THE UNIVERSITY OF CHICAGO

ESSAYS ON ECONOMIC GEOGRAPHY AND DEVELOPMENT

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

KENNETH C. GRIFFIN DEPARTMENT OF ECONOMICS

BY

SHANON HSUAN-MING HSU

CHICAGO, ILLINOIS

JUNE 2024

TABLE OF CONTENTS

LIST OF FIGURES . . . v

LIST OF TABLES . . . vii

ACKNOWLEDGMENTS . . . ix

ABSTRACT . . . xi

1 COERCIVE GROWTH: FORCED RESETTLEMENT, AGGLOMERATION, AND ECONOMIC DEVELOPMENT IN MALAYSIA . . . 1
- 1.1 Introduction . . . 1
- 1.2 Historical Context . . . 9
  - 1.2.1 Chinese in British Malaya . . . 10
  - 1.2.2 The Squatter Problem in the Malayan Emergency . . . 11
  - 1.2.3 The Briggs Plan: Emergency Resettlement . . . 13
- 1.3 Data . . . 17
  - 1.3.1 Resettlement Density . . . 17
  - 1.3.2 Population Census, 1931-2000 . . . 18
  - 1.3.3 Other Data Sources . . . 19
- 1.4 Empirical Strategy . . . 20
  - 1.4.1 Empirical Specification . . . 20
  - 1.4.2 Pre-characteristic Balance . . . 29
- 1.5 Results . . . 30
  - 1.5.1 Population Growth and Changes in Ethnic Composition . . . 30
  - 1.5.2 Economic Structure . . . 33
  - 1.5.3 Household Income . . . 35
  - 1.5.4 Discussion . . . 37
- 1.6 Model . . . 38
  - 1.6.1 Production . . . 38
  - 1.6.2 Consumption . . . 39
  - 1.6.3 Sectoral labor supply . . . 39
  - 1.6.4 Migration . . . 42
  - 1.6.5 Trade . . . 44
  - 1.6.6 Static equilibrium . . . 45
- 1.7 Identification and Estimation . . . 46
  - 1.7.1 Identification Strategy . . . 47
  - 1.7.2 Estimation . . . 56
- 1.8 Conclusion . . . 61
- 1.A Figures . . . 63
- 1.B Tables . . . 70
- 1.C Theoretical Results . . . 79

1.C.1 Properties of Fréchet Distribution . . . 79
1.C.2 Sectoral Labor Supply . . . 80
1.C.3 Migration . . . 83
1.D Appendix Figures . . . 85
1.E Appendix Tables . . . 95

2 GAINING STEAM: INCUMBENT LOCK-IN AND ENTRANT LEAPFROGGING 103
2.1 Introduction . . . 103
2.2 Context and Data Construction . . . 111
2.2.1 Water and Steam Power in US Mills . . . 111
2.2.2 County Waterpower Potential . . . 114
2.2.3 Census of Manufactures, Establishment level Data . . . 117
2.2.4 Data Linking . . . 120
2.3 Estimating Differences by County Waterpower Potential . . . 121
2.3.1 Waterpower Potential, Power Use, and Mill Growth . . . 125
2.3.2 Non-Mill Manufacturing, Steam-Use, and Backward Linkages to Steam Production . . . 127
2.3.3 Potential Other Forces Driving Steam Adoption . . . 129
2.3.4 Robustness to Linkage Error . . . 131
2.4 Key Empirical Patterns . . . 133
2.4.1 Cost Structures for Steam and Water . . . 135
2.4.2 Operating Costs . . . 137
2.4.3 Barriers to Switching Technologies . . . 138
2.4.4 Alternative Reasons for Lower Steam Adoption among Incumbents . . . 138
2.5 A Model of Steam Adoption . . . 142
2.5.1 Static Choices: Production and Demand . . . 142
2.5.2 Dynamic Choices: Firm Entry and Power Choice . . . 144
2.5.3 Equilibrium . . . 145
2.5.4 The Arrival of Steam . . . 146
2.5.5 Parametric Assumptions . . . 147
2.5.6 Solution Algorithms . . . 150
2.6 Structural Estimation . . . 151
2.6.1 Estimation Strategy . . . 151
2.6.2 Estimation Results . . . 157
2.6.3 Model Validation . . . 160
2.7 Counterfactual Experiments . . . 163
2.7.1 Local Waterpower Potential, Switching Barriers, and the Incidence of Steam Power . . . 163
2.7.2 "Cash for Clunkers" Policy Counterfactual . . . 167
2.7.3 Fixed Costs and the Speed of Technology Adoption . . . 169
2.8 Conclusion . . . 171
2.A Figures . . . 173
2.B Tables . . . 184

2.C Establishment-Level Manuscripts from the Census of Manufactures 197
2.C.1 Variable Coverage 198
2.C.2 Digitization and Processing of the Census Manuscripts 201
2.C.3 Adjustment for County Border Changes 204
2.C.4 Creating a Linked Panel of Mills 205
2.D Measuring County Waterpower Potential 212
2.D.1 NHDPlusV2 Data 212
2.D.2 Theoretical Water Power 213
2.D.3 Practical Water Power. 214
2.D.4 Aggregating to County Waterpower Potential 215
2.E Other County-Level Data 216
2.F Switching Case Studies from Historical Society Records 218
2.G Alternative Specifications 221
2.H Solution Algorithms 226
2.H.1 Dynamic Programming 226
2.H.2 Dynamic Equilibrium 227
2.I Structural Estimation 234
2.I.1 Estimation Procedure 234
2.I.2 Identification of Structural Parameters 236
2.J Counterfactual Experiments 237
2.J.1 Option Value Decomposition 237
2.J.2 Consumer Surplus 238
2.K Appendix Figures 240
2.L Appendix Tables 259

REFERENCES 280

LIST OF FIGURES

1.1 The New Villages and Transportation Network 63
1.2 Permutation Procedure: Counterfactual Site Selection 64
1.3 County Resettlement Density, Expected and Residualized 65
1.4 County Population Growth from 1947 to 1957, by Ethnic Group 66
1.5 Changes in Population Distribution from 1931 to 2000, by County Resettlement Density 67
1.6 Distribution of Marginal Productivity Spillovers, by Ethnic Group 68
1.7 Distribution of Marginal Amenity Spillovers, by Ethnic Group 69
1.8 Employment Share in 1947, by Ethnic Group 86
1.9 Population Distribution in 1947 87
1.10 The Black Areas in 1957 88
1.11 Population Distribution of the Squatters 89
1.12 County Resettlement Density, compared to Expected Resettlement Density Predicted by Gravity Model 90
1.13 Village Resettled Population, compared to Expected Resettled Population Predicted by Gravity Model 91
1.14 Built-up Capital in 1990 92
1.15 Chinese Employment Share Within Manufacturing in 1947, by Industry 93
1.16 Convexity of the Loss Function in Estimating Migration Costs 94

2.1 Components of County Waterpower Potential 174
2.2 County Waterpower Potential, Measured and Residualized 175
2.3 Power Source By Industry 176
2.4 Share of Mills using Steam Power, by Decade and County Waterpower Potential 177
2.5 Steam-use Share, for Entrants and Water Incumbents 178
2.6 Mill Size Distribution, by Power Source 179
2.7 Mill Size Distribution, by County Waterpower Potential 180
2.8 Water Technology and the Impacts of Steam Power 181
2.9 Water-to-Steam Switching Subsidies: Steam Adoption and Annual Costs 182
2.10 Technology Adoption and Fixed Costs 183
2.11 River Segment Flow Rates, in the 1880 Water Census Compared to NHDPlusV2 241
2.12 Selected Coverage in the 1880 Water Census, Compared to Comprehensive NHDPlusV2 Data 242
2.13 Example Census Images: The Rogers' Lumber Mill 243
2.14 Distribution of County-Level Manufacturing Revenue, in County Tabulations and Aggregated Establishment-level Data 244
2.15 Unreported Data in County-Industry Tabulations, for Flour and Lumber Mills, Compared to Aggregated Establishment-Level Data 245
2.16 Geographic Concentration of Production in 1850, by Industry 246
2.17 Distribution of Total Horsepower Installed, by Power Source 247
2.18 Distribution of Hand-Links' ML-Model Probability, by Type and Waterpower Potential 248

2.19 "False Match Rate" and "Found Match Rate" of Machine-Learning Model, Compared to Hand-Links, by ML-Model Cutoff Value 249
2.20 Growth in Mill Revenue, by Steam Switching Choice 250
2.21 Mill Size by Power Source, Within-County 251
2.22 Initial Mill Size, for Exiters and Survivors 252
2.23 Mill Size for Entrants and Incumbents, within Power Source 253
2.24 Mill Growth for Incumbents and Successive Generations of Entrants 254
2.25 Estimated Relationship between Water Powered Mills in 1850 and County Waterpower Potential, Excluding Rivers with Widths Above Different Cutoffs 255
2.26 Water and Steam Adoption Costs: Structural Estimates 256
2.27 Water Technology and the Impacts of Steam Power (with Partial Reversibility of Water Power) 257
2.28 Water-to-Steam Switching Subsidies: Steam Adoption and Annual Costs 258

LIST OF TABLES

1.1 Balance of County Geography, Amenity, and Pre-resettlement Characteristics . . 71
1.2 Post-Resettlement Population Distribution, by County Resettlement Density . . 72
1.3 Sectoral Employment, by County Resettlement Density 73
1.4 Sectoral Employment Share, by County Resettlement Density 74
1.5 Manufacturing Activity, by County Resettlement Density 75
1.6 Human Capital, by County Resettlement Density 76
1.7 Household Income, by County Resettlement Density 77
1.8 Parameter Estimates 78
1.9 Population in British Malaya from 1911 to 1957, by Ethnic Group 96
1.10 Migration Outcomes in 1980, by County Resettlement Density 97
1.11 Manufacturing Activity in 1970, by County Resettlement Density and Industry Employment Share of Chinese 98
1.12 Educational Attainment in 1980 for Different Age Cohorts, by County Resettlement Density 99
1.13 Household Asset Ownership, by County Resettlement Density 100
1.14 Household Income, Controlling for Household characteristics, by County Resettlement Density 101
1.15 Housing Elasticity in 1989 102

2.1 Composition of Milling 185
2.2 Mill Activity in 1850, by County Waterpower Potential 186
2.3 Steam Adoption and Mill Growth, by County Waterpower Potential 187
2.4 Entry Rates and Survival Rates, by County Waterpower Potential 188
2.5 Steam Adoption Shares for Entrants and Water Incumbents, by County Waterpower Potential 189
2.6 Non-Mill Manufacturing Establishments, Non-Mill Steam-Use, and Steam Manufacturing, by County Waterpower Potential 190
2.7 Model Fit to Target Moments 191
2.8 Parameter Estimates 192
2.9 Non-Targeted Differences between Lower Waterpower and Baseline Regions . . . 193
2.10 The Impact of Steam on Mill Revenue 1830-1900 (PDV in %) 194
2.11 The Impact of Steam Power on Firm Values in 1830 (in Percentage Points) . . . 195
2.12 Costs and Benefits of Steam Switching Subsidies 196
2.13 Coverage Rates 260
2.14 Survival Rates, by County Waterpower Potential and Initial Power Source . . . 261
2.15 Per Capita Manufacturing Growth and Steam Adoption, by Waterpower Potential 262
2.16 Steam Adoption and *Flour* Mill Growth, by County Waterpower Potential . . . 263
2.17 *Flour* Mill Entry Rates and Survival Rates, by County Waterpower Potential . . 264
2.18 Steam Adoption of Entrants and Water *Flour* Mills, by County Waterpower Potential 265
2.19 Robustness to Alternative Drivers of Steam Use 266
2.20 Robustness to Alternative Drivers of County Growth 267

2.21 Steam use, by Distance to Railroad Station 268
2.22 Confusion Matrix: Hand Links vs. Predicted Links 269
2.23 Robustness to Measurement and Linking Error: Entry and Survival 270
2.24 Robustness to Measurement and Linking Error: Steam Use 271
2.25 Robustness to Sample 272
2.26 Survival Rates 273
2.27 Incumbency, Size, and Steam Use 274
2.28 Steam Use and Characteristics of Owners 275
2.29 Lumber and Flour Mill Activity in 1850, by County Waterpower Potential, by Different River Classifications 276
2.30 Model Fit without Agglomeration 277
2.31 Jacobian: Effect of Parameter on Moments, $\frac{dM}{d\theta_k}$ 278
2.32 Sensitivity: Effect of Moment on Parameters, $\frac{d\theta}{dM_k}$ 279

ACKNOWLEDGMENTS

I am extremely grateful to my advisors, Richard Hornbeck and Oeindrila Dube, for their guidance and support throughout my doctoral studies. Joining Rick's research team was one of the best decisions I made. I learned a great deal from Rick's approach to research in economics and the social sciences, from conceptualizing ideas to executing research and interpreting results. Rick has always been positive and supportive, yet his keen insights always push me to be better. Thank you, Rick, for teaching me how to be a professional economist, a lesson I will carry forward in my career.

Another pivotal event in my Ph.D. was choosing Oeindrila as a reader for my third-year paper and subsequently holding weekly meetings with her. Under Oeindrila's guidance, I learned to critically evaluate and refine the quality and direction of research projects—a skill that ultimately shaped my job market paper. Her advice and encouragement were crucial, especially in the early stages of the project, and her patience and support helped me through challenging times. Thank you, Oeindrila, for being such an amazing advisor and mentor.

I am deeply thankful to my committee members, Michael Dinerstein and Jonathan Dingel. Their meticulous attention to detail has significantly improved my work. The empirical parts of the paper greatly benefited from discussions with Michael, while the structural model and estimation would not have been possible without Jonathan's class and his guidance. Thank you both for your effort and commitment, and for being such a great committee.

The research community at the University of Chicago has been tremendously supportive. I have benefited from discussions with Rodrigo Adão, Christopher Blattman, Fiona Burlig, Joshua Dean, Tomás Domínguez-Iino, Rachel Glennerster, Anders Humlum, Michael Kremer, and James Robinson. I am particularly grateful to Anders Humlum for his early guidance when I first embarked on the steam project. His kindness and dedication have set a high standard that I strive to uphold in my own career. I also want to express special thanks to James Robinson for his encouragement and stimulating discussions on research ideas. Robinson's erudition and openness in research have inspired me to pursue novel and challenging topics.

I want to extend my gratitude to several individuals who have been instrumental in my academic journey over the past decade, including Melissa Dell, Dave Donaldson, Dina Pomeranz, Brendan O'Flaherty, Eric Verhoogen, and Tsong-Min Wu. I am particularly grateful to Melissa Dell for inspiring my interest in development and economic history, and for encouraging me to pursue a Ph.D. Your belief in me when I knew so little has kept me going, especially after the setbacks in the application process. I am also extremely grateful to Dave Donaldson and Dina Pomeranz for their guidance and mentorship during my predoctoralship.

My friends and classmates, particularly Scott Behmer, Hazen Eckert, Furkan Kilic, Chanwool Kim, Nadav Kunievsky, Keven Lee, Ricardo Quineche, Estéfano Rubio, and Walter Zhang, have made this journey both enlightening and enjoyable. I cherish our extended discussions and the laughter we shared. Special thanks to Furkan Kilic for his companionship and productivity spillovers since our first year. This experience would not have been the same without you.

Lastly, I owe a deep gratitude to my parents for their constant presence and support as I pursued my passions. Their unwavering love has been the bedrock of this journey. To my wife, Yu-Ting, thank you for being by my side throughout it all. Your patience and support have meant a lot to me, especially during tough times. I could not have reached this milestone without you.

Chapter 1 is individual work. I thank Sarah Ling Francis for her extensive research assistance, and Hugh Alexander and Alexander Poole for their help with archival research. I also thank the data entry teams at Digital Divide Data. Chapter 2 is joint work with Richard Hornbeck, Anders Humlum, and Martin Rotemberg, supported by the National Science Foundation under Grant Number SES-1757050/1757051. The views expressed are those of the authors and do not necessarily reflect those of the National Science Foundation.

