Building State Capacity through Public Land Disposal: An Application of Matrix Completion for Counterfactual Prediction

Jason Poulos

University of California, Berkeley

Abstract

How would the frontier have evolved in the absence of mid-nineteenth century homestead acts? I propose using recurrent neural networks (RNNs) to predict the counterfactual time-series of frontier state capacity had there been no homestead acts. Time-specific estimates signify that homestead acts positively impacted state government finances 50 years following their implementation. Exploiting variation in the intensity of homestead entries aggregated from 1.46 million individual land patents, difference-in-differences estimates imply that homesteads significantly increased state government revenue and education spending over a period extending into the twentieth century.

Address for correspondence: Department of Political Science, 210 Barrows Hall #1950, Berkeley, CA 94720-1950. *Email:* poulos@berkeley.edu. I thank Sean Gailmard, Eric Schickler, Ross Mattheis, and Shom Mazumder for helpful comments. I acknowledge support of the National Science Foundation Graduate Research Fellowship (DGE 1106400). This work used the computer resources of Stampede2 at the Texas Advanced Computing Center (TACC) under an Extreme Science and Engineering Discovery Environment (XSEDE) startup allocation (TG-SES180010).

Political scientists are increasingly interested in patterns of state development across time and place. Several scholars (e.g., Bensel, 1990; Murtazashvili, 2013; Frymer, 2014) theorize a relationship between mid-nineteenth century public land laws and the development of the national government. It is argued that laws designed to transfer public land to private individuals increased the bureaucratic capacity of the federal government to administer land and also reduced enforcement costs.

I argue that public land laws had long-lasting impacts on state capacity, or the ability of state governments to finance and implement policies (Besley and Persson, 2010). I explore the role of two major land policies in shaping state capacity: the Homestead Act (HSA) of 1862, which opened for settlement hundreds of millions of acres of western frontier land, and the Southern Homestead Act (SHA) of 1866, which opened over 46 million acres of land for homesteading.[1] I provide evidence that homesteads authorized under these laws had long-run positive impacts on the capacity of frontier state governments.

The view that the western frontier had long-lasting impacts on the evolution of democratic institutions can be traced to Turner (1956). Turner's "frontier thesis" posited that homestead laws acted as a "safety valve" for relieving pressure from congested urban labor markets in eastern states. The view of the frontier as a "safety valve" has been explored by Ferrie (1997), who finds evidence in a linked census sample of substantial migration to the frontier by unskilled workers and considerable gains in wealth for these migrant workers. Homestead laws not only offered greater economic opportunities to eastern migrants, but also the scarcity of people on the western frontier meant that state and local governments competed with each other to attract migrants in order to lower labor costs for businesses and to increase land values and tax revenues (Engerman and Sokoloff, 2005). Frontier governments offered migrants broad access to cheap land and property rights, unrestricted voting rights, and a more generous provision of schooling and other public goods. As Engerman and Sokoloff

[1]I use the terminology of "frontier" states interchangeably with "public land states" throughout the paper. Public land states are states created out of the public domain. In the South, these states include Alabama, Arkansas, Florida, Louisiana, and Mississippi. Western public land states include the 25 states that comprise the Midwestern, Southwestern, and Western U.S. (except Hawaii).

(2005) notes, the commitment of frontier state governments to offer greater political and economic opportunities for migrants was more prevalent in western states than in the South.

Garcia-Jimeno and Robinson (2009) test the frontier thesis in a global context and conclude that the economic effect of the frontier depends on the quality of political institutions at the time of frontier expansion: frontier expansion promoted equitable outcomes only when societies were initially democratic; however, when institutional quality is weak, the existence of frontier land can yield worse development outcomes because non-democratic political elites can consolidate frontier lands for themselves. Historical institutional quality as a function of fiscal capacity and newspaper entry was considerably lower in southern public land states compared to southern border states and northern states (Grosjean, 2014). Regional differences in institutional quality may explain the empirical findings of the present paper: estimates from a difference-in-differences (DID) model that leverages variation in both the timing and intensity of homesteads show that homesteads significantly lowered land inequality in western frontier counties, but had no significant impact on land inequality in southern counties.

The political economy literature is largely in agreement that inequality and state capacity are inversely related. State capacity refers not only to the ability to raise revenue, but also the state's ability to implement policies such as public education through redistributive spending (Besley and Persson, 2010). The canonical model of Meltzer and Richard (1981) predicts a positive relationship between inequality and redistribution because greater inequality implies the median voter is poorer than the average voter, which in turn increases demand for redistribution in majority-rule elections. However, models that allow for differences in political influence across economic groups predict an inverse relationship. In Benabou's (2000) model, the pivotal voter is wealthier than the median and has the power to block redistribution as inequality increases. In Besley and Persson's (2009) framework, greater economic power of the ruling class reduces investment in state capacity. Landed elites might choose an inefficient organization of the state in order to create inefficiencies in

tax collection (Acemoglu et al., 2011) or "hollow-out" tax institutions in order to constrain the state's ability to tax in the future (Suryanarayan, 2017). Similarly, Galor et al. (2009) propose a model where wealthy landowners block education reforms because education favors industrial labor productivity and decreases the value in farm rents. Inequality in this context can be thought of as a proxy for the amount of *de facto* political influence elites have to block reforms and limit the capacity of the state (Acemoglu and Robinson, 2008).[2]

The paper makes a methodological contribution in applying an alternative to the synthetic control method (SCM) (Abadie et al., 2010) for inferring the effect of a policy intervention on observational time-series data. Building on a new literature that uses machine learning algorithms such as L1-regularized linear regression (Doudchenko and Imbens, 2016) or deep neural networks (Poulos, 2017) for counterfactual prediction, the proposed method applies the matrix completion method (MCM) proposed by Athey et al. (2017) to predict the treated unit time-series in the absence of the intervention. I perform placebo tests and find that the matrix completion method outperforms the SCM in terms of minimizing prediction error.

The paper proceeds as follows: in the section below, I overview the historical context of homestead laws and their relationship to state capacity and two potential causal mechanisms: land inequality and railroad development; Section 3 describes the proposed method of RNN-based counterfactual time-series prediction. Section **??** proposes a method of statistical inference and evaluates the proposed method against the SCM. In Section 4, I use the proposed method to estimate the long-run impacts of homestead acts on state capacity. Section 5 reports DID estimates of the effect of cumulative homesteads on state capacity, land inequality, and railroad access. Section 6 concludes.

[2]The inverse relationship between land inequality and state capacity has been empirically demonstrated in the context of taxes, revenues, and public school spending at the county-level in 1890 and 1930 ((Ramcharan, 2010); (Vollrath, 2013)).

2 Historical background

The 1862 HSA opened up hundreds of millions of acres of western public land for settlement. The HSA provides that any adult citizen — including women, immigrants who had applied for citizenship, and freed slaves following the passage of the Fourteenth Amendment— could apply for a homestead grant of 160 acres of frontier land. Applicants were required to live and make improvements on the land for five years before filing to claim a homestead land grant. Under the HSA, the bulk of newly surveyed land on the western frontier was reserved for homesteads, although the law did not end sales of public land. The explicit goal of the HSA was to liberalize the homesteading requirements set by the Preemption Act of 1841, which permitted individuals already inhabiting public land to purchase up to 160 acres at $1.25 per acre before the land was put up for sale. The implicit goal was to promote rapid settlement on the western frontier and reduce federal government's enforcement costs (Allen, 1991).

