State-Building 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 policies? I propose using matrix completion — a machine learning method commonly used for recommendation tasks — to predict the counterfactual time-series of frontier state capacity had there been no homesteading. The matrix completion method outperforms several linear regression-based estimators in placebo tests. Time-specific causal estimates signify that homestead policies had significant and long-lasting negative impacts on state government expenditure and revenue. These results are consistent with difference-in-difference estimation that exploits variation in the timing and intensity of homestead entries aggregated from 1.46 million individual land patents.

Keywords: Counterfactual Prediction; Difference-in-Differences; Matrix Completion; State Capacity

---

Postal address: Department of Political Science, 210 Barrows Hall #1950, Berkeley, CA 94720-1950. *E-mail:* 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 policies and the development of the U.S. government, arguing that policies designed to transfer public land to private individuals increased the bureaucratic capacity of the federal government to administer land.

Public land policies 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. 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 policies 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 policies 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 local labor costs 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.

García-Jimeno and Robinson (2008) 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 scholars have noted that public land policies were often exploited by land speculators, ranchers, miners, and loggers, to accumulate public land and extract natural resources during the early stages of capitalist development (Gates, 1942; Murtazashvili, 2013). According to this view, homesteading laws were *de jure* social polices but *de facto* corporate welfarism.

The paper makes a methodological contribution in applying an alternative method for estimating causal impacts of policy interventions on time-series cross-sectional 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, I apply a matrix completion method 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 synthetic control method and other regression-based estimators in terms of minimizing prediction error. In addition, I show how to evaluate the overall effect of the policy intervention using a randomization inference procedure in which approximately unbiased p-values are obtained under minimal assumptions by permuting the time-series dimension of the data under the null.

The paper proceeds as follows: in Section 2, I overview the historical context of homestead policies and its relationship to state capacity and land inequality; Section 3 describes the method of matrix completion for counterfactual prediction, benchmarks the method against the synthetic control method and alternative estimators, and describes the inferential procedure. In Section 4, I report the results of placebo tests to verify the consistency of the matrix completion estimator and present estimates of the long-run impacts of homestead policies on state capacity. Section 5 reports DID estimates of the effect of homesteads on state capacity and land inequality, and Section 6 concludes.

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. 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). 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.[1]

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 (PLS) (Lanza, 1999, pp. 13). Note that I use the terminology of PLS interchangeably with "frontier" states throughout the paper. PLS are states created out of the public domain. In the South, these states include Alabama, Arkansas, Florida, Louisiana, and Mississippi. Western PLS include the 25 states that comprise the Midwestern, Southwestern, and Western U.S. (except Hawaii). 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.

[1] Congress repealed the cash entry restriction in 1876, and sharply reversed policy in 1889 by ending cash entry in all PLS except for Missouri (Gates, 1940).

Homestead policies 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) argues that speculators benefited disproportionately from public land policies because the economic balance of power tilted toward the wealthy. Gates (1942) characterizes western speculators who bought land in bulk prior to the 1889 restriction as being influential in state and local governments, resistant to paying taxes, and opposed to expenditures except for transportation facilities close to their land.

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 ("post-period") outcomes of affected ("treated") units against the outcomes of unaffected ("control") units. In the current application, PLS 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 homestead policies — serve as control units.[2] A common approach to the problem is the synthetic control method, 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

[2]This group includes states of the original 13 colonies, Maine, Tennessee, Texas, Vermont, and West Virginia.

outcomes. The synthetic control method predicts patterns across units that are assumed to remain constant over time.

This paper applies the method of matrix completion via nuclear norm minimization (MC-NNM) 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.[3] Matrix completion methods (e.g., Mazumder et al., 2010) exploit correlations within and across units, but ignore the temporal dimension of the data and typically assume missing values are sampled uniformly at random (Yoon et al., 2018). In contrast, the MC-NNM estimator allows for patterns of missing data to have a time-series dependency structure that arise from simultaneous or staggered adoption.