ABSTRACT

This dissertation consists of two chapters. The first chapter studies the agglomeration effects resulting from the forced relocation of 600,000 rural Chinese to compact villages in 1950s British Malaya. I find that counties receiving more resettlement experienced persistently higher population densities and a greater share of Chinese, with increased internal migration over time. These areas saw a shift in employment from agricultural to non-agricultural sectors. Residents in the more resettled counties had higher incomes, particularly among local Chinese, while other ethnic groups benefit mainly through entering the non-agricultural sector. I estimate a spatial general equilibrium model with occupational choice, migration, and heterogeneous agglomeration forces across sectors and groups, exploiting the resettlement program as a population shifter. I find substantial barriers to productivity spillovers between different ethnic groups and a larger external economy for the non-agricultural sector.

In the second chapter, we examine how access to local waterpower delayed the transition from water to steam power in 19th-century US manufacturing, focusing on early users of mechanical power: lumber and flour mills. Digitizing Census of Manufactures manuscripts for 1850-1880, we show that as steam costs declined, manufacturing activity grew faster in counties with less waterpower potential. This growth was driven by steam powered entrants and agglomeration, as water powered incumbents faced barriers to switching technologies primarily from sunk costs. Despite substantial entry and exit, these switching barriers remained influential for aggregate steam adoption throughout the 19th century, as water power required lower fixed costs and therefore was attractive to relatively low productivity entrants. These entrants were later locked-in to water power, even if their productivity grew. Estimating a dynamic model of firm entry and steam adoption, we find that the interaction of switching barriers and high fixed costs is a quantitatively important drag on technology adoption.

CHAPTER 1

COERCIVE GROWTH: FORCED RESETTLEMENT, AGGLOMERATION, AND ECONOMIC DEVELOPMENT IN MALAYSIA

1.1 Introduction

The clustering of people and economic activity can increase productivity through shared inputs, labor market pooling, and knowledge spillovers. These agglomeration forces—typically manifested in urban environments—are closely linked to growth and economic development. Local interactions are fundamental to these benefits. However, in settings with cultural and social tensions—a reality in many countries—these productive interactions could be impeded. Knowing the magnitude of these barriers and their economic impact is essential for designing placed-based and nation-building policies, yet empirical evidence on this is rare.

A major challenge for examining local productivity spillovers across individuals is self selections—people choosing locations based on unobserved productivity or amenities. For instance, in the presence of ethnic tensions, the decision by individuals to settle in areas with a higher concentration of their ethnic group could be due to better location fundamentals, increased productivity from interacting with local population, or a preference for being close to others of the same ethnicity. Identifying productivity spillovers within and across ethnic groups requires separating these various forces.

This paper analyzes an ethnic-based resettlement program during the 1950s Malayan Emergency in British Malaya, which forcibly relocated 600 thousand rural Chinese into roughly 500 compact villages. This historical setting offers a rare opportunity to examine agglomeration forces by sector and the ethnic share in a sector. The ethnic-based forced resettlement shifted both the population size and ethnic composition across regions, while limiting self selection. I leverage the program to study its economic consequences by ethnic group over the next 50 years. I develop a spatial general equilibrium model with occupation and migration choices, while allowing the agglomeration forces to depend on local ethnic share. I use the resettlement as an exogenous population shifter to identify key model parameters related to the agglomeration forces.

Chinese settlements in Malaya substantially expanded during the British colonization, especially after the 19th century, due to the growth of labor-intensive industries of tin mining and rubber plantation. A large portion of the Chinese immigrants were employed in these exporting industries and residing in towns or cities, despite agriculture being the main economic activity throughout the colonial regime. However, there has been a growing community of rural Chinese living near the forest after the Great Depression due to the disruption of the tin and rubber industries. The Japanese occupation during 1942-1945 further pushed many Chinese to the forest fringe. These Chinese "squatters"—who typically occupied lands without titles—became a security issue in the post-war Malayan Emergency (1948-1960) as some of them were a source of food and information for the communists.

With the goal of controlling the squatters and cutting off their support to the communists, the British initiated a large-scale resettlement program (the "Briggs Plan") that forcibly relocated the squatters into compact, guarded villages in more accessible areas. By the end of the Emergency, approximately 600,000 individuals were relocated to around 500-600 New Villages across Malaya, almost 90% of them were ethnic Chinese. The resettled population amounted to one-tenth of the total population and one-third of all the Chinese in Malaya.

I exploit knowledge of the program to isolate exogenous variation of the population distribution caused by the forced resettlement and examine its long-run impacts in the receiving areas. The policy was implemented in two stages: first, sites were selected for accommodating the squatters. This process was primarily driven by proximity to main roads or rivers to ensure better security access, with many similarly suitable locations scattering along the transportation network. Second, the squatters were relocated to these "New Villages" in a manner that minimized dislocation.

Guided by this two-stage procedure, I specify counterfactual resettlement in two steps. First, I randomly permute counterfactual locations of the New Villages along the road and river network, conditioning on the type of land use and proximity to initial squatter settlements. Next, I overlay historical maps on land-use pattern, communist activities, and population settlement to measure the squatter distribution, and use a gravity model to specify the dislocation-minimizing plan, assuming that the resettlement officers incurred a cost proportional to the distance over which the squatters were moved. I repeat this permutation procedure for a thousand times and average the resettlement density in a county, defined as a function of total number resettled into a county per unit area, to obtain the expected resettlement density.

The estimation compares areas with varying resettlement density while controlling for the expected density (Borusyak and Hull, 2023) and other covariates pertaining to initial access to roads and population distribution. The identifying variation comes from the exact locations of the New Villages relative to the average location along the major transportation routes, and deviations from the dislocation-minimizing plan predicted by a gravity model. These residual variations were driven by wartime security and the British' limited information on the squatter distribution, preventing them from fine-tuning the site selection and relocation process based on unobserved economic factors. I provide supporting evidence that they were orthogonal to prewar economic conditions.

Indeed, key geographic characteristics, including topography and agricultural suitability, are balanced across locations of varying residual resettlement density. Important pre-period characteristics are also balanced, including proximity to public goods and amenities, population density, shares of land use by main exporting industries, distance to major industrial facilities, and distance to major towns or cities.

The resettlement substantially altered the population distribution both in the short run and over a longer horizon. Chinese population in counties with one SD higher resettlement increased roughly 40% from 1947 to 1957, a mechanical effect of the policy, whereas there was no change of non-Chinese population in these counties during this period. After the mobility restriction was lifted in 1960, the more resettled counties attracted additional migrants from other counties. By 2000, counties with one SD higher resettlement had a 60% increase in Chinese population and 20% in total population. Overall, areas with higher resettlement experienced an immediate and sustained rise in the share of Chinese.

These endogenous population changes in the more resettled areas after 1960 can result from a preference for locating near people in the same ethnic group, but the increase of non-Chinese over time suggests that it was also driven by changes in the underlying economic structure. Since the Chinese predominantly employed in the secondary or tertiary industries during the colonial period, a higher Chinese population thus serves as an skill-biased labor supply shock that can induce entry of firms that benefit from the local abundance of industrial human capital.

Indeed, I find that while the more resettled counties experienced a larger overall employment size, growth of employment in the secondary and tertiary sectors—such as manufacturing, trade, and services—outweighs that in the primary sector. This is driven both by a higher presence of Chinese who were more likely to employ outside the agricultural sector and by the lower share of non-Chinese working in agriculture in these areas.

The lower share of agricultural employment in the resettled areas is consistent with a higher returns to scale for non-agricultural sectors. Considering that Chinese had a comparative advantage in non-agricultural works prior to the resettlement, local abundance of Chinese workers tend to lower the relative wages in non-agricultural industries in the absence of external economies of scale. That the resettled areas saw a higher employment share in these industries suggests that local spillovers between workers make them more productive, which further attracts a larger employment.

Additionally, I find suggestive evidence of knowledge spillovers from Chinese to non-Chinese entrepreneurs within the manufacturing sector. Counties with higher resettlement density saw a surge in non-Chinese owned establishments entering industries that initially employed a higher share of Chinese prior to resettlement. This is consistent with transfers of industry know-how between Chinese and non-Chinese entrepreneurs.

Besides industry spillovers, I also find evidence of within-ethnicity productivity spillovers, with a larger Chinese income premium in the more resettled counties. Indicators of household asset ownership from the Census data in 1980 show that households in the more resettled areas were more likely to own durable assets—including vehicle, television, refrigerator—and have higher income, consistent with an agglomeration economy. However, the benefit of the concentration of Chinese mainly accrues to Chinese households. Non-Chinese households overall did not earn higher income in the more resettled areas and only those employed in the non-agricultural sector saw a marginal income gain.

Measures of human capital were also higher in the more resettled areas. I find that in 1980, individuals in these areas had higher educational attainment in years of schooling, completion rate of primary and secondary education, and the ability to speak English. These effects are stronger for the Chinese individuals, who showed a much higher completion rate for secondary education. The education effect shows up only for younger cohorts below age 50 by 1980, a group of people who had not finished their education by the time the resettlement occurred. I also find better health outcomes for people living in the more resettled areas, as indicated by a larger birth weight of the first child.

The larger effect of resettlement on human capital and income for Chinese suggests that an influx of Chinese incurs a different level of productivity externality to local Chinese population as compared to other ethnic groups, and that the within-group spillovers are stronger than cross-group spillovers. In fact, cross-group spillovers can even be negative in sectors with a small external economy. This aligns with anecdotal accounts of tensions between Chinese and

Malays inhibiting interactions and knowledge sharing. Additional barriers, such as language differences and an ethnically segregated education system, could further isolate the social networks of different ethnic groups and contribute to reduced sharing of ideas and knowledge across groups.

The evidence so far provides a qualitative sense of the sign of the spillover elasticities. However, the effects of resettlement on local wages and population changes includes both the direct effect of spillover and an indirect effect of occupational choice and migration due to the general equilibrium channels. For example, an influx of Chinese with industrial human capital can increase the wage in manufacturing relative to agriculture because of positive productivity spillovers, thereby attracting more people into manufacturing that further amplifies the effect. On the other hand, agglomeration raises the average wage of that region and induces in-migration from other places, which generates even more externality. Therefore, simply regressing wages on sectoral population can over-estimate the elasticity.

I develop a spatial general equilibrium model with occupation choice and migration, and one that allows the agglomeration forces to vary by sector and with the ethnic share in a sector. The model takes as given the initial population distribution right after the resettlement, as measured by the 1957 census, and generates the equilibrium outcomes in 1980 census.

I identify key model parameters related to the agglomeration forces using the plausibly exogenous variation of the resettlement as instruments for equilibrium population distribution. Consistent with the empirical evidence, the model rationalizes the higher manufacturing employment share in the more resettled areas with a larger external economy for manufacturing than agriculture. Moreover, I estimate that the elasticity of productivity with respect to the ethnic share in an industry to be 0.1, implying that within-group spillovers are stronger than cross-group spillovers.

A rich literature has documented the advantages of density and geographic concentration of economic activity (Glaeser et al., 1992; Duranton and Puga, 2004; Rosenthal and Strange, 2004; Ahlfeldt et al., 2015; Davis and Dingel, 2019; Heblich et al., 2020; Smith and Kulka, 2023). The positive spillovers have been studied across firms (Greenstone et al., 2010) as well as individuals (Moretti, 2004b; Ciccone and Peri, 2006). This paper contributes to the relatively understudied area of spillovers across social groups that exhibit frictions in interactions, highlighting the uneven benefits from agglomeration that could exacerbate inequality across groups. One exception is Ananat et al. (2013, 2018), who argue the lower levels of cross-race social interactions can be an important driver for the increasing black-white wage gap in city size. My paper contributes by studying a natural experiment that exogenously changes the ethnic composition of a place, and using a model to quantify the economic loss due to frictions of cross-group interactions. It also complements works that document the distributional consequences from agglomeration (Ahlfeldt and Pietrostefani, 2019; Fajgelbaum and Gaubert, 2020).

A burgeoning body of work has attributed stronger productivity spillovers among workers in specific occupations or industries as an important driver of innovative activities (**?**). This paper emphasizes the differential spillovers among people of different demographic or social backgrounds in an early development context, where industrialization and the increasing returns of manufacturing plays a crucial role. In addition, allowing ethnic composition to be part of the amenity of a location, this paper also relates to the literature examining how local amenities, including residents' characteristics, shape residential sorting patterns (Bayer et al., 2004, 2007; Diamond, 2016).

This paper contributes to works on the economic consequences of forced migration and villagization (Hilhorst and Leeuwen, 2000; Whittaker, 2012; Bazzi et al., 2016; Abel, 2019; Becker et al., 2020; Carlitz et al., 2022; Peters, 2022; Sarvimäki et al., 2022). The nature of coercive population movement generally leads to welfare loss of the relocated individuals. The literature has mainly focused on the economic or political consequences of these historical events and the longer-run impacts on the descendants. The closest work to mine is Peters

(2022), who studies the refugee settlement in postwar Germany and, consistent with the increasing returns to scale for manufacturing, also finds that the increased population persisted and spurred industrialization later. My work highlights that the economic consequences on the receiving areas can vary by local population composition.

Forced villagization has been commonly used in many countries as a development policy (cf. (Hilhorst and Leeuwen, 2000)) or nation/state-building policy (cf. Whittaker (2012); Carlitz et al. (2022). The literature has typically found that villagization has either neutral or negative economic impacts on the villagers, presumably because inter-group conflicts intensified as people from different social or historical backgrounds were forced to co-locate (Dippel, 2014). Consistent with my theoretical framework, when the within-ethnicity productivity elasticity is large enough, the cross-group spillover can even be negative. Another reason that villagization often fails in the past might be also due to the non-market, economic planning aspect in those settings, where villagers were forced to grow certain crops required by the government. In contrast, people in the New Villages after the end of the Emergency were in a market-driven economy and can choose any productive activity or move out of the village at their will.

There is a large literature on the economic implications of ethnic diversity and conflicts (Abadie and Gardeazabal, 2003; Alesina and La Ferrara, 2005; Montalvo and Reynal-Querol, 2005; Besley and Mueller, 2012; Rohner and Thoenig, 2021; Eberle et al., 2020). Much of this literature has been largely descriptive or focused on the economic aftermath of violent conflicts, overlooking the costs stemming from latent tensions between groups. This paper underscores that even if the underlying social tensions are not severe enough to cause violent conflicts, they may still lead to economic losses through lower productivity spillovers. Relatedly, Ashraf and Galor (2013) show a U-shaped relationship between ethnic diversity and economic performance across countries, consistent with the idea that diversity can be beneficial via positive spillovers from knowledge exchange, yet excessive diversity may prove detrimental due to amplified inter-group frictions and reduced productivity spillovers. My paper contributes to this strand of work by studying a within-country context and using a general equilibrium model to quantify the economic loss due to inter-group frictions.

Finally, this paper is related to the literature on Asia's rapid economic development, particularly the role of state intervention in the growth experience (Haggard, 1990; Amsden, 1992; Mundial, 1993; Wade, 2004; Dell et al., 2018; Lane, 2022). The higher returns to scale for manufacturing than agriculture I find suggests that industrial policies that promote structural change from agriculture to manufacturing, common in many East/Southeast Asian countries, may spur an economic takeoff with self-reinforcing dynamics stemming from productivity spillovers. In addition, I estimate from the resettlement of Chinese a higher within-group productivity spillover, which, if allowed to vary by group, may differ from what one would obtain from variations of other groups. The positive ethnic-share elasticity indicates that the productivity spillover for the Chinese is more substantial in sectors or regions dominated by Chinese. This might be attributed in part to a higher local collective action among Chinese community due to their historical exposure to centralized state in the theme of Dell et al. (2018).

The rest of the paper is organized as follows. Section 1.2 discusses the historical context, and Section 1.3 describes the data. Section 1.5 discusses the empirical strategy and examines the reduced-form effects of resettlement. Section 2.5 lays out the model, and Section 1.7 discusses the structural estimation. Section 1.8 concludes.

1.2 Historical Context

This section first provides a brief introduction of the historical origins of Chinese in Malaysia. Next, I will discuss the resettlement of Chinese squatters during the Malayan Emergency and factors that determine the number of people relocated to which locations.

