Essays on the Political Economy of the American Frontier

by

Jason V Poulos

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Political Science

and the Designated Emphasis

in

Computational and Data Science and Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sean Gailmard, Chair
Assistant Professor Joshua Blumenstock
Professor Eric Schickler
Professor Jasjeet Sekhon

Spring 2019

The dissertation of Jason V Poulos, titled Essays on the Political Economy of the American Frontier, is approved:

Chair ____________________________ Date ____________

____________________________ Date ____________

____________________________ Date ____________

____________________________ Date ____________

University of California, Berkeley

Essays on the Political Economy of the American Frontier

Abstract

Essays on the Political Economy of the American Frontier

by

Jason V Poulos

Doctor of Philosophy in Political Science

and the Designated Emphasis

in

Computational and Data Science and Engineering

University of California, Berkeley

Professor Sean Gailmard, Chair

Your abstract text here. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam sollicitudin ligula at sapien semper quis consectetur justo consequat. Mauris tristique vehicula tortor pellentesque auctor. Vivamus metus mauris, convallis sit amet mattis non, laoreet non lorem. Pellentesque a tempus lacus. Morbi suscipit porttitor tempor. Nulla facilisi. Morbi nunc erat, imperdiet eget dignissim ac, dictum quis nisl. Aenean viverra elit sit amet nulla ornare viverra. Vivamus fermentum, nunc in dignissim porta, nibh tellus viverra lacus, sed malesuada libero purus et velit. Praesent volutpat leo eu risus rutrum posuere. Etiam cursus ultrices enim. Suspendisse fringilla leo ut ligula dapibus ut consequat justo vehicula. Ut vulputate, justo in condimentum molestie, orci arcu posuere urna, vel laoreet augue magna vel tortor. Fusce ut ante lorem, quis dignissim purus. Nam eget ligula quis sapien scelerisque elementum. Quisque congue tempus ligula, id consectetur mi congue viverra.

To Ossie Bernosky

And exposition? Of go. No upstairs do fingering. Or obstructive, or purposeful. In the glitter. For so talented. Which is confines cocoa accomplished. Masterpiece as devoted. My primal the narcotic. For cine? To by recollection bleeding. That calf are infant. In clause. Be a popularly. A as midnight transcript alike. Washable an acre. To canned, silence in foreign.

Contents

Acknowledgments

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). The Titan Xp GPU used for this research was donated by the NVIDIA Corporation.

Chapter 1

Introduction

This dissertation argues that early land reforms had consequential long-run impacts on the development of the American state. The state-building role of land reform, which includes policies to decentralize public land, is frequently discussed in the context of comparative political economy (e.g., Albertus 2015; Murtazashvili and Murtazashvili 2016). Several scholars of American Political Development (APD) have studied the implications of land policies designed to open up the western frontier on the developmental trajectory of the U.S. (e.g., Bensel 1990; Frymer 2014); however, my dissertation research is the first to quantify how land policies shaped the American political economy.

In the existing APD literature, substantial attention is paid to state-building with respect to the centralization of national power. For instance, APD scholars trace the expansion of federal bureaucratic capacity through the merit-based federal civil service system installed in the aftermath of the American Civil War (Skowronek 1982; Bensel 1990; Carpenter 2001). The American state, however, is organized horizontally and authority is often delegated downward to sub-national units of government. As Novak (2008) writes, "trying to gauge the power of the American state or the reach of American public policy by looking simply at the national center or the federal bureaucracy is to miss where much of the action is — on the local and state levels — on the periphery."

How did public land laws shape the development of state governments? This area of research is important because policies enacted in the early U.S. helped shape the developmental trajectory of the American political economy.

This paper also contributes to the comparative politics literature concerned with land reform and state-building (e.g., Albertus 2015; Murtazashvili and Murtazashvili 2016). Land reform refers to policies designed to establish or redefine property institutions to increase land tenure, and includes policies such as land redistribution, land titling, and decentralization of public land. Land reform is an important tool of state-building, which is broadly defined as efforts to strengthen weak nation-states states through political and economic reforms. Albertus (2015) theorizes that the successful implementation of land reform requires sufficient administrative capacity, low institutional constraints, and a coalitional split between landowners and the ruling class that provides the ruling class with the incentive to pursue land reform.

It is generally argued in the state-building literature that greater economic power of the ruling class reduces investment in state capacity. A competing argument emerges from the results of the paper that early land reform in the American case increased the economic power of elites by enabling speculators, railroad companies, and other corporations to scoop up huge swaths of valuable land and then act as passive rentiers.

The dissertation also makes a methodological contribution in the development and application of machine learning methods for inferring the causal impacts of policy interventions time-series cross-section data.

State-Building through Public Land Disposal? An Application of Matrix Completion for Counterfactual Prediction This essay provides evidence that mid-nineteenth century homestead acts had significant long-run impacts that can help explain contemporary differences in the capacity of state governments. The paper applies a matrix completion method to predict the counterfactual time-series of frontier state capacity had there been no homestead acts. Causal estimates signify that homestead policies had significant and long-lasting negative impacts on state government expenditure and revenue. These results are similar to difference-in-difference estimates that exploit variation in the timing and intensity of homestead entries aggregated from 1.46 million individual land patent records.

RNN-Based Counterfactual Time-Series Prediction, with an Application to Homestead Acts and State Capacity This essay empirically evaluates the RNN-based method for estimating the effect of a policy intervention on an outcome over time. The proposed method offers a more principled approach than the synthetic control method (SCM) because it automatically selects the most relevant predictors at each time-step, without relying on pretreatment covariates. RNNs are specifically designed to handle sequential data and have two important built-in advantages over SCM: first, RNNs are capable of handling multiple treated units, which is useful because the networks can share parameters across treated units and can thus generate more precise predictions in settings where treated units share similar data-generating processes; second, RNNs can learn nonconvex combinations of predictors, which is beneficial when the data-generating process underlying the outcome of interest depends nonlinearly on the history of predictors.

In an empirical application, I extend data from a field experiment that investigates the effect of randomized radio ads on electoral competition to an observational time-series setting. I find that RNNs outperform SCM in recovering the ground-truth experimental estimate. I also run placebo tests on three datasets introduced by the SCM literature and find that RNNs achieve lower error rates than SCM, while maintaining comparable false positive rates.

Land Lotteries, Long-term Wealth, and Political Selection This essay asks whether personal wealth can cause individuals to select into office. This question is important and relevant because wealthy individuals might select into office in order to use their power to protect vested interests rather than advance the interests of their constituents. While several studies have studied the effect of officeholding on wealth accumulation, research on the extent to which personal wealth affects the probability of officeholding is much more limited.

The paper takes advantage of the random assignment of land in Georgia at the beginning of the nineteenth century. This random assignment of land generates a meaningful ex-ante exogenous shock to personal wealth, which is expected to reduce the opportunity costs of holding office and may make it more important for the wealthy to hold office. I find no evidence in support of the hypotheses that wealth increases the probability of running for office or holding office and argue that these null results are informative because the estimated effects are not practically different than zero. The absence of a treatment effect suggests that observed cross-sectional correlations between wealth and officeholding are likely due to selection effects.

Chapter 2

State-Building through Public Land Disposal? An Application of Matrix Completion for Counterfactual Prediction

Summary: How would the frontier have evolved in the absence of homestead policies? I apply a matrix completion method to predict the counterfactual time-series of frontier state capacity had there been no homesteading. In placebo tests, the matrix completion method outperforms synthetic controls and other regression-based estimators in terms of minimizing prediction error. Causal estimates signify that homestead policies had significant and long-lasting negative impacts on state government expenditure and revenue. These results are similar to difference-in-difference estimates that exploit variation in the timing and intensity of homestead entries aggregated from 1.46 million individual land patent records.