Let Y denote a $\mathrm{N} \times \mathrm{T}$ matrix of outcomes for each unit $i = 1, \ldots, \mathrm{N}$ at time $t = 1, \ldots, \mathrm{T}$. 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 $\mathrm{N} \times \mathrm{T}$ complete matrix M, where $M_{it} = 1$ if $(it) \in \mathcal{M}$ and $M_{it} = 0$ if $(it) \in \mathcal{O}$ is nonmissing.[4] 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 Y for treated units with $M_{it} = 1$.

In an observational setting, units are part of the assignment mechanism that generates 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, 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.

[3] Schnabel et al. (2016) first connected the matrix completion problem with causal inference in an observational setting in the context of recommender systems under selection bias.

[4] The process that generates 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).

Selection bias in the staggered adoption setting can occur in the current application if PLS 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 data subject to selection bias.

3.1 Matrix completion estimator

Matrix completion methods attempt to impute missing entries in a low-rank matrix by solving a convex optimization problem via nuclear norm minimization, even when relatively few values are observed in the full matrix Y (Candès and Recht, 2009; Candes and Plan, 2010). Low-rank matrices arise in the present context when only a few factors contribute to the outcomes. The MC-NNM estimator is

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

where L^* a low-rank matrix to be estimated, X is a $\mathrm{N} \times \mathrm{P}$ matrix of normalized unit-specific covariates, and $\boldsymbol{\gamma}^*$ and $\boldsymbol{\delta}^*$ are vectors of unit and time effects, respectively. The identifying condition is that, conditional on L^*, the error vector $\boldsymbol{\epsilon}$ is independent across rows (units) and $\mathrm{E}[\boldsymbol{\epsilon}|L^* + \boldsymbol{\beta}^* + \boldsymbol{\gamma}^* + \boldsymbol{\delta}^*] = 0$. Estimating L^* involves minimizing the sum of squared errors via nuclear norm regularized least squares:

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

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 (2) iteratively replaces missing values with those recovered from a singular value decomposition (SVD) (Mazumder et al., 2010).[5]

Athey et al. (2017) note two drawbacks of the MC-NNM estimator: first, it penalizes

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

the errors for each value with $M_{it} = 0$ equally without regard to the fact that $\Pr(M_{it}) = 1$ (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 $\boldsymbol{\epsilon}$ 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 synthetic control literature, with the actual treated unit removed from each dataset: Abadie and Gardeazabal's (2003) study of the economic impact of terrorism in the Basque Country during the late 1960s (N = 16, T = 43); Abadie et al.'s (2010) study of the effects of a large-scale tobacco control program implemented in California in 1988 (N = 38, T = 31); and Abadie et al.'s (2015) study of the economic impact of the 1990 German reunification on West Germany (N = 16, T = 44). 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, $\mathrm{RMSE} = \sum_{it} |L^* - \hat{L}|^2 / \sqrt{\mathrm{NT}}$.

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

DID Horizontal regression of Y on unit and time 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 (Ilin and Raiko, 2010)

SC-ADH Synthetic control 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 Y is transposed.

Figure 1 reports the average prediction error of the estimators in a staggered treatment adoption setting, across different ratios $\mathrm{T_0/T}$, where $\mathrm{T_0}$ denotes the number of pre-periods. Across all estimators, the average RMSE decreases and confidence bands narrow as $\mathrm{T_0/T}$ approaches unity because the estimators have more information to generate counterfactual predictions. The MC-NNM estimator generally outperforms all other estimators in terms of average RMSE across different ratios $\mathrm{T_0/T}$. The strong performance of the MC-NNM estimator 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-intervention ("pre-period") observations of control units to predict counterfactuals.[6]