1.2.1 Chinese in British Malaya

Chinese settlements have been in Malaya since the 16th century and greatly expanded in the 19th century during the British colonization, spurred by new discoveries of tin deposits in Malaya and the deteriorating economic and political conditions in China.[1] Most Chinese immigrants were brought in by labor brokers to work on labor-intensive tin mining and rubber plantation, the two main exporting industries in the colonial period. The influx of Chinese labor into Malaya continued to rise until the 1930s when the Malayan government implemented immigration quotas for male labor and banned direct emigration from China, after which the population composition in Malaya remained stable, with roughly 50% Malays, 35% Chinese, and 15% Indians and others (Appendix Table 1.9).[2]

The Chinese and Malays have historically specialized in different industries, with the former primarily working in mining, rubber plantation, and the industrial/commercial sectors, and the latter mostly employed in agriculture (Appendix Figure 1.8). Specifically, tin mining was dominated by the Chinese up to the early 20th century until European mining companies entered with more advanced technology and out-competed Chinese miners, many of which then transitioned to rubber growing.[3]. Even though British Malaya was still a predominately agricultural economy, the capital accumulated from tin mining and rubber plantations allowed some Chinese to venture into commercial banking, manufacturing, and other secondary/tertiary industries.[4] In contrast, the Malays mainly specialized in coconut cultivation and padi rice planting.[5]

1. Lim (1964, p. 44).

2. Humphrey (1971, p. 34).

3. Humphrey (1971, p. 36), Lee and Tan (2000, pp. 96-98)

4. The secondary industries mainly involved the processing of rubber and tin for export and consumer goods for the domestic market, such as food processing and pineapple canning (Lee and Tan, 2000, p. 19-20).

5. Ginsburg et al. (1958, p. 244), Lee and Tan (2000, p. 100).

1.2.2 The Squatter Problem in the Malayan Emergency

Chinese communities were primarily located in urban areas along the west coast or mining towns within tin-rich states due to their commercial and industrial roles until the 1930s, after which there had been a growing community in the rural areas near the forest fringe.[6] By the end of World War II, about one third of the Chinese in Malaysia lived near the forest.[7]. As these Chinese agriculturalists often occupied lands without formal title—typically forest reserve lands, state lands, or Malay Reservation lands—the government referred to them as "squatters."[8]

The large squatter population resulted from several factors, most notably shifts in the mining and rubber industries and the Japanese occupation.[9] Labor demand for Chinese workers in tin mining shrank due to increased mechanization, exhaustion of existing tin deposits, displacement of smaller Chinese miners by European mining companies, and a drop in prices during the Great Depression.[10] The rubber industry also experienced significant price volatility on the international market, prompting Chinese laborers to revert to food cultivation. In addition, the Japanese occupation from 1942 to 1945 contributed significantly to the surge of the squatter population. The British employed a scorched earth policy to hinder their use by the Japanese during their retreat at the end of 1941, causing widespread destruction to the industrial sectors and unemployment among Chinese populations. The ongoing war between Japan and China, along with the Japanese' imposition of forced labor, exacerbated the exodus of urban Chinese to rural areas.[11]

6. Robinson (1956, p. 76), Sandhu (1964, p. 2), Lee and Tan (2000, p. 96)

7. It was initially estimated to be around 300,000 and was updated in 1952 to around 500,000, which was further revised post-war to more than 600,000. See The National Archives of the UK (hereafter "TNA"), CO 1022/29, pp. 63, 71-72.

8. TNA: CO 1022/29, p. 71, Humphrey (1971, p. 39).

9. TNA: CO 717/178.

10. Humphrey (1971, p. 43), Loh (1988, pp. 23, 27-29).

11. Humphrey (1971, p. 39, 47), Loh (1988, pp. 57-60).

The squatters participated in a range of occupations, including fishing, cultivation of food and cash crops, and livestock rearing. Some undertook commercial farming and became the primary source of fresh vegetables for urban markets.[12]

After the British returned in 1945, they realized the security concern with the squatter community, especially after the conflicts erupted with communists during the Malayan Emergency (1948-1960). Many Chinese squatters either sympathized with or directly participated in the activities of the communists because of the general animosity between ethnic Chinese and the Japanese, and that the Malayan Communist Party (MCP) was involved in the previous battle against the Japanese during the occupation.[13] Some of the Chinese squatters were part of the non-military association of the communists that was in charge of collecting information and supplies.[14] The lack of administrative control in the isolated locations where the squatters resided made it difficult for the British to prevent the squatters from, voluntarily or forcibly, providing support to the Communists, who were conducting guerrilla warfare in the jungle.[15]

A committee was appointed by the end of 1948 to examine the circumstances and it concluded that squatters should ideally be settled in areas they were already occupying, and where this was not feasible, they should be resettled in suitable alternative locations.[16]

12. Sandhu (1964, p. 4).

13. The MCP had strong ties to the Malayan Period's Anti-Japanese Army (MPAJA), a group that previously fought against the Japanese. After the Japanese invasion, the MCP retreated into the jungle and built up its fighting force, MPAJA, which was disbanded after the British returned (Sandhu, 1964, p. 4). Many of the members of MPAJA later became the fighters of the communist guerrilla in the Malayan Emergency.

14. The Communist movement comprised two organizations: the fighting forces—Malayan Races Liberation Army (MRLA)—and the non-military association—Min Yuen ("Masses Movement")—that primarily provided information and supplies. See Sandhu (1964, p. 6), Humphrey (1971, p. 63), Loh (1988, p. 121).

15. Humphrey (1971, p. 49), Loh (1988, pp. 106-107).

16. TNA: CO 717/178.

1.2.3 The Briggs Plan: Emergency Resettlement

The resettlement or squatters was delegated to the state governments but due to the high cost involved and the limited finance of the states, it was only after General Briggs was appointed the new Director of Operations in 1950 that the program started to scale up.[17] General Briggs viewed the resettlement—later known as the "Briggs Plan"—as central to the counterinsurgency strategy and ensured the necessary resources were allocated from the federal government to the states.

The main objectives of the Briggs Plan were twofold: to create secure, populated areas to facilitate intelligence gathering; and to cut the Communists' logistical support by isolating them and forcing them to eventually attack the British on unfavorable terms.[18] The resettlement was implemented with a focus on speed: starting from the southern states in June 1950, it moved fast toward the north and was largely completed by the end of 1952.[19]

The procedure in each state varied to some degree, but typically followed the order of site selection, land clearing, marking house plots and roads, and issuing removal notice to the squatters before the actual relocation.[20] Site selection was mainly based on security concerns which I will discuss in more detail later. After a site was selected, lands had to be cleared for setting up a "New Village", often involved cutting down rubber trees as many sites were on state-owned rubber estates. To prevent escape, the length of notice was usually less than 14 days, and shorter when the perceived risk of escape or resistance was higher.[21] After the relocation, the original settlement was burned down.

17. In early 1950, just prior to Briggs' arrival, only about 7,000 squatters had been resettled (Loh, 1988, pp. 123-124).

18. TNA: AIR 20/7777.

19. Sandhu (1964, p. 11), Humphrey (1971, p. 106).

20. TNA: CO 1022/29, p.125.

21. If the target squatters were suspected to have links to the Communists and/or might resist resettlement, the British military would have appeared at dawn without notice to forcibly resettle them (Humphrey, 1971, p. 102)

The relocation was fairly short-range, normally involved a distance of fewer than 15 miles.[22] When the distance relocated was larger than 3 miles, the squatters typically had to abandon their homes, crops, and potentially their previous occupation. For other cases where the resettlement were close enough to the original settlement, the squatters were concentrated around a central site that enabled them to maintain their holdings and jobs but required them to give up their homes unless they were within the new settlement's perimeter.[23]

In addition to the general mobility regulation in the so-called "Black Areas"—high-risk areas with communists activity that were subject to curfews, food restrictions, and travel bans—the movement into and out of the resettlement areas was further controlled.[24] The New Villages were encircled by double barbed wire with checks and searches by the police at the entrance to prevent food smuggling for the communists.[25]

By the end of the Emergency, approximately 573,000 individuals were resettled to about 500 New Villages. The resettled population was 86 percent Chinese, 9 percent Malays, and the remaining 5 percent consisting of Indians and others.

Determinants of Resettlement Density

Understanding why the New Villages were sited in certain areas and, given these sites, what determines the number resettled in specific locations, is critical for the identification of the impacts of resettlement. As briefly discussed above, the policy was implemented in two stages: the sites were first selected and then the squatters were relocated to these sites in a way that minimized their dislocation. Therefore, the site selection criteria and the initial distribution of Chinese squatters play a crucial role in determining the variation of population resettlement.

22. Sandhu (1964, p. 14).

23. Humphrey (1971, p. 27).

24. Corry (1954, pp. 19-20), Nyce (1973, p. 180).

25. Humphrey (1971, p. 358), Humphrey (1971, pp. 118).

The key principles for the placement of resettlement areas outlined by General Briggs are: (i) proximity to main transportation routes and distance from elevated observation points for security reasons; (ii) availability and cost-effectiveness of land; (iii) sustainability and potential for economic development, factors such as proper terrain for drainage, sufficient agricultural suitability, and proximity to water sources; (iv) proximity to squatter populations to minimize dislocation.[26] I will discuss these factors in turn.

First, the main objective of the program was security. This led to Chinese squatters being relocated near major transportation network for ease of police access and away from elevated observation points for defense purposes.[27] In instances where a village had to be situated away from a main road, an all-weather connecting road was constructed. In more remote areas lacking major roads, sites were chosen to be in proximity to navigable waterways[28]. Indeed, most of the New Villages were situated along the roads or in proximity to a navigable river (Figure 1.1). For defensibility, resettlement areas should also be more elevated though the data does not show a meaningful correlation.

Second, there needed to be available lands for squatter settlement. Considering the land acquisition costs, there was a preference for State lands or lands of little town value if alienated.[29] Indeed, as I will show from the digitized land use map prior to the resettlement, many New Villages were located on rubber estates, a large number of which were state-owned. I thus control for land use patterns in the analysis and I will revisit this when discussing the empirical strategy.

Third, factors for future sustainability was considered, including drainage, agricultural suitability, and proximity to water sources. The village should ideally be sited on rolling terrain to promote drainage and prevent flooding. As a large portion of the squatters

26. Humphrey (1971).

27. TNA: CO 717/177, Sandhu (1964), dhu Renick (1965), Humphrey (1971).

28. TNA: CO 717/201, Humphrey (1971)

29. Humphrey (1971).

were initially vegetable farmers, it was preferable to have high-quality soil and abundant agricultural land in the vicinity. There should also be sufficient water supply either from adjacent towns or from wells within the village.[30]

While information about these economic aspects were meant to be collected through field surveys prior to resettlement, in practice this was often not feasible due to the security risks, staff shortages, and urgent situation.[31] Numerous cases have shown that economic considerations were overlooked in favor of security and speed.[32] I will show that geography related to the productivity of a place, such as ruggedness, agricultural suitability, and proximity to rivers are balanced once conditional on the transportation network, suggesting that the colonial government was not able to fine-tune the locations based on economic considerations.

Finally, the squatters were resettled in a way that minimized their dislocation as much as possible conditional on meeting the security needs because the closer the distance relocated, the lower the moving costs and the economic disturbance.[33] In practice, the squatters were relocated to the closest site until the capacity of that site was filled, but due to the lack of field survey, the British was poorly informed about the squatter distributions during site selection.[34] As a result, if an area had more squatters requiring resettlement than previously expected, additional squatters would need to be relocated to other sites further away.

As the squatters self-selected to where they were prior to the resettlement and that initial distribution played an important role in the number resettled into a region, it is crucial to

30. Humphrey (1971).

31. Humphrey (1971).

32. A colonial liaison officer noted that "...a major mistake in site selection was that some sites became flooded in heavy rains."—Notes on Planning and Housing Aspects of Resettlement and the Development of New Villages (Arkib Negara Malaysia, hereafter, "ANM", 1953).

33. Sandhu (1964), Tsou (2007).

34. A newspaper clip in 1952 wrote: "The Government had only the haziest idea of the numbers [of the squatters]: it was first believed that there were 318,500, but the total was nearer 500,000." (TNA: CO 1022/29, p. 63). In fact, by the end of the Emergency, around 600,000 were resettled.

condition on that in the empirical analysis. I will revisit this in Section 1.4.

1.3 Data

1.3.1 Resettlement Density

I measure counties' resettlement density—the number of people relocated to a county per county area, with the inverse hyperbolic sine transformation—as a shifter for population distribution by the end of the Emergency. The total number of relocated individuals per county is obtained by adding the populations of each New Village within the county.

I measure the resettled population by village and the location of the village from the "Corry report" (Corry, 1954), an official 1954 colonial report to the High Commissioner of the Federation of Malaya. It offers a detailed list of 430 New Villages shortly after the primary resettlement phase, with the village names and the estimated populations of the villages.[35] Village locations are drawn from Baillargeon (2021), who georeferenced the locations of the New Villages documented in the Corry report.

I cross-check resettlement figures with a 1959 Malayan Christian Council survey, which shows consistent village populations for those documented in both sources. It also shows additional villages compared to the Corry report, suggesting that there were around 150 villages built between 1954 and 1959.

I use the Corry report as my baseline measure of resettled populations as it was an official report that meant to provide a comprehensive coverage of the early resettlement outcomes during the Briggs Plan, where speed and security were the top priority. The Corry report may be less biased compared to the Christian survey, which could have selectively targeted places based on evangelistic motivations. I assess robustness of the results using the 1959 survey.

35. The report aimed to assess agricultural land sufficiency in the villages, investigate land titles to prevent unlawful reoccupation, and estimate the number of rural Chinese needing resettlement for security reasons.

1.3.2 Population Census, 1931-2000

I collected and digitized the tabulated Census of Population of Malaysia at the county level for the years 1931, 1947, 1957, 1970, 1980, and 2000.[36] County (or "Mukim") is the smallest administrative unit consistently tabulated over time, where 1931 was the first year that documents population by county.

To accommodate changes in county boundaries over time, I created time-consistent borders based on the boundaries in 1947, right before the Emergency, grouping together counties with overlapping geographies across years. My baseline county boundaries consist of a sample of 777 counties. For panel regressions starting from 1931, I generate another set of 614 grouped counties based on the 1931 borders.

The tabulated data provides population count by ethnic group—Malays, Chinese, Indians, and Others—in each county from 1931 to 2000. My main analysis focuses on outcomes in 1980, where the census report tabulates additional information such as employment by industry and occupation for each county. The 1980 census also tabulates, at the district level, population by place of last previous residence for each current residence.[37] I use these bilateral migration flows to estimate migration costs.

I use the 2% microdata from the 1980 census to measure labor market outcomes, asset ownership, migration status, and educational attainment at the individual or household level. The microdata allows me to measure key economic outcomes by county-ethnicity, which is typically not available in the tabulation data. For example, I draw on data regarding the ownership of various household assets to measure the consumption of durable goods, which can serve as a proxy for household income.[38]

36. Unified British Malaya censuses covering the whole British Malaya—including the Straits Settlements, Federated Malay States, and Unfederated Malay States— began in 1921.

37. Each administrative district contains several counties.

38. If one assumes that durable assets enter the household utility function in a Cobb-Douglas form, these assets would constitute a fixed share of household income. This implies that consumption on these assets would scale linearly with household income.

1.3.3 Other Data Sources

I measure manufacturing activity from the Directory of Manufacturing in 1970, which lists all registered manufacturing firms in Peninsular Malaysia with a total of around 12,000 establishments.[39] It includes the establishments' name, address, main products and industry, and employment size. I digitized and georeferenced the establishments to the county using the provided addresses, and classified the ownership of the establishments into Chinese-owned and others, exploiting the distinct patterns between Chinese names and other ethnic groups.[40]

Since the census lacks direct productivity measures like income or wages, I turn to the Second Malaysian Family Life Survey (MFLS-2) from 1988-1989 for household income information.[41] Using this survey, I estimate a linear model to predict household earnings and apply the model to the census sample for a more comprehensive coverage of household income.[42] The model includes district fixed effects, household size, and indicators of various assets ownership—including automobile, motorcycle, bicycle, phone, refrigerator, and TV—as well as pairwise interactions of these dummies, with an R squared of 0.34.

In addition to the census data, I also use satellite-based, built-up volumes from the Global Human Settlement Layer (GHSL) project to measure the persistence effect of resettlement on population distribution. The built-up volumes are calculated using the surface and height data at a 100-meter resolution from the Sentinel-2 and Landsat satellite images.