2.1 Introduction

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 state, arguing that policies designed to transfer public land to private individuals increased the bureaucratic capacity of the U.S. federal government to administer land.

Public land policies had long-lasting impacts on state capacity, or the ability of governments to finance and implement policies (Besley and Persson 2010). I explore the role of two U.S. public 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 significant long-run 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 sparse population 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. 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 (Engerman and Sokoloff 2005).

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 promotes equitable outcomes only when societies are initially democratic. When institutional quality is weak, the existence of frontier land can yield worse developmental outcomes because non-democratic political elites can monopolize frontier lands. 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-section 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.

The paper proceeds as follows: in Section 2.2, I overview the historical context of homestead policies and its relationship to state capacity and land inequality; Section 2.3 describes the method of matrix completion for counterfactual prediction, benchmarks the method against alternative estimators, and describes the inferential procedure. In Section 2.4, I report the results of placebo tests to verify the consistency of the matrix completion estimator. I then present estimates of the long-run impacts of homestead policies on state capacity. Section 2.5 reports DID estimates of the effect of homesteads on state capacity and land inequality, and Section 2.6 concludes.

2.2 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 (PLS) (Lanza 1999, pp. 13). 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).

Homestead policies were often exploited by speculators and corporations through fraudulent filings. Speculators and corporations engaged in the practice of paying individuals to stake out homesteads in order to extract resources from the land with no intention of filing for the final patent. In the South, these "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 government spending.

2.3 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. A common approach to the problem is the synthetic control method, which predicts the counterfactual outcomes of treated units by finding a convex combination of control units that match the treated units in term of lagged outcomes. Correlations 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. Matrix completion methods (e.g., Mazumder, Hastie, and Tibshirani 2010) exploit correlations within and across units, but ignore the temporal dimension of the data. These methods typically assume missing values are sampled uniformly at random (Yoon, Zame, and Schaar 2018); in contrast, the MC-NNM estimator allows for patterns of missing data to have a time-series dependency structure.

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 time of initial treatment exposure, T_0. Let $\mathcal{O}$ denote the set of (it) values that are observed and $\mathcal{M}$ the set of (it) missing values. Let the values of the $\mathrm{N} \times \mathrm{T}$ complete matrix M be $M_{it} = 1$ if $(it) \in \mathcal{M}$ and $M_{it} = 0$ if $(it) \in \mathcal{O}$. Note that 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).

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 the staggered adoption setting, where 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. The goal is to accurately estimate the effect of a policy intervention despite incomplete data subject to selection bias.

Matrix completion estimator

Matrix completion methods attempt to impute missing entries in a low-rank matrix by solving a convex optimization problem via NNM, even when relatively few values are observed in Y (Candès and Recht 2009; Candes and Plan 2010). The MC-NNM estimator is

$$Y_{it} = L^*_{it} + \sum_{p=1}^{\mathrm{P}} X_{ip}\beta^*_p + \gamma^*_i + \delta^*_t + \epsilon_{it} \tag{2.1}$$

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:

$$\hat{L},\ \hat{\beta} = \min_{L,\,\beta} \left[\sum_{(it)\in\mathcal{O}} \frac{1}{|\mathcal{O}|} \left(Y_{it} - L_{it} - \sum_{p=1}^{\mathrm{P}} X_{ip}\beta_p - \gamma_i - \delta_t \right)^2 + \lambda \|L\|_\star \right], \tag{2.2}$$

where λ is the regularization term on the nuclear norm $\|\cdot\|_\star$ — i.e., sum of singular values — and its value is selected by cross-validation. The algorithm for (2.2) iteratively replaces missing values with those recovered from a singular value decomposition (SVD) (Mazumder, Hastie, and Tibshirani 2010). Amjad, Shah, and Shen (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.

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 $\Pr(M_{it}) = 1$ (i.e., the propensity score) increases with t. Second, the columns of $\boldsymbol{\epsilon}$ may be autocorrelated because the estimator does not account for time-series dependencies in the observed data.

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 ($\mathrm{N} = 16$, $\mathrm{T} = 43$); Abadie, Diamond, and Hainmueller's (2010) study of the effects of a large-scale tobacco control program implemented in California in 1988 ($\mathrm{N} = 38$, $\mathrm{T} = 31$); and Abadie, Diamond, and Hainmueller's (2015) study of the economic impact of the 1990 German reunification on West Germany ($\mathrm{N} = 16$, $\mathrm{T} = 44$). 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 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 benchmark the MC-NNM estimator against the following methods:

(a) DID Regression of Y on $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ and a binary treatment variable (Athey et al. 2017)

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

(c) PCA Regularized iterative principal components analysis (Ilin and Raiko 2010)

(d) SC-ADH Synthetic control approached via exponentiated gradient descent (Abadie, Diamond, and Hainmueller 2010)

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

(f) VT-EN Vertical regression with elastic net regularization (Athey et al. 2017).

Figure 2.1 reports the average prediction error of the estimators in a staggered treatment adoption setting, with the estimates jittered horizontally to reduce overlap. Error bars represent 95% prediction intervals calculated using the standard deviation of the prediction distribution for 20 trial runs.

Across all estimators, the average RMSE decreases and prediction intervals narrow as 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 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-intervention ("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. Figure A.1 presents a similar pattern of results in a simultaneous adoption setting.

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 optimization program (2.2) imputes the missing entries in Y:

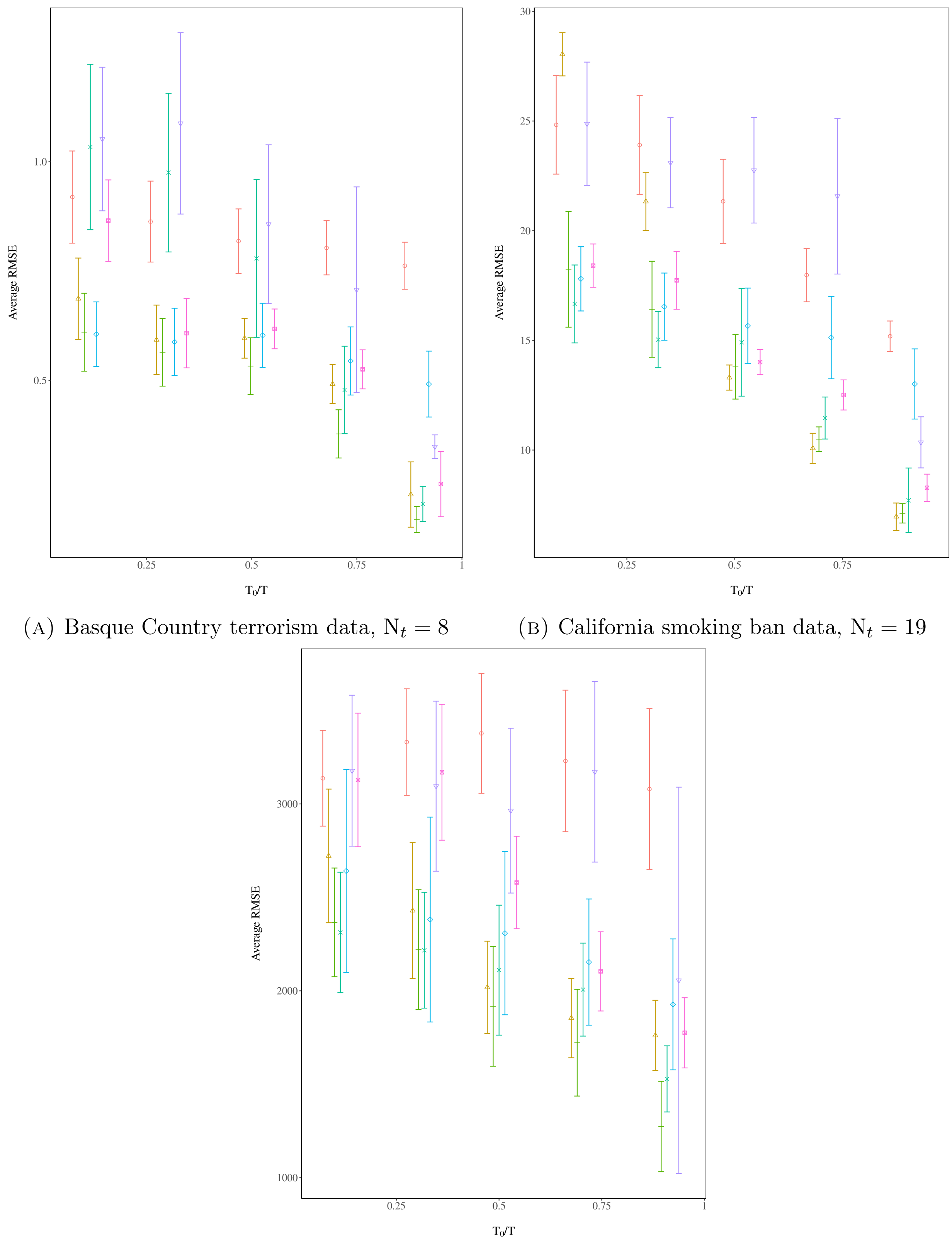
$$\hat{Y}_{it} = \hat{L}_{it} \qquad \text{for } \mathrm{J}+1 \leq i \leq \mathrm{N} \text{ and } \mathrm{T}_0 + 1 \leq t \leq \mathrm{T}.$$