In the pre-Reconstruction South, public land was not open to homestead but rather unrestricted cash entry, which permitted the direct sale of public land to private individuals of 80 acres or more for at least $1.25 an acre. The 1866 SHA restricted cash entry and reserved for homesteading over 46 million acres of public land, or about one-third of the total land area in the five southern public land states (Lanza, 1999, pp. 13). Similar to the HSA, homesteaders could patent up to 160 acres after five years of inhabiting and improving the land, but unlike the HSA, could not commute homestead entries to cash entry after six months. Congress repealed the cash entry restriction in 1876, and sharply reversed policy in 1889 by ending cash entry in all public land states except for Missouri (Gates, 1940). In sum, the SHA followed the same application procedures as the HSA but differed in that it restricted cash sales of public land for the decade after its passage.

2.1 Land monopolization and inequality

About 150 million acres of public land, or about 7% of the total area of frontier states, had already been sold by the time of the passage of the HSA.[3] By the turn of the twentieth century, 250 million acres (11% of total land area) had been sold, while 100 million acres (4% of total land area) had been claimed by homestead. In the South, about 50 million acres of public land, or about 31% of the states' total acreage, had already been sold by before the passage of the SHA. A substantial rise in the number and total acreage, respectively, of homestead entries in the South and West occurred after the 1889 cash-entry restriction.

Homestead policy may have failed to create a more equitable land distribution in part due to the accumulation of public land by speculators and corporations through corrupt practices, such as the use of dummy entry-men, which is the practice of paying individuals to stake out a homestead in order to extract resources from the land with no intention of filing for the final patent. In the South, dummy entry-men were used by timber and mining companies to extract resources while the cash entry restriction of the SHA was in effect. When the restriction was removed, there was no need for fraudulent filings because the larger companies could buy land in unlimited amounts at a nominal price (Gates, 1940, 1979). The same pattern of fraudulent filings existed in the West, where Murtazashvili (2013, pps. 216-218) argues that speculators benefited disproportionately from land laws because the economic balance of power tilted toward the wealthy.

2.2 What the railroad will bring us: More revenue

Frontier states prioritized spending on banking and transportation projects in order to raise land values and attract more settlers (Sylla and Wallis, 1998). The promise of increasing land values and future tax revenues led frontier states to sharply increase borrowing in the mid-1830s by selling long-term bonds to finance transportation and banking investments. Frontier states in the Midwest (e.g., Illinois, Indiana, and Michigan) borrowed to invest

[3]Source: author's estimates using land patent data described in Section 4.

in canals and railroads, while those in the South (e.g, Arkansas, Florida, and Mississippi) borrowed in order to charter state banks. Indiana, for instance, passed an 1836 act that added $10 million in debt spending for transportation projects such as the Wabash and Erie Canal. The state also changed its property tax structure from a flat tax on land to an *ad valorem* tax on all wealth in order to capture the expected increase in land values that would result from the projects (Wallis et al., 2004).

Because railroads expanded commerce by making it cheaper to trade, railroad access is theoretically expected to increase returns to farm land, and in turn increase the property tax base. Donaldson and Hornbeck (2016), for instance, find that average farm values increased substantially as the railroad network expanded from 1870 to 1890, and estimate that the absence of railroads would have decreased farm land values by 60%. Atack and Margo (2011) attribute two-thirds of the increase in improved farm acreage in Midwestern states to the expansion of railroad access in the decade prior to the Civil War. As evidence that railroad access increases the property tax base through higher land values, Atack et al. (2012) find school attendance rates increased in counties that gained access to the rail network between 1850 and 1860.

3 Matrix completion for counterfactual prediction

An important problem in the social sciences is estimating the effect of a binary intervention on an outcome over time. When interventions take place at an aggregate level (e.g., a state), researchers make causal inferences by comparing the post-intervention outcomes of affected ("treated") units against the outcomes of unaffected ("control") units. In the current application, public land states are the treated units and state land states — i.e., states that were not crafted from the public domain and were therefore not directly affected by the homestead acts — serve as control units.[4] A common approach to the problem is the SCM,

[4]The control group includes states of the original 13 colonies, Maine, Tennessee, Texas, Vermont, and West Virginia.

which predicts the counterfactual outcomes for treated units by finding a convex combination of control units that match the treated units in term of lagged outcomes. The SCM predicts patterns across units that are assumed to remain constant over time.

This paper applies the method of matrix completion proposed by Athey et al. (2017) to predict counterfactual outcomes in a setting where multiple treated units are exposed to a binary intervention and the date of initial exposure to treatment may vary between treated units. Let $\mathbf{Y}$ denote a $N \times T$ matrix of outcomes for each unit $i, \ldots, N$ at time $t = 1, \ldots, T$. $\mathbf{Y}$ is incomplete because we observe each element Y_{it} for only the control units and the treated units prior to first treatment exposure. Let $\mathcal{O}$ denote the set of (it) values that are observed and $\mathcal{M}$ the set of missing values. Define the $N \times T$ complete matrix $\mathbf{M}$, where $M_{it} = 1$ if $(it) \in \mathcal{M}$ and $M_{it} = 0$ if $(it) \in \mathcal{O}$ is nonmissing.[5] This setup is motivated by the fundamental problem of causal inference (Holland, 1986) in that we cannot directly observe counterfactual outcomes and we instead wish to impute missing values in $\mathbf{Y}$ for treated units with $M_{it} = 1$.

In an observational setting, units are part of the assignment mechanism that generates $\mathbf{M}$ and patterns of missing data follow one of two specific structures. In the case of simultaneous adoption of treatment, a subset of units are exposed to treatment at time T_0 and every subsequent period. The second structure arises from staggered adoption (Athey and Imbens, 2018), which differs from simultaneous adoption in that T_0 may vary across treated units. In either case, there are selection biases because the probability of missingness may depend on the unobserved data. Selection bias in the staggered adoption setting can occur, for example, if public land states are exposed to treatment (i.e., settled homesteads) earlier because they have higher quality land. The goal is to accurately estimate the effect of a policy intervention despite incomplete and biased data due to selection bias.

Schnabel et al. (2016) first connected the matrix completion problem with causal inference

[5]The process that generates $\mathbf{M}$ is referred to the assignment mechanism in the causal inference literature (Imbens and Rubin, 2015) and the missing data mechanism in missing data analysis (Little and Rubin, 2014, Chap. 1).

in an observational setting in the context of recommender systems under selection bias and propose a weighted matrix factorization method based on propensity scoring under the assumption that the assignment mechanism is probabilistic.