To measure key geographical covariates, I collected and digitized various historical maps. These include a road and railway map from 1942; a land utilization map from 1943; topographical maps from 1945 showing the locations of buildings by type (e.g., mosques,

39. All establishments are required to register under the Registration of Business Ordinance 1957.

40. Most Malaysian establishments at the time were sole proprietorships or partnerships and were often named after the owner(s).

41. Conducted by RAND and Malaysia's National Population and Family Development Board, the survey offers comprehensive demographic and socioeconomic data on nearly 3,000 households.

42. Although the sample of MFLS-2 is meant to be representative of Peninsular Malaysia, the geographic coverage is limited with less than 200 counties covered.

Chinese temples, post offices, and railway stations); and a map from 1945 showing prewar industrial facilities.[43] Additional covariates were sourced from publicly accessible data: elevation data was obtained from the Shuttle Radar Topography Mission (SRTM) in 2000; suitability for padi rice, coconut, and palm oil cultivation comes from the Food and Agriculture Organization's Global Agro-Ecological Zones (GAEZ) database; and terrain ruggedness is based on Nunn and Puga (2012).

1.4 Empirical Strategy

In this section, I first discuss how I isolate plausibly exogenous variations from the Briggs Plan to construct a shifter for population distribution and the identification assumptions. Next, I discuss the estimation procedure. Lastly, I examine balance of geography and pre-period characteristics as supporting evidence of the identifying assumptions.

1.4.1 Empirical Specification

The empirical analysis aims to examine whether places that became more densely populated and ethnically homogeneous due to the Briggs Plan, have experienced a different development trajectory. The thought experiment compares two initially similar locations, one of which received greater resettlement density than the other, and examines whether they have different economic outcomes over time and whether such differences vary by ethnic group. Reduced form effects at the county level thus encompasses responses from both local residents and migrants, including those who initially resettled and those who subsequently sorted endogenously.

43. The road/railway map and industrial facilities map originate from the U.S. Office of Strategic Services (U.S. Office of Strategic Services, 1942, 1944). The topographical maps are from the HIND 1076 map series Survey of India Offices (P.Z.O.) (1944), and the land use maps are from the GSGS 4474 series War Office (1943).

Consider a reduced-form model of county c as follows.

$$Y_c = \beta ResettleDensity_c + \gamma X_c + \varepsilon_c, \tag{1.1}$$

where Y_c includes county population density, population share of Chinese, the (log) number of manufacturing establishments, employment shares by sector, and average household income. These outcomes are examined separately for Chinese and non-Chinese populations to inform the relative strength of productivity spillovers among people of the same ethnicity versus those across ethnic lines.

The baseline controls X_c include state fixed effects, (log) county area, an indicator for whether a county underwent any resettlement, distance to the coast, pre-period road density, distance to roads, distance to rail stations, 1947 Chinese population share, and land use shares for rubber plantation and mining.[44] I control for the extensive margin of resettlement to compare only within counties that received some resettlement. This is to account for potential unobservables that might differ between areas with and without any resettlement.[45]

I define county resettlement density from the Briggs Plan, which serves as a population shifter post resettlement, as the inverse hyperbolic sine of the total number of resettled population in a county per unit area; i.e.,

$$ResettleDensity_c \equiv \operatorname{asinh}\left(\frac{\sum_{i \in c} g_{1i} \times g_{2i}}{area_c}\right), \tag{1.2}$$

where i denotes spatial units (or "sites") more granular than counties; g_{1i} is an indicator of whether site i is chosen as a resettlement area; g_{2i} denotes the number resettled to site i; and

44. It is assumed without loss of generality that ε_c is orthogonal to X_c as one can always project the term $\gamma X_c + \varepsilon_c$ onto the space of X_c to decompose it into two orthogonal terms.

45. In one extreme, counties along the coast were mostly alienated and too populated to be feasible for resettlement. Areas deep in the jungle were also not feasible for resettlement due to their remoteness and lack of state capacity. The non-resettled counties were included only to improve efficiency and I show in the Appendix that results are robust to using only the resettled counties in the analysis.

$area_c$ denotes the area of county c.[46]

Based on the historical record summarized in Section 1.2.3, I make two assumptions about $g_1 \equiv \{g_{1i}\}_{i=1}^{I(s)}$ and $g_2 \equiv \{g_{2i}\}_{i=1}^{I(s)}$. First, I assume that the exact locations of the New Villages are orthogonal to the economic fundamentals or amenities of a place, conditional on the vector $w_1 \equiv \{w_{1i}\}_{i=1}^{I(s)}$ which includes distance to the transportation network, land-use type, and the number of nearby Chinese squatters. This assumption is motivated by the British objective to resettle squatters near the transportation network and that, given the urgency of the situation and lack of field survey, the British were unable to fine-tune the location selection based on other unobserved economic factors.

Second, I assume that the number resettled to a new village is exogenous, given the chosen locations and the initial distribution of squatters ($w_2 \equiv \{w_{2i}\}_{i=1}^{I(s)}$). This is motivated by the goal of minimizing dislocation when relocating squatters and the fact that the British were poorly informed about the spatial distribution of squatters when the sites were selected. For instance, if the British realized that an area had more squatters with security concerns than previously expected, the additional squatters would have to be relocated to other nearby counties as there weren't sufficiently capacity to accommodate them locally. This generates idiosyncratic variation of the realized resettlement density.

These assumptions are formalized as follows.

Assumption 1. (Resettlement Exogeneity)

(i) Site selection: $g_1 \perp\!\!\!\perp \varepsilon \mid w_1$; that is, conditional on the transportation network, land-use patterns, and the number of nearby Chinese squatters, the location of a New Village is considered exogenous.[47]

46. Without loss of generality, let sites be small enough that each site contains at most one New Village. The random variables g_{1i} and g_{2i} are not independent since the number of people resettled to site i would be zero if it is not selected as a resettlement area.

47. A weaker assumption of mean independence between g_1 and ε, conditional on w, suffices for identification. Here, ε represents error terms across all counties, denoted as $\varepsilon \equiv \{\varepsilon_c\}_{c=1}^{C(s)}$, where $C(s)$ is the number of counties in state s.

(ii) Number resettled: $g_2 \perp\!\!\!\perp \varepsilon \mid (g_1, w_2)$; that is, conditional on the locations of the New Villages and the initial distribution of Chinese squatters, the number of individuals resettled to a village is exogenous.

Assumption 1(i) would be violated if, for example, squatters bribed resettlement officers to influence their final destination. This is less of a concern in this context, as the resettlement officers were largely British or white Commonwealth citizens who generally lacked trust towards the squatters due to the difficulty of distinguishing them from the communists. In addition, historical work documented that corruption among the British civil servants in Malaya were rare (Humphrey, 1971). Indeed, I will show that there is no meaningful correlation between resettlement density and desirable geography of a place, nor with pre-period economic conditions.

Assumption 1(ii) might be problematic if the unobserved communist risk that affected resettlement density also itself impacted long-term economic outcomes. However, this is unlikely given the sporadic nature of the communist attacks, which were mainly conducted as guerrilla warfare in the jungle, whereas most of the economic activities happened in towns or cities controlled by the British. Anecdotal evidence shows that despite the risk, people were able to travel and work daily, suggesting that communist activities during the Emergency were unlikely to influence post-Emergency economic outcomes.[48]

Under Assumption 1, the remaining challenge to identifying β comes from the fact that the number resettled to a county depends on the initial layout of transportation network and Chinese squatter settlement not only within the county, but also other counties in the state. For example, a county would likely receive more resettled squatters if nearby counties lacked sufficient roads or suitable sites to accommodate the squatters. This means that two counties with identical road density could experience vastly different resettlement density depending

48. A resettlement officer wrote in a memorandum: "...it is not then the situation that the bandits have paralyzed the country...people do travel freely by road and rail between Singapore and Kuala Lumpur, people do go to work every day and come home again." (TNA: CO 1022/29).

on the road density of their neighboring counties, a factor that can directly influence county outcomes through channels such as market access (Donaldson and Hornbeck, 2016). Similarly, counties with a large Chinese population in nearby counties were also likely to receive more resettlement, even when conditioning on the Chinese population within the county itself. These exposures, if not captured by X_c, would show up in the error term and lead to omitted variable bias.

To address this issue, I directly model and control for the omitted variable using the knowledge of the program. Specifically, the potential omitted variable in (1.1), under Assumption 1, is the conditional expectation of resettlement density given w_1 and w_2, denoted $\mathbb{E}[f_c(g_1, g_2)|w_1, w_2]$. Therefore, β can be identified once this expected resettlement density is controlled for (Borusyak and Hull, 2023).[49]

To measure the expected resettlement density, I make two further assumptions regarding the distribution of g_1 and g_2, drawing upon the design of the Briggs Plan. Let the probability distribution of g_1 and g_2 be $G_1(\cdot)$ and $G_2(\cdot)$, respectively, the assumptions are formalized as follows.

Assumption 2. (Resettlement Design)

(i) Equally suitable sites: $G_1(g_1|w_1)$ is uniform; i.e., conditional on the transportation network, land use patterns, and the number of nearby Chinese squatters, areas are equally suitable for establishing a New Village.

(ii) Minimized dislocation: $\mathbb{E}[f_c(g_1, g_2; w_1, w_2)|g_1, w_2] = f_c(g_1, \bar{g}_2(g_1, w); w)$; that is, conditional on the locations of the New Villages and the initial distribution of Chinese squatters, the expected resettlement density aligns with a dislocation-minimizing number $\bar{g}_2$, as predicted by a gravity model.

49. As shown in Borusyak and Hull (2023), controlling for or re-centering $ResettleDensity_c$ by $\mathbb{E}[f_c(g_1, g_2)|w_1, w_2]$ can both purge the omitted bias. One application of this method is Dell and Olken (2020), who identify the impact of proximity to sugar plants by contrasting actual proximity to sugar plants with that to counterfactual sugar plants to eliminate omitted variable bias.

Assumption 2(i) imposes that sites with identical covariates w_1 had an equal probability of being selected as a resettlement area. Assumption 2(ii) models the dislocation-minimizing resettlement $\bar{g}_2$ using a gravity migration equation. The micro-foundation is that while resettlement officers might have idiosyncratic preferences over where to relocate the squatters due to security concerns during the Emergency, they faced a resettlement cost that increases with the distance. Otherwise, they considered potential sites with the same covariates as equally attractive. The "dislocation-minimizing" number of people resettled to village i is given by

$$\bar{g}_2(g_1, w) \equiv \sum_{j=1}^{J} n_j \times \frac{d_{ji}^{-\theta}}{\sum_{s=1}^{I} d_{js}^{-\theta}}, \tag{1.3}$$

where n_j denotes the Chinese squatter population at the original settlement $j = 1, \ldots, J$; d_{ji} denotes the distance between origin j and destination $i = 1, \ldots, I$; and θ is the resettlement cost elasticity that captures how costly it was to relocate a person to a greater distance.[50]

To measure the initial population of Chinese squatters, $\{n_j\}_{j=1}^{J}$, I digitize and overlay three historical maps: (i) land-use maps from 1944; (ii) population census map from 1947 (Appendix Figure 1.9); and (iii) a map during the early years of Malayan Emergency delineating the "Black Areas", which were regions with communist activities (Appendix Figure 1.10.[51] I classify a cluster of Chinese population as "squatters" susceptible to resettlement if it was located within the Black Areas and within a 5 kilometer radius of a forest (Appendix Figure 1.11).

50. The resettlement cost elasticity θ is calibrated to be 0.65 using the observed number of people resettled. The calibration minimizes the sum of squared residuals when regressing the actual number resettled on the ideal plan; i.e., $\hat{\theta} \equiv \arg\min_\theta \sum_i (g_{2i} - \bar{g}_{2i}(\theta))^2$.

51. From 1953, the British refer to high-risk environments with communist activities as the "Black Areas" (as opposed to "White Areas"). These areas were subject to various Emergency regulations, including curfews, food restrictions and travel bans (Baillargeon, 2021).

I estimate the conditional expected resettlement density using a permutation procedure.[52] Each permutation $s = 1, 2, \ldots, S$ consists of two steps and is performed independently for each state:

(i). Randomly (and uniformly) permute counterfactual New Village locations $g_1^{(s)}$, conditional on (i) distance to roads; (ii) land-use type; and (iii) the number of nearby Chinese squatters in the county.[53]

(ii). Calculate the "dislocation-minimizing" number resettled $\bar{g}_2^{(s)}$ according to (1.3), following which the permuted resettlement density is calculated as $f_c(g_1^{(s)}, \bar{g}_2^{(s)})$.

The expected resettlement density is then approximated by averaging $f_c(g_1^{(s)}, \bar{g}_2^{(s)})$ across permutations of a thousand times ($S = 1000$):

$$\widehat{ResettleDensity}_c = \frac{1}{S} \sum_{s=1}^{S} f_c(g_1^{(s)}, \bar{g}_2^{(s)}). \tag{1.4}$$

Figure 1.2 illustrates this procedure with a single covariate—distance to road—for the state of Johor. The black dot is one actual New Village in the state. The dashed lines show the pre-existing transportation network. The plausibly suitable areas, shaded in gray, are those with roughly the same distance to the network as the actual village.[54] A counterfactual location for that particular village, marked by triangle, is drawn randomly from the suitable areas.

52. Under Assumption 2, the expected resettlement density can be expressed as $\mathbb{E}\left[f_c(g_1, g_2) \mid w\right]$ as

$$\int_{G_1} \int_{G_2} f_c\left(g_1, g_2\right) dG_2\left(g_2 \mid g_1, w\right) dG_1\left(g_1 \mid w\right) = \int_{G_1} f_c\left(g_1, \bar{g}_2\right) dG_1\left(g_1 \mid w\right).$$

53. When no roads were accessible within a 5-kilometer buffer but a river falls within that range, the permutation is conditional on the same distance to the nearest river. In terms of the number of Chinese squatter, I order counties within a state according to the quantile of Chinese squatters and block by counties within the same quantile.

54. In practice, the suitable areas for that particular village would be a subset of the depicted gray areas because of other covariates. For example, only areas with equal distance to the road, under the same land-use type, and have a similar number of Chinese squatters nearby are equally suitable.

A comparison of the actual resettlement pattern with the conditional expected resettlement density suggests that Assumption 2(ii) is reasonable. The actual county resettlement density is strongly correlated with and centered around the expected resettlement density calculated according to the permutation procedure (Appendix Figure 1.12). The underlying resettled population at the village level also aligns well with the prediction of the gravity model (Appendix Figure 1.13).

Figure 1.3 maps the New Villages on top of the expected county resettlement density. Panel A exhibits spatial clustering and overlap between the resettlement areas and the expected resettlement density, consistent with the British relocating people to places with a denser road network and/or a larger pre-existing Chinese settlement.

My preferred specification for estimation controls for the expected resettlement density, in addition to the baseline controls:

$$Y_c = \beta ResettleDensity_c + \lambda \widehat{ResettleDensity}_c + \gamma X_c + \varepsilon_c. \tag{1.5}$$

The identifying variation—the residualized resettlement density (Figure 1.3, panel B)—mainly comes from the exact location of the New Villages, relative to the average location along the transportation network, as well as the more distant resettlement that are not explained by pulling in nearby squatter populations.

To estimate differences in industrial activities at the county-industry pair level with varying resettlement density and Chinese employment share, I estimate the following regressions:

$$\begin{aligned} Y_{cj} = &\beta_1 ResettleDensity_c + \beta_2 ResettleDensity_c \times ChiEmpShare_{cj} + \alpha ChiEmpShare_{cj} \\ &+ \lambda_1 \widehat{ResettleDensity}_c + \lambda_2 \times \widehat{ResettleDensity}_c ChiEmpShare_{cj} + \gamma_{cj} X_{cj} + \delta_j + \varepsilon_{ct}, \end{aligned} \tag{1.6}$$

where $ChiEmpShare_{cj}$ is the state-by-industry employment share of Chinese individuals measured from the 1947 Population Census; X_{cj} fully saturates $ChiEmpShare_{cj}$ and baseline controls X_c from equation (1.5); and δ_j denotes industry fixed effects that absorb nationwide industry-specific shocks.