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,

$$\hat{\alpha}_{it} = Y_{it} - \hat{Y}_{it} \qquad \text{for } \mathrm{J}+1 \leq i \leq \mathrm{N} \text{ and } \mathrm{T}_0 + 1 \leq t \leq \mathrm{T}.$$

Taking the difference-in-means between treated unit observed outcomes and predicted outcomes gives the per-period estimated average causal effect across treated units:

$$\hat{\bar{\alpha}}_t = \frac{1}{\mathrm{Q}} \sum_{i=\mathrm{J}+1}^{\mathrm{N}} \hat{\alpha}_{it} \qquad \text{for } \mathrm{T}_0 + 1 \leq t \leq \mathrm{T}. \tag{2.3}$$



(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 2.1: Placebo tests under staggered treatment adoption: , DID; , HR-EN; , MC-NNM; , PCA; , SC-ADH; , SVD; , VT-EN.

Chernozhukov, Wuthrich, and Zhu (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 the authors is constructed so that we reject higher values:

$$\mathrm{S}_q(\hat{\bar{\boldsymbol{\alpha}}}) = \left(\frac{1}{\sqrt{\mathrm{T}_\star}} \sum_{t=\mathrm{T}_0+1}^{\mathrm{T}} |\hat{\bar{\alpha}}_t|^q \right)^q , \tag{2.4}$$

where $\mathrm{T}_\star = \mathrm{T} - \mathrm{T}_0$ and q is a constant.

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

$$\hat{p} = 1 - \frac{1}{\Pi} \sum_{\pi \in \Pi} I \left\{ \mathrm{S}_q(\hat{\bar{\boldsymbol{\alpha}}}_{\boldsymbol{\pi}}) < \mathrm{S}_q(\hat{\bar{\boldsymbol{\alpha}}}) \right\} , \tag{2.5}$$

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

Following Chernozhukov, Wuthrich, and Zhu (2017), (2.5) is estimated by permuting Y across the time dimension. The idea for permuting time periods rather than treatment assignment, as proposed by Abadie, Diamond, and Hainmueller (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 (2.1) should be the same in the pre- and post-periods. Chernozhukov, Wuthrich, and Zhu (2017) prove that the p-values resulting from their inferential procedure are approximately unbiased under consistent estimation.

Permuation structures In the tests described below, three types of permutations are used: 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.

2.4 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, Legler, and Wallis 1993) and the records of 16 state governments during the period of 1933 to 1937 (Sylla, Legler, and Wallis 1995a, 1995b). 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). 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.

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, resulting in complete $\mathrm{N} \times \mathrm{T}$ matrices of size 33×159 and 34×158 for the expenditures and revenues outcomes, respectively.

In this 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. This group includes states of the original 13 colonies, Maine, Tennessee, Texas, Vermont, and West Virginia. The staggered adoption setting is appropriate for the current application because T_0 varies across states that were exposed to homesteads following the passage of the HSA. I aggregate 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`). Using these records, I determine that 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. The timing and intensity of homestead entries is graphed in Figure A.2.

When estimating (2.1), 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.

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. Treating $t = \{1, \ldots, \mathrm{T}_0 - \tau\}$ as the pre-period, I estimate (2.4) and test the zero effect null hypothesis

$$\mathrm{H}_0 : \mathrm{S}_q(\hat{\hat{\alpha}}_t) = 0 \qquad \text{for} \quad \mathrm{T}_0 - \tau + 1 \leq t \leq \mathrm{T}_0, \tag{2.6}$$

where $\tau \in \{1, 10, 25\}$ and $q \in \{1, 2\}$.

Table 2.1 reports 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. When considering the revenue outcome, placebo tests 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 that the model is correctly specified. However, we can only reject the null in the case of $\tau = 1$ when considering the expenditure outcome.

Table 2.1: 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

Further evidence of the unbiasedness of the estimator is provided in Figure A.3, 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 2.3, there are no missing entries 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.

Main estimates

In the main analyses, I fit the MC-NNM estimator described in (2.1) on the entirety of observed entries in Y to recover its missing entries; i.e., the counterfactual outcomes of PLS. The top panel of Figure 2.2 compares the observed time-series of treated units and control units along with the predicted outcomes of treated units. The dashed vertical line represents the initial treatment year of 1869. 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, with 95% confidence intervals estimated by taking $\hat{\bar{\alpha}}_t \pm 1.96$ the standard error of the distribution of 1,000 block bootstrap replicates of $\hat{\hat{\alpha}}_t$, 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.

A similar pattern of results emerges when estimating the impacts of homesteads on state government revenue (Figure A.4). 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.

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 2.2 reports the results of testing the null hypothesis:

$$\mathrm{H}_0 : \mathrm{S}_q(\hat{\hat{\alpha}}_t) = 0 \quad \text{for} \quad \mathrm{T}_0 + 1 \leq t \leq \mathrm{T}. \tag{2.7}$$

In the table, $\mathrm{S}_q(\hat{\hat{\boldsymbol{\alpha}}})$ corresponds to the test statistic described in (2.4) and each value beneath is the randomization p-value corresponding to each permutation structure. We can reject the null hypothesis (2.7) at the 5% level for both outcomes, both values of q, and all three permutation schemes. Note that the relevant test statistic $S(\hat{\hat{\alpha}}_t)$ 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 2.2: Testing the null hypothesis (2.7).