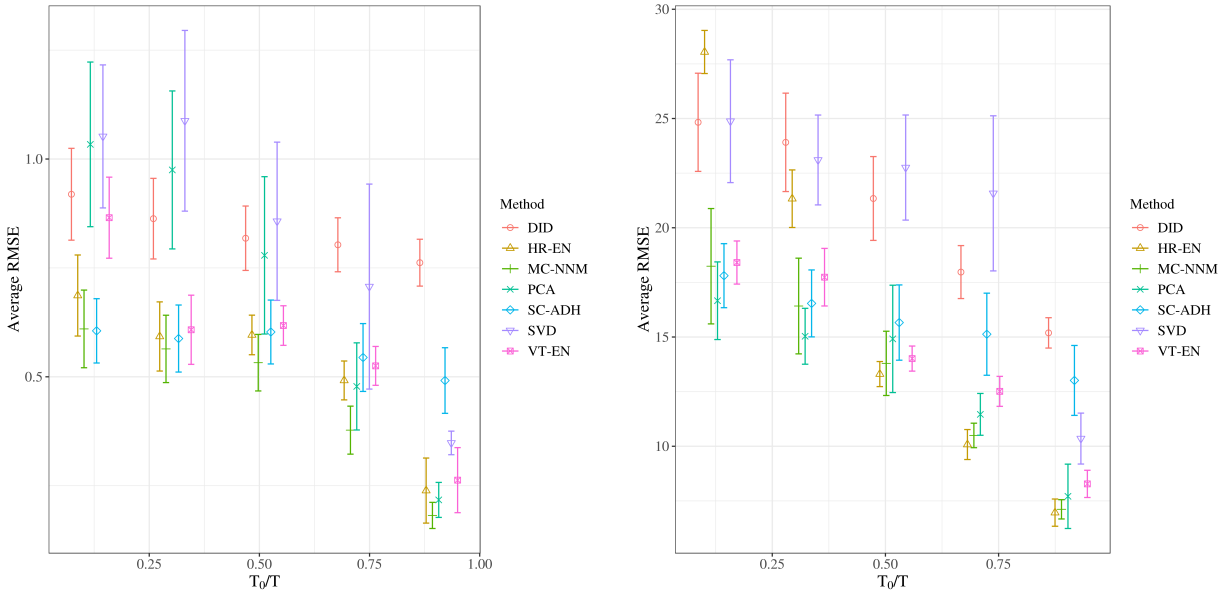
3.3 Hypothesis testing

The method proposed by Athey et al. (2017) focuses estimation and does not provide guidance on hypothesis testing. Consider a setup with J control units indexed by $i = 1, \ldots, \mathrm{J}$ and Q treated units indexed by $i = \mathrm{J}+1, \ldots, \mathrm{N}$. The MC-NNM estimator imputes the missing post-period treated unit outcomes

$$\hat{Y}_{it} = \hat{L}_{it}, \qquad \mathrm{J}+1 \leq i \leq \mathrm{N}, \qquad \mathrm{T_0}+1 \leq t \leq \mathrm{T}. \tag{3}$$

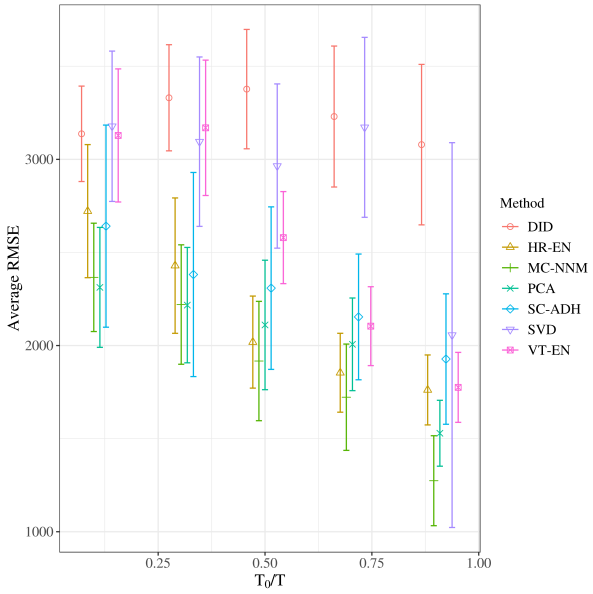
The inferred causal effect of the intervention on the treated group is the difference between the observed outcomes of the treated units and the counterfactual outcomes that would have been observed in the absence of the intervention, $\hat{\alpha}_{it} = Y_{it} - \hat{Y}_{it}$ for $\mathrm{J}+1 \leq i \leq \mathrm{N}$ and

[6]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.

