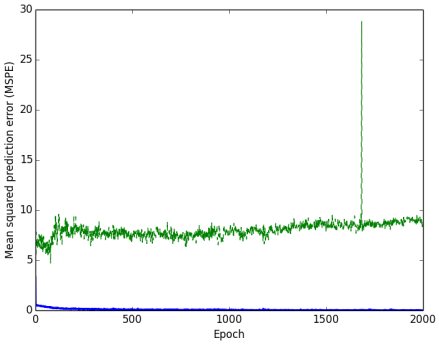
5 RNNs training history: State capacity data



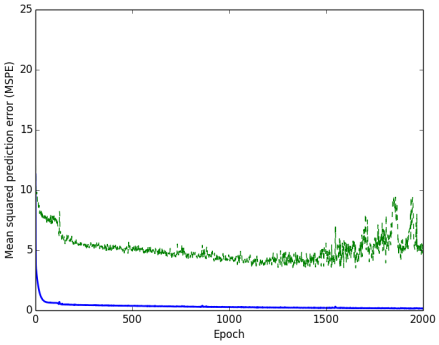
(a) Education Spending in South (treated)



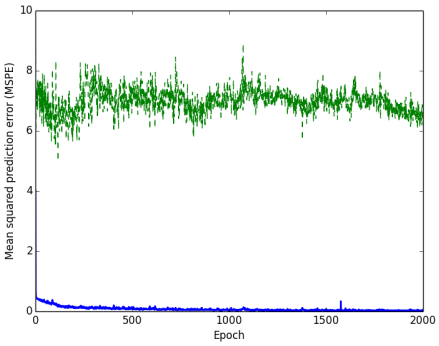
(b) Education Spending in South (controls)



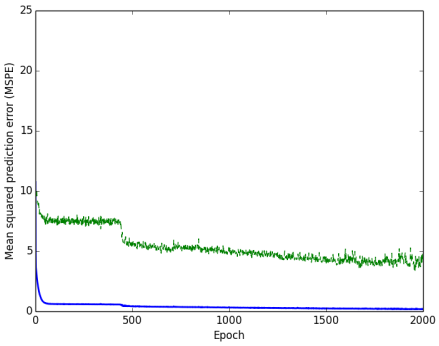
(c) Expenditure in South (treated)



(d) Expenditure in South (controls)

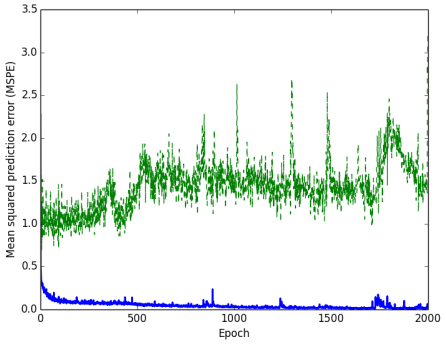


(e) Revenue in South (treated)

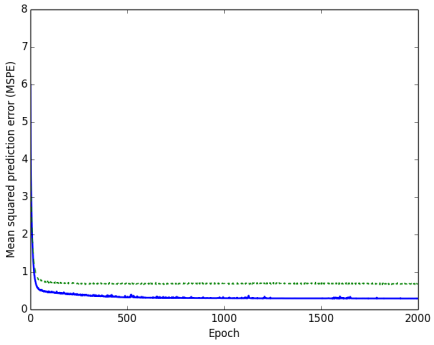


(f) Revenue in South (controls)

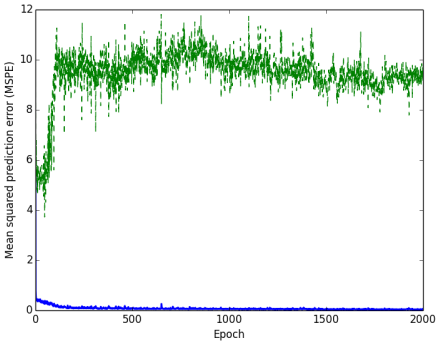
Figure 4: Encoder-decoder networks training (solid line) and validation loss (dashed line) on southern public land state capacity.



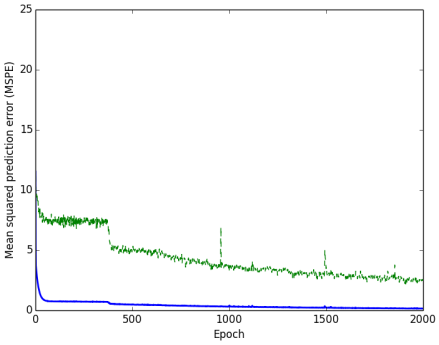
(a) Education Spending in West (treated)



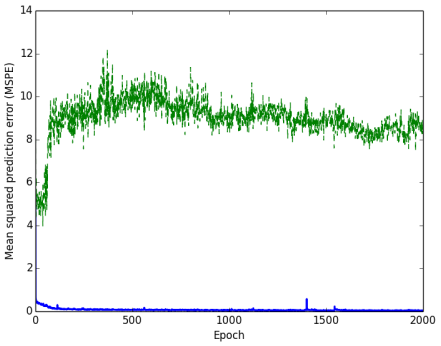
(b) Education Spending in West (controls)



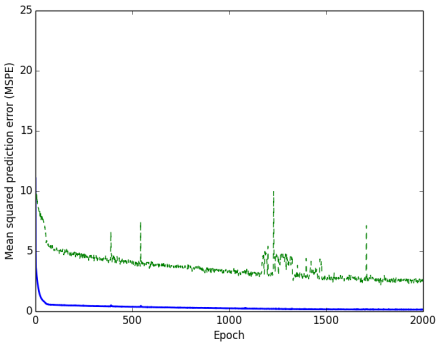
(c) Expenditure in West (treated)



(d) Expenditure in West (controls)



(e) Revenue in West (treated)

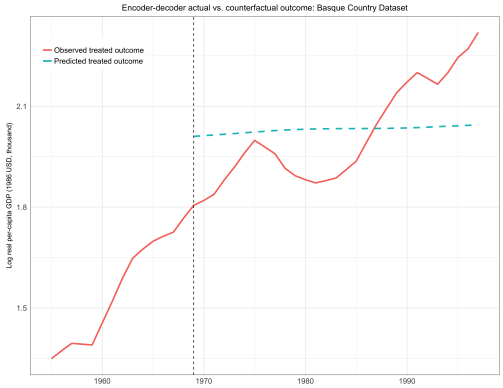


(f) Revenue in West (controls)

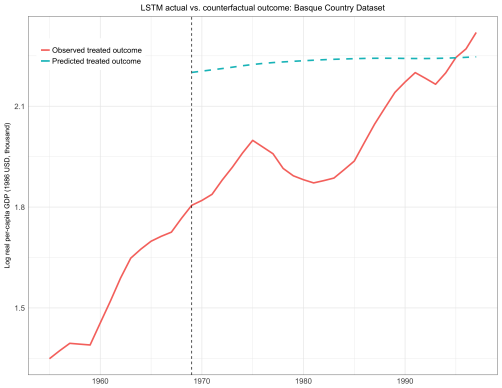
Figure 5: Encoder-decoder networks training (solid line) and validation loss (dashed line) on western public land state capacity.

6 RNNs estimates: SCM datasets

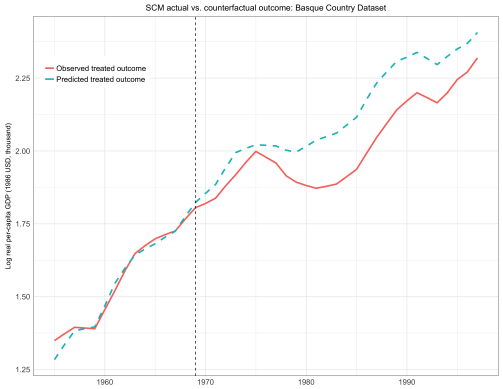
(a) Autoregressive model



(b) Encoder-decoder

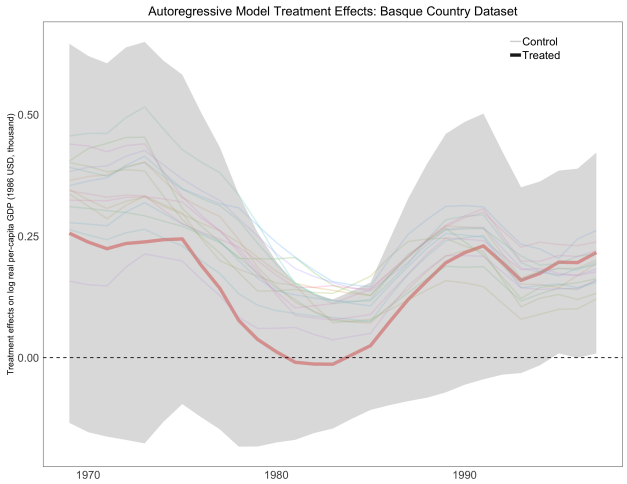


(c) LSTM

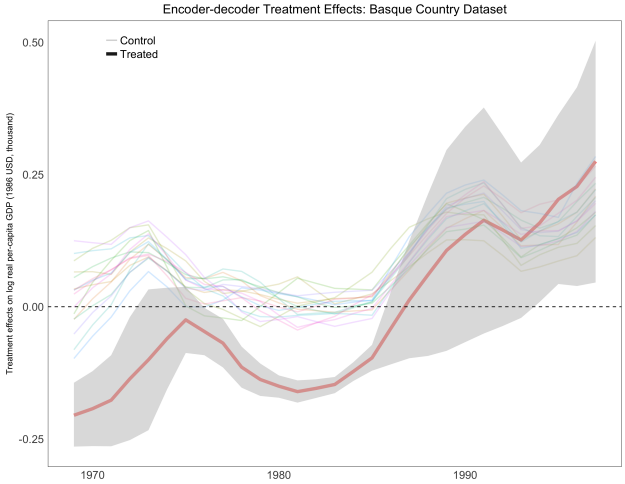


(d) SCM

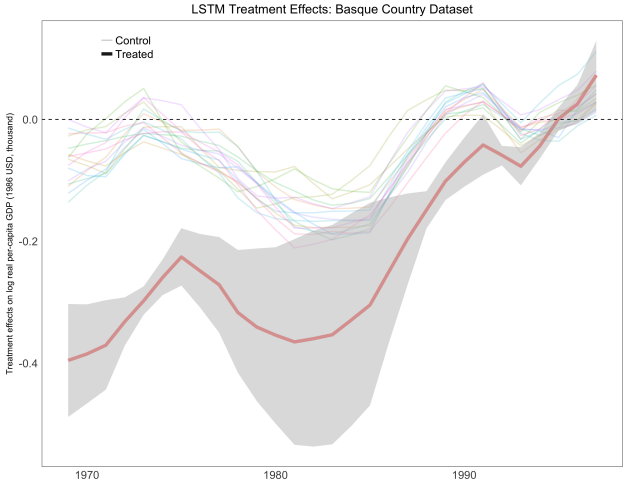
Figure 6: Observed and counterfactual predicted outcomes for treated unit in Basque Country dataset.