The identifying variations for β_1 and β_2 come from within-state differences in county resettlement density and its interaction with pre-period industry-specific shares. Specifically, the estimated β_1 corresponds to the impact of a one percent increase in resettlement density on county-industry outcomes Y_{cj} for industries initially having zero Chinese employment share; and the estimated β_2 gives the additional effect for industries fully employed by Chinese, relative to those with zero Chinese share.

To estimate elasticity with respect to county resettlement density for variables like the total number of establishments and the employment share of ethnic groups in specific industries, I face a challenge due to zeros in the data—for instance, counties without any manufacturing establishments. To estimate elasticities that reflect differences on both the extensive and intensive margins in a way not affected by the unit of outcomes, I use the Poisson Pseudo Maximum Likelihood (PPML) estimator (Silva and Tenreyro, 2006). This estimates the effect of county resettlement density as a percentage of the baseline mean (Chen and Roth, 2023).

I report the Conley standard errors that are robust to spatial correlation within a 30-kilometer radius (Conley, 1999). The distance cutoff is chosen based on the localized nature of resettlement shock, which was predominantly within 15 kilometers, beyond which the treatment can be regarded as independent. This is in line with Figure 1.3 (panel B), where there is no significant spatial correlation in the residual resettlement density across counties, with a median county width of approximately 8 kilometers. In the Appendix, I consider robustness of the results to a set of distance cutoffs up to 50 kilometers and clustered standard errors by district (with a total of 66 districts).

1.4.2 Pre-characteristic Balance

In this section, I explore whether various county characteristics measured before resettlement were balanced. Certain characteristics, such as road density and initial Chinese settlement, should naturally correlate with county resettlement density given the objective of the program. However, one would expect that the residual variation—after controlling for the expected resettlement and baseline controls—to be orthogonal to those characteristics, if indeed the procedure successfully purged any remaining omitted variable bias.

Table 1.1 reports the relationship between county resettlement density and various location characteristics. For attributes related to geographic productivity (Columns 1-4), I consider elevation, ruggedness, and the suitability for rice and coconut cultivation—the main food crops in Malaysia. For location amenities and public goods (Columns 5-8), I consider distance to the nearest police station, distance to the nearest post or telegraph office, distance to the nearest hospital, and distance to the nearest Chinese temple. I also examine characteristics related to pre-period economic activities, including (log) population density, the land use shares for rubber and mining—the two major export industries in British Malaya—and proximity to industrial facilities and major cities like Singapore, George Town, Malacca, Ipoh, and Kuala Lumpur—the main commercial and administrative centers of British Malaya.[55]

The raw correlations between these characteristics and county resettlement density are consistent with the resettlement plan, which relocated squatters along the road network. Indeed, Panel A shows that when controlling only for state fixed effects, an indicator of any resettlement in the county, and (log) county area, counties more densely resettled were closer to other public goods provided by the state (Columns 5-7), more densely populated (Column 9), and closer to industrial factories and major cities (Columns 12-13). The higher land use for rubber plantation also aligns with historical accounts that many resettlement areas were

55. Industrial facilities include major strategic industries—airplane and automotive repair facilities, engineering and machine shops, shipbuilding and repair facilities, chemical plants, construction plants, coal mines, power plants, storage facilities, rubber plants, tin plants, and food and clothing manufacturers.

on state-owned rubber estates (Column 10).

However, the examined characteristics are generally balanced once I additionally control for key covariates from the Briggs Plan—including proximity to roads and rail stations, distance to coastline, initial Chinese population share, and the expected resettlement density. Moreover, the magnitudes of the estimates are small: one standard deviation higher resettlement density is associated with a 14-meter increase in elevation, which is about one-ninth of the county mean. These suggest that the identification assumptions are plausible.

Notably, counties with higher resettlement density are not more suitable for agricultural production—if anything, they are less so—even though agricultural suitability was considered in selecting resettlement sites (Columns 3 and 4). This is consistent with historical documentation that resettlement areas often lacked sufficient agricultural land and that economic factors were secondary to security and expedience in the resettlement process.[56]

1.5 Results

This section examines the effects of higher resettlement density in the receiving areas over the next five decades. I first show how the resettlement persistently altered the population distribution. I then present the reduced-form impacts of resettlement for Chinese and non-Chinese individuals separately.

1.5.1 Population Growth and Changes in Ethnic Composition

There was a substantial change in population distribution of the Chinese during the Emergency. Figure 1.4 maps the population growth by Chinese and non-Chinese from 1947 to 1957, during which most resettlement were completed. Counties that experienced resettlement saw substantial Chinese population growth, while neighboring counties witnessed a decline, consistent with the relocation of Chinese squatters from surrounding areas into compact

56. Humphrey (1971), Short (1975, p. 399), Lee and Tan (2000, p. 261).

villages (Panel A). In contrast, there were no significant changes in the population of other ethnic groups (Panel B).

Indeed, regression results show that the Emergency resettlement has persistently shaped the population distribution in Malaysia. Figure 1.5 shows regression (1.5) estimates on county population growth over time (Panel A) and changes in Chinese population share (Panel B) since 1947.[57] By 1957, shortly after most resettlement was completed, counties with one standard deviation higher resettlement density witnessed a 40% rise in Chinese population and negligible changes in other ethnic groups, leading to a 6 percentage point increase in the Chinese population share. This can be seen as the mechanical outcome of the resettlement program.

Post-1960, after mobility restrictions were lifted, these counties not only retained their Chinese population but also saw a steady increase in non-Chinese population over time. From 1957 to 2000, counties with one standard deviation higher resettlement density had an additional 20% increase in both the Chinese and non-Chinese population.

Before resettlement, there were no significant pre-existing population differences in counties later experiencing higher resettlement density, consistent with the identification assumption that counties with different (residualized) resettlement density are comparable in unobserved characteristics. For instance, had the British targeted areas already experiencing growth due to higher labor demand that I do not observe, an influx of population from 1931 to 1947 would have been expected. Moreover, any non-resettlement related factors affecting post-period population shifts would need to disproportionately affect the Chinese. This rules out concerns related to general locational advantages that would have affected both populations equally.

Despite initial similarities, counties with varying resettlement density eventually differed in population density and ethnic composition (Table 1.2). By 1980, counties with one

57. This specification is equivalent to a Two-Way Fixed Effects regression with year and county fixed effects, and interactions of year dummies with county resettlement density, excluding year 1947. The identification assumption is a parallel trend in population size and composition for counties with varying resettlement density, and is weaker than assuming exogeneity in level as in section 1.4.

standard deviation higher resettlement density saw an 11% population density increase and a 5 percentage point rise in Chinese share. By 2000, these figures reached 19.4% and 4.1 percentage points, respectively (Columns 3 and 6).

The denser population in counties with higher resettlement were reflected in a 30.2% more build-up volumes in 1990 (Column 7). The larger point estimate of build-up capital than population suggests that housing supply is relatively elastic. The greater amount of build-up capital also shows up at a more granular scale on satellite images, with a clustering of built-up volumes precisely at the location of the New Village despite a relatively uniform settlement pattern in surrounding areas prior to the resettlement for many cases (Appendix Figure 1.14). This is consistent with a concentration of nearby population in the resettlement areas.

From 1957 to 2000, the population shifts suggest counties with higher resettlement density attracted migrants from other regions. Indeed, Appendix Table 1.10 documents that these counties had a higher share of internal migrants by 1980, particularly among the Chinese. Chinese residents in higher resettlement density counties are 13% more likely to be internal migrants (Panel A, Column 2) and this higher share of Chinese migrants was driven more by endogenous migration after 1960 rather than the initial forced resettlement (Panel B, Column 2).

The post-1960 sorting of Chinese into more densely resettled counties may be attributed to economic benefits, such as higher wages, or homophily—the preference for living near others with similar ethnic backgrounds—or both. However, the rise in the non-Chinese population in these areas was unlikely due to social incentives given the ethnic tensions and more likely driven by economic incentives, which I examine next.[58]

58. For example, a violent ethnic conflict (the "13 May Incident") between Chinese and Malays erupted following an election in 1969.

1.5.2 Economic Structure

I now turn to examine whether the increased population density and a higher concentration of Chinese in the more densely resettled counties impacted the local economic structure by 1980, two decades after the Emergency ended. Table 1.3 documents that counties with greater resettlement density exhibited a larger overall employment size, with heterogeneous effects across sectors. Employment in the primary sector, which includes agriculture and mining, saw a 10% increase in counties with one standard deviation higher resettlement density, similar in scale to that of the population (Table 1.2, Column 2). In contrast, employment in the secondary and tertiary sectors expanded by over 20% in these counties.

This significant growth in employment outside the primary sector stems from a combination of factors. First, the Chinese population, which had a higher propensity to work in the secondary and tertiary sectors (Appendix Figure 1.8), would contribute to a disproportionally higher employment increase in these sectors even if they continue to choose occupation by sector with the same probability as before the resettlement. Indeed, I find that although the Chinese were slightly more likely to sort into manufacturing and services, this mostly reflects changes in occupation within the non-agricultural sector and not across, with a similar share of Chinese working in the primary sector in the more densely resettled counties (Table 1.4, Column 1). The other source of changes in the employment structure in the more resettled counties is a 14% reduction in the share of non-Chinese working in the primary sector (Panel A, Column 2).

The shift towards non-agricultural employment in these counties suggests that local wages in these industries rose following the influx of Chinese, who had a comparative advantage in the industrial and services sector historically. This right-shift in labor supply would theoretically lower relative wages in these sectors in a neoclassic model with downward-sloping demand. However, the observed relative employment growth implies that labor demand may be flat or even upward-sloping. This is consistent with the non-agricultural sectors exhibiting a larger external economy of scale due to agglomeration forces such as those proposed by Marshall (1890).[59]

The substantial increase in employment share of the non-Chinese in sectors such as utility, construction, trade, and transportation, suggest that these industries might indirectly benefit from their input-output linkages with sectors like manufacturing and services, which directly benefited from a positive Chinese labor supply shock.

Additionally, evidence on the manufacturing sector suggests knowledge spillovers from Chinese to non-Chinese within the manufacturing sector. Table 1.5 documents that counties with one standard deviation higher resettlement density saw an average increase of 13 manufacturing establishments, driven by both Chinese and non-Chinese ownership (Panel A). In particular, there was a surge in non-Chinese entrepreneurs entering the manufacturing sector, particularly in sub-industries initially dominated by Chinese employment. This is consistent with transfers of industry know-how between Chinese and non-Chinese entrepreneurs (Appendix Table 1.11).

Considering the varying skill demands across industries, with the industrial sector typically demanding higher skilled labors compared to agriculture, one might expect that the different economic structures in the more densely resettled counties could lead to a higher human capital. Indeed, Table 1.6 shows that residents in counties with more resettlement generally have better education, particularly for the Chinese, with an additional 0.42 years of schooling (Panel A, Column 2); 6.9% more likely to complete primary education (Panel B, Column 2); and 17.8% more likely to complete secondary education (Panel C, Column 2). The larger estimate from secondary education completion is consistent with a higher valued placed on skilled labor.

Younger individuals, particularly those under 50 in 1980 who were still in or had not finished their educational years during the resettlement, showed the most significant educational gains

59. Agglomeration economies include benefits like labor market pooling, reduced transportation costs, and knowledge spillovers.

(Appendix Table 1.12). In contrast, the resettlement is not correlated with the cohorts that had completed their education by 1950. This provides an additional supporting evidence that these counties were not ex-ante more industrialized and populated with a group of more educated individuals during the colonial period.

Health outcomes also improved in these counties, as evidenced by higher birth weights reported in the Malaysian Family Survey of the late 1980s. This improvement was observed across both Chinese and non-Chinese mothers, with a more substantial effect for Chinese (Table 1.6, Panel D).

1.5.3 Household Income

The analysis so far suggests that resettlement led to a shift in the local economic structure of recipient counties towards the non-agricultural sector. This section examines how this change might have translated into different levels of household income by 1980 for Chinese and non-Chinese households, respectively.

I first document that households in counties with higher resettlement density were more likely to own durable assets, such as automobile, refrigerator, TV, etc., by late 1980s (Appendix Table 1.13), especially for Chinese households. This suggests that in the more resettled counties, households were generally richer and there was a larger income premium between Chinese and other ethnic groups.

Indeed, Table 1.7, Panel A, documents that households in counties with one standard deviation higher resettlement density had, on average, a 6.9% higher household income (Column 1). The effect is notably stronger for Chinese households, which saw an 11.1% income increase in these counties compared to those in other areas (Column 2). In contrast, non-Chinese households in higher resettlement density counties had a marginally higher income of 3.7%, which was not statistically significant (Column 3). This leads to a 7.4% Chinese income premium in more resettled counties (Column 4).

When splitting households based on the employment sector of the household head—primary versus secondary/tertiary—the income differential in more resettled counties is smaller for the primary sector at 2.9% (Panel B, Column 1). Even so, Chinese households involved in the primary sector still earn 7.3% more in counties with higher resettlement density (Column 2), whereas non-Chinese households' incomes remained comparable to those in other areas (Column 3).

The larger income differential in the more resettled counties was primarily driven by households in the non-agricultural sector, who, on average, earned 7.7% more than non-agricultural households in other areas (Panel C, Column 1). The effect is again stronger among Chinese households, at 12.1% (Column 2). There was a smaller, 4.4% (and not statistically significant) increase in income for non-Chinese households (Column 3). Despite the overall gains for non-Chinese households in these areas, a 7.8% Chinese income premium persists (Column 4).

These income differences across counties, sectors, and ethnic groups remain even after controlling for the educational level of the household head (Appendix Table 1.14). Given the higher average educational attainment in more resettled counties, especially among Chinese, the observed income disparities could partly stem from the private returns to education rather than from productivity spillovers due to local agglomeration. Indeed, controlling for the years of schooling of the household head reduces the income differential in the more resettled counties, especially among the Chinese households. However, the general pattern of a Chinese income premium in both the primary and non-primary sectors persists, suggesting that educational disparities alone cannot fully explain the economic benefits experienced by Chinese in these areas relative to other groups.

1.5.4 Discussion

The evidence put together above suggests that the resettlement of Chinese squatters during the Emergency stimulated an agglomeration economy. Counties with higher resettlement densities saw an increase in population density, income levels, and a shift toward non-agricultural sectors, consistent with external economies of scale. Improved labor market conditions then attracted further migration from other regions.

The Chinese population benefited substantially from this agglomeration, achieving significant income gains, even after accounting for industry and educational attainment, suggesting a stronger benefit from population concentrations of similar ethnic backgrounds. This is consistent with stronger within-group social interactions that generate larger productivity spillovers among the Chinese population.

Non-Chinese ethnic groups also gained from agglomeration, albeit to a lesser degree, mainly through transitioning into non-agricultural sectors. In the more densely resettled counties, they were more likely to work outside agriculture. Manufacturing firm data suggests that industry-specific expertise from the Chinese spilled over to non-Chinese entrepreneurs, facilitating their entry into sectors traditionally dominated by Chinese. Nonetheless, non-Chinese households saw lesser income increases, hinting at more limited cross-ethnic productivity spillovers. This is presumably due to barriers of social interactions between groups, such as ethnic tensions and language barriers, that hindered cross-ethnic spillovers.

In a counterfactual world without these barriers, Malays might have seen greater benefits from interacting with the Chinese population, potentially encouraging more Malays to move to Chinese-dense areas and speed up the structural change towards the non-agricultural sector. Similarly, as a minority constituting less than 40% of Malaysia's total population, the Chinese could have gained from interacting with a broader group of people.

To quantify the aggregate economic impact of these barriers, I develop a spatial general equilibrium model with migration and occupational choices, which allows the agglomeration forces to vary with sector and local ethnic composition.

1.6 Model

The model consists of N locations (or counties) and two sectors $k \in \{A, M\}$: Agriculture (A) and Manufacturing (M). Individuals are characterized by ethnicity $e \in \{c, m\}$: Chinese (c) and Malays (m), and each is endowed with an initial county. After drawing a regional taste shock, individuals make their migration decision. After they move, they draw idiosyncratic efficiency units for each sector and choose which sector to work. Lastly, consumption and production takes place.