$\mathrm{T}_0+1 \leq t \leq \mathrm{T}$. Taking the difference-in-means between treated unit observed outcomes and predicted outcomes gives the per-period estimated average causal effect across treated units:

$$\hat{\bar{\alpha}}_t = \frac{1}{\mathrm{Q}} \sum_{i=\mathrm{J}+1}^{\mathrm{N}} \hat{\alpha}_{it} \qquad \mathrm{T}_0 + 1 \leq t \leq \mathrm{T}. \tag{4}$$

Chernozhukov et al. (2017) propose a randomization inference approach for testing the sharp null hypothesis $\mathrm{H}_0 : \hat{\bar{\boldsymbol{\alpha}}} = \bar{\boldsymbol{\alpha}}^{\boldsymbol{o}}$, where $\{\bar{\alpha}_t^o\}_{t=\mathrm{T}_0}^{\mathrm{T}}$ is a trajectory of per-period average effects under the null. The test statistic suggested by Chernozhukov et al. (2017) is constructed so that we reject higher values:

$$\mathrm{S}_q(\hat{\bar{\boldsymbol{\alpha}}}) = \left(\frac{1}{\sqrt{\mathrm{T}_\star}} \sum_{t=\mathrm{T}_0+1}^{\mathrm{T}} |\hat{\bar{\alpha}}_t|^q \right)^q, \tag{5}$$

where $\mathrm{T}_\star = \mathrm{T} - \mathrm{T}_0$. I estimate p-values by permuting Y across the time dimension. Letting $\hat{\bar{\boldsymbol{\alpha}}}_{\boldsymbol{\pi}}$ denote the vector of per-period average causal effects estimated for each permutation $\pi \in \Pi$, the randomization p-value is

$$\hat{p} = 1 - \frac{1}{\Pi} \sum_{\pi \in \Pi} I\left\{ \mathrm{S}_q(\hat{\bar{\boldsymbol{\alpha}}}_{\boldsymbol{\pi}}) < \mathrm{S}_q(\hat{\bar{\boldsymbol{\alpha}}}) \right\}, \tag{6}$$

where $I(\cdot)$ denotes the indicator function.

The idea for permuting time periods rather than treatment assignment, as proposed by Abadie et al. (2010), is that 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 $\boldsymbol{\epsilon}$ in (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.

Permuation structures I rely on three types of permutations: i.i.d. random permutations of the time index t; i.i.d. block random permutations of $\mathrm{K} = \mathrm{T}/b$ 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 $\mathrm{T}-1$ 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 policies on state capacity

In this section, I estimate the causal impacts of homestead policies on state capacity, as measured by state government spending and revenue. I create 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).[7] The expenditure measure includes state government spending on education, social welfare programs, and transportation. The revenue measure incorporates state government income streams such as tax revenue and non-tax revenues such as land sales.[8]

The 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). Missing values are imputed separately in the pre- and -post-periods by carrying the last observation forward and remaining missing values are imputed by carrying the next observation backward. The raw outcomes data are log-transformed to alleviate exponential effects. Lastly, I remove states with no variance in the pre-period outcomes results in complete $\mathrm{N} \times \mathrm{T}$ 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

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

[8] Tax revenue is commonly used as a measure of fiscal capacity (Lieberman, 2002), which is the bureaucratic ability of governments to raise taxes from multiple sources in order to finance policies. Fiscal capacity is strongly correlated with state capacity (Besley and Persson, 2010).

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.[9] The earliest homestead entries occurred in 1869 in about half of the western frontier states, about seven years following the enactment of the HSA. In 1872, the first homesteads were filed in southern PLS.[10]

When estimating (1), unit-specific covariates include state-level average farm sizes measured in the 1860 and average farm values measured in the 1850 and 1860 censuses. In theory, we should expect that homesteaders migrate to more productive land and thus excluding these pre-period measures of agricultural productivity may result in overestimating the actual impact of homestead policies. To control for selection bias arising from differences in 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 zero effect null hypothesis

$$\mathrm{H}_0 : \mathrm{S}_q(\hat{\bar{\alpha}}_t) = 0 \qquad \text{for} \;\; \mathrm{T}_0 - \tau + 1 \leq t \leq \mathrm{T}_0, \tag{7}$$

where $\tau \in \{1, 10, 25\}$ and $q \in \{1, 2\}$. This placebo null hypothesis is tested by treating $t = \{1, \ldots, \mathrm{T}_0 - \tau\}$ as the pre-period.