3.1 Matrix completion estimator

Matrix completion methods can attempt to impute missing entries in a low-rank matrix by solving a convex optimization problem via nuclear norm minimization (NNM), even when relatively few values are observed in the full matrix $\mathbf{Y}$ (e.g., Candès and Recht, 2009; Candes and Plan, 2010).[6] These methods exploit relationships within and across outcomes by treating $\mathbf{Y}$ as static and missing values are sampled uniformly at random. The method of matrix completion via NNM (MC-NNM) proposed by Athey et al. (2017) allows for patterns of missing data in $\mathbf{Y}$ to have a time-series dependency structure that arise from simultaneous or staggered adoption. The model to be estimated is

$$Y_{it} = L^*_{it} + \sum_{p=1}^{P} X_{ip}\beta^* + \gamma^*_i + \delta^*_t + \epsilon_{it} \tag{1}$$

where $\mathbf{L}^*$ a low-rank matrix to be estimated, $\mathbf{X}$ is a $N \times P$ matrix of normalized unit-specific covariates, and γ^* and δ^* are unit and time fixed effects, respectively. The identifying condition is that the error term ϵ is independent across rows (units), conditional on $\mathbf{L}^*$, and $\mathrm{E}[\epsilon|\mathbf{L}^* + \beta^* + \gamma^* + \delta^*] = 0$. The MC-NNM estimator for $\mathbf{L}^*$ minimizes the sum of squared reconstruction errors via nuclear norm regularized least squares:

$$\min_{\mathbf{L},\beta} \left[\sum_{(i,t)\in\mathcal{O}} \frac{1}{|\mathcal{O}|} \left(Y_{it} - L_{it} - \sum_{p=1}^{P} X_{ip}\beta - \gamma_i - \delta_t \right)^2 + \lambda \|\mathbf{L}\|_\star \right], \tag{2}$$

where λ is the regularization term on the nuclear norm $\|\cdot\|_\star$ (i.e., sum of singular values) that is chosen by cross-validation. The algorithm for calculating the estimator iteratively replaces

[6] Low-rank matrices arise in the present context when only a few factors contribute to the outcomes. Candès and Recht (2009) note that the observed values in $\mathbf{Y}$ cannot be mostly equal to zero.

missing values with those recovered from a singular value decomposition (SVD) (Mazumder et al., 2010).[7]

Athey et al. (2017) note two drawbacks of the MC-NNM estimator: first, it penalizes the errors for each value with $M_{it} = 0$ equally without regard to the fact that $P(M_{it}) = 1$ (i.e., the propensity score) increases with t. Second, the estimator does not account for time-series dependencies in the observed data and therefore it is likely that the columns of ϵ are autocorrelated.

3.2 Simulations

In this section, I evaluate the accuracy of the MC-NNM estimator on the following three datasets common to the SCM literature, with the actual treated unit removed from each dataset: Abadie and Gardeazabal's (2003) study of the economic impact of terrorism in the Basque Country during the late 1960s ($N = 16$, $T = 43$); Abadie et al.'s (2010) study of the effects of a large-scale tobacco control program implemented in California in 1988 ($N = 38$, $T = 31$); and Abadie et al.'s (2015) study of the economic impact of the 1990 German reunification on West Germany ($N = 16$, $T = 44$). For each trial run, I randomly select half of the control units to be treated and predict their counterfactual outcomes for periods following a randomly selected initial treatment time T_0. I compare the predicted values to the observed values by calculating the root-mean squared error (RMSE), $\sum_{it} |\mathbf{L}^* - \hat{\mathbf{L}}|^2/\sqrt{NT}$.

I benchmarks the MC-NNM estimator against the following previously used estimators:

DID Horizontal regression of $\mathbf{Y}$ on unit and time fixed effects and a binary treatment variable (Athey et al., 2017)

HR-EN Horizontal regression with elastic net regularization (Athey et al., 2017)

PCA Regularized iterative Principal Components Analysis (PCA) (Ilin and Raiko, 2010; Josse and Husson, 2012)

[7] Amjad et al. (2018) propose an alternative approach of approximating $\mathbf{L}^*$ via SVD, and then using linear regression on the "de-noised" matrix, rather than relying on matrix norm regularizations.

SC-ADH SCM approached via exponentiated gradient descent (Abadie et al., 2010)

SVD Low-rank SVD approximation estimated by expectation maximization (Troyanskaya et al., 2001)

VT-EN The same as HR-EN, but $\mathbf{Y}$ is transposed.

Figure 1 reports that the MC-NNM estimator tends to outperform all other estimators in terms of average RMSE across different ratios T_0/T in the staggered adoption setting. Across all estimators, the average RMSE decreases and confidence bands narrow as T_0/T approaches unity because the estimators have more information to generate counterfactual predictions. The MC-NNM performs comparatively well against the regression-based estimators, which can be attributed to the fact that it is capable of using additional information in the form of pre-period observations of the treated units, whereas the regression-based estimators rely only on the pre-treatment observations of control units to predict counterfactuals.[8]

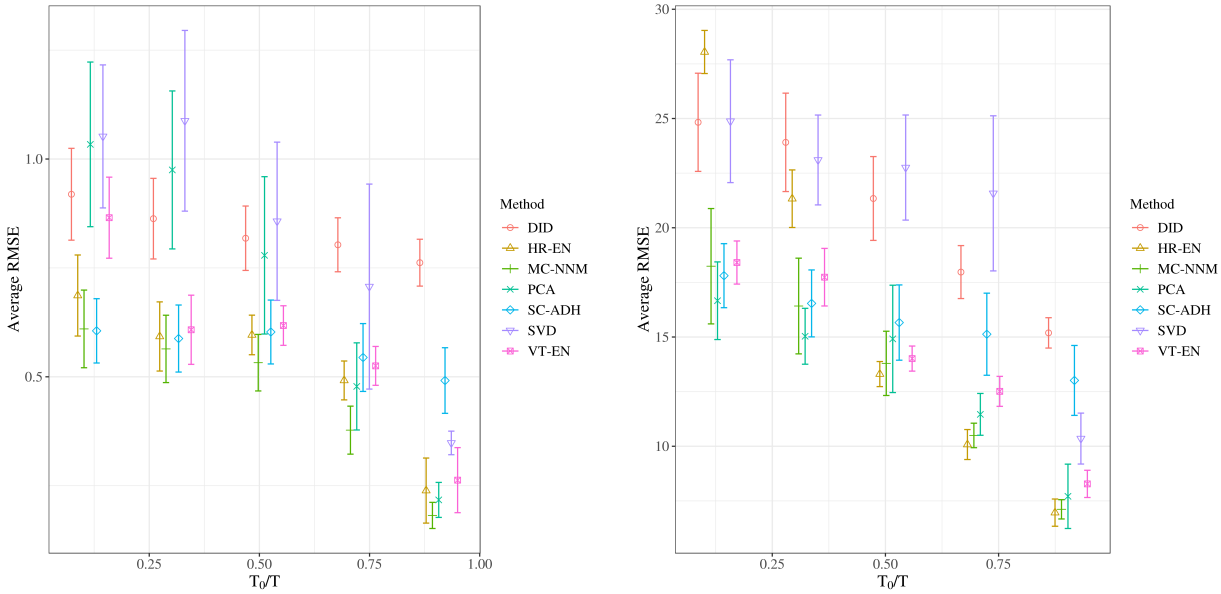
3.3 Hypothesis testing

Consider a setup with J control units indexed by $i = 1, \ldots, J$ and L treated units indexed by $i = J+1, \ldots, L+N$. The MC-NNM estimator imputes the missing post-period treated unit outcomes $\hat{Y}_{it} = \hat{L}_{it}$, for $J+1 \leq i \leq N$ and $T_0+1 \leq t \leq T$, where T_0 denotes the number of pre-periods. The inferred causal effect of the intervention on the treated group is the difference between the observed outcomes of the treated units and the counterfactual outcomes that would have been observed in the absence of the intervention:

$$\hat{\alpha}_{it} = Y_{it} - \hat{Y}_{it}, \qquad J+1 \leq i \leq N, \qquad T_0+1 \leq t \leq T \tag{3}$$