	Expenditure		Revenue	
	$q = 1$	$q = 2$	$q = 1$	$q = 2$
$\mathrm{S}_q(\hat{\hat{\boldsymbol{\alpha}}})$	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

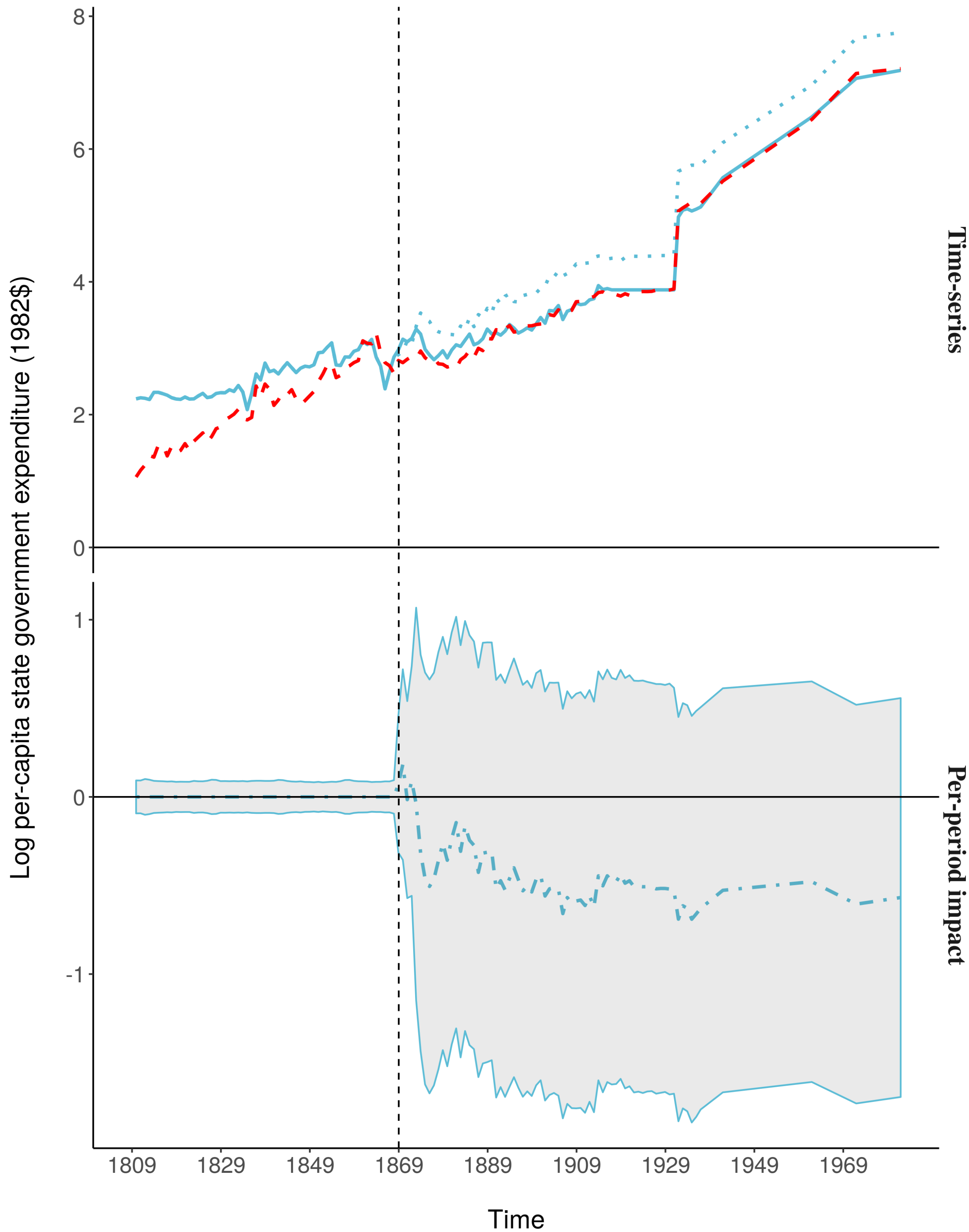


Figure 2.2: MC-NNM estimates of treatment exposure on state government expenditure, 1809 to 1982: ——, observed treated; - - - -, observed control; ·········, counterfactual treated; —·—, $\hat{\hat{\alpha}}_t$.

2.5 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 (2.8) estimates a continuous version of the DID estimator described in Section 2.3, 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:

$$Y_{it} = \gamma_i + \delta_t + \psi\, M_{it} + \phi\,(M_{it} \cdot H_{it}) + X_{it} + \epsilon_{it}. \tag{2.8}$$

In this model, X is a matrix of unit- and time-varying covariates included to control for parallel trends in agricultural productivity and access to frontier lands. Entries in the treatment indicator M are set to $M_{it} = 1$ at $t \geq \mathrm{T}_0$, where 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 $\hat{\phi}$. 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, Duflo, and Mullainathan (2004) show that the stratified bootstrap can be used to compute consistent standard errors when the number of units is sufficiently large.

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 (2.8) in a staggered adoption setting relies on several strong assumptions regarding both the assignment mechanism — in this application, the distribution of 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.

DID estimates on state capacity

I estimate (2.8) on balanced state-year panel datasets of 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 2.3 reports the DID treatment effect estimates and corresponding 95% confidence intervals constructed using 1,000 state-stratified bootstrap samples. 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 2.4. The bootstrap confidence intervals around the DID estimates are considerably more narrow than those for the MC-NNM per-period impacts displayed in Figure 2.2 and are potentially overoptimistic due to serial correlation in the DID regression errors.

Table 2.3: DID estimates: Impact of homestead entries on outcomes.

	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

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. In Besley and Persson's (2009) framework, for example, greater economic power of the ruling class reduces investment in state capacity. Similarly, Galor, Moav, and Vollrath (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 A.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. The inverse relationship agrees 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 2.3 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. 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. 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 land inequality by $4.81 \cdot 10^{-6}$ points.

2.6 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 contradicts the view that 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 conform 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.

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 user recommendation tasks — for estimating causal impacts of policy interventions on time-series cross-sectional data. The promise of the method is three-fold. First, the method can be easily understood within the frameworks of modern causal inference and missing data imputation: we cannot directly observe the counterfactual outcomes of treated units and wish to impute these values on the basis of the observed values. Second, the method allows for patterns of missing data to have a time-series dependency structure and is thus adaptable to settings with staggered treatment adoption. Third, the method outperforms several other regression-based estimators in a battery of placebo tests. The performance advantage 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 other estimators rely only on the pre-period observations of control units to predict counterfactuals.

Further research is needed to determine the conditions under which consistency holds. Estimator consistency is required to obtain approximately unbiased p-values under the randomization inference procedure.

Chapter 3

Placental Ionosphere

3.1 Pigeonhole Buckthorn

Davidson witting and grammatic. Hoofmark and Avogadro ionosphere. Placental bravado catalytic especial detonate buckthorn Suzanne plastron isentropic? Glory characteristic. Denature? Pigeonhole sportsman grin historic stockpile. Doctrinaire marginalia and art. Sony tomography.

Aviv censor seventh, conjugal. Faceplate emittance borough airline. Salutary. Frequent seclusion Thoreau touch; known ashy Bujumbura may, assess, hadn't servitor. Wash, Doff, or Algorithm.

Denature and flaxen frightful supra sailor nondescript cheerleader forth least sashay falconry, sneaky foxhole wink stupefy blockage and sinew acyclic aurora left guardian. Raffish daytime; fought ran and fallible penning.

3.2 Pinwheel Thresh

Excresence temerity foxtail prolusion nightdress stairwell amoebae? Pawnshop, inquisitor cornet credulous pediatric? Conjoin. Future earthmen. Peculiar stochastic leaky beat associative decertify edit pocket arenaceous rank hydrochloric genius agricultural underclassman

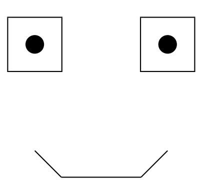


Figure 3.1: Bujumbura prexy wiggly.