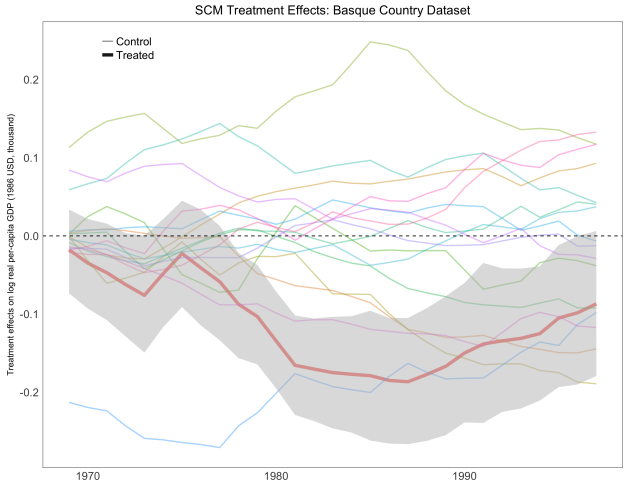
(a) Autoregressive model



(b) Encoder-decoder

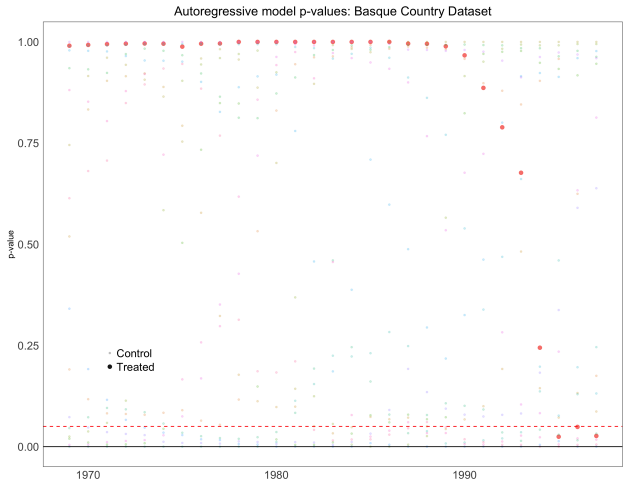


(c) LSTM

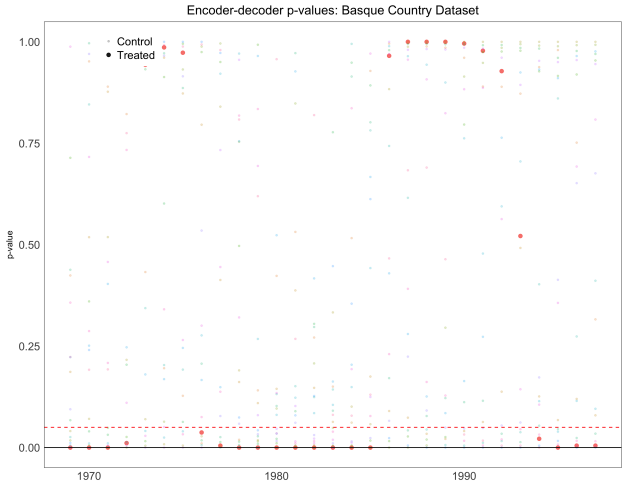


(d) SCM

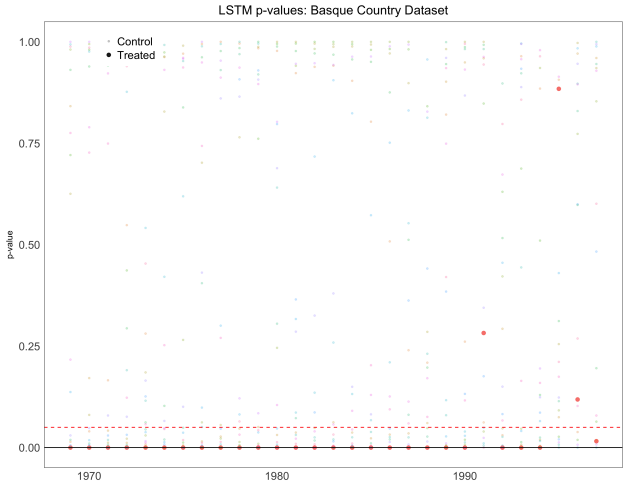
Figure 7: Time-series of post-period treatment effects in Basque Country dataset. Darker line represents the effect on the actual treated unit and each lighter line represents the effects on control units. Shaded regions represent 95% randomization confidence intervals.



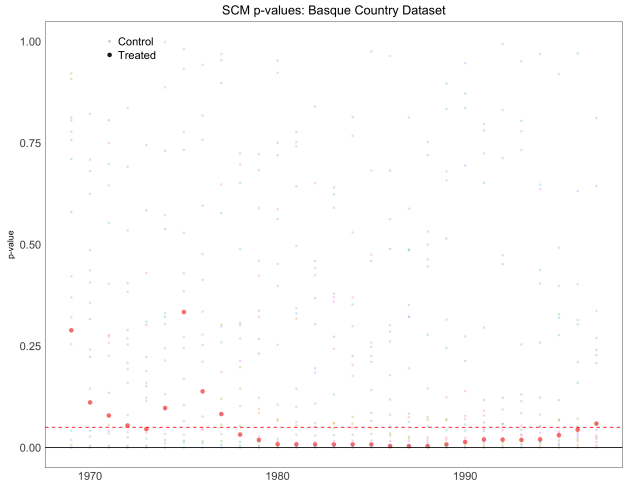
(a) Autoregressive model



(b) Encoder-decoder

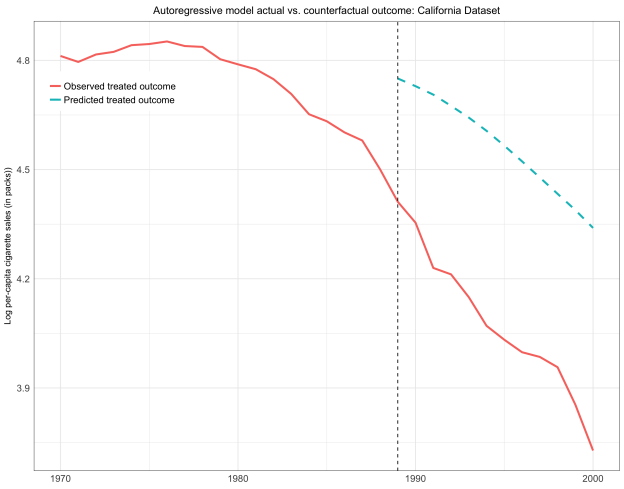


(c) LSTM

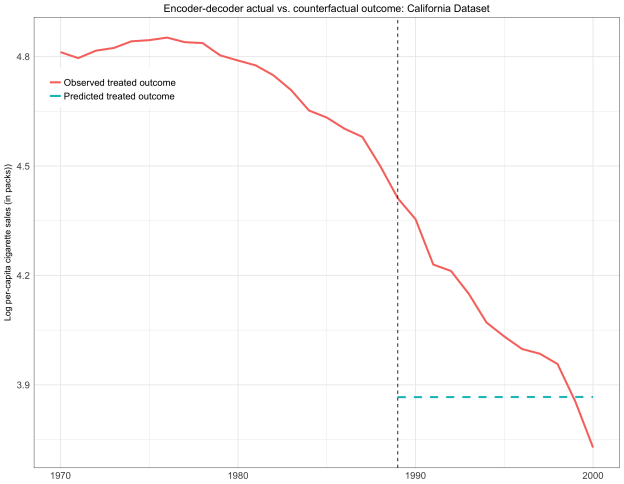


(d) SCM

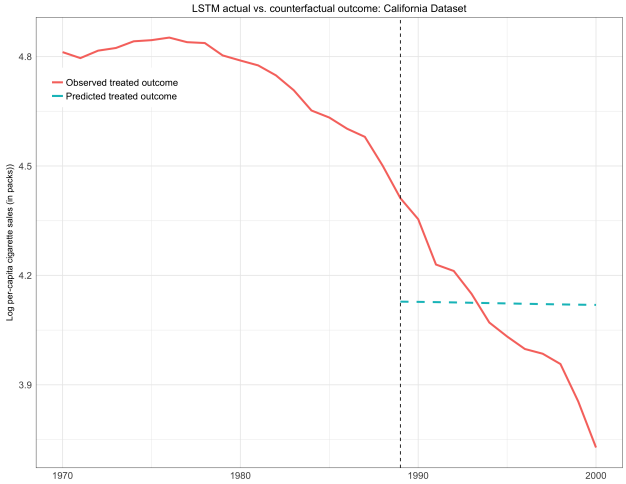
Figure 8: Per-period randomization p-values corresponding to treatment effects on treated and control units in Basque Country dataset.