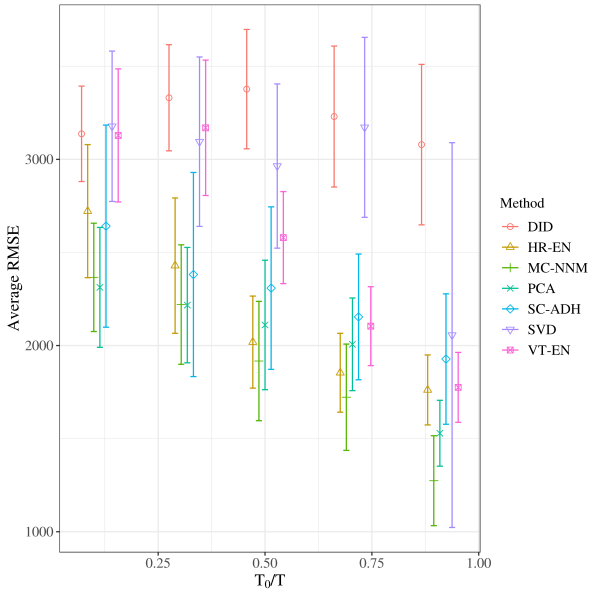
$$\hat{\bar{\alpha}}_t = \frac{1}{L} \sum_{i=J+1}^{N} \hat{\alpha}_{it}, \tag{4}$$

[8]Figure SM-3 in the Supporting Materials (SM) present a similar pattern of results in a simultaneous adoption setting.



(a) Basque Country terrorism data, $N_t = 8$

(b) California smoking ban data, $N_t = 19$



(c) West German reunification data, $N_t = 8$

Figure 1: Placebo tests under staggered treatment adoption. Error bars represent 95% confidence intervals calculated using the standard deviation of the prediction distribution for 20 trial runs. Note that the estimates are jittered horizontally to reduce overlap.

where $\hat{\bar{\alpha}}_t$ corresponds to the per-period estimated average causal effect across treated units.

Chernozhukov et al. (2017) propose a randomization inference approach for testing the sharp null hypothesis $\mathrm{H}_0 : \hat{\bar{\alpha}}_t = \bar{\alpha}_t^o$ for $T_0 + 1 \leq t \leq T$, where $\{\bar{\alpha}_t^o\}_{t=T_0}^T$ is a trajectory of per-period average effects under the null. The test statistic suggested by Chernozhukov et al. (2017) summarizes the trajectory of causal effects over the post-period:

$$S(\hat{\bar{\alpha}}_t) = \left(\frac{1}{\sqrt{T_\star}} \sum_{t=T_0+1}^{T} |\hat{\bar{\alpha}}_t|^q \right)^q, \tag{5}$$

where $T_\star = T - T_0$ and $q \in \{1, 2\}$ in the applications. I estimate p-values by creating $\pi \in \Pi$ permutations of $\mathbf{Y}$ across the time dimension and calculating the proportion of test statics calculated on the permuted data under the null that are more extreme than the observed test statistic:

$$\hat{p} = \frac{1}{\Pi} \sum_{\pi \in \Pi} \mathbf{1} \left\{ S(\hat{\bar{\alpha}}_\pi) > S(\hat{\bar{\alpha}}_t) \right\}, \tag{6}$$

where $\hat{\bar{\alpha}}_\pi$ is the average causal effect estimated on the permuted data $\mathbf{Y}_\pi$. The idea for permuting time periods rather than treatment assignment is if the data are stationary and weakly dependent, which is often the case in an aggregate time-series setting, then the distribution of the error term ϵ in Equation (1) should be the same in the pre- and post-periods. Chernozhukov et al. (2017) prove that the p-values resulting from their inferential procedure are approximately unbiased assuming that the MC-NNM estimator is consistent, which can be verified by placebo tests such as those described in Section 1. I rely on three types of permutations: *i.i.d.* random permutations of the time index t; *i.i.d. block* random permutations of $K = T/b$ non-overlapping blocks, where b is selected according to the optimal block length for the dependent bootstrap (Politis and White, 2004); and *moving block* permutations that circularly shift t by one period, resulting in $T - 1$ permutations. The latter two permutations are capable of preserving the dependence structure of the data and are thus appropriate for weakly dependent data.

4 Impact of homestead acts on state capacity

Did homestead acts impact state capacity over the long-run? In this section, I estimate the causal impact of homestead acts on frontier state government finances. I use measures of total expenditure and revenue collected from the records of 48 state governments during the period of 1783 to 1932 (Sylla et al., 1993) and the records of 16 state governments during the period of 1933 to 1937 (Sylla et al., 1995a,b). Comparable measures for 48 states are drawn from U.S. Census special reports for the years 1902, 1913, 1932, 1942, 1962, 1972, and 1982 (Haines, 2010).[9]

The outcomes data pre-processing steps are as follows. Each measure is inflation-adjusted according to the U.S. Consumer Price Index (Williamson, 2017) and scaled by the total free population in the decennial census (Haines, 2010). I impute missing values separately in the pre- and -post-periods by carrying the last observation forward and impute remaining missing values by carrying the next observation backward. The raw outcomes data are log-transformed to alleviate exponential effects. Removing states with no variance in the pre-intervention outcomes results in complete $N \times T$ matrices of size 33×159 and 34×158, for the expenditures and revenues outcomes, respectively.

The staggered adoption setting is appropriate for the current application because the year of initial treatment exposure T_0 varies across states, about half of which are exposed to homesteads following the passage of the HSA. I determine the years of initial exposure to homesteads by aggregating to the state level approximately 1.46 million individual land patent records authorized under the HSA. Land patent records provide information on the initial transfer of land titles from the federal government and are made accessible online by the U.S. General Land Office (`https://glorecords.blm.gov`). The earliest successful homestead filings occurred in 1869 in about half of the western frontier states, about seven years following the enactment of the HSA. The first homesteads were filed in southern states in 1872. Figure SM-1 visualizes the timing and intensity of homestead entries.

[9] I take the mean of duplicate state-year observations, which arise for the years 1902, 1913, and 1932.