1.6.1 Production

Each region n produces a unique good in sector A and M (Armington, 1969). In each sector $k \in \{A, M\}$ of a region, there is a continuum of perfectly competitive firms producing this homogeneous regional variety. Each firm's production technology has constant returns to scale and uses labor as the only input, leading to the region production function $Q_{nk} = H_{nk}$, where H_{nk} is the total labor summed across ethnic groups (in efficiency unit, defined later) employed in region n, sector k. I assume that efficient units from Chinese and Malays are perfect substitutes in the production function.

Firms in sector k, region n choose labor H_{nk} to maximize profit, taking local sectoral wages (per efficiency unit) w_{nk} and prices $\{p_{nrk}\}$ as given, where p_{nrk} is the price of goods produced in sector k and region n when sold in region r. In equilibrium, no arbitrage condition implies that $p_{nrk} = (\tau_{nr}/\tau_{nn})p_{nnk}$ for regions n, r, and sector k, with $\tau_{nr} \geq 1$ for all n and r being the iceberg trade cost from n to r. Perfect competition implies that in equilibrium firms earn zero profit, with $w_{nk} = p_{nnk}/\tau_{nn}$, where p_{nnk} is the price of sector-k goods sold locally under the trade cost τ_{nn} within region n.[60]

60. Notice that I allow within-region trade to be costly (when $\tau_{nn} > 1$). Local wage would decline as τ_{nn}

1.6.2 Consumption

Individuals of ethnicity e living in location n derive utility from consuming agricultural and manufacturing goods and enjoying the amenity in location n:

$$U_n^e(C_A, C_M) = a_n^e \left(\frac{C_A}{\alpha}\right)^{\alpha} \left(\frac{C_M}{1-\alpha}\right)^{1-\alpha}$$
$$C_k = \left(\sum_{r=1}^{N} c_{rk}^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}},$$

where $\sigma > 1$ is the constant elasticity of substitution across regional varieties, which is assumed to be the same for both sectors; and a_n^e is the valuation of individuals of ethnicity e for amenities in location n, which I will discuss in more detail in the migration decision.

Utility maximization implies that the indirect utility of a group-e individual in region n with income y_n^e is $a_n^e y_n^e / P_n$, where $P_n \equiv P_{nA}^{\alpha} P_{nM}^{1-\alpha}$ is the ideal price index in region n; and $P_{nk} \equiv (\sum_{l=1}^{N} \tau_{ln}^{1-\sigma} w_{lk}^{1-\sigma})^{1/(1-\sigma)}$ is the price index of sector k goods in region n.

1.6.3 Sectoral labor supply

Individuals with heterogeneous productivity earn their income by inelastically supplying one unit of labor. Each person is characterized by a vector of efficiency units in sectors A and M, $\Lambda_i = (\Lambda_{iA}, \Lambda_{iM})$, where Λ_{ik} denotes the effective labor individual i provides if he/she works in sector k. I assume that individuals of ethnicity e, region n, draw their gross efficiency units in sector k independently from the following Fréchet distribution:

$$F_{nk}^e(\Lambda) = \exp\left(-\phi_{nk}^e \Lambda^{-\theta}\right),$$

increases, which plays the same role of worse location fundamental.

where scale ϕ^e_{nk} parameterizes the average productivity of ethnicity e in sector k and region n. The ϕ^e_{nk} captures the absolute and comparative advantage of different ethnic groups, as well as fixed location fundamentals that make a place more productive in specific sectors. The dispersion of efficiency units is governed by shape parameter θ, with a higher value of θ corresponding to smaller dispersion.

Due to human capital externalities, an individual i in ethnic group e's net efficiency unit in region n and sector k, denoted by λ^e_{ink}, depends not only on its own skill Λ^e_{ink} but also local population distribution:

$$\lambda^e_{ink} = \Lambda^e_{ink} f_\lambda(L^e_{nk}, L^{e'}_{nk}),$$

where I parameterize $f_\lambda(\cdot)$ as a function of sectoral population size and ethnic composition:

$$f_\lambda(L^e_{nk}, L^{e'}_{nk}) \equiv (L_{nk})^{\gamma_k} \left(\frac{L^e_{nk}}{L_{nk}}\right)^{\gamma^e}.$$

Parameters γ_k and γ^e govern the strength of productivity spillovers from local interactions, which are allowed to depend not only on the number of workers in the sector (γ_k), but the ethnic composition (γ^e) of that sector. Specifically, the elasticity of productivity with respect to the concentration of ethnic group e in region n is given by

$$\frac{\partial \ln \lambda^e_{nk}}{\partial \ln L^e_n} = \underbrace{\left(\gamma_k \frac{L^e_{nk}}{L_{nk}} + \gamma^e \left(1 - \frac{L^e_{nk}}{L_{nk}}\right)\right)}_{\text{Direct effect}} \underbrace{\left(1 + \frac{\partial \ln \pi^e_{nk}}{\partial \ln L^e_n}\right)}_{\text{Indirect/GE effect}}, \tag{1.7}$$

where π^e_{nk} is the share of ethnicity e in region n working in sector k.

A higher population of ethnicity e has a direct effect on group e's productivity in sector k, holding fixed the occupation structure of group e, and an indirect effect due to endogenous changes in occupation share. Let's first consider the direct effect. The firm term in equation (1.7) is a weighted average of γ_k and γ^e, weighted by the share of workers in sector k from ethnicity e. That is, if sector k in region n is dominated by group e (with a large L^e_{nk}/L_{nk}), the agglomeration elasticity for group e is close to γ_k. If instead, the sector is dominated by group e', the agglomeration elasticity for group e would be close to γ^e. The indirect effect then scales the agglomeration elasticity, depending on how a larger size of ethnic group e affects the occupation share of e in sector k due to changes in the relative wage in the general equilibrium.

There are also cross-ethnicity productivity spillovers. For ethnic $e' \neq e$, the elasticity of group e's efficiency with respect to the concentration of the other group e' is given by

$$\frac{\partial \ln \lambda^e_{nk}}{\partial \ln L^{e'}_n} = \underbrace{(\gamma_k - \gamma^e)\,\frac{L^{e'}_{nk}}{L_{nk}}}_{\text{Direct effect}} \underbrace{\left(1 + \frac{\partial \ln \pi^{e'}_{nk}}{\partial \ln L^{e'}_n}\right)}_{\text{Indirect/GE effect}}. \tag{1.8}$$

Again, the cross-ethnicity spillover from e' to e also has a direct and indirect component. But overall, when $\gamma_k > \gamma^e$, the cross-ethnicity spillover for sector k tends to be positive, and vice versa. Moreover, the effect is proportional to the share of group e' in sector k.

Given the Frećhet distributed efficiency units Λ^e_k, the share of individuals of ethnicity e in region n who work in sector k can be expressed as

$$\pi^e_{nk} = \phi^e_{nk}\left(\frac{w^e_{nk}}{\bar{w}^e_n}\right)^{\theta}, \tag{1.9}$$

where

$$w^e_{nk} = w_{nk}\,(L_{nk})^{\gamma_k}\left(\frac{L^e_{nk}}{L_{nk}}\right)^{\gamma^e}, \tag{1.10}$$

and the average wage (up to a scale) for ethnicity e in region n is given by

$$\bar{w}^e_n = \left(\phi^e_{nA}\left(w^e_{nA}\right)^{\theta} + \phi^e_{nM}\left(w^e_{nM}\right)^{\theta}\right)^{1/\theta}.$$

1.6.4 Migration

Each individual is endowed with an initial location, who then decides where to migrate at a cost after drawing a regional taste shock. When making the migration decision, individuals know their ethnicity but have not learned their skill realizations. Individual i of ethnicity e draws an idiosyncratic taste for region n, denoted u^e_{in}, from the following Fréchet distribution:

$$F^e_n(u) = \exp\left(-\bar{a}^e_n u^{-\nu}\right),$$

where the scale parameter $\bar{a}^e_n$ captures any exogenous, ethnicity-specific amenity of location n; and the shape parameter ν describes the dispersion of u, with a higher ν corresponding to a smaller taste dispersion across regions.

Individual i of group e has a valuation for amenities in region n, denoted a^e_{in}, that depends on her idiosyncratic taste u^e_{in} and local population distribution:

$$a^e_{in} = u^e_{in} f_a(L^e_n, L^{e'}_n),$$

where, again, I parameterize $f_a(\cdot)$ as a function of population size and ethnic composition:

$$f_a(L^e_n, L^{e'}_n) = (L_n)^{\beta} \left(\frac{L^e_n}{L_n}\right)^{\beta^e}.$$

Parameters β and β^e capture congestion and amenity spillovers that are allowed to depend not only on the size of local population but also on the ethnic composition. Specifically, β^e allows for homophily, a taste for living near other people of the same ethnicity, which might also reflect ethnic tension across groups.

Indirect utility for individual i of ethnicity e from origin r living in destination n, who has an idiosyncratic preference a^e_{in}, is given by

$$V^e_{irn} = \eta^{-1}_{rn} a^e_{in} \Gamma_\theta \bar{w}^e_n P^{-1}_n,$$

where η_{rn} is the iceberg migration cost from r to n, and $\Gamma_\theta \bar{w}^e_n P^{-1}_n$ is the real wage in region n, with $\Gamma_\theta \equiv \Gamma(1 - 1/\theta)$ and $\Gamma(\cdot)$ being the Gamma function.

Since the indirect utility is equal to a Fréchet random variable u^e_{in} multiplied by a constant $\eta^{-1}_{rn} L^\beta_n (L^e_n/L_n)^{\beta^e} \Gamma_\theta \bar{w}^e_n P^{-1}_n$, it is itself Fréchet distributed. The distribution of V^e_{irn} implies that the share of ethnicity e initially residing in region r who choose to migrate to n is

$$m^e_{rn} = \frac{\left(\eta^{-1}_{rn} V^e_n\right)^\nu}{\sum_{l=1}^N \left(\eta^{-1}_{rl} V^e_l\right)^\nu},$$

where the mean value of residing in region n for ethnicity e is

$$V^e_n = \left(\bar{a}^e_n\right)^{1/\nu} \left(L_n\right)^\beta \left(\frac{L^e_n}{L_n}\right)^{\beta^e} \bar{w}^e_n P^{-1}_n. \tag{1.11}$$

We can then write the bilateral migration flow of ethnic group e from r to n as

$$L^e_{rn} = \eta^{-\nu}_{rn} \times \frac{\check{L}^e_r}{(\Pi^e_r)^\nu} \times \frac{L^e_n/\bar{L}}{(\mathcal{V}^e_n)^{-\nu}}, \tag{1.12}$$

where, following the terminology of the trade literature, I define two migration market access terms:

$$\Pi^e_r \equiv \left(\sum_{l=1}^N (\eta^{-1}_{rl} V^e_l)^\nu\right)^{1/\nu}, \tag{1.13}$$

$$\mathcal{V}^e_n \equiv V^e_n \left(L^e_n/\bar{L}\right)^{-1/\nu}, \tag{1.14}$$

and I use $\bar{L}$ to denote the total population in the country, which I normalize to be one.

The term Π_r^e captures the overall value for group e to move *out of* region r. On the other hand, $\mathcal{V}_n^e$ captures group e's overall value of moving *into* region n. These terms are referred to as the outward or inward migration market access in the trade literature (Anderson and Van Wincoop, 2003).

1.6.5 Trade

Bilateral trade flows from region n to region r incur an exogenous iceberg trade cost, $\tau_{nr} \geq 1$, with $\tau_{nr} = 1$ corresponding to the case of friction-less trade. Given this and consumer preferences, trade flow expenditures on sector-k goods from r to n (with goods flowing from n to r), denoted X_{nrk}, have the standard gravity form:

$$X_{nrk} = X_{rk} \frac{\tau_{nr}^{1-\sigma} (w_{nk})^{1-\sigma}}{\sum_{l=1}^{N} \tau_{lr}^{1-\sigma} (w_{lk})^{1-\sigma}}, \tag{1.15}$$

where the total expenditure of region r on sector-k goods is given by $X_{rk} = \alpha_k Y_r$, with $\alpha_A = \alpha$ and $\alpha_M = 1 - \alpha$; and $Y_n = w_{rA} H_{rA} + w_{rM} H_{rM}$.

It is useful to rewrite the above equation as

$$X_{nrk} = \alpha_k \tau_{nr}^{1-\sigma} \times \frac{Y_n / \bar{Y}}{\mathcal{P}_{nk}^{1-\sigma}} \times \frac{Y_r}{P_{rk}^{1-\sigma}}, \tag{1.16}$$

where, similar to migration flows, I define two terms of the trade market access:

$$P_{rk} \equiv \left(\sum_{l=1}^{N} \tau_{lr}^{1-\sigma} w_{lk}^{1-\sigma} \right)^{1/(1-\sigma)}, \tag{1.17}$$

$$\mathcal{P}_{nk} \equiv w_{nk}^{-1} \left(Y_n / \bar{Y} \right)^{1/(1-\sigma)}. \tag{1.18}$$

I use $\bar{Y} \equiv \sum_r Y_r$ to denote the total income of the economy, which is normalized to one as the numeraire.

As with the migration flows described above, (the inverse of) P_{rk} captures the inward trade market access to sector-k goods of region r and (the inverse of) $\mathcal{P}_{nr}$ captures the outward trade market access to sector-k goods of region n.

1.6.6 Static equilibrium

For any strictly positive initial population vector $\{\check{L}_l^e\}$ and a vector of ethnicity- or industry-specific location fundamentals $\{\phi_{nk}^e, \bar{a}_n^e, \tau_{ln}, \eta_{ln}\}$, an equilibrium is a vector of prices $\{w_{nk}, p_{nk}\}$ and quantities $\{L_{nk}^e, H_{nk}\}$, such that (i) firms and consumers behave optimally and (ii) goods and labor markets clear for all regions.

The goods market clearing condition can be written as

$$w_{nk}H_{nk} = \sum_{r=1}^{N} \alpha_k(w_{rA}H_{rA} + w_{rM}H_{rM}) \frac{\tau_{nr}^{1-\sigma} w_{nk}^{1-\sigma}}{\sum_{l=1}^{N} \tau_{lr}^{1-\sigma} w_{lk}^{1-\sigma}} \tag{1.19}$$

which embeds two underlying conditions: (i) total sectoral sales of a region are equal to payments to labor and (ii) a region's total income is fully spent on goods from all locations.

The labor market clearing condition can be written as

$$H_{nk} = \sum_e H_{nk}^e = \sum_e L_n^e \pi_{nk}^e \left(\Gamma_\theta \bar{w}_n^e w_{nk}^{-1}\right) \tag{1.20}$$

$$L_n^e = \sum_r \check{L}_{rn}^e \frac{\left(\eta_{rn}^{-1} V_n^e\right)^\nu}{\sum_{l=1}^{N} \left(\eta_{rl}^{-1} V_l^e\right)^\nu}. \tag{1.21}$$

Equation (1.20) says that a region's total efficiency units in sector k is given by the sum of efficiency units contributed by the two ethnic groups, and the part coming from group e is the product of their sectoral employment ($L_n^e \pi_{nk}^e$) multiplied by their average efficiency units ($\Gamma_\theta \bar{w}_n^e w_{nk}^{-1}$). Equation (1.21) follows from the migration flow identity, where a region's equilibrium population of ethnicity e is equal to the sum of migration flows of ethnicity-e individuals from all regions.

Using equations (1.9), (1.10), and (1.11), one can substitute out π^e_{nk}, $\bar{w}^e_n$, and V^e_n and replace them with exogenous parameters and endogenous outcomes $\{w_{nk}, L^e_n\}$. As a result, the equilibrium is characterized by a system of $6 \times N$ equations (1.19–1.21) in $6 \times N$ unknowns $\{w_{nk}, H_{nk}, L^e_n\}$ with $k \in \{A, M\}$ and $e \in \{c, m\}$.