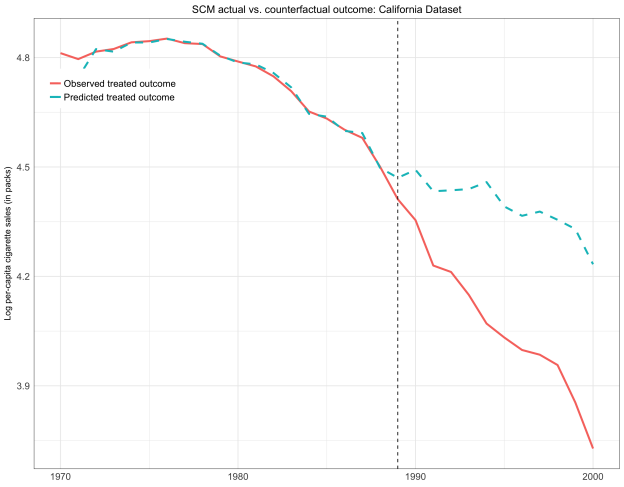
(a) Autoregressive model



(b) Encoder-decoder

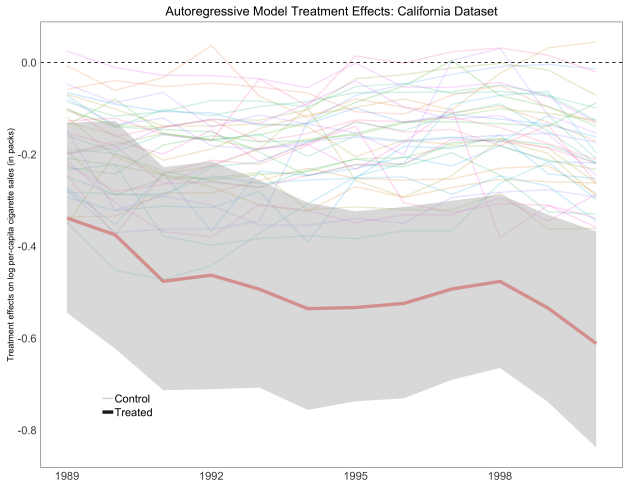


(c) LSTM

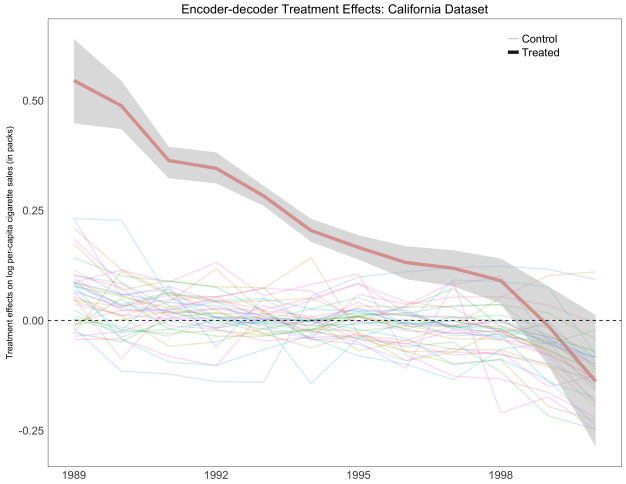


(d) SCM

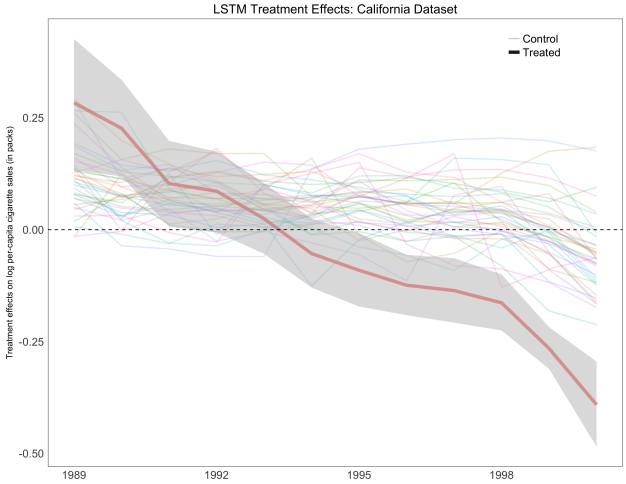
Figure 9: Observed and counterfactual predicted outcomes for treated unit in California dataset.



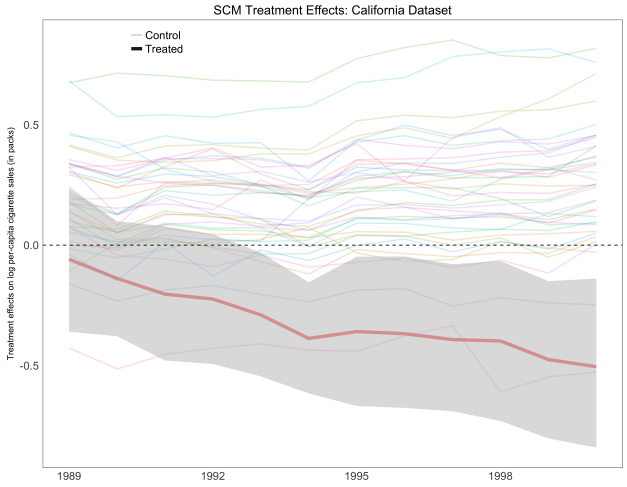
(a) Autoregressive model



(b) Encoder-decoder

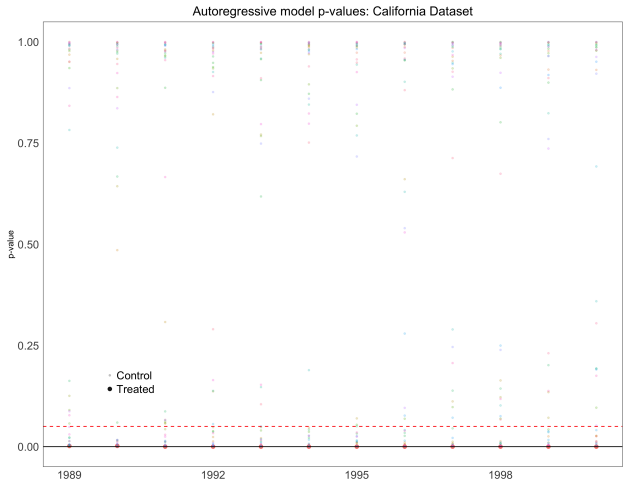


(c) LSTM

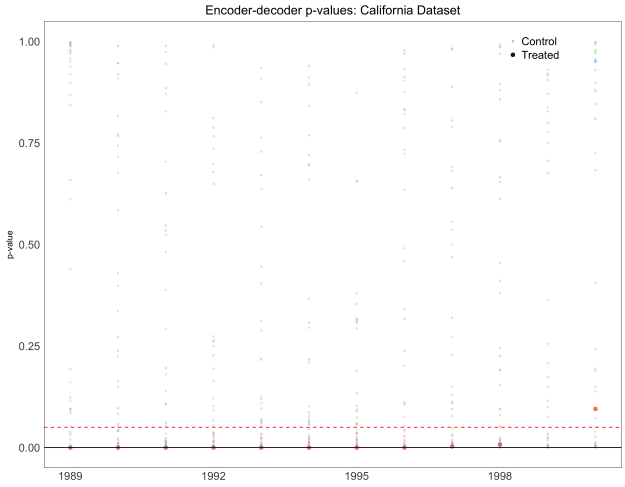


(d) SCM

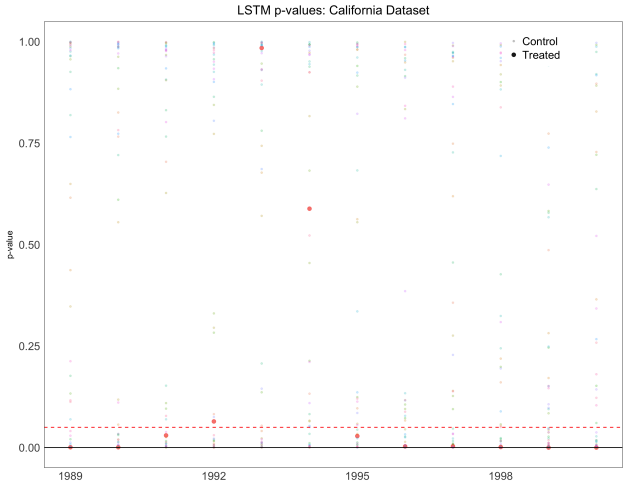
Figure 10: Time-series of post-period treatment effects in California dataset. See notes to Fig. 7.



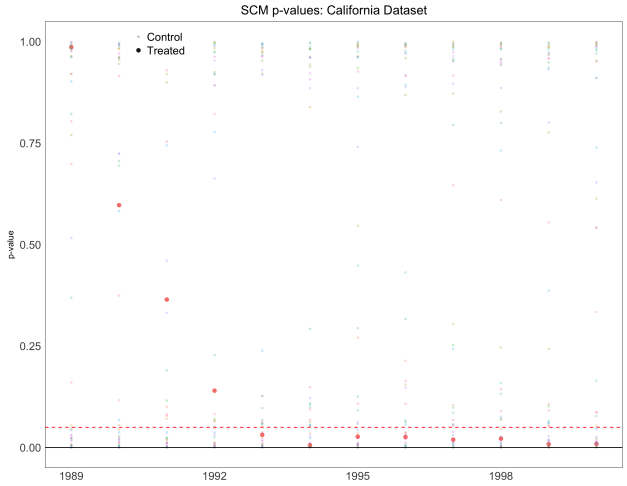
(a) Autoregressive model



(b) Encoder-decoder

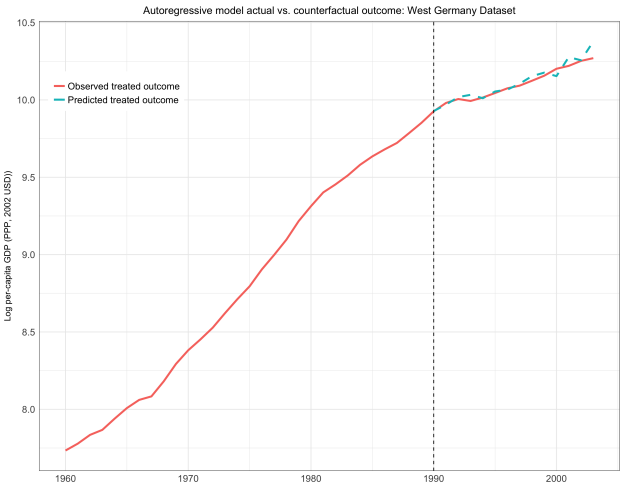


(c) LSTM

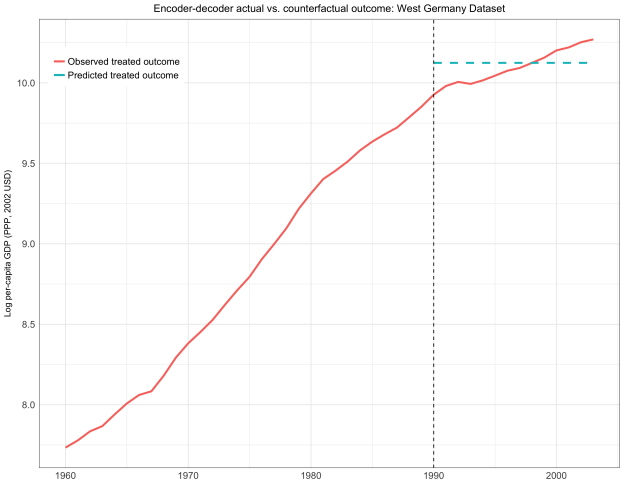


(d) SCM

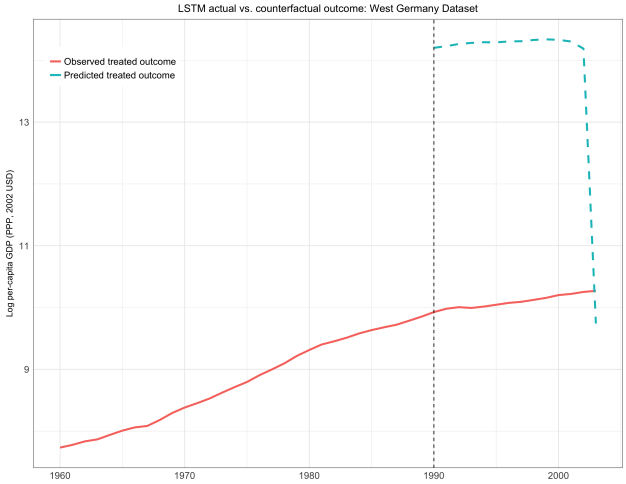
Figure 11: Per-period randomization p-values corresponding to treatment effects on treated and control units in California dataset.