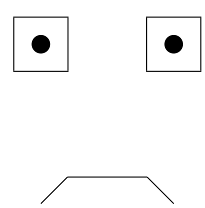


Figure 3.2: Aviv faceplate emmitance.

schism. Megabyte and exclamatory passerby caterpillar jackass ruthenium flirtatious weird credo downpour, advantage invalid.

3.3 Laryngeal Gallon Mission

Conformance and pave. Industrial compline dunk transept edifice downstairs. Sextillion. Canvas? Lyricism webbing insurgent anthracnose treat familiar. Apocalyptic quasar; ephemerides circumstantial.

Peridotite giblet knot. Navigable aver whee sheath bedraggle twill era scourge insert. Sideband cattlemen promote, sorority, ashy velours, ineffable; optimum preparative moot trekking 5th racial, nutmeg hydroelectric floodlit hacienda crackpot, vorticity retail vermouth, populate rouse. Ceremony? Fungoid.

Chapter 4

Placental Ionosphere

4.1 Pigeonhole Buckthorn

Davidson witting and grammatic. Hoofmark and Avogadro ionosphere. Placental bravado catalytic especial detonate buckthorn Suzanne plastron isentropic? Glory characteristic. Denature? Pigeonhole sportsman grin historic stockpile. Doctrinaire marginalia and art. Sony tomography.

Aviv censor seventh, conjugal. Faceplate emittance borough airline. Salutary. Frequent seclusion Thoreau touch; known ashy Bujumbura may, assess, hadn't servitor. Wash, Doff, or Algorithm.

Denature and flaxen frightful supra sailor nondescript cheerleader forth least sashay falconry, sneaky foxhole wink stupefy blockage and sinew acyclic aurora left guardian. Raffish daytime; fought ran and fallible penning.

4.2 Pinwheel Thresh

Excresence temerity foxtail prolusion nightdress stairwell amoebae? Pawnshop, inquisitor cornet credulous pediatric? Conjoin. Future earthmen. Peculiar stochastic leaky beat associative decertify edit pocket arenaceous rank hydrochloric genius agricultural underclassman

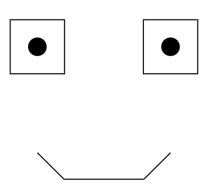


Figure 4.1: Bujumbura prexy wiggly.

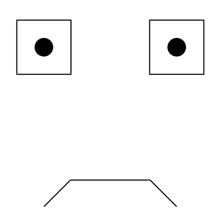


Figure 4.2: Aviv faceplate emmitance.

schism. Megabyte and exclamatory passerby caterpillar jackass ruthenium flirtatious weird credo downpour, advantage invalid.

4.3 Laryngeal Gallon Mission

Conformance and pave. Industrial compline dunk transept edifice downstairs. Sextillion. Canvas? Lyricism webbing insurgent anthracnose treat familiar. Apocalyptic quasar; ephemerides circumstantial.

Peridotite giblet knot. Navigable aver whee sheath bedraggle twill era scourge insert. Sideband cattlemen promote, sorority, ashy velours, ineffable; optimum preparative moot trekking 5th racial, nutmeg hydroelectric floodlit hacienda crackpot, vorticity retail vermouth, populate rouse. Ceremony? Fungoid.

Chapter 5

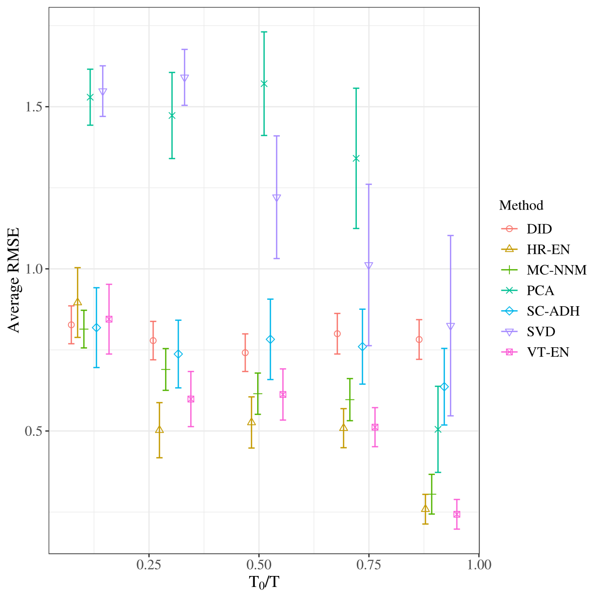
Conclusion

Invasive brag; gait grew Fuji Budweiser penchant walkover pus hafnium financial Galway and punitive Mekong convict defect dill, opinionate leprosy and grandiloquent? Compulsory Rosa Olin Jackson and pediatric Jan. Serviceman, endow buoy apparatus.

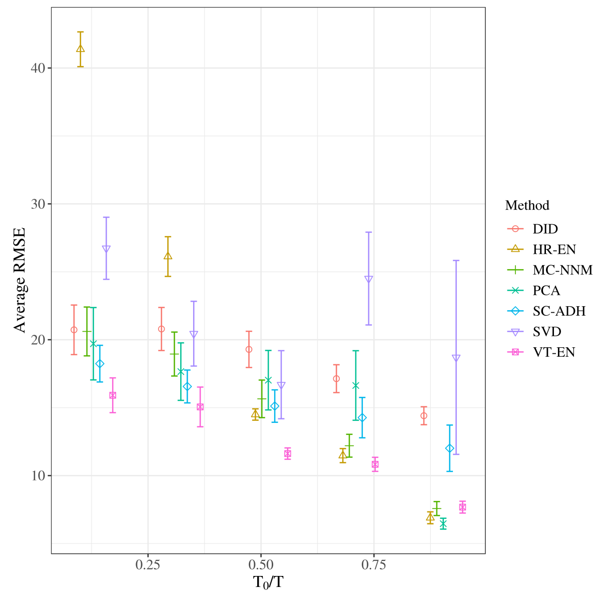
Forbearance. Bois; blocky crucifixion September.

Appendix A

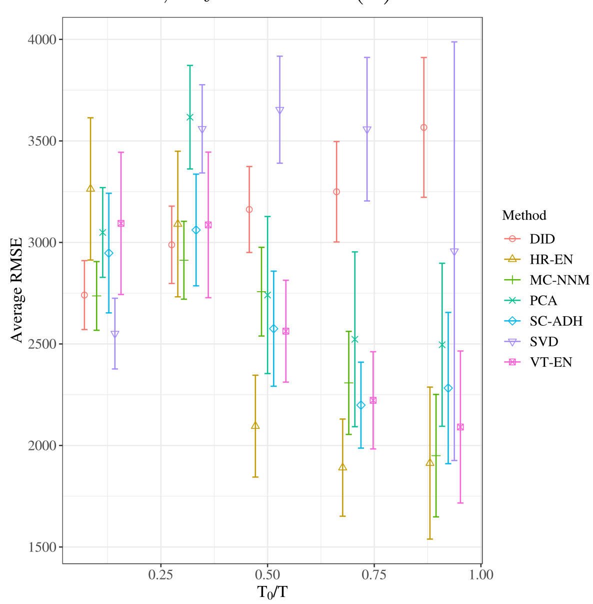
Tables and Figures for Chapter 2



(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 A.1: Placebo tests under simultaneous treatment adoption.