[9] 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`).

[10] Figure SM-1 visualizes the timing and intensity of homestead entries.

[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.

Table 1 reports the average treatment effect over the placebo post-period and randomization p-values calculated by (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.

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 $|\Pi| = \mathrm{T} - 1$ permutations.

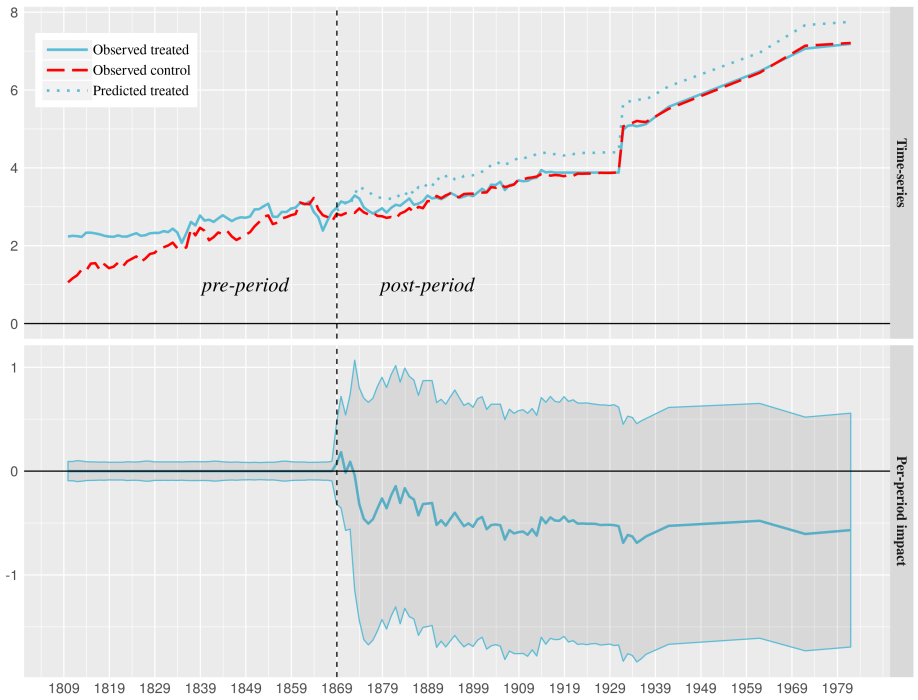
Further evidence of the consistency of the MC-NNM estimator is provided in Figure SM-4, which presents the results of placebo tests on control units using both pre- and post-period observations. Similar to the simulations on the synthetic control datasets discussed in Section 3.2, there are no missing entries in each outcome 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 $\mathrm{T}_0/\mathrm{T} \geq 0.5$, the estimator generally yields comparable error rates to PCA, synthetic control, and vertical regression estimators.