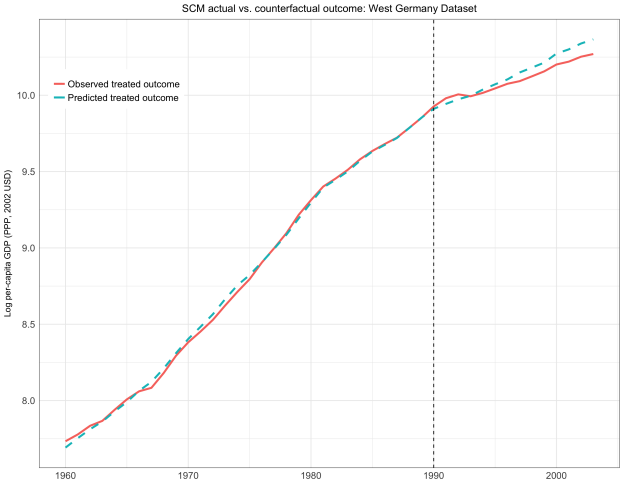
(a) Autoregressive model



(b) Encoder-decoder

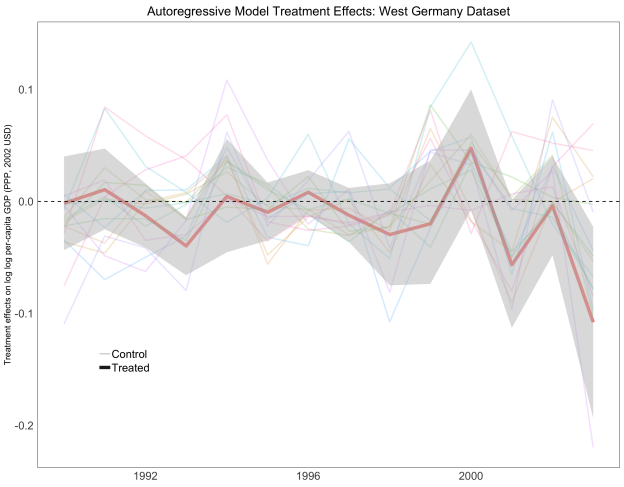


(c) LSTM

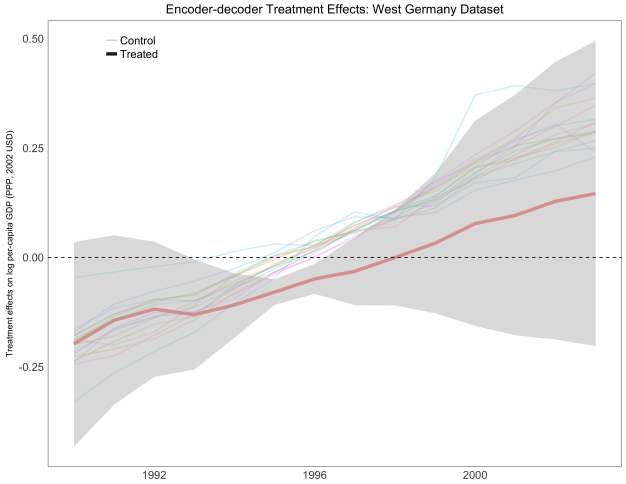


(d) SCM

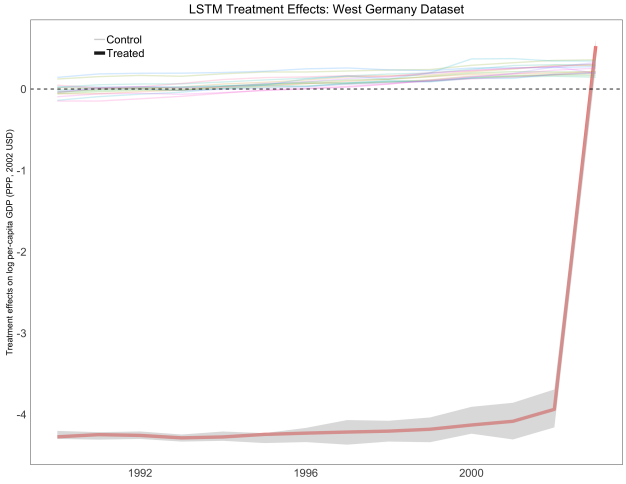
Figure 12: Observed and counterfactual predicted outcomes for treated unit in West Germany dataset.



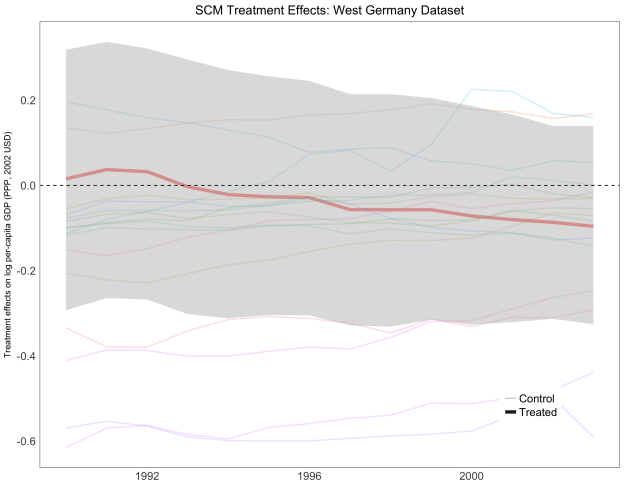
(a) Autoregressive model



(b) Encoder-decoder

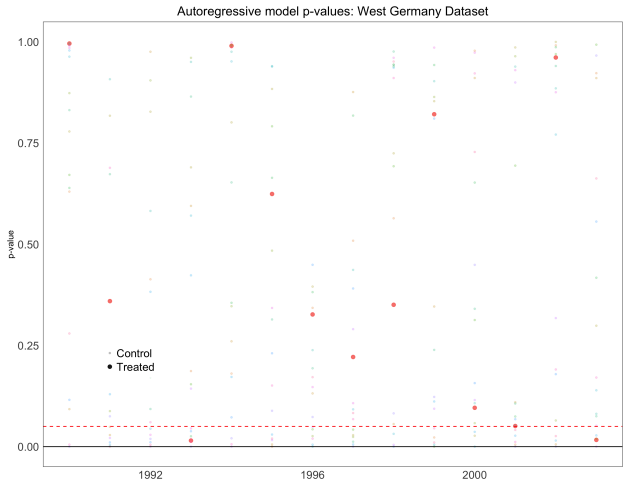


(c) LSTM

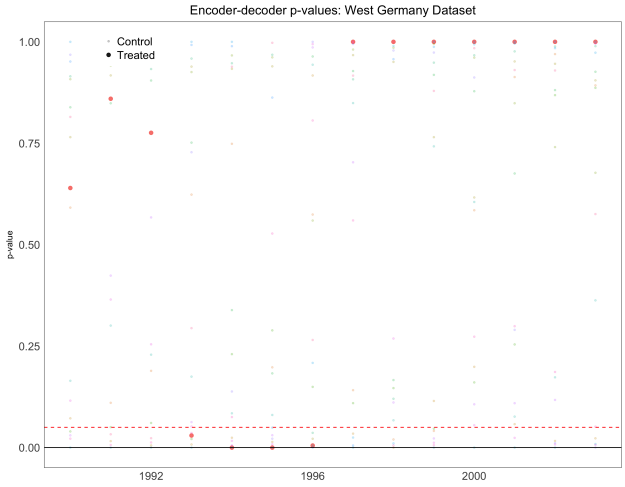


(d) SCM

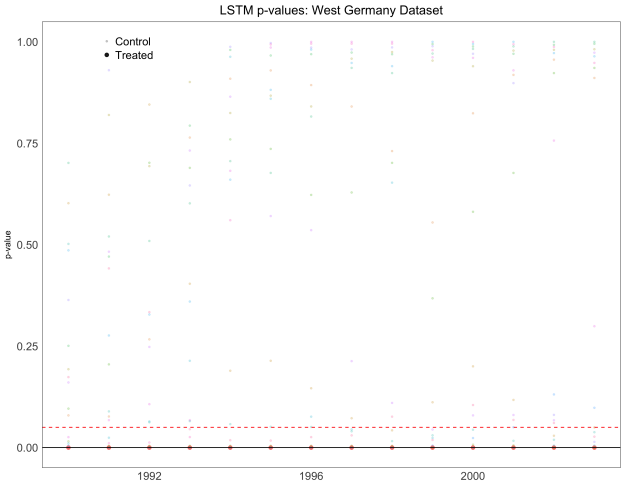
Figure 13: Time-series of post-period treatment effects in West Germany dataset. See notes to Fig. 7.



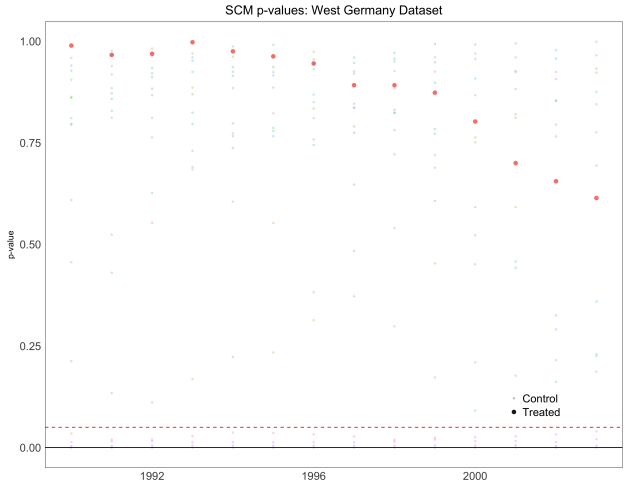
(a) Autoregressive model



(b) Encoder-decoder



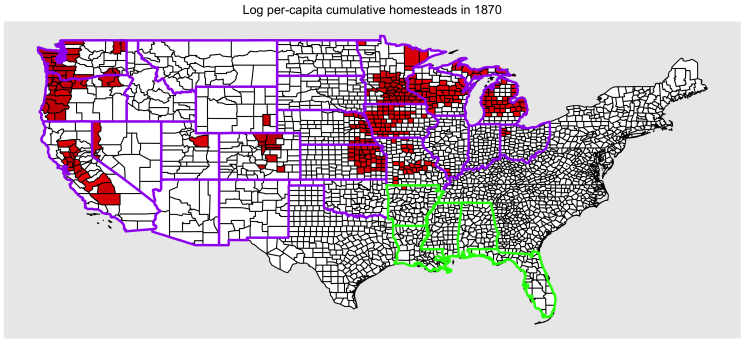
(c) LSTM



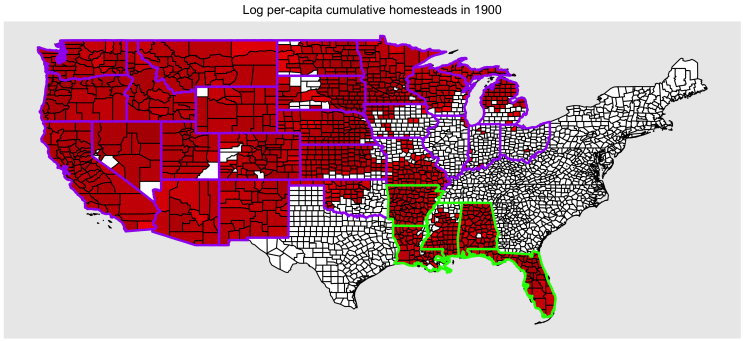
(d) SCM

Figure 14: Per-period randomization p-values corresponding to treatment effects on treated and control units in West Germany dataset.