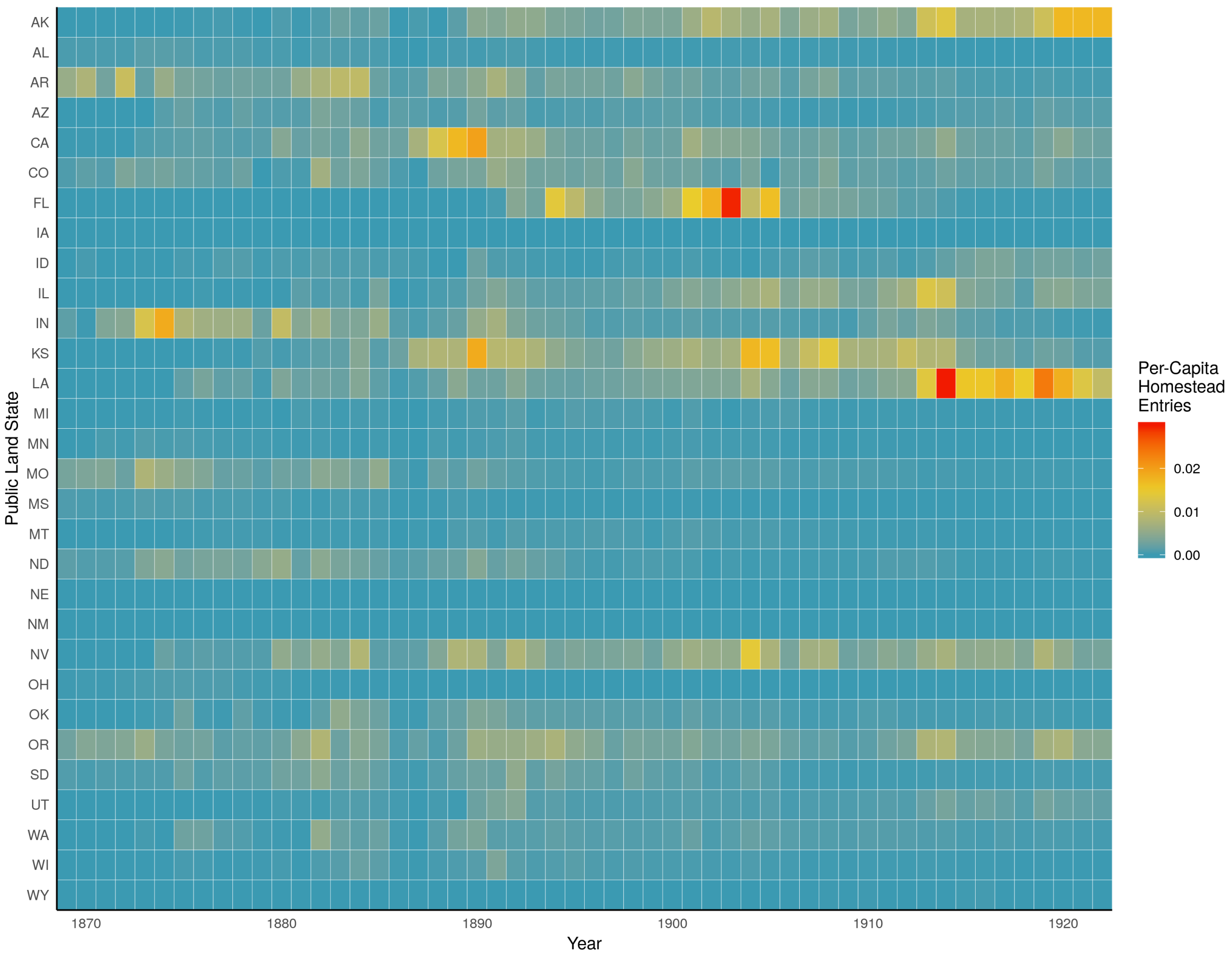
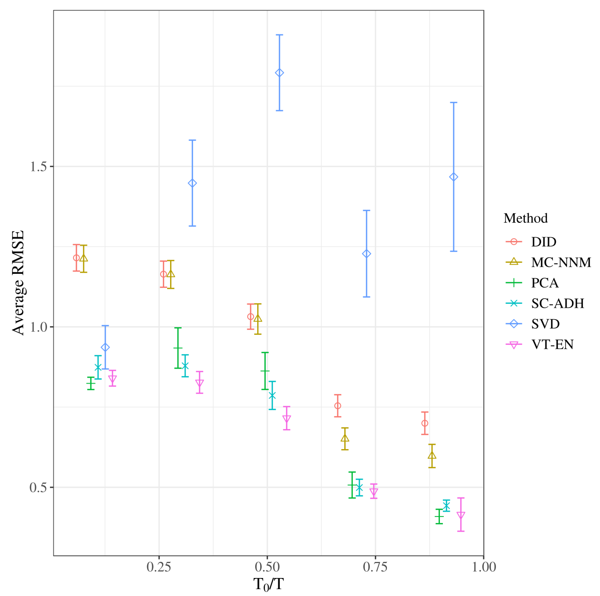
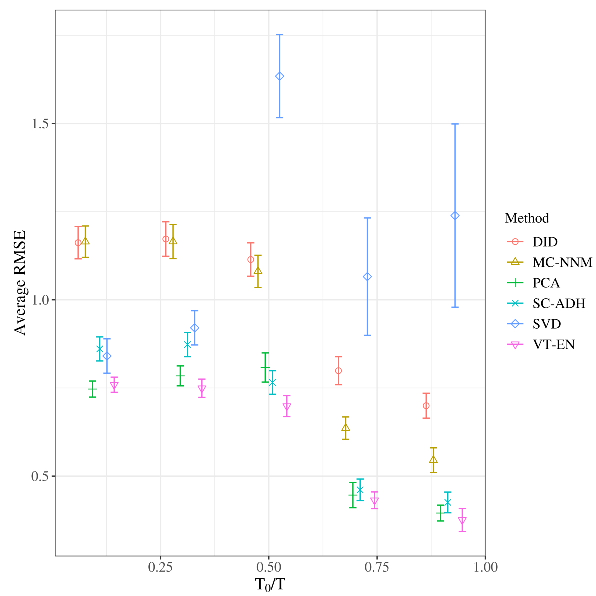


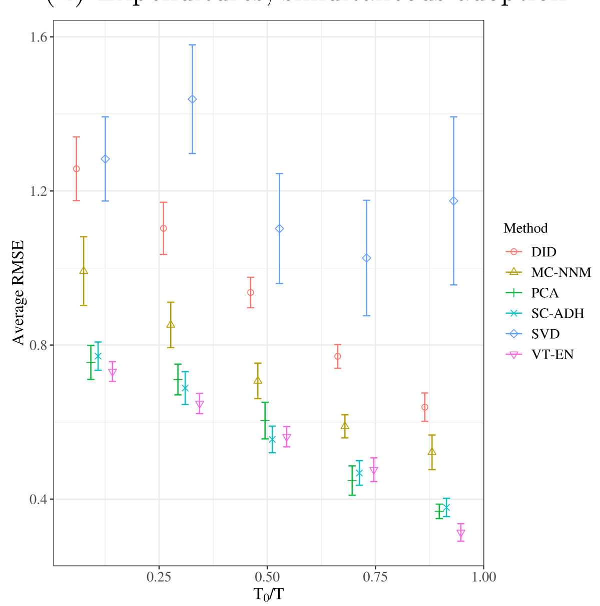
Figure A.2: Per-capita homestead entries in state i and year t, 1869-1922.