4.2 Main estimates

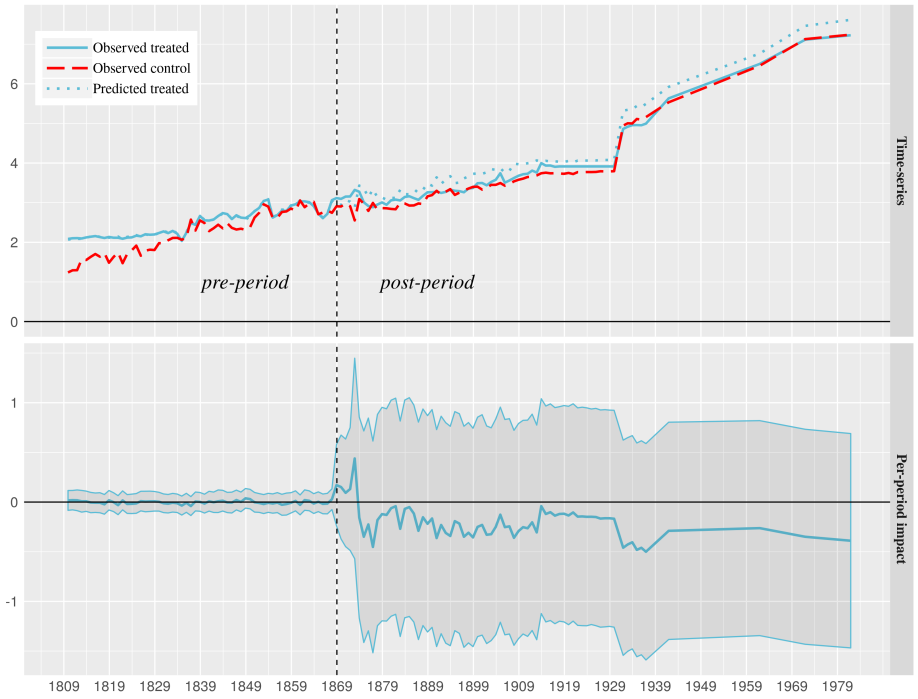
In the main analyses, I fit the MC-NNM estimator described in (1) on the entirety of observed entries in Y to recover its missing entries; i.e., the counterfactual outcomes of PLS. The value of the regularization term λ is optimally selected on the value that yields the lowest average RMSE calculated 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 observed time-series of treated units and control units along with the 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 $\hat{\bar{\alpha}}_t$, which corresponds to the estimated per-period average causal effect of treatment exposure on the treated units. These per-period causal impacts are plotted in the bottom panels. Bootstrap confidence intervals for $\hat{\bar{\alpha}}_t$ are calculated by block resampling with optimal block lengths selected by the procedure described by Politis and White (2004).

The per-period impact time-series for both outcomes are essentially zero during the pre-period and within the bounds of the bootstrap confidence intervals, which demonstrates that the model is closely fitting the pre-period observations. Per-period impacts on state government spending peak in 1870, at the same time most PLS were first exposed to homesteads, representing a 0.18 [-0.35, 0.71] log increase in per-capita expenditure. By 1876, after most PLS had been exposed to homesteads, homestead exposure decreases expenditure by 0.51 [-1.67, 0.66] log points, and the trajectory of causal impacts remains negative for the rest of the time-series. Similarly, per-period impacts on revenue peak in 1873, representing a 0.43 [-0.57, 1.44] log increase in per-capita revenues, at the same time southern PLS are exposed to homesteads. The causal impacts on revenue quickly decrease and remain negative for the remaining time-series; in 1877, exposure to homesteads confer a 0.45 [-1.51, 0.61] log point decrease in per-capita revenue.



(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 and mean observed outcomes (dashed time-series) among control units, 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 $\hat{\delta}_t$ in (4). Shaded regions represent 95% confidence intervals estimated by taking $\hat{\delta}_t \pm 1.96$ the standard error of the distribution of 1,000 bootstrap replicates of $\hat{\delta}_t$.

The estimated bootstrap confidence intervals are useful for evaluating per-period causal impacts but are not helpful in evaluating the overall effect of homestead policies. Table 2 reports randomization p-values from testing the null hypothesis of a zero effect:

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

The null hypothesis (8) can rejected at the 5% level for both outcomes, both values of q, and all three permutation schemes. Note that the relevant test statistic $S(\hat{\tilde{\alpha}}_t)$ measures the trajectory of average causal effects in absolute terms and thus does not provide information on the direction or evolution of the causal effects over time.

Table 2: Testing the null hypothesis (8).

	Expenditure		Revenue	
	$q = 1$	$q = 2$	$q = 1$	$q = 2$
$\mathrm{S}_q(\hat{\tilde{\boldsymbol{\alpha}}})$	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: $\mathrm{S}_q(\hat{\tilde{\boldsymbol{\alpha}}})$ corresponds to the test statistic described in (5) and each value beneath is the randomization p-value corresponding to each permutation structure. See footnotes to Table 1.