Unit-specific covariates $\mathbf{X}$ include state-level average farm sizes measured in the 1860 and average farm values measured in the 1850 and 1860 censuses.[10] In theory, we should expect that homesteaders migrate to more productive land and thus excluding these pre-intervention measures of agricultural productivity may result in overestimating the actual impact of homestead policy. To control for selection biases in terms of access to frontier lands, I create a measure of railroad access using digitized railroad maps provided by Atack (2013), which contain information on the year that each rail line was built. Overlaying the railroad track map over historical county borders, I calculate the total miles of operational track per square mile and aggregate the measure to the state-level.[11]

4.1 Placebo tests

Prior to presenting the main results, I assess the validity of the key assumption underlying the approach by discarding post-period observations from the data and testing the null hypothesis

$$\mathrm{H}_0 : S(\hat{\tilde{\alpha}}_t) = 0 \quad \text{for } T_0 - \tau + 1 \leq t \leq T_0, \tag{7}$$

where $\tau \in \{1, 10, 25\}$. This placebo null hypothesis is tested by the same procedure described in Section 3.3, treating $\{1, \ldots, T_0 - \tau\}$ as the pre-period.

Table 1 reports the average treatment effect over the placebo post-period and randomization p-values calculated by Equation (6). Placebo tests on the revenue outcome yield two-sided p-values greater than the significance level of $\alpha = 0.05$, regardless of the value of q or permutation structure. These results provide evidence in favor of the validity of the consistency assumption. However, we can only reject the null in the case of $\tau = 1$ when considering the expenditure outcome, which suggests that the consistency may only hold

[10] Table SM-1 provides data sources and definitions.

[11] Using these data, I estimate that 29% of counties had railroad access in 1862 and 91% had access by 1911 (Fig. SM-2). The railroad access measure defines access with respect to county boundaries, which Atack et al. (2012) point out has limitations because a county without access might be adjacent to one with access and county boundaries frequently changed over time.

when the ratio T_0/T is very high.

Table 1: Placebo test p-values.

	Expenditure						Revenue					
	i.i.d.		i.i.d. Block		Moving Block		i.i.d.		i.i.d. Block		Moving Block	
τ \ q	1	2	1	2	1	2	1	2	1	2	1	2
1	0.051	0.056	0.098	0.099	0.047	0.047	0.469	0.499	0.488	0.511	0.482	0.494
10	0.028	0.027	0.034	0.033	0.012	0.024	0.543	0.575	0.548	0.582	0.565	0.600
25	0.022	0.024	0.042	0.042	0.024	0.024	0.581	0.594	0.627	0.653	0.635	0.634

Notes: randomization p-values corresponding to each permutation structure and value of τ and q. i.i.d. block and i.i.d. block p-values are calculated using $|\Pi| = 1,000$ permutations. Moving block p-values are based on $|\Pi| = T - 1$ permutations.

Further evidence of the consistency of the MC-NNM estimator is provided in Figure SM-4, which presents the results of placebo tests on control units using both pre- and post-period observations for each outcome. Similar to the simulations on the SCM datasets discussed in Section 3.2, there are no missing entries in each outcome $\mathbf{Y}$ because the actual treated units are removed prior to the placebo tests. I randomly choose about half of the remaining control units as hypothetical treated units and predict their values for time periods following a randomly selected T_0.

The MC-NNM estimator outperforms DID and SVD estimators in terms of minimizing RMSE for each ratio T_0/T. At $T_0/T \geq 0.5$, the estimator generally yields comparable error rates to PCA, SCM, and vertical regression estimators.

4.2 Main estimates

In the main analyses, I fit the MC-NNM estimator described in Equation 1 on the entirety of observed entries in $\mathbf{Y}$ to recover its missing entries; i.e., the counterfactual outcomes of public land states. The value of the regularization term λ is optimally selected on the value that yields the lowest average RMSE calculated during cross-validation. The average RMSE calculated on the validation sets during the λ selection process are 0.41 and 0.48 for the

expenditures and revenues outcomes, respectively.

The top panels of Figure 2a and 2b compare the means of the observed time-series of treated units and control units along with the mean predicted outcomes of treated units. The observed means of the treated and control units are essentially identical in the post-period. However, we are interested primarily in the difference in the observed and predicted treated unit outcomes, which is the quantity $\hat{\tilde{\alpha}}_t$ in Equation (4). This quantity corresponds to the estimated per-period average causal effects of treatment exposure on the treated units and the per-period effects are plotted in the bottom panels. For display purposes, bootstrap confidence intervals for $\hat{\tilde{\alpha}}_t$ are calculated by block resampling with optimal block lengths selected by the procedure described by Politis and White (2004).

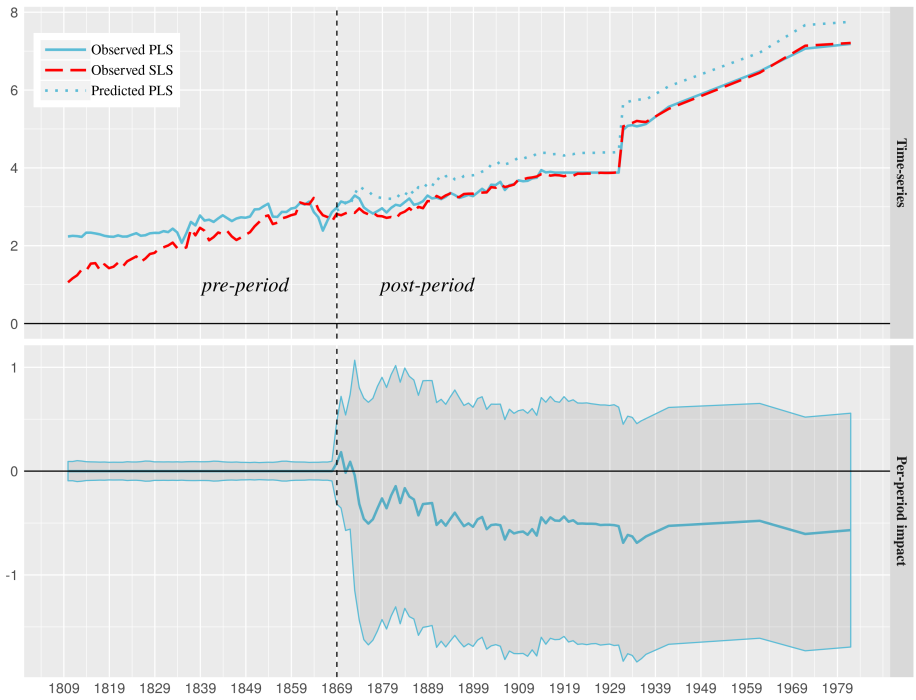
Table 2 reports randomization p-values from testing the null hypothesis of a zero effect:

$$\mathrm{H}_0 : S(\hat{\tilde{\alpha}}_t) = 0 \quad \text{for} \;\; T_0 + 1 \leq t \leq T. \tag{8}$$