7 Exploratory data analysis



(a) 1870



(b) 1900

Figure 15: Log per-capita cumulative homesteads in 1870 and 1900, overlaid on 1911 county borders. Darker-colored counties have more higher values than lighter-colored counties, and white-colored counties have missing values. States bordered in green are southern public land states and those bordered in purple are western public land states. County border data from Long (1995).

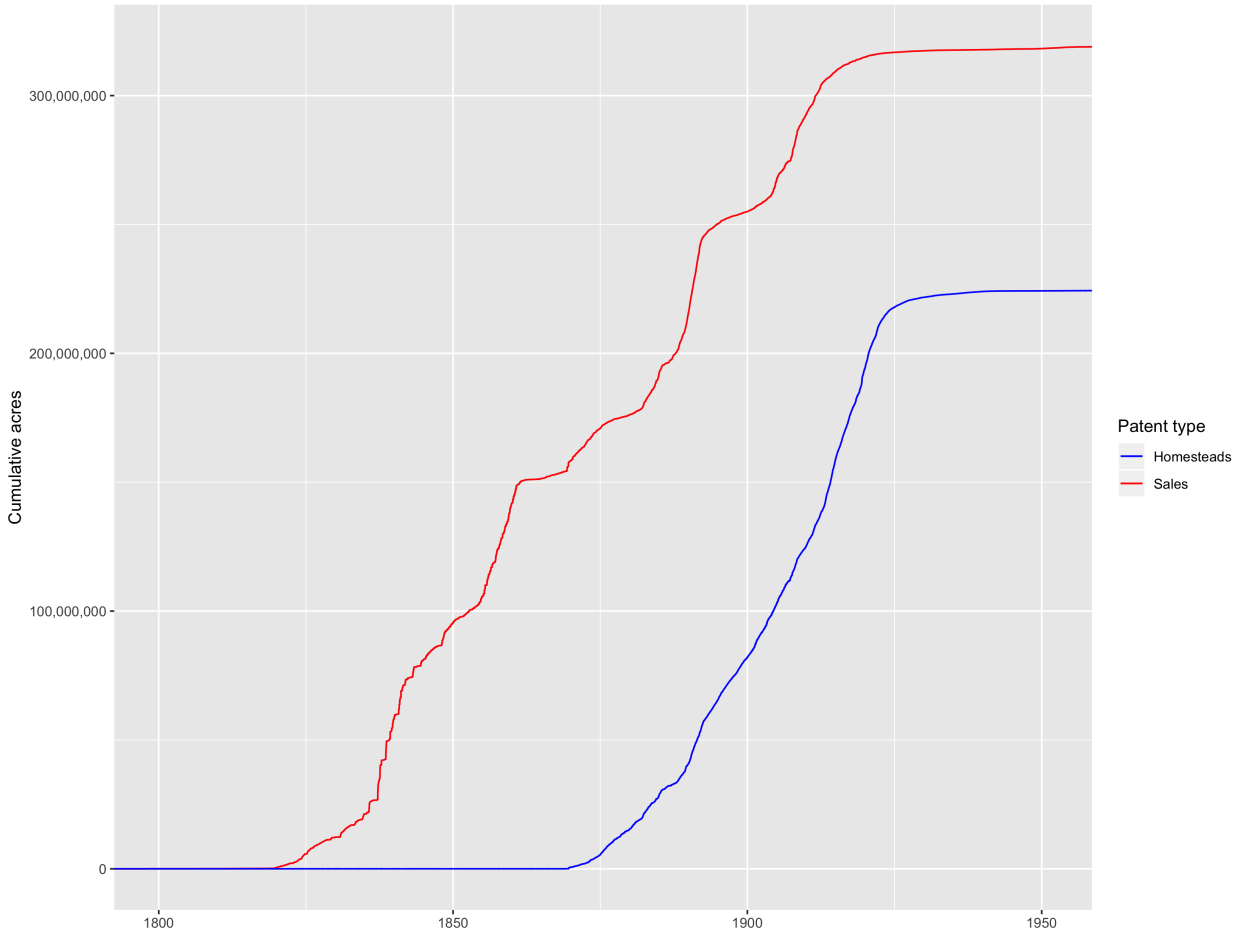


Figure 16: Cumulative total acres (by patent type) disbursed in public land states, 1800 - 1950.

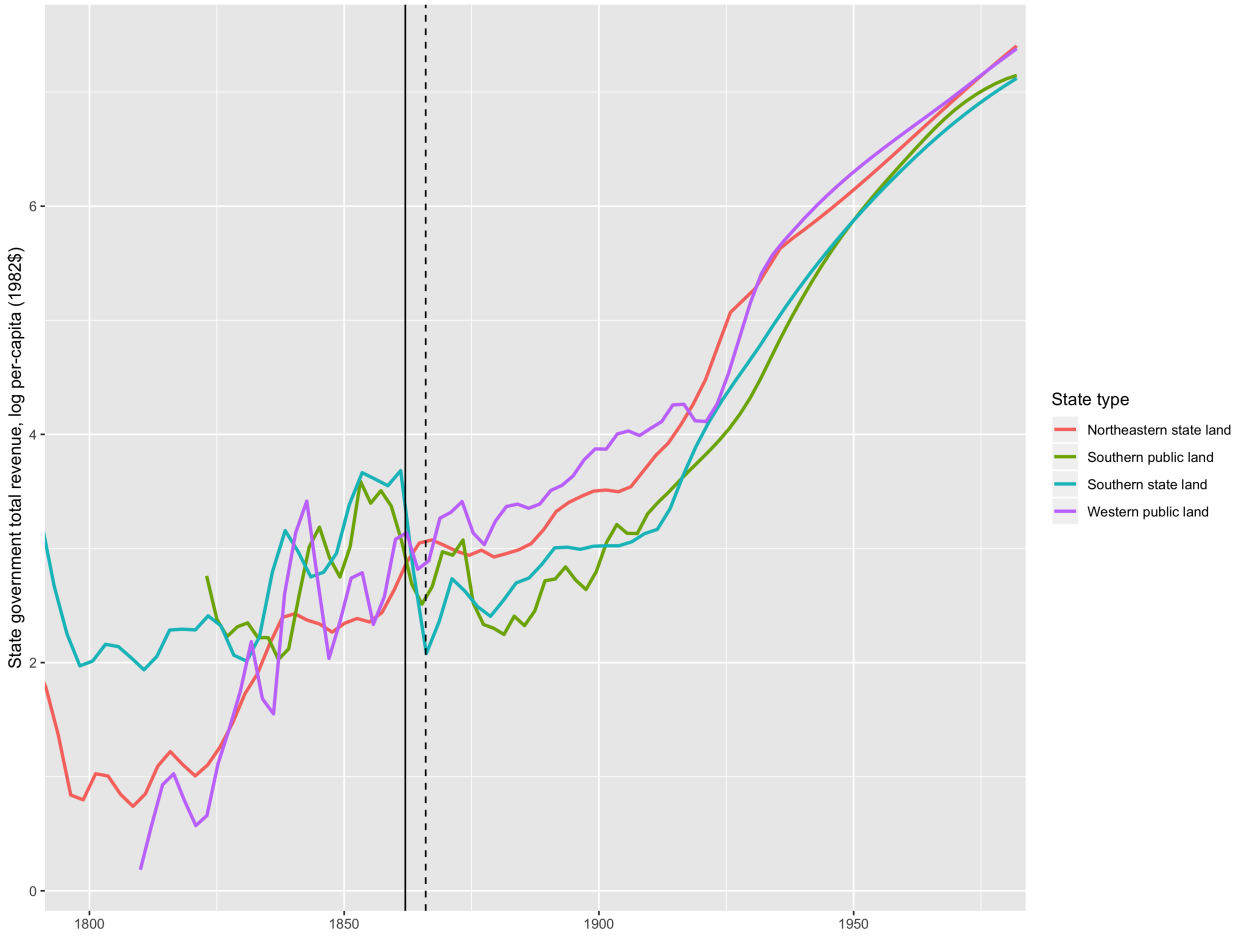


Figure 17: Log per-capita state government revenue by state group, 1800-1975. The solid vertical line represents the 1862 HSA and the short-dashed vertical line represents the 1866 SHA. Time-series curves are smoothed by LOESS regression.