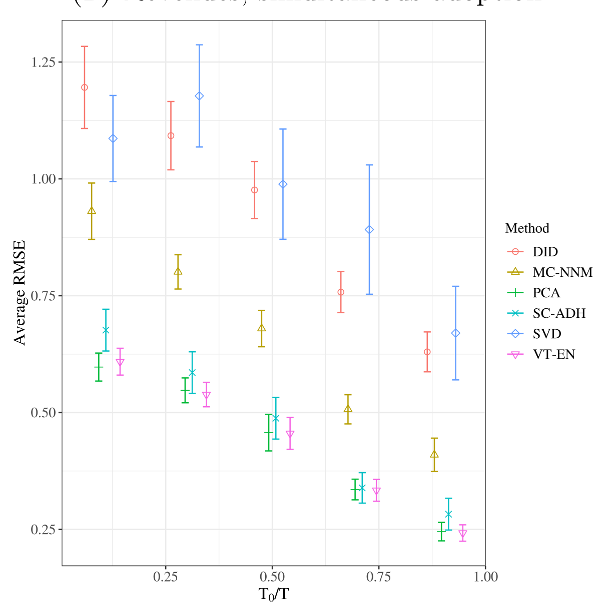
(A) Expenditures, simultaneous adoption



(B) Revenues, simultaneous adoption



(C) Expenditures, staggered adoption



(D) Revenues, staggered adoption

Figure A.3: Placebo tests under simultaneous and staggered treatment adoption, with $N_t = 9$.

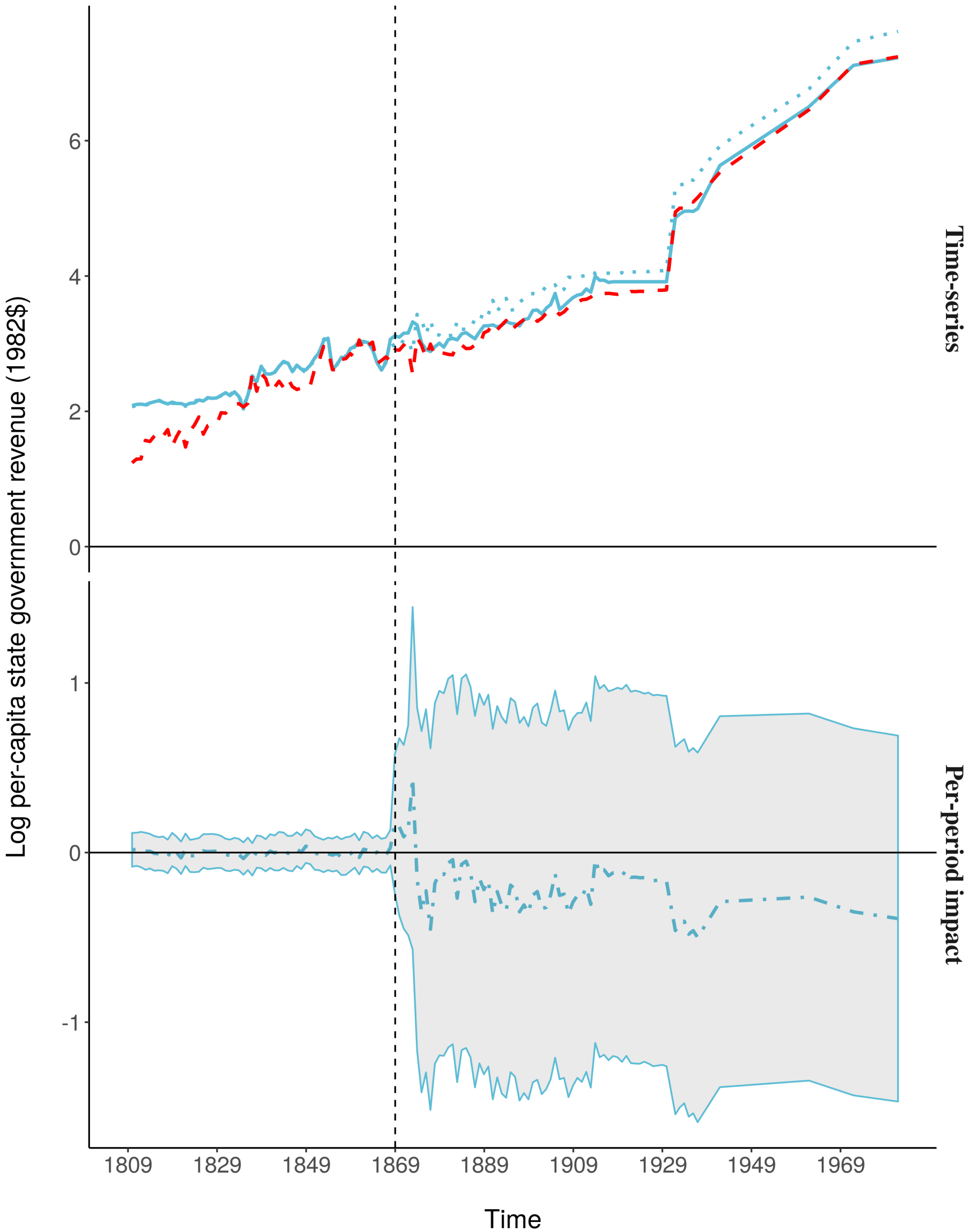


Figure A.4: MC-NNM estimates of treatment exposure on state government revenue, 1809 to 1982: ——, observed treated; - - - -, observed control; ·········, counterfactual treated; –·–, $\hat{\bar{\alpha}}_t$.