1.7 Identification and Estimation

In this section, I discuss how I identify and estimate the model parameters. I assume that the bilateral migration and trade costs are symmetric and scale proportionally with distance. Specifically, bilateral migration costs are modeled as $\eta_{rn} = (d_{rn}/d_{min})^\kappa$, where d_{min} is the minimum within-county distance, and $\kappa \geq 0$ represents the distance elasticity of migration costs. Similarly, bilateral trade costs are represented by $\tau_{nr} = (d_{rn}/d_{min})^\xi$, where $\xi \geq 0$ is the distance elasticity for trade costs.[61] The model is characterized by a tuple of location fundamentals $\{\phi^e_{nk}, \bar{a}^e_n\}$ and 11 structural parameters:

$$\Theta \equiv \{ \underbrace{\alpha, \sigma}_{\text{Preference}}, \underbrace{\xi, \kappa}_{\text{Trade/Migration}}, \underbrace{\theta, \gamma_A, \gamma_M, \gamma^e}_{\text{Productivity}}, \underbrace{\nu, \beta, \beta^e}_{\text{Amenity}} \}.$$

I take 3 parameters externally: the elasticity of substitution across regional variety σ, the migration elasticity ν, and the distance elasticity of trade cost ξ. I estimate the remaining 8 parameters.

61. I measure cross-county distances d_{rn} for any $r \neq n$ with the Euclidean distance between the centroid of r and n; and within-county distances d_{rr} with the distance between the centroid and boundary of county r. I allow the within-county migration or trade costs to be greater than one (except for the county with the smallest distance) to accommodate the fact that counties vary in size. This is without loss of generality as costly migrations within a county lowers the utility in a way isomorphic to having worse amenity fundamentals $\bar{a}^e_n$. Similarly, costly trades within a county lowers the productivity in a way isomorphic to having worse productive fundamentals ϕ^e_{nk}.

1.7.1 *Identification Strategy*

In this section, I first introduce a proposition that shows the identification of the agricultural expenditure share α and the market access terms. Next, I discuss my strategy for identifying the remaining model parameters and recovering the location fundamentals. Finally, I discuss the estimation procedure and results.

Market Access Terms

I derive four underlying conditions involving the trade and migration market access terms from the equilibrium conditions (1.19)–(1.20).

(i). Total sales equals payments to labor: $w_{nk}H_{nk} = \sum_r X_{nrk}$. Using equation (1.16), this can be written as

$$\mathcal{P}_{nk}^{1-\sigma} = \frac{\alpha_k}{\Omega_{nk}} \sum_r \tau_{nr}^{1-\sigma} Y_r P_{rk}^{\sigma-1},$$

where $\Omega_{nk} \equiv w_{nk}H_{nk}/Y_n$ denotes the share of income in region n generated from sector k.

(ii). Total income equals total expenditure: $Y_r\alpha_k = \sum_n X_{nrk}$. This can be written as

$$P_{rk}^{1-\sigma} = \sum_n \tau_{nr}^{1-\sigma} Y_n \mathcal{P}_{nk}^{\sigma-1}.$$

(iii). Final population equals total in-migrations: $L_n^e = \sum_{r=1}^N L_{rn}^e$. Using equation (1.12), this can be written as

$$(\mathcal{V}_n^e)^{-\nu} = \sum_r \eta_{rn}^{-v} \check{L}_r^e \left(\Pi_r^e\right)^{-v}$$

(iv). Initial population equals total out-migrations: $\check{L}_r^e = \sum_{n=1}^{N} L_{rn}^e$. This can be written as

$$(\Pi_r^e)^{v} = \sum_n \eta_{rn}^{-v} L_n^e \left(\mathcal{V}_n^e\right)^{\nu}.$$

Putting these together, the derivation above yields a system of four equations:

$$\mathcal{P}_{nk}^{1-\sigma} = \frac{\alpha_k}{\Omega_{nk}} \sum_r \tau_{nr}^{1-\sigma} Y_r P_{rk}^{\sigma-1}, \tag{1.22}$$

$$P_{rk}^{1-\sigma} = \sum_n \tau_{nr}^{1-\sigma} Y_n \mathcal{P}_{nk}^{\sigma-1}, \tag{1.23}$$

$$\left(\mathcal{V}_n^e\right)^{-\nu} = \sum_r \eta_{rn}^{-v} \check{L}_r^e \left(\Pi_r^e\right)^{-v}, \tag{1.24}$$

$$(\Pi_r^e)^{v} = \sum_n \eta_{rn}^{-v} L_n^e \left(\mathcal{V}_n^e\right)^{\nu}. \tag{1.25}$$

Proposition 1. *Given observed data on* $\{Y_n, \Omega_{nk}, \check{L}_n^e, L_n^e\}$ *and parameter values* $\{\tau_{nr}^{1-\sigma}, \eta_{nr}^{-\nu}\}$*, there exists a unique (up to scale) set of values of* $\{\mathcal{P}_{nk}^{\sigma-1}, P_{rk}^{\sigma-1}, \left(\mathcal{V}_n^e\right)^{\nu}, \left(\Pi_r^e\right)^{v}\}$ *that satisfy equations* (1.22)–(1.25).

Proof. Given data on total income $\{Y_n\}$ and the sectoral income share Ω_{nk}, the agricultural expenditure share α is identified. This follows from the fact that since every region spends α_A of income on agricultural goods, the whole economy must also spend that same share on agricultural goods in aggregate:

$$\alpha = \frac{\sum_n w_{nA} H_{nA}}{\sum_n w_{nA} H_{nA} + w_{nM} H_{nM}} = \frac{\sum_n Y_n \Omega_{nA}}{\bar{Y}} = \sum_n Y_n \Omega_{nA}.$$

The remainder of the proof follows closely from Allen and Donaldson (2022), Proposition 3. □

Notice that equations (1.22)–(1.25) can be split into two distinct system of equations (1.22)–(1.23) and (1.24)–(1.25). In fact, identification of the trade market access terms

$\{P_{nk}^{\sigma-1}, P_{rk}^{\sigma-1}\}$ only requires $\{Y_n, \Omega_{nk}, \tau_{nr}^{1-\sigma}\}$; and, similarly, to identify the migration market access terms $\{(\mathcal{V}_n^e)^\nu, (\Pi_r^e)^\upsilon\}$ one only requires $\{\check{L}_n^e, L_n^e, \eta_{nr}^{-\nu}\}$. This Proposition also implies that I can determine the market access terms without taking a stand on the functional form of the agglomeration forces or the values of parameters $\{\gamma_A, \gamma_M, \gamma^e, \beta, \beta^e\}$.

Migration Cost Elasticity

As the migration cost elasticity κ enters the migration cost multiplicatively with the taste dispersion ν, I estimate the product of $\tilde{\kappa} \equiv \kappa\nu$. Specifically, I proceed with non-linear least squares that minimizes the difference between model-predicted (county-to-county) migration flows and the observed (district-to-district) migration flows data. The estimation proceeds as follows.

(i). Guess an initial $\tilde{\kappa}$ and calculate the associated migration costs $\eta_{rn}^{\nu} = (d_{rn}/d_{min})^{\tilde{\kappa}}$.

(ii). Given data on the initial and final population distribution $\{\check{L}_r^e, L_n^e\}$, solve for the migration market access terms $\{(\mathcal{V}_n^e)^\nu, (\Pi_n^e)^\nu\}$ using Proposition 1.

(iii). Calculate the implied bilateral migration flows:

$$L_{rn} = \sum_e L_{rn}^e = \sum_e d_{rn}^{-\tilde{\kappa}} \times \frac{\check{L}_r^e}{(\Pi_r^e)^\nu} \times \frac{L_n^e}{(\mathcal{V}_n^e)^{-\nu}}.$$

(iv). Aggregate the model-implied migration flows to the level of data, which is district-by-district, and calculate the bilateral migration shares of each destination by origin:

$$m_{jh} = \frac{\sum_{r\in j(r)} \sum_{n\in h(n)} L_{rn}}{\sum_{r\in j(r)} \sum_n L_{rn}},$$

where $j(r)$ and $h(n)$ denote the district that county r and n belongs to, respectively.

(v). Calculate the loss objective as the sum of squared differences between the model-predicted (log) migration shares and the data counterparts:

$$loss \equiv \frac{1}{N_d^2} \sum_{j,h} \left(\ln m_{jh} - \ln \hat{m}_{jh}\right)^2, \tag{1.26}$$

where N_d is the total number of districts and $\hat{m}_{jh}$ denotes the bilateral migration shares in the data.

(vi). Search over the space of $\tilde{\kappa}$ until the loss function is minimized.

The identification assumption underlying this approach is that differences between observed migration flows and the model predicted migration is due to classical measurement errors that are uncorrelated with geography and other unobservables that enter the migration market access terms. As sample size approaches infinity, measurement errors vanish and the observed migration flows would exactly equal the model predicted migration flows under the true $\tilde{\kappa}$. Identification thus requires that there exists such a unique $\tilde{\kappa}$. Although this is hard to prove directly, I provide suggestive evidence in Figure 1.16 that, at least in the sample, the loss function appears to be convex and there is a unique $\tilde{\kappa}$ that attains the minimum.

Another challenge is that the observed migration flows may not have the same frequency of 24 years as in the model. This is because the 1980 census from which I measure migration flows tabulates population by place of last previous residence by each place of current residence without restricting to a period. Indeed, the 2% census microdata suggests that the average person spends around 12 years in the current location. Given this, I consider that the tabulated migration flows are measured at a frequency of 12 years and, assuming that the migration shares are stable over time, I convert the migration shares matrix to a frequency of 24 years, following the procedure of Artuç et al. (2010) and Caliendo et al. (2019).[62] This

62. Specifically, I first calculate the 12-years migration shares matrix from the observed bilateral flows, such that each row sums to 1. Then, under the assumption that the migration shares matrix, which can be interpreted as a probability transition matrix, is constant over the two 12-years period, I take the square of it to obtain the 24-years migration shares matrix $\hat{m}_{jh}$.

procedure also ensures that there are no zeros in the 24-years migration shares, which I can take log as in equation (1.26).

Shape of Fréchet Skills

The shape parameter θ governs the dispersion of Fréchet distributed productivity draws of the individuals. A higher value of θ implies a smaller dispersion of productivity. Notice that each individual's potential earning is also Fréchet distributed. Let y^e_{ink} denote the earning of individual i of ethnicity e who works in sector k and resides in region n. Then, the assumed distribution implies

$$\frac{\text{Var}[y^e_{ink}]}{\mathbb{E}[y^e_{ink}]^2} = \frac{\Gamma(1-\frac{2}{\theta}) - \Gamma(1-\frac{1}{\theta})^2}{\Gamma(1-\frac{1}{\theta})^2}. \tag{1.27}$$

Specifically, the variance of y^e_{ink}, normalized by the squared expectation, is a function of θ that goes to infinity as θ approaches 2 from above, and decreases monotonically toward zero as θ increases.[63] This implies that there exists a unique solution of θ for any normalized variance; i.e., θ is identified from this moment.[64]

Productivity Spillovers

The parameters governing the strength of productivity spillovers $\{\gamma_A, \gamma_M, \gamma^e\}$ affect the expected earning in a sector and hence people's occupational choice. I rewrite the occupation

63. For the variance of Fréchet distributed y^e_{ink} to exist, θ must be larger than 2.

64. As θ is assumed to identical across locations, sectors, and ethnic groups, it is over-identified from the data, where the normalized variance can differ across locations, sectors, and ethnic groups.

choice equation (1.9) in terms of the trade market access (1.18) as

$$\ln \bar{w}_n^e = \gamma_k \ln L_{nk} + \gamma^e \ln \left(\frac{L_{nk}^e}{L_{nk}} \right) - \frac{1}{\theta} \ln \pi_{nk}^e - \left(\frac{1}{\sigma - 1} \right) \ln (\mathcal{P}_{nk})^{\sigma - 1} - \left(\frac{1}{\sigma - 1} \right) \ln Y_n + \underbrace{\frac{1}{\theta} \ln \phi_{nk}^e}_{\text{error term}}, \quad \forall k \in \{A, M\}. \tag{1.28}$$

The outcome variable is the average wage of ethnic group e in location n, which is observed in the data.[65] Local employment size in sector k shifts the average wage of group-e via γ_k. The ethnic composition of local employment shifts local wage via γ^e.

To see the key variation that identifies γ^e, we can subtract equation (1.28) of one group from the other to eliminate any region-industry specific terms:

$$\ln \left(\frac{\bar{w}_n^c}{\bar{w}_n^m} \right) = \gamma^e \ln \left(\frac{L_{nk}^c}{L_{nk}^m} \right) - \frac{1}{\theta} \ln \left(\frac{\pi_{nk}^c}{\pi_{nk}^m} \right) + \underbrace{\frac{1}{\theta} \ln \left(\frac{\phi_{nk}^c}{\phi_{nk}^m} \right)}_{\text{error term}}. \tag{1.29}$$

This expression is a relative (inverse) demand curve of sector k, where the negative $1/\theta$ term reflects the neoclassical force that gives rise to a downward sloping demand when the within-group agglomeration force disciplined by γ^e is not too strong.[66]

The unobserved productivity for ethnicity e in sector k and region n enters as the error term, which typically correlates positively with local population due to the selection of people into more productive places. This tends to bias the OLS estimate of γ_k upward. Similarly, ethnic group e who is more productive in a specific location-industry pair would sort to take

65. The reason that the left-hand-side of equation (1.28) doesn't vary with k, while the right-hand-side does, is due to a Fréchet property and the fact that the shape parameter is constant across industries (see Appendix 1.C.1). If the Fréchet skills have different shape parameters across sectors, that would generate sector-specific average wages by ethnicity. In that sense, one can view equation (1.28) as a limiting case where the sector-specific shape parameters converge to the same value.

66. Only the relative share matters here and not the relative quantity because I assume that Chinese and Malays are perfect substitutes. If Chinese and Malays have equal probability of choosing to work in sector k, then the relative abundance of ethnic populations would not influence relative wages, except through γ^e.

advantage of the better fundamentals, biasing the OLS estimate of γ^e upward. On the other hand, classical measurement errors of the population distribution can also bias γ_k and γ^e downward due to attenuation.

I address these issues with an instrumental variable strategy, exploiting the exogenous resettlement shocks that shifted the initial population distribution in 1957 and ended up affecting equilibrium population in 1980. Specifically, I use the residualized resettlement density as the population shifter, which I have shown in section 1.5 that persistently increased both the total population size and the Chinese share in 1980. Let it be denoted by Z_n^r.

Since the British did not relocate people in a way that correlates with underlying productivity (conditional on covariates), this population shifter is plausibly orthogonal to location fundamentals ϕ_{nk}^e, leading to my identifying moment conditions:

$$\mathbb{E}[Z_n^r \ln \phi_{nk}^e] = 0, \ \forall k, e. \tag{1.30}$$

These four moment conditions identify the three parameters related to productivity spillovers: $\{\gamma_A, \gamma_M, \gamma^e\}$. To see this, notice that γ^e can be identified from equation (1.29), using Z_n^r as an instrument for $\ln(L_{nk}^e/L_{nk})$, given θ. Then, moving all the terms in equation (1.28) to the left, except for $\ln L_{nk}$, I identify γ_k with the same instrument for $\ln L_{nk}$.