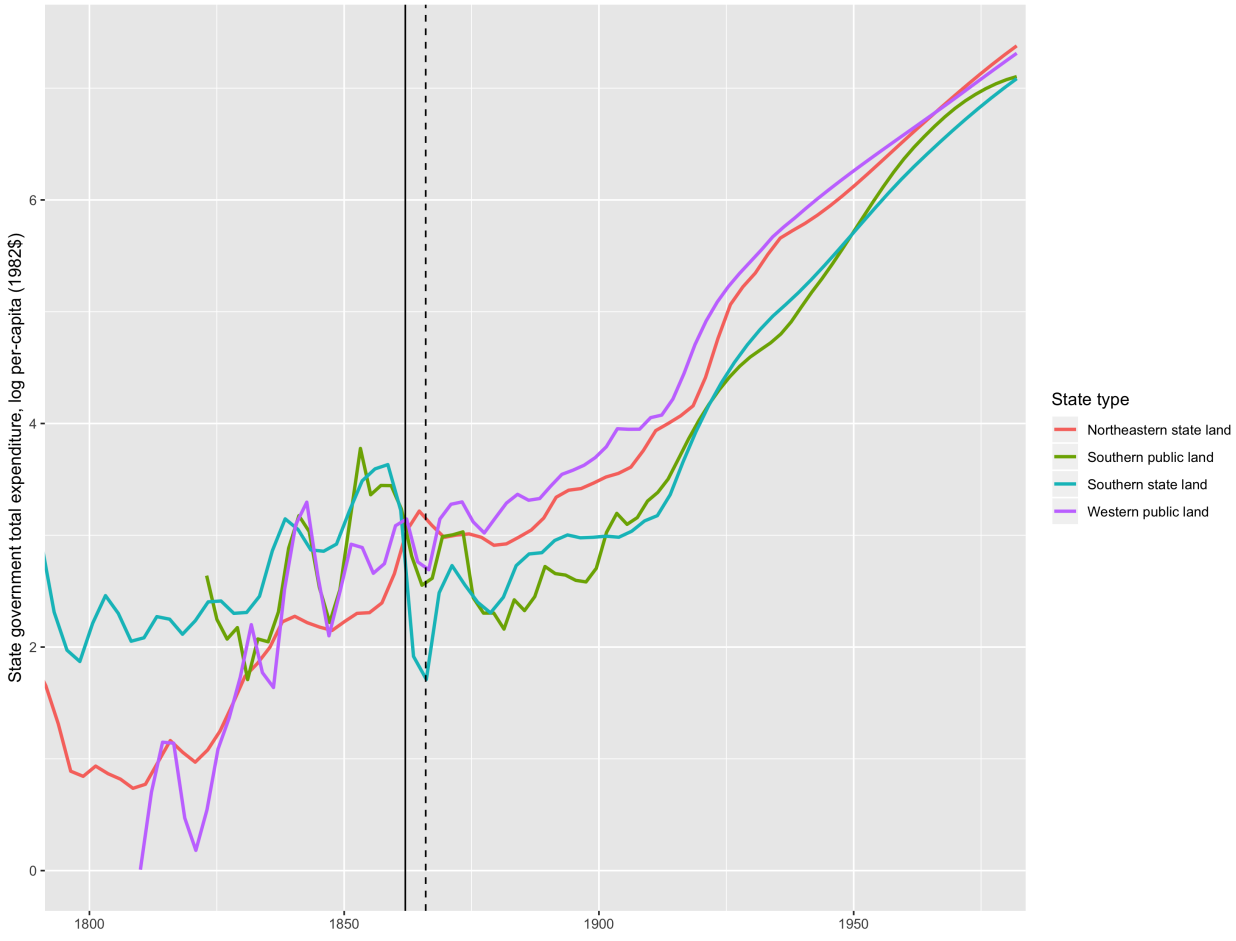


Figure 18: Log per-capita state government expenditures by state group, 1800-1975. See notes to Fig. OA-17.

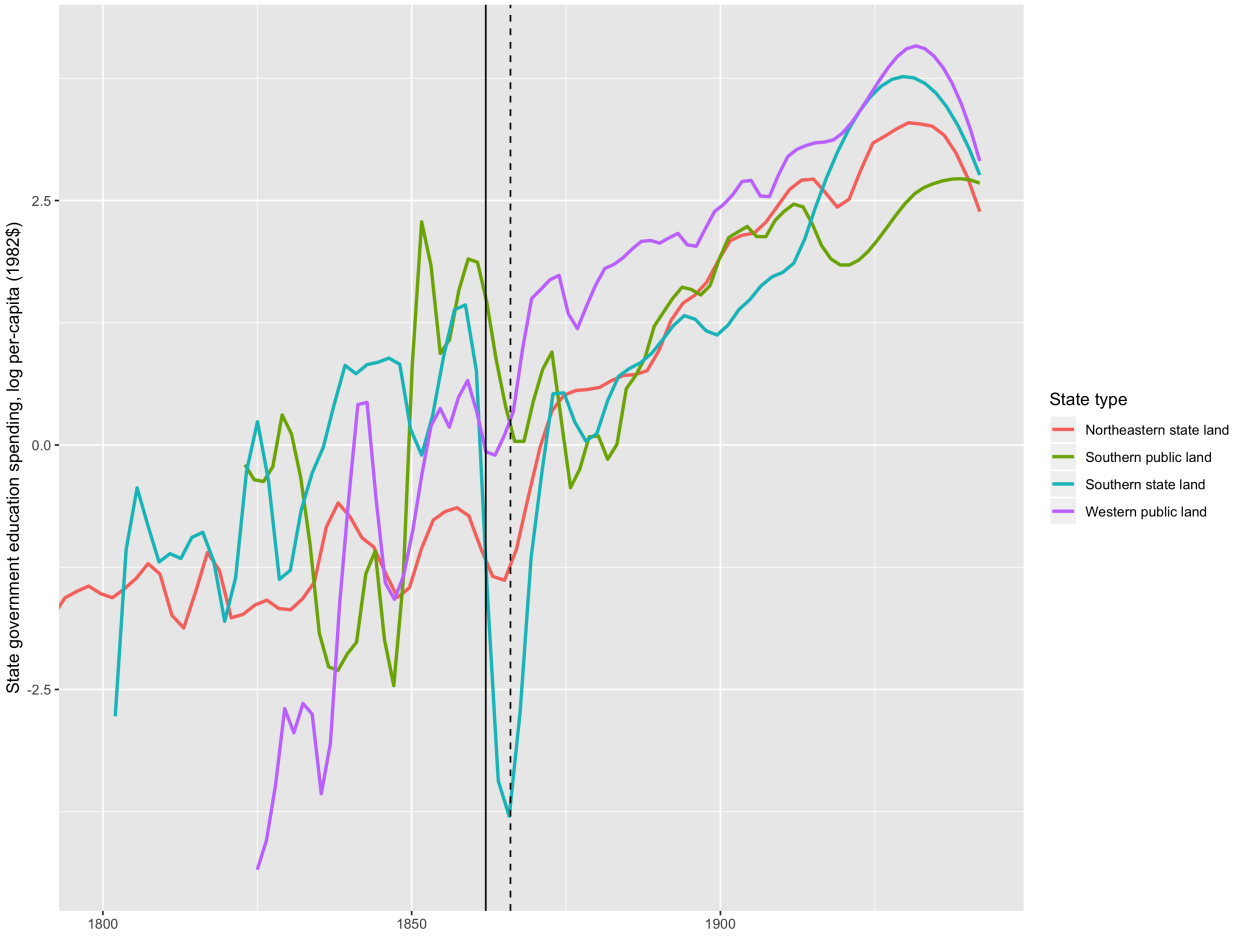


Figure 19: Log per-capita state government expenditures by state group, 1800-1942. See notes to Fig. OA-17.

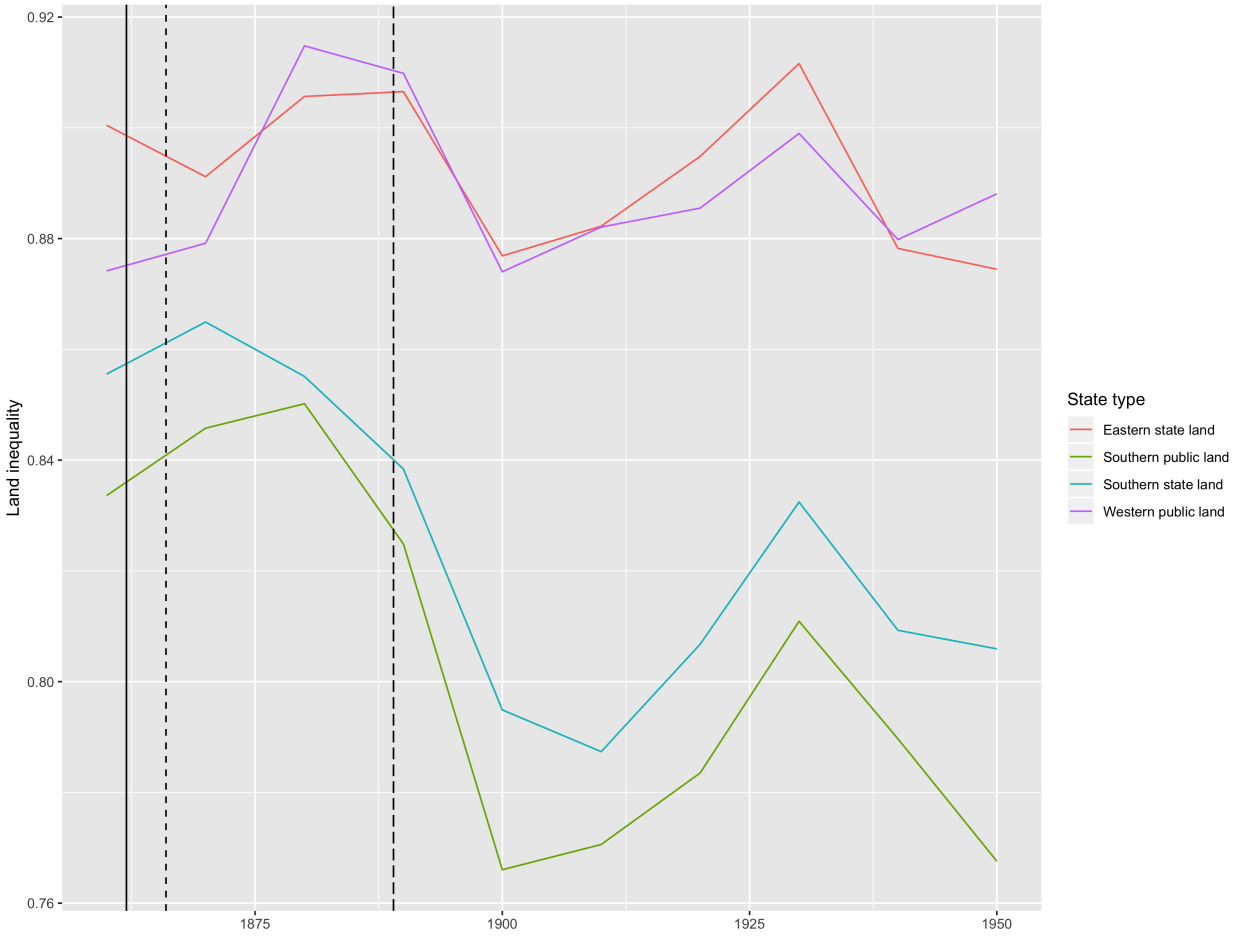


Figure 20: Land inequality by state group, 1860-1950. The solid vertical line and short-dashed line represents the passage of the 1862 HSA and 1866 SHA, respectively. The long-dashed vertical line represents the 1889 cash-entry restriction.