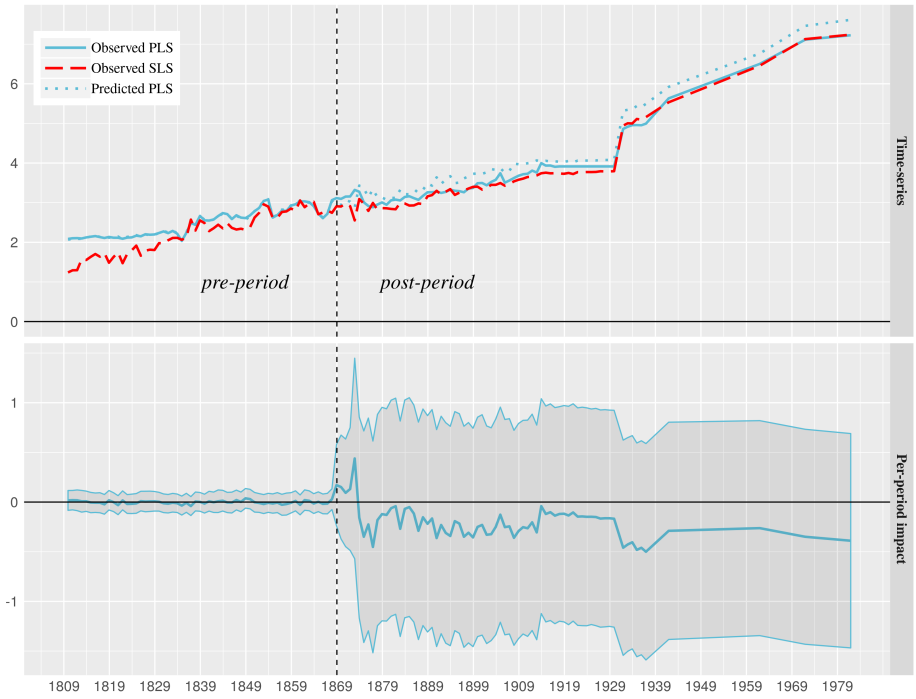
Table 2: Causal estimates on state capacity datasets.

	Expenditure		Revenue	
	$q = 1$	$q = 2$	$q = 1$	$q = 2$
$S(\hat{\tilde{\alpha}}_t)$	3.87	1.40	1.97	0.76
i.i.d.	0.002	0.003	0.001	0.001
i.i.d. Block	0.001	0.002	0.001	0.001
Moving Block	< 0.001	< 0.001	< 0.001	< 0.001

Notes: $S(\hat{\tilde{\alpha}}_t)$ corresponds to the test statistic described in Equation (5) and each value beneath is the randomization p-value corresponding to each permutation structure. See footnotes to Table 1.



(a) Log per-capita state government expenditure (1982$)



(b) Log per-capita state government revenue (1982$)

Figure 2: Top panel: Mean observed (solid time-series) and counterfactual predicted (dotted time-series) outcomes of treated units (PLS) and mean observed outcomes (dashed time-series) among control units (SLS), displayed for the time period of 1809 to 1982. Dashed vertical line represents the initial treatment year of 1869. Bottom panel: Per-period average causal impacts of homestead exposure on PLS, or $\hat{\tilde{\alpha}}_t$ in Equation (4). Shaded regions represent 95% confidence intervals estimated by taking $\hat{\tilde{\alpha}}_t \pm 1.96$ the standard error of the distribution of 1,000 bootstrap replicates of $\hat{\tilde{\alpha}}_t$.

5 DID estimation

The matrix completion approach estimates the impact of a discrete event on a continuous outcome. However, in this application a continuous form of treatment is available in the form of homestead entries. Equation (9) estimates a continuous version of the DID estimator described in Section 3.2. Similar to the case of binary treatment, the continuous DID estimator is adapted to a setting of staggered adoption because the initial date of exposure to homesteads varies across public land states. The first difference comes from variation in the date of initial exposure to homesteads, and the second difference comes from variation in the intensity of homestead entries.

$$Y_{i,t} = a_i + b_t + \gamma\, M_{it} + \delta\, (M_{it} \cdot \text{homesteads}_{it}) + X'_{it} + \epsilon_{it}. \tag{9}$$

In this model, a_i and b_t are vectors of unit and time fixed effects, respectively, included to control for unobserved heterogeneity across units. In the staggered adoption setting, entries in the missing value indicator M_{it} assume the value of 1 at $t \geq T_0$, where the initial exposure time T_0 varies across units. The continuous treatment exposure variable $\text{homesteads}_{i,t}$ measures the per-capita statewide sum of homestead entries in state i and year t.

The coefficient corresponding to the interaction term, $\hat{\delta}$, is the estimated average causal effect of exposure to homesteads. I use unit-stratified bootstrapped samples to construct nonparametric standard errors for $\hat{\delta}$.[12] $X'_{i,1860}$ is the average value of farm land in 1860, which is included to ensure that $\hat{\delta}$ is not biased by parallel trends in pre-intervention agricultural productivity.

It should be emphasized that the DID estimator in a staggered adoption setting relies

[12]The model assumes i.i.d. errors, which understates the standard errors for $\hat{\delta}$ when the regression errors are serially correlated, or $\text{Corr}(\epsilon_{s,t}, \epsilon_{s,t-1}) \neq 0$, which can arise when the time-series lengths are not sufficiently long to reliably estimate the data generating process. Bertrand et al. (2004) show that the stratified bootstrap can be used to compute consistent standard errors when the number of units is sufficiently large.

on several strong assumptions regarding both the assignment mechanism, which in this application is the distribution of initial treatment times T_0, and the counterfactual outcomes of the treated units. The framework of Athey and Imbens (2018), for instance, assumes the distribution of T_0 is completely random conditional on $X'_{i,1860}$. In the current application, this assumption ignores the possibility that initial exposure to homesteads might be determined by unobserved factors. The framework also that the counterfactual outcomes at time t does not depend on the future date of treatment exposure if $t < T_0$ or the history of treatment exposure if $t > T_0$. Violations of these assumptions would arise if the homestead policy is anticipated prior to T_0 or if the size of frontier state government is determined by whether the state was exposed early or late to homesteads.

5.1 DID estimates

I apply the DID estimator on two unbalanced state-year panel datasets: the panel of southern public land states spans from 1823 to 1982 and has 38 observations prior to the enactment of the SHA. The panel of western public land states spans from 1810 to 1982 and has 52 pre-intervention years. The treatment effect estimates, summarized in Fig. 3, indicate that per-capita cumulative homesteads significantly increase per-capita revenue in western states by 0.02 log points [0.0007, 0.04] and has a nonsignificant impact on revenue in southern states. Estimates on per-capita expenditure in the South and West are also nonsignificant. Per-capita cumulative homesteads significantly raise education spending in the South by 0.17 log points [0.03, 0.28], while having no impact on education spending in the West. As expected, withholding average farm values from the DID specification biases treatment effect estimates upwards (Fig. SM-3).