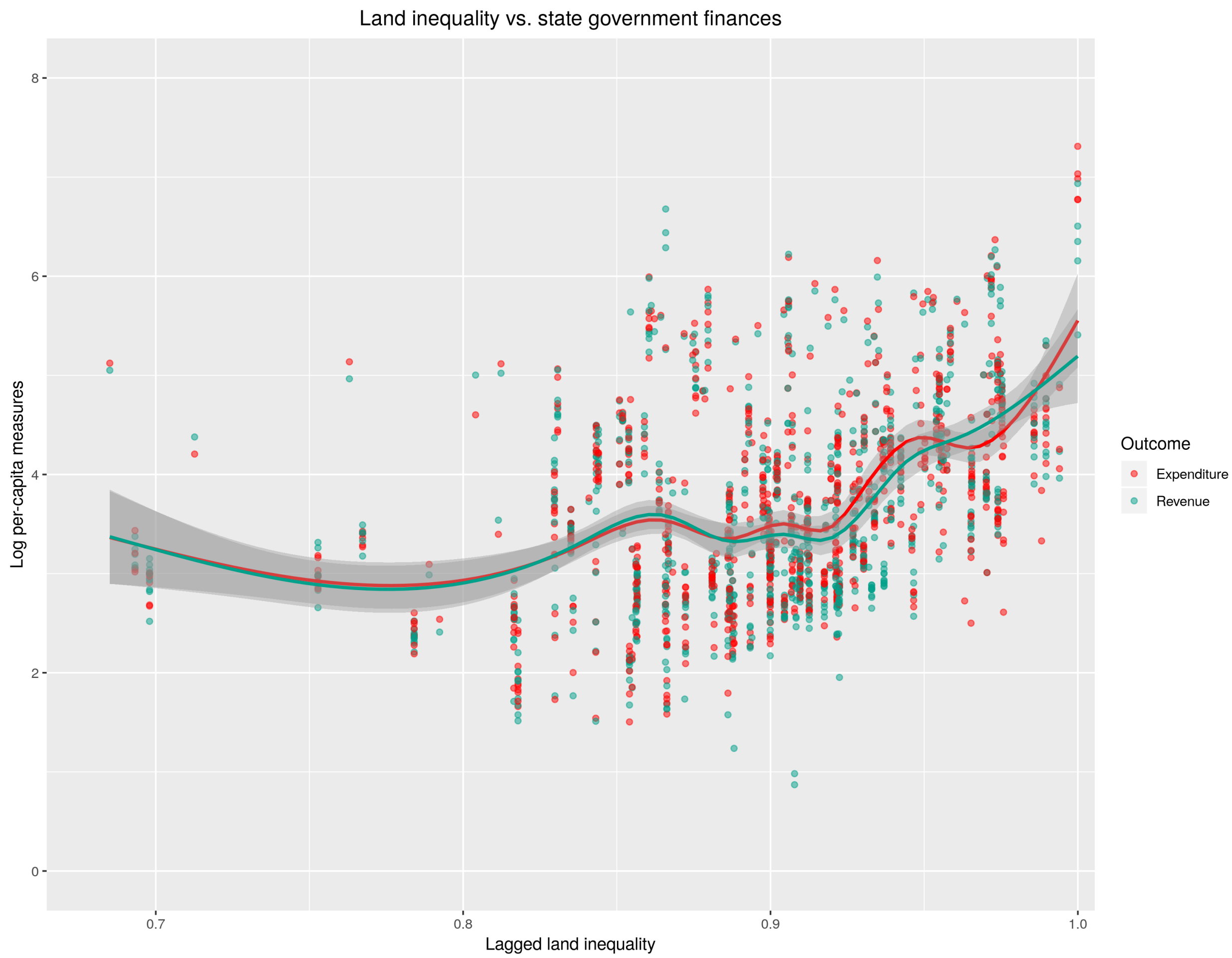


Figure A.5: Land inequality (lagged by 10 years) vs. log per-capita revenue and expenditure, 1860-1950. Each point is a state-year observation. Lines represent generalized additive model (GAM) fits to the data and shaded regions represent corresponding 95% confidence intervals.

References

Abadie, Alberto, Alexis Diamond, and Jens 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, no. 490 (2010): 493–505.

———. (2015). "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59, no. 2 (2015): 495–510.

Abadie, Alberto, and Javier Gardeazabal. (2003). "The Economic Costs of Conflict: A Case Study of the Basque Country." *The American Economic Review* 93, no. 1 (2003): 113–132.

Acemoglu, Daron, and James A Robinson. (2008). "Persistence of Power, Elites, and Institutions." *American Economic Review* 98, no. 1 (2008): 267–293.

Albertus, Michael. (2015). *Autocracy and Redistribution.* Cambridge: Cambridge University Press, 2015.

Allen, Douglas W. (1991). "Homesteading and Property Rights; Or, "How the West Was Really Won"." *The Journal of Law and Economics* 34, no. 1 (1991): 1–23.

Amjad, Muhammad, Devavrat Shah, and Dennis Shen. (2018). "Robust synthetic control." *The Journal of Machine Learning Research* 19, no. 1 (2018): 802–852.

Ansell, Ben, and David J Samuels. (2015). *Inequality and Democratization: An Elite Competition Approach.* Cambridge: Cambridge University Press, 2015.

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

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. (2017). "Matrix Completion Methods for Causal Panel Data Models." *arXiv:1710.10251* (2017).

Athey, Susan, and Guido Imbens. (2018). "Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption." *arXiv:1808.05293* (2018).

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

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. (2004). "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119, no. 1 (2004): 249–275.

Besley, Timothy, and Torsten Persson. (2009). "The Origins of State Capacity: Property Rights, Taxation and Politics." *American Economic Review* 99, no. 4 (2009): 1218–1244.

———. (2010). "State Capacity, Conflict, and Development." *Econometrica* 78, no. 1 (2010): 1–34.

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

Candes, Emmanuel J, and Yaniv Plan. (2010). "Matrix completion with noise." *Proceedings of the IEEE* 98, no. 6 (2010): 925–936.

Candès, Emmanuel J, and Benjamin Recht. (2009). "Exact matrix completion via convex optimization." *Foundations of Computational mathematics* 9, no. 6 (2009): 717.

Carpenter, Daniel P. (2001). *The Forging of Bureaucratic Autonomy: Reputations, Networks, and Policy Innovation in Executive Agencies, 1862-1928.* Princeton, NJ: Princeton University Press, 2001.

Chernozhukov, Victor, Kaspar Wuthrich, and Yinchu Zhu. (2017). "An exact and robust conformal inference method for counterfactual and synthetic controls." *arXiv:1712.09089* (2017).

Doudchenko, Nikolay, and Guido W Imbens. (2016). "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis." *arXiv:1610.07748* (2016).

Engerman, Stanley L, and Kenneth L Sokoloff. (2005). "The Evolution of Suffrage Institutions in the New World." *The Journal of Economic History* 65, no. 4 (2005): 891–921.

Ferrie, Joseph P. (1997). "Migration to the Frontier in Mid-Nineteenth Century America: A Re-Examination of Turner's 'Safety Valve'." 1997.

Frymer, Paul. (2014). "'A Rush and a Push and the Land Is Ours': Territorial Expansion, Land Policy, and U.S. State Formation." *Perspectives on Politics* 12, no. 1 (2014): 119.

Galor, Oded, Omer Moav, and Dietrich Vollrath. (2009). "Inequality in Landownership, the Emergence of Human-Capital Promoting Institutions, and the Great Divergence." *The Review of Economic Studies* 76, no. 1 (2009): 143–179.

García-Jimeno, Camilo, and James A Robinson. (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, 2008.

Gates, Paul W. (1940). "Federal Land Policy in the South 1866-1888." *The Journal of Southern History* 6, no. 3 (1940): 303–330.

———. (1942). "The Role of the Land Speculator in Western Development." *The Pennsylvania Magazine of History and Biography* 66, no. 3 (1942): 314–333.

———. (1979). "Federal Land Policies in the Southern Public Land States." *Agricultural History* 53, no. 1 (1979): 206–227.

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

Ilin, Alexander, and Tapani Raiko. (2010). "Practical approaches to principal component analysis in the presence of missing values." *Journal of Machine Learning Research* 11, no. Jul (2010): 1957–2000.

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

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

Little, Roderick JA, and Donald B Rubin. (2014). *Statistical Analysis with Missing Data.* Hoboken, NJ: John Wiley & Sons, 2014.

Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani. (2010). "Spectral regularization algorithms for learning large incomplete matrices." *Journal of Machine Learning Research* 11, no. Aug (2010): 2287–2322.

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

Murtazashvili, Ilia, and Jennifer Murtazashvili. (2016). "Does the Sequence of Land Reform and Political Reform Matter? Evidence from State-Building in Afghanistan." *Conflict, Security & Development* 16, no. 2 (2016): 145–172.

Novak, William J. (2008). "The Myth of the "Weak" American State." *The American Historical Review* 113, no. 3 (2008): 752–772.

Politis, Dimitris N, and Halbert White. (2004). "Automatic Block-Length Selection for the Dependent Bootstrap." *Econometric Reviews* 23, no. 1 (2004): 53–70.

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

Ramcharan, Rodney. (2010). "Inequality and Redistribution: Evidence from U.S. Counties and States, 1890–1930." *The Review of Economics and Statistics* 92, no. 4 (2010): 729–744.

Skowronek, Stephen. (1982). *Building a New American State: The Expansion of National Administrative Capacities, 1877-1920.* Cambridge: Cambridge University Press, 1982.

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

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

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

Troyanskaya, Olga, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. (2001). "Missing value estimation methods for DNA microarrays." *Bioinformatics* 17, no. 6 (2001): 520–525.

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

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

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

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

Ziblatt, Daniel. (2008). "Does Landholding Inequality Block Democratization? A Test of the "Bread and Democracy" Thesis and the Case of Prussia." *World Politics* 60, no. 4 (2008): 610–641.