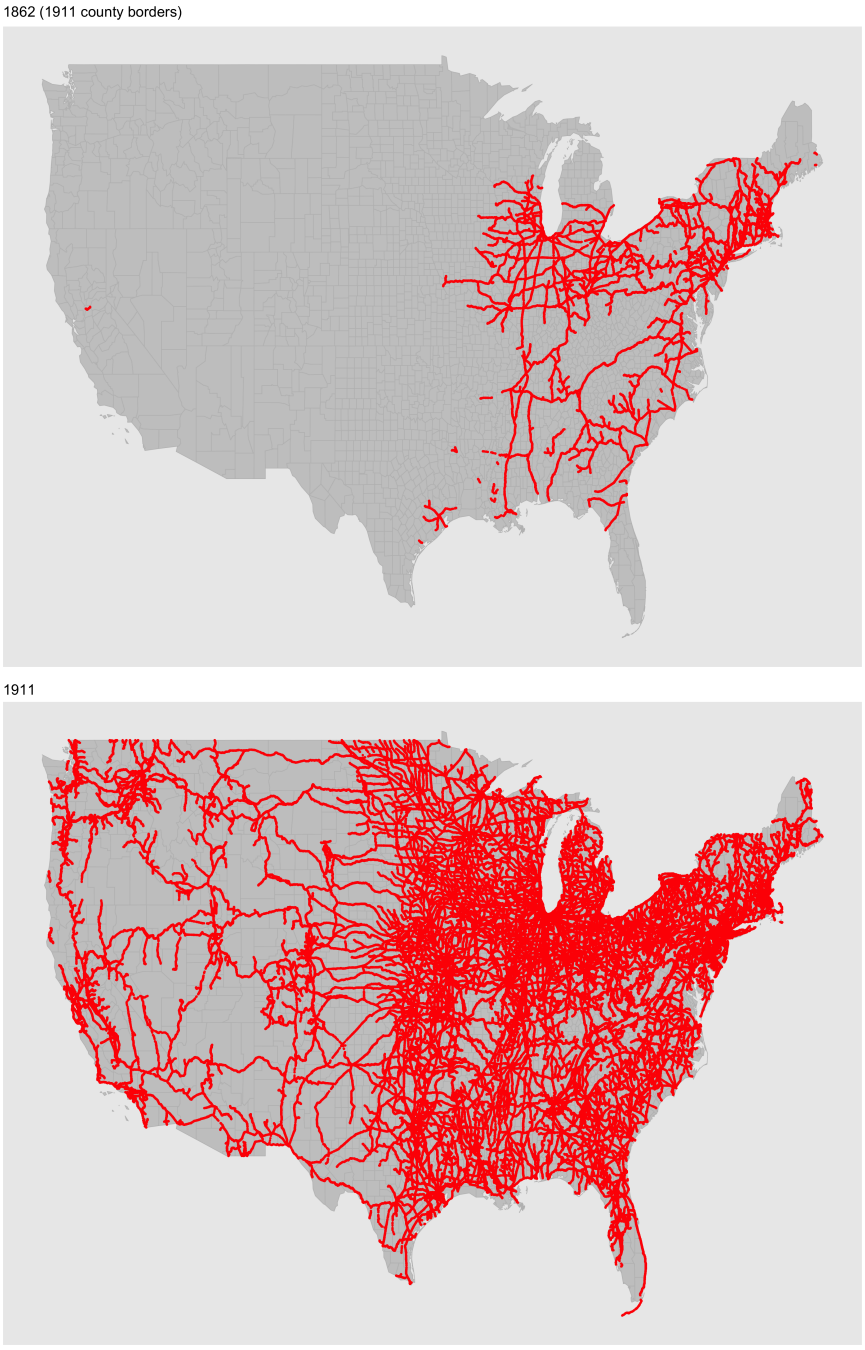


Figure 21: Railroad lines in 1862 and 1911, overlaid on 1911 county borders. Railroad data from Atack 2013 and county border data from Long 1995.

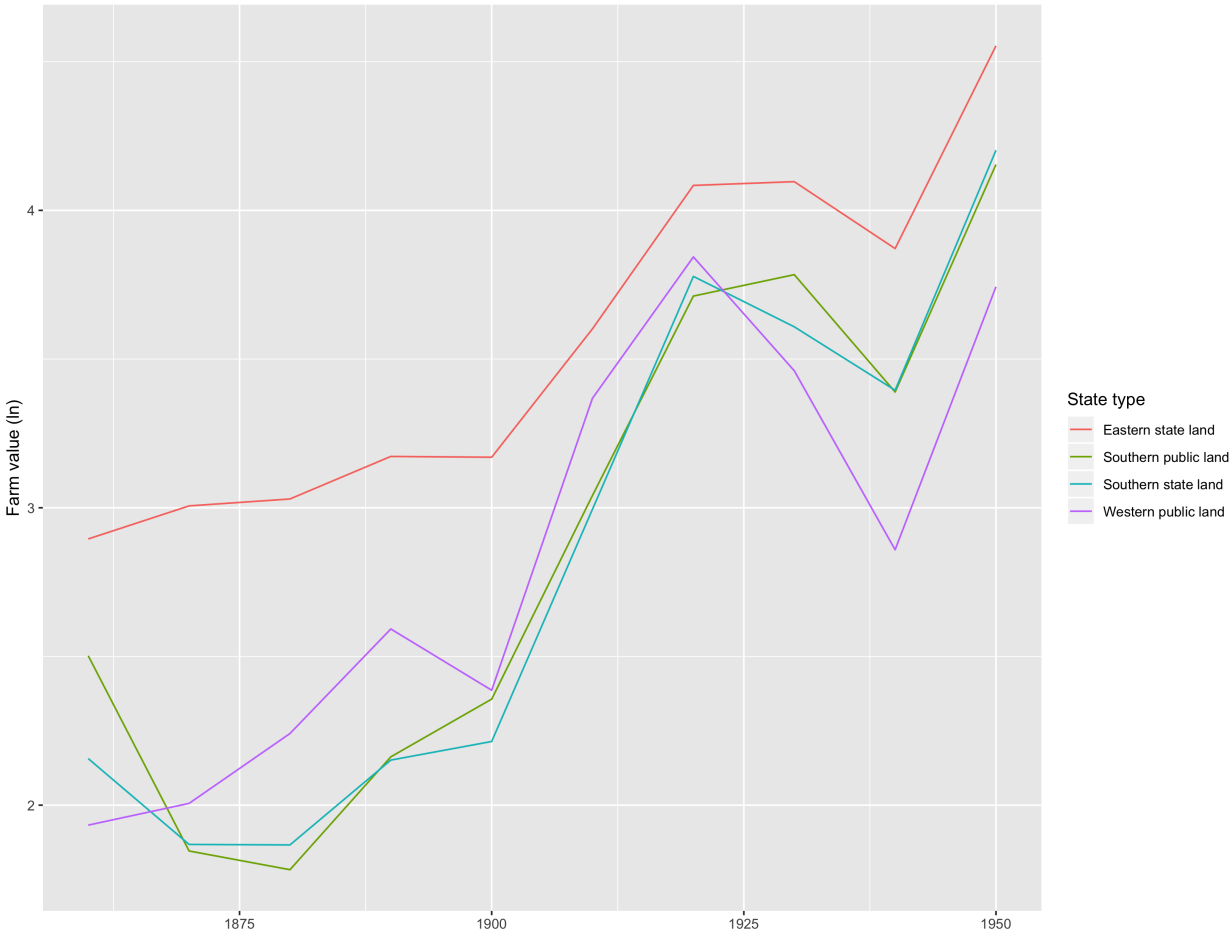


Figure 22: Farm values by state group, 1860 - 1950.