5.2 Mechanisms: Inequality and railroads

What are the channels through which homesteads affect state capacity? Land inequality is expected to influence state capacity, although the direction of influence is theoretically

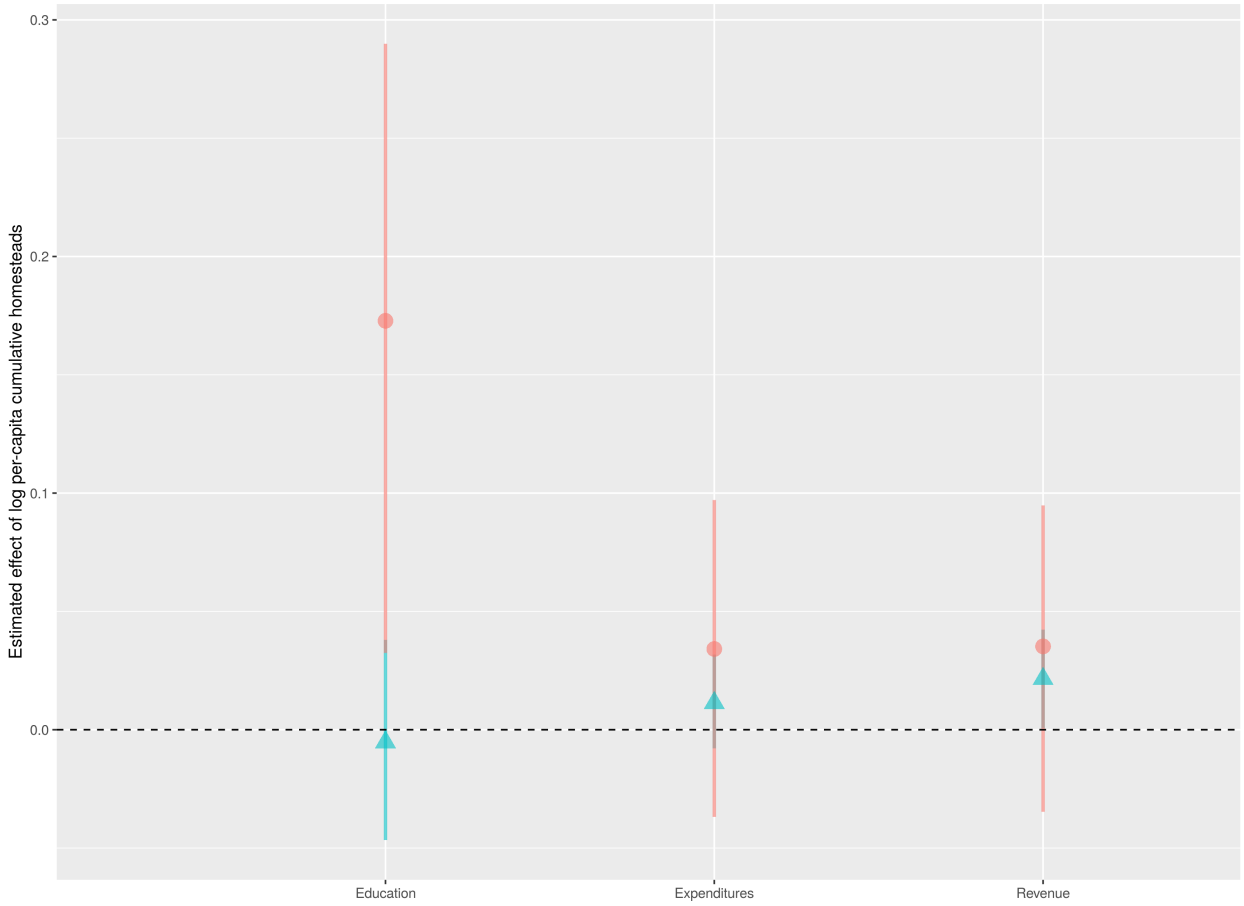


Figure 3: DID estimates of log per-capita cumulative homesteads on log per-capita state government finances. Lines represent 95% confidence intervals are constructed using 10,000 state-stratified bootstrap samples. Lines with triangles represent western public land state estimates and lines with circles represent southern public land state estimates.

ambiguous. To test whether homesteads affected future land inequality in frontier counties, I calculate a commonly-used measure of land inequality based on the Gini coefficient of census farm sizes. Gini-based land inequality measures are commonly used as proxy for the *de facto* bargaining power of landed elites (e.g., Boix, 2003; Ziblatt, 2008; Ansell and Samuels, 2015).[13]

Table SM-2 presents DID estimates of the impact of log per-capita cumulative homesteads on land inequality or railroad access during the period of 1860 to 1950.[14] Farm values are included in the regression as a proxy for agricultural productivity, which might be associated with farm sizes approaching ideal scale and therefore land inequality. I estimate that homesteads significantly decreased land inequality in western frontier state counties: a 10% increase in log per-capita cumulative homesteads is expected to lower land inequality by -0.0004 [-0.0005, -0.0002] points. The estimated impact on land inequality in the South is in the same direction, but not significant.

Railroad access is theoretically expected to increase the capacity of county and state governments by increasing the returns to farm land. DID estimates of the effect of log per-capita cumulative homesteads on railroad access (value between 0 and 1) can be interpreted as follows: a 10% increase in log per-capita cumulative homesteads is expected to increase railroad access in southern counties by 0.003 [0.001, 0.005] points and by 0.009 [0.007, 0.01] points in western counties. I include farm values in the specification as a proxy for economic development, which is expected to increase state capacity and correlate with other measures of development like railroad access.

[13]The Gini coefficient will underestimate land inequality in counties with high shares of propertyless farmers because tenant farms are included in the farm size data, which is problematic because farms can be operated by different tenants but owned by the same landlord. Following the procedure of Vollrath (2013), I correct for this problem by adjusting the farm Gini coefficient G by the ratio of farms to adult males, p. The adjusted coefficient is calculated as $G^A = pG + (1 - p)$.

[14]Since railroad access is measured every year, I take the mean of railroad access to the nearest decennial year; e.g., $y_{s,1870}$ is the mean of the access measure between 1862 and 1870 in county s.

6 Conclusion

Which historical processes are responsible for present-day differences in the capacity of state governments? For example, there exists considerable variation in both the amount and revenue sources of state and local government funding for public education: New York spent almost twice the national average per-pupil, primarily using local (54%) and state (41%) revenue sources, while Idaho spent about 60% of the national average from a combination of state (63%), local (26%) and federal (11%) sources.[15]

The findings of this paper signify that mid-nineteenth century homestead acts had positive impacts on frontier state government finances that can help explain contemporary differences in state capacity. RNN estimates imply that the HSA had a significant and positive impact on western state government expenditure about 50 years following its implementation. The delayed impacts can be explained by the facts that settlers were required to make improvements on land for five years before filing a grant and also homesteads did not substantially accumulate until after the 1889 cash-entry restriction. I find no evidence that the homestead acts had a significant impact on the state capacity of frontier states on average over the entire-post period that extends into the twentieth century. The inability to identify a significant average impact can be attributed to the progressive widening of confidence intervals over the post-period: the uncertainty of making counterfactual predictions based on the previous histories of (placebo) treated units increases as we move farther from the intervention year.