5 DID estimation

The matrix completion approach estimates the impact of a binary exposure to treatment 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, where 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:

$$Y_{it} = \gamma_i + \delta_t + \psi\, M_{it} + \phi\,(M_{it} \cdot H_{it}) + X_{it} + \epsilon_{it}, \tag{9}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are vectors of state and year dummies, respectively, and X is a matrix of unit- and time-varying covariates included to control for parallel trends in agricultural productivity and access to frontier lands. In the staggered adoption setting, entries in the treatment indicator $M_{it} = 1$ at $t \geq \mathrm{T}_0$, where the initial exposure time T_0 varies across units. The continuous treatment exposure variable H_{it} measures the per-capita statewide sum of homestead entries in state i and year t. The coefficient corresponding to the interaction term, $\hat{\phi}$, is the estimated average causal effect of exposure to homesteads. I use unit-stratified bootstrapped samples to construct nonparametric standard errors for $\hat{\phi}$.[12]

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 PLS. It should be emphasized that estimating (9) in a staggered adoption setting relies 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 the covariates. 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 $t < \mathrm{T}_0$ or the history of treatment exposure if $t > \mathrm{T}_0$. Violations of these assumptions would arise if the homestead policies 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.

[12]The model assumes i.i.d. errors, which understates the standard errors for $\hat{\delta}$ when the regression errors are serially correlated, or $\mathrm{Corr}(\epsilon_{it}, \epsilon_{i,t-1}) \neq 0$, 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.

5.1 DID estimates on state capacity

I estimate (9) on balanced state-year panel datasets covering state government finances from the years 1783 to 1982. The covariate matrix X_{it} includes measures of railroad access, farm sizes, and farm values. Missing values in X_{it} are imputed separately in the time periods before and after 1868, carrying the last observation forward and impute remaining missing values by carrying the next observation backward.

Table 3 reports the treatment effect estimates corresponding to the interaction term $\hat{\phi}$. The estimates indicate that a 10% increase in log per-capita homesteads is expected to significantly decrease log per-capita stage government finances by about 0.1%. The point estimates are considerably smaller in magnitude – albeit in the the same direction– as the per-period MC-NNM estimates presented in Section 4.2. The bootstrap confidence intervals around the DID estimates are considerably more narrow than those for the MC-NNM per-period impacts displayed in Figure 2 and most likely overoptimistic due to serial correlation in the DID regression errors.

Table 3: DID estimates: Impact of homestead entries on per-capita state government finances and land inequality.

	Expenditure	Revenue	Land inequality
Treatment effect ($\hat{\phi}$)	-0.013	-0.012	$-4.81 \cdot 10^{-4}$
	[-0.018, -0.009]	[-0.017, -0.008]	[$-9.756 \cdot 10^{-4}$, $-4.636 \cdot 10^{-5}$]
Adjusted r^2	0.74	0.73	0.84
n	5,247	5,372	463
Includes farm size & railroad access	Yes	Yes	No
Includes farm values	Yes	Yes	Yes
Includes state & year effects	Yes	Yes	Yes

Notes: Values in brackets represent 95% confidence intervals constructed using 1,000 state-stratified bootstrap samples.

5.2 Land inequality as a causal mechanism

Through which channels do homesteads affect state capacity? The political economy literature is largely in agreement that inequality and state capacity are inversely related. 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).

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). Note that 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. I correct for this problem by adjusting the farm Gini coefficient by the ratio of farms to adult males, as recommended by Vollrath (2013).

In Figure SM-5, a bivariate regression model yields a positive relationship between land inequality and state government finances during the period of 1860 to 1950, especially at

higher levels of inequality. This relationship points to inequality as a potential causal mechanism underlying the relationship between homesteads and state capacity.[13] The inverse relationship is consistent with the findings of Ramcharan (2010) and Vollrath (2013) in the context of taxes, revenues, and public school spending at the county-level in 1890 and 1930.