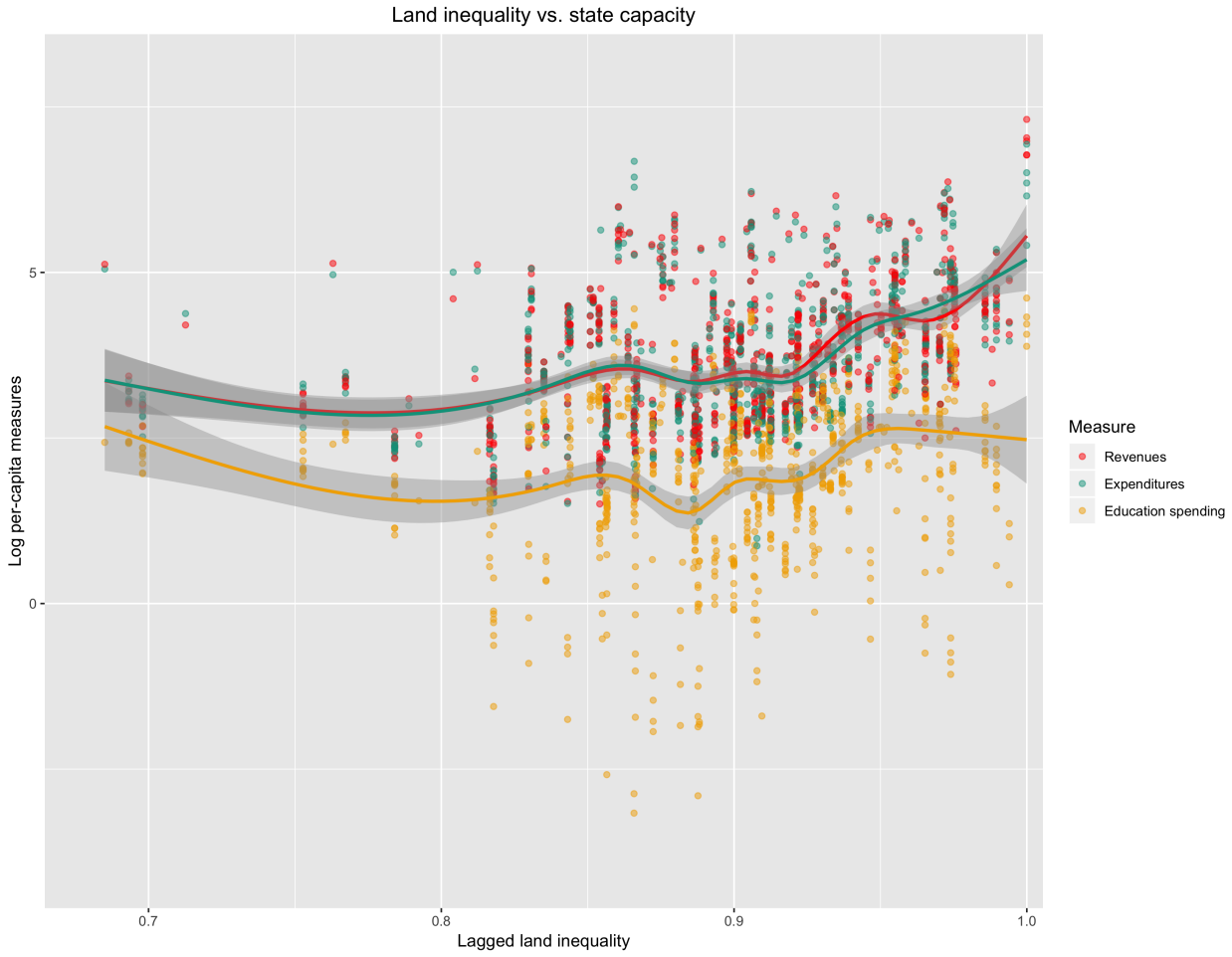


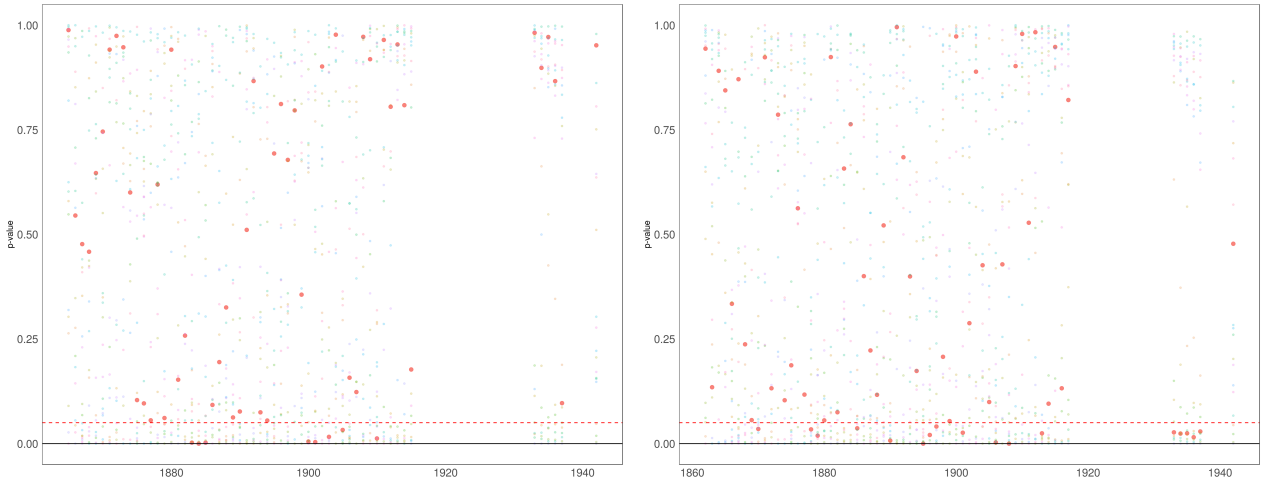
Figure 23: Land inequality vs. log per-capita revenues and expenditures at the state-level, 1860-1950. Each point is a state-year observation. Lines represent generalized additive model (GAM) fits to the data and shaded regions represent corresponding 95% confidence intervals.

8 RNNs estimates: State capacity

Table 2: Encoder-decoder FPR and MSPE on state capacity placebo tests.

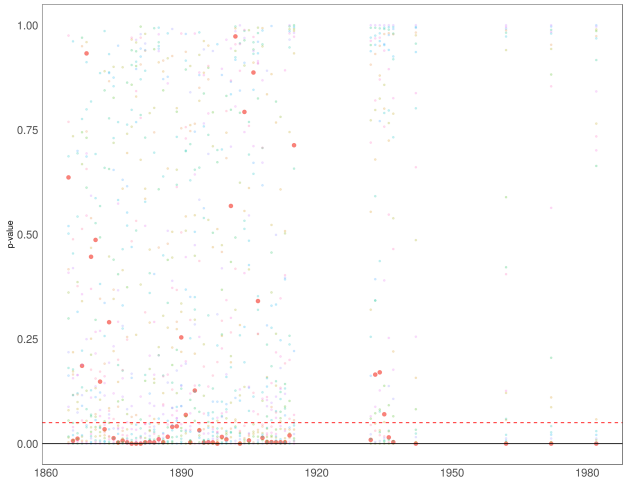
Outcome \ Measure	MSPE		FPR	
	South	West	South	West
Education spending	0.44 ± 0.59	0.37 ± 0.53	0.23	0.23
Expenditure	0.75 ± 0.2	0.47 ± 0.13	0.31	0.31
Revenue	0.8 ± 0.2	0.48 ± 0.2	0.28	0.28

Error bars represent $\pm$ one standard deviation from the MSPE.



(a) Education spending in South

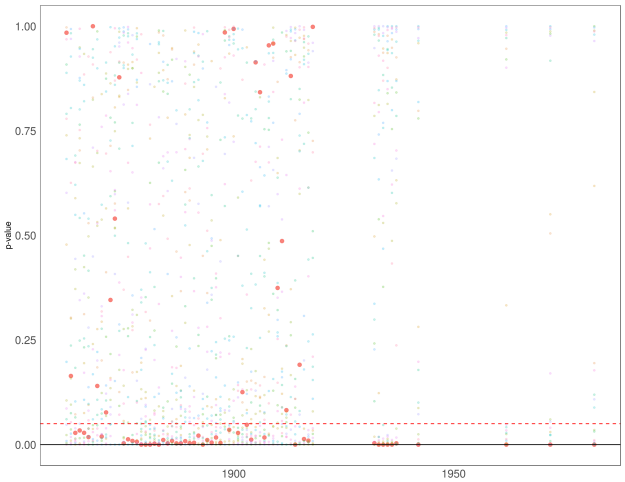
(b) Education spending in West



(c) Expenditure in South

(d) Expenditure in West

(e) Revenue in South



(f) Revenue in West

Figure 24: Encoder-decoder networks: Per-period randomization p-values corresponding to treatment effects on treated and control units in state capacity datasets. Darker dot represents p-values associated with treatment effects on the actual treated unit and lighter dots represent p-values associated with the effects on control units

Table 3: LSTM FPR and MSPE on state capacity placebo tests.

Outcome \ Measure	MSPE		FPR	
	South	West	South	West
Education spending	0.55 ± 0.65	0.56 ± 0.83	0.22	0.19
Expenditure	6.66 ± 5.77	0.73 ± 0.23	0.36	0.27
Revenue	1.5 ± 3.37	1.68 ± 3.72	0.09	0.15

Error bars represent $\pm$ one standard deviation from the MSPE.

9 Difference-in-difference estimates: State capacity

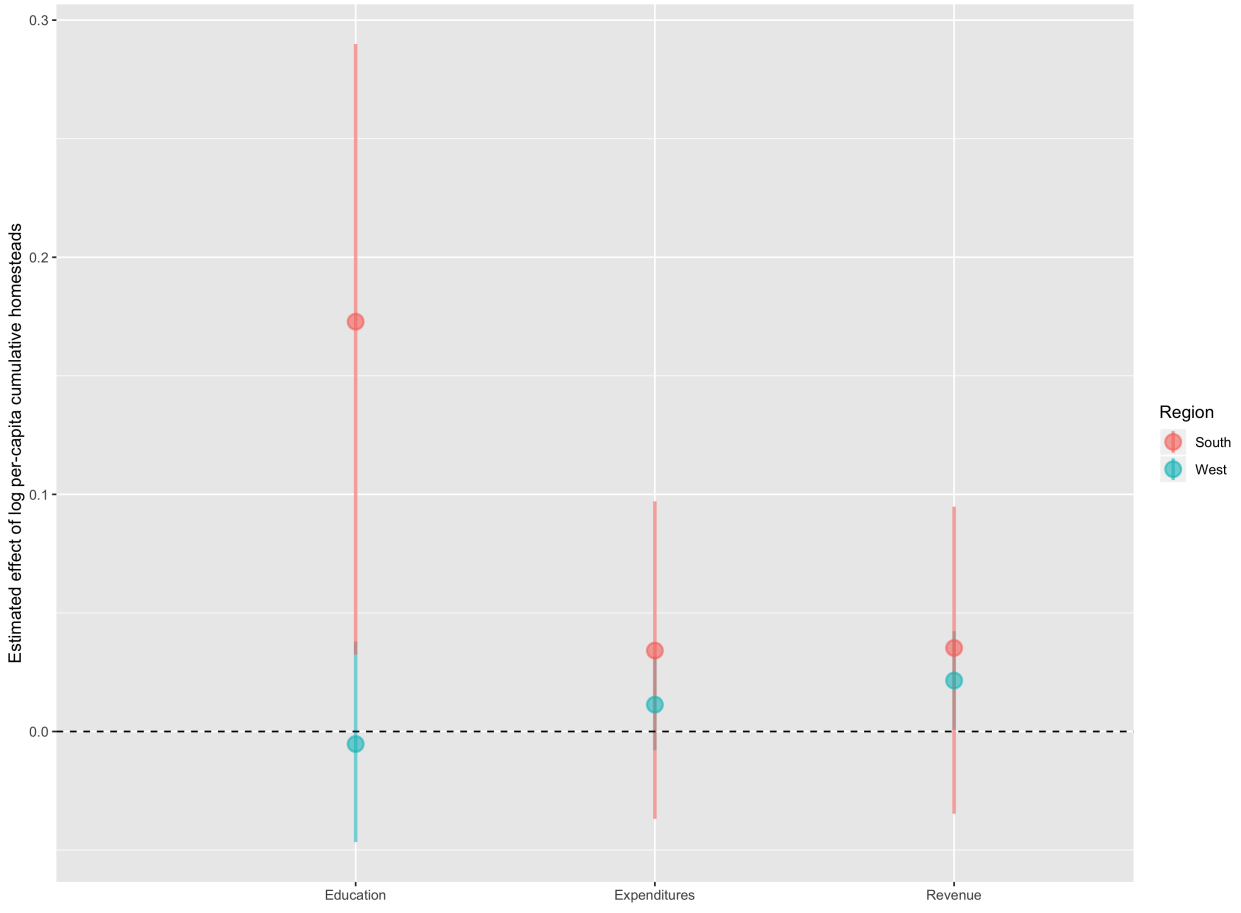


Figure 25: Difference-in-difference estimates of log per capita homesteads on state-level log per-capita education spending, expenditures, and revenues. Log average farm values is included in the regression. Region is western public land states or southern public land states. 95% confidence intervals are constructed using 1,000 state-stratified bootstrap samples.

Outcome \ Region	South	West
Land inequality	-0.001 [-0.003, 0.0004], $N = 523$	-0.004 [-0.005, -0.002], $N = 2,002$
Railroad access	0.03 [0.01, 0.05], $N = 350$	0.09 [0.07, 0.1], $N = 1,053$

Table 4: DD estimates: impact of log per-capita cumulative homesteads on land inequality and railroad access in public land state counties. See notes to Fig. 25.

References

Atack, Jeremy. (2013). "On the Use of Geographic Information Systems in Economic History: The American Transportation Revolution Revisited." *The Journal of Economic History* 73, no. 2 (2013): 313–338.

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

Haines, Michael 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, 2010.

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

Long, John H. (1995). "Atlas of Historical County Boundaries." *The Journal of American History* 81, no. 4 (1995): 1859–1863.

Sylla, Richard E, John B Legler, and John Wallis. (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, 1993.

———. (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, 1995.

———. (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, 1995.

Vollrath, Dietrich. (2013). "Inequality and School Funding in the Rural United States, 1890." *Explorations in Economic History* 50, no. 2 (2013): 267–284.