I also estimate a DID model that leverages variation in both the timing and the intensity of cumulative homesteads across public land states and find significant positive impacts on state capacity. I include in the DID specification average farm values in order to control for homesteaders seeking more productive lands. The DID estimates imply that per-capita cumulative homesteads significantly increase per-capita revenue in western states by 0.02 log

[15]Source: 2014 Annual Survey of School System Finances, U.S. Census Bureau. `https://www.census.gov/programs-surveys/school-finances.html`.

points and raise education spending in the South by 0.17 log points. The DID estimates are of similar magnitude and direction than the RNN estimates averaged over the post-period, although the confidence intervals around the DID estimates are considerably more narrow and possibly overoptimistic due to serial correlation in the DID regression errors.

I explore land inequality and railroad access as possible causal mechanisms underlying the relationship between land reform and state capacity. DID estimates reveal that per-capita cumulative homesteads lowered land inequality in western counties, but did not significantly alter the distribution of land ownership in southern counties. Railroad access is theoretically expected to expand the property tax bases of state governments by increasing the returns to agricultural land. I find that cumulative homesteads significantly increased railroad access in frontier counties over a period extending into the twentieth century.

This paper makes a methodological contribution in proposing a novel alternative to SCM for estimating the effect of a policy intervention on an outcome over time in settings where appropriate control units are unavailable. In placebo tests, RNN-based models outperform SCM in terms of predictive accuracy while yielding a comparable proportion of false positives. RNNs have advantages over SCM in that they are structured for sequential data and can learn nonconvex combinations of predictors, which is useful when the data-generating process underlying the outcome depends nonlinearly on the history of predictors.

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association 105*(490), 493–505.

Abadie, A., A. Diamond, and J. Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science 59*(2), 495–510.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the basque country. *The American Economic Review 93*(1), 113–132.

Acemoglu, D. and J. A. Robinson (2008). Persistence of power, elites, and institutions. *American Economic Review 98*(1), 267–293.

Acemoglu, D., D. Ticchi, and A. Vindigni (2011). Emergence and persistence of inefficient states. *Journal of the European Economic Association 9*(2), 177–208.

Allen, D. W. (1991). Homesteading and property rights; or, "how the west was really won". *The Journal of Law and Economics 34*(1), 1–23.

Amjad, M., D. Shah, and D. Shen (2018). Robust synthetic control. *The Journal of Machine Learning Research 19*(1), 802–852.

Ansell, B. and D. J. Samuels (2015). *Inequality and Democratization: An Elite Competition Approach*. New York, NY: Cambridge University Press.

Atack, J. (2013). On the use of geographic information systems in economic history: The american transportation revolution revisited. *The Journal of Economic History 73*(2), 313–338.

Atack, J., R. Margo, and E. Perlman (2012). The impact of railroads on school enrollment in nineteenth century america.

Atack, J. and R. A. Margo (2011). The impact of access to rail transportation on agricultural improvement: The american midwest as a test case, 1850-1860. *Journal of Transport and Land Use 4*(2).

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

Athey, S. and G. Imbens (2018). Design-based analysis in difference-in-differences settings with staggered adoption. *arXiv preprint arXiv:1808.05293*.

Benabou, R. (2000). Unequal societies: Income distribution and the social contract. *American Economic Review*, 96–129.

Bensel, R. F. (1990). *Yankee Leviathan: the Origins of Central State Authority in America, 1859-1877*. Cambridge University Press.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Besley, T. and T. Persson (2009). The origins of state capacity: Property rights, taxation and politics. *American Economic Review 99*(4), 1218–1244.

Besley, T. and T. Persson (2010). State capacity, conflict, and development. *Econometrica 78*(1), 1–34.

Boix, C. (2003). *Democracy and Redistribution*. New York, NY: Cambridge University Press.

Candes, E. J. and Y. Plan (2010). Matrix completion with noise. *Proceedings of the IEEE 98*(6), 925–936.

Candès, E. J. and B. Recht (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics 9*(6), 717.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2017). An exact and robust conformal inference method for counterfactual and synthetic controls. *arXiv preprint arXiv:1712.09089*.

Donaldson, D. and R. Hornbeck (2016). Railroads and american economic growth: A "market access" approach. *The Quarterly Journal of Economics 131*(2), 799–858.

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

Engerman, S. L. and K. L. Sokoloff (2005). The evolution of suffrage institutions in the new world. *The Journal of Economic History 65*(4), 891–921.

Ferrie, J. P. (1997). Migration to the frontier in mid-nineteenth century america: A re-examination of turner's 'safety valve;.

Frymer, P. (2014). 'a rush and a push and the land is ours': Territorial expansion, land policy, and u.s. state formation. *Perspectives on Politics 12*(1), 119.

Galor, O., O. Moav, and D. Vollrath (2009). Inequality in landownership, the emergence of human-capital promoting institutions, and the great divergence. *The Review of Economic Studies 76*(1), 143–179.

Garcia-Jimeno, C. and J. A. Robinson (2009). The myth of the frontier.

Gates, P. W. (1940). Federal land policy in the south 1866-1888. *The Journal of Southern History 6*(3), 303–330.

Gates, P. W. (1979). Federal land policies in the southern public land states. *Agricultural History 53*(1), 206–227.

Grosjean, P. (2014). A history of violence: The culture of h=honor and homicide in the US South. *Journal of the European Economic Association 12*(5), 1285–1316.

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

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association 81*(396), 945–960.

Ilin, A. and T. Raiko (2010). Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research 11*(Jul), 1957–2000.

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

Josse, J. and F. Husson (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique 153*(2), 79–99.

Lanza, M. L. (1999). *Agrarianism and Reconstruction Politics: The Southern Homestead Act*. Baton Rouge, LA: LSU Press.

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

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

Meltzer, A. H. and S. F. Richard (1981). A rational theory of the size of government. *Journal of Political Economy 89*(5), 914–927.

Murtazashvili, I. (2013). *The Political Economy of the American Frontier*. New York, NY: Cambridge University Press.

Politis, D. N. and H. White (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews 23*(1), 53–70.

Poulos, J. (2017, December). RNN-Based Counterfactual Time-Series Prediction. *ArXiv e-prints*.

Ramcharan, R. (2010). Inequality and redistribution: Evidence from u.s. counties and states, 1890–1930. *The Review of Economics and Statistics 92*(4), 729–744.

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

Suryanarayan, P. (2017). Hollowing out the state: Franchise expansion and fiscal capacity in colonial india.

Sylla, R. and J. J. Wallis (1998). The anatomy of sovereign debt crises: Lessons from the american state defaults of the 1840s. *Japan and the World Economy 10*(3), 267–293.

Sylla, R. E., J. B. Legler, and J. 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.

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

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

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics 17*(6), 520–525.

Turner, F. J. (1956). *The Significance of the Frontier in American History*. Ithaca, NY: Cornell University Press.

Vollrath, D. (2013). Inequality and school funding in the rural united states, 1890. *Explorations in Economic History 50*(2), 267–284.

Wallis, J. J., R. E. Sylla, and A. Grinath III (2004). Sovereign debt and repudiation: The emerging-market debt crisis in the us states, 1839-1843.

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

Ziblatt, D. (2008). Does landholding inequality block democratization?: A test of the "bread and democracy" thesis and the case of prussia. *World Politics 60*(4), 610–641.