Table 3 presents DID estimates of the impact of log per-capita homesteads on land inequality at the state-level during the period of 1870 to 1950.[14] Average 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 frontier states: a 1% increase in log per-capita homesteads is expected to lower the land inequality Gini coefficient by $4.81 \cdot 10^{-6}$ points.

6 Conclusion

The findings of this paper signify that mid-nineteenth century homestead policies had long-lasting impacts that can potentially explain contemporary differences in state government capacity. MC-NNM and DID estimates imply that homestead policies — or the homestead entries authorized by those policies — had significant and negative impacts on state government expenditure and revenue that lasted a century following its implementation. The direction of these estimates is inconsistent with the observation of Engerman and Sokoloff (2005), frontier state governments sought to increase public investments in order to attract eastern migrants following the passage of the HSA, and that homesteads would increase state and local tax bases. Instead, the results are more consistent with the view that homestead policies were exploited by land speculators and natural resource companies and that the rents from public land were appropriated by the private sector.

[13] However, this relationship is subject to reverse causality because state policies determining expenditures and revenue can also shape the distribution of landownership.

[14] Since land inequality is measured every decennial, I aggregate homesteads to the next decennial year; e.g., the number of homesteads measured in 1880 is the total for the years 1871 to 1880.

I explore land inequality as a possible causal mechanism underlying the relationship between land reform and state capacity. First, I provide evidence of a positive relationship between land inequality and state government finances and that the slope of correlation increases at higher levels of inequality. A nonlinearity in the relationship between inequality and state capacity can arise in theoretical models that incorporate economic differences in political influence: greater income inequality reduces investments in fiscal capacity when elites have a monopoly on political power, however when inequality gets too high, the poor can impose redistribution through majority voting. Second, I present DID estimates that reveal per-capita homesteads significantly lowered land inequality in frontier states; although, the magnitude of the effect is negligible.

This paper makes a methodological contribution in applying matrix completion — a machine learning method commonly used for recommendation tasks — for estimating causal impacts of policy interventions on time-series cross-sectional data. In placebo tests, the matrix completion method outperforms several other 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-period observations of control units to predict counterfactuals.

In addition, I show how to evaluate the overall effect of the policy intervention using a randomization inference procedure in which p-values are obtained by permuting the time-series dimension of the data under the null. The p-values resulting from the procedure are approximately unbiased assuming that the MC-NNM estimator is consistent, which can be verified by placebo tests.

References

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

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

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

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

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

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

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

Ansell, B. and Samuels, D. J. (2015) *Inequality and Democratization: An Elite Competition Approach*. Cambridge: 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**, 313–338.

Atack, J., Margo, R. and Perlman, E. (2012) The impact of railroads on school enrollment in nineteenth century America. Available from `http://elisabethperlman.net/papers/PerlmanMargoAtack_SchoolRR.html`.

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

Athey, S. and Imbens, G. (2018) Design-based analysis in difference-in-differences settings with staggered adoption. *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: Cambridge University Press.

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

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

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

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

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

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

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

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

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

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**, 119.

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

García-Jimeno, C. and Robinson, J. A. (2008) The myth of the frontier. In *Understanding Long-Run Economic Growth: Geography, Institutions, and the Knowledge Economy*, 49–88. Chicago, IL: University of Chicago Press.

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

— (1942) The role of the land speculator in western development. *The Pennsylvania Magazine of History and Biography*, **66**, 314–333.

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

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**, 945–960.

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

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

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

Lieberman, E. S. (2002) Taxation data as indicators of state-society relations: Possibilities and pitfalls in cross-national research. *Studies in Comparative International Development*, **36**, 89–115.

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

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

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

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

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

Poulos, J. (2017) RNN-based counterfactual time-series prediction. *arXiv:1712.03553*.

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

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

Suryanarayan, P. (2017) Hollowing out the state: Franchise expansion and fiscal capacity in colonial India. Available from `https://ssrn.com/abstract=2951947`.

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

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

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

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 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**, 267–284.

Williamson, S. H. (2017) Seven ways to compute the relative value of a US dollar amount, 1774 to present. Available from `http://MeasuringWorth.com`.

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

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