I require bilateral trade costs $\tau_{nr}^{1-\sigma} = (d_{rn}/d_{min})^{\xi(1-\sigma)}$ to solve for the trade market access terms $\{\mathcal{P}_{nk}^{\sigma-1}, P_{nk}^{\sigma-1}\}$, which are used in the estimating equation. However, since I do not observe data on trade flows, I target $\xi(1-\sigma) = -1.29$ as the elasticity of trade flows with respect to distance, estimated by Monte et al. (2018). In the literature, estimates of σ—the elasticity of substitution between goods produced in different regions—often lie within the range of 4 and 9.[67] I set $\sigma = 8$ based on recent work of Vietnam (Balboni, 2024) and assess

67. For example, Donaldson and Hornbeck (2016) estimates a value of 9.22 in the U.S. during the late 19th century; Peters (2022) estimates 5.02 in post-war Germany; and Balboni (2024) estimates 7.92 in Vietnam in 2009.

the robustness of the results to a set of other values.[68]

Amenity Spillovers

The parameters of amenity spillovers, β and β^e, affect people's migration choices. The value of residing in region n (1.11) can be rewritten in terms of the migration market access (1.14) as

$$\ln \bar{w}_n^e = (-\beta + \beta^e) \ln L_n + \left(\frac{1}{\nu} - \beta^e\right) \ln L_n^e + \frac{1}{\nu} \ln (\mathcal{V}_n^e)^\nu + \left(\frac{\alpha}{\sigma - 1}\right) \ln P_{nA}^{\sigma-1} + \left(\frac{1-\alpha}{\sigma - 1}\right) \ln P_{nM}^{\sigma-1} - \underbrace{\frac{1}{\nu} \ln \bar{a}_n^e}_{\text{error term}} . \tag{1.31}$$

To gain more insights of the within-group amenity spillover β^e, we can again write the log Chinese wage premium as

$$\ln\left(\frac{\bar{w}_n^c}{\bar{w}_n^m}\right) = \left(\frac{1}{\nu} - \beta^e\right) \ln\left(\frac{L_n^c}{L_n^m}\right) + \frac{1}{\nu} \ln\left(\frac{(\mathcal{V}_n^c)^\nu}{(\mathcal{V}_n^m)^\nu}\right) - \underbrace{\frac{1}{\nu} \ln\left(\frac{\bar{a}_n^c}{\bar{a}_n^m}\right)}_{\text{error term}}. \tag{1.32}$$

This expression is a relative (inverse) labor supply curve across space, where the first term shows that the neoclassical force, disciplined by $1/\nu$, tends to predicts a upward sloping supply when the amenity spillover term β^e is not too strong. Intuitively, if β^e is positive and large, a higher Chinese share is an attractive amenity to Chinese that makes them willing to accept a lower wage. The inward migration market access term $\mathcal{V}_n^e$ captures the potential migrants of group e from other counties, which is a shifter of the labor supply. The error term captures all unobserved characteristics that make a county more attractive to one group

68. In theory, equation (1.28) could be used to estimate σ (and θ) if one can exogenously shift the relevant independent variable (e.g., the market access term $\mathcal{P}_{nk}^{\sigma-1}$) using an instrument that is uncorrelated with location fundamentals. In practice, however, my cross-sectional resettlement instruments lack enough variation for this purpose. This is unsurprising given the high spatial correlation typically observed in right-hand-side variables like trade market access and local income.

versus the other. This means that OLS estimate of β^e tends to be biased upward.

To address the endogeneity issue, I again exploit population shifters Z_n^r for identification, as the resettlement program did not target places with varying amenity fundamentals. One concern that remains is that even though resettlement was plausibly independent of the pre-existing amenities, it might change local amenities ex-post by making it more or less attractive for reasons not related to population distribution, and hence not captured by β and β^e. For example, if the British built more schools per unit area in places with a higher resettlement density, the better school access would show up in the error term, violating the exogeneity assumption. In light of this, I also employ productivity shifters as the model-implied instruments for local population that are plausibly uncorrelated with local amenity. I use measures of agricultural suitability from FAO, such as suitability of padi rice (Z_n^p), coconut (Z_n^c), and palm oil (Z_n^l)—major crops in Malaysia during the studied period—as the instruments. The moment conditions are:

$$\mathbb{E}[Z_n \ln \bar{a}_n^e] = 0, \ \forall e, \ Z_n \in \{Z_n^r, Z_n^p, Z_n^c, Z_n^l\}. \tag{1.33}$$

I require the migration elasticity ν to identify both β^e and β. Estimates of this parameter are rare in the literature, particularly in the developing countries. Existing studies generally estimate a value between 2 and 4 (Monte et al., 2018; Morten and Oliveira, 2018; Bryan and Morten, 2019; Tombe and Zhu, 2019). I assume $\nu = 3$ as my baseline and consider robustness of the results to alternative values between 2 and 4.[69]

69. Again, I could potentially use equation (1.32) to estimate the migration elasticity. However, my instruments do not provide enough independent variations for both terms on the right, as the value of residing in a place for group-e is highly correlated with its population distribution.

Recovery of Location Fundamentals

The exogenous location fundamentals can be recovered as the residuals from equations (1.28) and (1.31). Specifically, I recover ϕ^e_{nk} as the residuals of (1.28) after estimating γ_k and γ^e. Notice that the average of ϕ^e_{nk} within sector k and group e across regions captures the absolute advantage of ethnicity e in sector k. Any deviation from that average across regions reflects idiosyncratic reasons that make group e more productive in a sector-region pair.

Similarly, I recover $\bar{a}^e_n$ as the residuals of (1.31), which captures both the location amenities that are neutral to all ethnicities and features of a place that are particularly attractive to a specific ethnic group.

1.7.2 Estimation

In this section, I discuss my estimation procedure and the results. I also discuss how my estimates relate to the literature.

Estimation Procedure

The estimation proceeds as follows. I first estimate the migration cost elasticity with respect to distance, $\tilde{\kappa}$, and use it to calculate the migration cost matrix $\eta_{nr}^{-\nu}$. Next, I iteratively solve for the market access terms and agricultural expenditure share α, based on Proposition 1. Then, I estimate the shape parameter of Fréchet skills θ by targeting the average of the normalized variance of wages within a (n, k, e) cell, weighted by the number of population in the cell. Lastly, I estimate $\{\gamma_A, \gamma_M, \gamma^e, \beta, \beta^e\}$ using a generalized method of moments (GMM) estimator based on the moment conditions of equations (1.30) and (1.33).

In implementing the procedure, I weight the estimations by the number of households to mitigate small sample biases as in section 1.5. I bootstrap this entire procedure to get the standard errors of the parameter estimates, where, in each bootstrap, I randomly sample with replacement individuals from the census microdata by district, and aggregate the outcomes to the county level.[70]

Estimation Results

This section presents my parameter estimates and discusses how they relate to the literature. Table 1.8 documents the parameter estimates.

Distance elasticity of migration costs Migration costs rise in distance with an elasticity κ of 0.52. This estimate falls within the range of existing estimates from the literature. For example, Bryan and Morten (2019) estimates around 0.37 in Indonesia between 1995 and 2012, and Peters (2022) estimates 1.09 in post-war Germany in 1955.[71]

Skill dispersion. The dispersion of productivity draws, governed by the shape parameter θ, determines the scope of selection into sectors based on individuals' comparative advantage, with a larger θ corresponding to a smaller dispersion. My estimate of 3.79 falls within the range of existing works. Lagakos and Waugh (2013) estimates a value of 5.3 for agriculture and 2.7 for the non-agricultural sector in the U.S. from 1996-2010. Hsieh et al. (2019) estimates a value between 1.5 and 2.6 in the U.S. from 1960-2012.[72]

Productivity spillovers. I estimate that the size of local employment in the non-agricultural sector increases labor productivity of that sector with an elasticity of $\gamma_M = 0.55$, whereas the elasticity for the agricultural sector is much smaller, with $\gamma_A = 0.23$. For the non-agricultural

70. Administrative districts are larger than the counties. There are 66 grouped districts with consistent boundary over 1957-1980.

71. As Bryan and Morten (2019) estimates the bilateral migration costs non-parametrically instead of assuming them to be proportional to distance, I translate the Figure 3 of their paper into my setting, with $1 - \eta_{nr}^{-1} \approx -0.5 + 0.147 \ln d_{nr}$. This implies that their distance elasticity of migration costs varies with distance, as opposed to a constant elasticity in my case. For the comparison exercise, I take the average log distance of 7.5 of their setting to arrive at $\partial \ln \eta_{nr} / \partial \ln d_{nr} \approx 0.37$.

72. One reason why their estimates might be smaller is that the variance of wages in their model can also be attributed to variations in (endogenous) educational attainment, and not solely come from idiosyncratic productivity draws.

sector, my estimate is larger than the values of 0.2 estimated by Kline and Moretti (2014) but smaller than the estimates obtained by Greenstone et al. (2010), in the range 1.25-3.1 .[73] Estimates of productivity spillover in the agricultural sector are rare but my estimate of a smaller elasticity aligns with the conventional wisdom that agglomeration forces in agriculture is more limited than the industrial sector.[74]

I estimate a sizable productivity spillover elasticity with respect to ethnic composition, with $\gamma^e = 0.31$. This suggests that holding fixed the total population size of a county, increasing the Chinese employment share would enhance the productivity of local Chinese workers. The effect on Malay workers is more nuanced and depends on the sector. Since $\gamma^e < \gamma_M$, equation (1.8) shows that Malay workers in the non-agricultural sector would still benefit from a rise of Chinese population. But since $\gamma^e > \gamma_A$, increased Chinese population can reduce the agricultural productivity of Malay workers. These predictions are consistent with my empirical findings that Malays working in the non-agricultural sector have higher income in the more resettled counties, while those working in agriculture did not see significant changes in in income.

Although my estimated composition-dependent spillover based on ethnicity lacks direct comparisons in existing literature, prior studies have examined similar externality based on other demographic characteristics such as education and occupations. For instance, Moretti (2004a) estimates wage elasticities of 0.14 and 0.21 for college and high school graduates, respectively, with respect to the share of college graduates in a city.[75] Rossi-Hansberg et al.

73. See a discussion in Kline and Moretti (2014).

74. A large literature estimates the overall productivity elasticity with respect to density though mainly focused on the more developed countries. They generally find a value between 0.02 and 0.09. Estimates for the developing countries are scarce, but tend to be larger than 0.1. See Combes and Gobillon (2015) and Melo et al. (2009) for a review.

75. Moretti (2004a) finds that a 1 percentage point increase in the share of college educated workers in a city leads to a 1.3% increase in wages. I convert this to elasticity by taking 0.25 as the average college share in 1990. Diamond (2016) obtains larger estimates, with a wage elasticity with respect to college share of 0.31 for college graduates and 0.93 for non-college workers. Although her estimates include the substitution effect across high- and low-skilled workers.

(2023) estimates the wage elasticities with respect to the share of "cognitive non-routine" occupations, and finds substantial elasticities of 1.3 for workers in cognitive occupations and 0.84 for those in non-cognitive occupations.[76]

Having estimated $\{\gamma_A, \gamma_M, \gamma^e\}$, I calculate the marginal spillover of a group-e worker on the efficiency of group-e' worker for each county n and industry k, denoted by $\gamma_{nk}^{e,e'}$, using equations (1.7) and (1.8).[77] Figure 1.6 illustrates their distributions across counties. Within-group ($e = e'$) spillovers are generally stronger than cross-group ($e \neq e'$) spillovers and the marginal spillovers tend to be larger than the values obtained by Fajgelbaum and Gaubert (2020) when examining within/cross-group spillovers among college and non-college graduates.[78]

Amenity spillovers. My baseline estimate of the amenity spillover elasticity with respect to local population size is $\beta = -0.05$.[79] The small estimate suggests that congestion forces—any disamenities due to density, such as a greater traffic or higher housing price—is weak. As discussed in Bryan and Morten (2019), an extension of the model to include housing as a non-traded good imply that my current endogenous amenity spillover with respect to population size can be decomposed as $L_n^{\beta} = L_n^{\beta_a - \delta\beta_r}$, where β_a is the pure amenity spillover

76. I assume perfect substitutability between Chinese workers and non-Chinese workers. If they are in fact imperfect substitutes, which seems plausible, the true γ^e should be even higher. The reason is that neoclassical forces would predict that an influx of Chinese would lead to higher wage gains among non-Chinese workers compared to Chinese workers due to the complementarity between groups. Hence, a stronger within-group productivity spillover would be required to justify the minimal wage increase among non-Chinese workers in the data.

77. By the Envelop Theorem, marginal spillovers evaluated at equilibrium outcomes only entail the direct effect (first term of (1.7) and (1.8)), as the indirect/general equilibrium effect would be zero. Intuitively, at equilibrium wages and prices, a marginal population increase doesn't change relative wages across industries and locations, thus occupation and migration choices remain unchanged.

78. They calibrate four constant elasticities using the estimates from Diamond (2016): $(\gamma_{UU}^P, \gamma_{SU}^P, \gamma_{US}^P, \gamma_{SS}^P)$, where γ_{SU}^P denotes the marginal productivity spillover of a college graduate (S) on the efficiency of a non-college graduate (U), and so forth. They obtain $(\gamma_{UU}^P, \gamma_{SU}^P, \gamma_{US}^P, \gamma_{SS}^P) = (0.003, 0.044, 0.020, 0.053)$.

79. My baseline estimate uses county resettlement density as the instrument. Using the FAO agricultural suitability indices as alternative instruments, which do not have the potential endogeneity concern discussed in section 1.7.1, I estimate $\beta = 0.09$ and $\beta^e = 0.19$.

capturing the preference for living in a more populated county and β_r is the inverse of housing supply elasticity, with δ being the share of income on housing. Using the Malaysian Family Life Survey in 1989, I estimate β_r to be around 0.3 using the resettlement shocks as demand shifter (Appendix Table (?)), which corresponds to an elasticity of 3.3 (1/0.3). This is a relatively high elasticity compared to estimates in the U.S, which are mostly between 1 and 3 (Gyourko et al., 2008; Saiz, 2010).[80] In addition, housing expenditure share δ in 1980 was 17.6% (Department of Statistics Malaysia, 1980).[81] Together, these imply that pure amenity spillover is around $\beta_a = \beta + \delta\beta_r = 0.003$.

There are few estimates of the (net) amenity spillover (β) in low-income countries that I can use for comparison with my estimate. Bryan and Morten (2019) estimates a value of 0.04, albeit with limited statistical power. Allen and Donaldson (2022) uses historical data from the U.S. spanning 1800 to 2000 to estimate contemporaneous and historical amenity spillovers jointly. They find a value of -0.26 for contemporaneous spillover and 0.31 for historical spillover. Their historical amenity spillover is generated from population 50 years ago. Considering that my two-period model on a 24-year time scale does not distinguish between contemporaneous and historical spillovers, it is reasonable that my estimate falls between the values of their contemporaneous and historical spillovers.

My baseline estimate of the amenity spillover elasticity with respect to ethnic composition is $\beta^e = 0.05$. Similar to the productivity spillover, I calculate the marginal amenity spillover of a group-e individual on the utility of group-e' individuals for each county n, denoted by $\beta_n^{e,e'}$, which varies with population composition across counties. Figure 1.7 shows the distribution of $\beta_n^{e,e'}$. Given that β^e is positive, an increase in group-e's population would increase the utility of people in group-e more than those in group $e' \neq e$. The stronger within-group amenity spillover is consistent with specialization and the economies of scale

80. In Indonesia, Bryan and Morten (2019) estimates a value of 4 (although underpowered).

81. The expenditure category is "gross rent, fuel and power". The same expenditure share in 1973 was 14.9%.

in the provision of urban amenities, such as restaurants or entertainment, as discussed in Duranton and Puga (2004). It is also in line with the presence of social frictions, as indicated by consumption segregation documented in Davis et al. (2019).

There are no existing estimates in the literature that my estimates of $\beta_n^{e,e'}$ can be directly compared to. The closest paper, Fajgelbaum and Gaubert (2020), considers the amenity spillover with respect to college share estimated by Diamond (2016) and obtains marginal elasticities ranging from -1.24 to 0.77. Notably, their numbers also indicate that within-group amenity spillovers tend to be more positive than cross-group spillovers.[82]

1.8 Conclusion

This paper studies the agglomeration economies resulting from a large-scale, forced relocation of rural Chinese to compact villages in 1950s British Malaya. This ethnicity-based resettlement substantially altered both population size and ethnic composition across Malaysia's landscape. Leveraging the resettlement program, I employ a design-based identification strategy and find that areas with higher resettlement densities experienced an increase in industrial activities and a higher share of non-agricultural employment, alongside higher income and education levels. Additionally, these areas attracted internal migration from other regions.

The observed higher income among Chinese households in the more resettled areas, along with limited income differences for other groups, is consistent with social frictions across ethnic lines hindering productive interactions that underlie the agglomeration forces discussed in Marshall (1890), such as labor market pooling, input-output linkages, and knowledge spillovers. Meanwhile, other ethnic groups primarily benefit from switching into the non-agricultural sector, which exhibits a larger external economy of scale than agriculture.

The finding that within-group productivity spillovers surpass cross-group spillovers

82. Similar to marginal productivity spillovers, the authors calibrate four constant amenity spillover elasticities: $(\gamma_{UU}^A, \gamma_{SU}^A, \gamma_{US}^A, \gamma_{SS}^A) = (-0.43, 0.18, -1.24, 0.77)$, where γ_{SU}^A denotes the marginal amenity spillover of a college graduate (S) on the utility of a non-college graduate (U), and so forth.

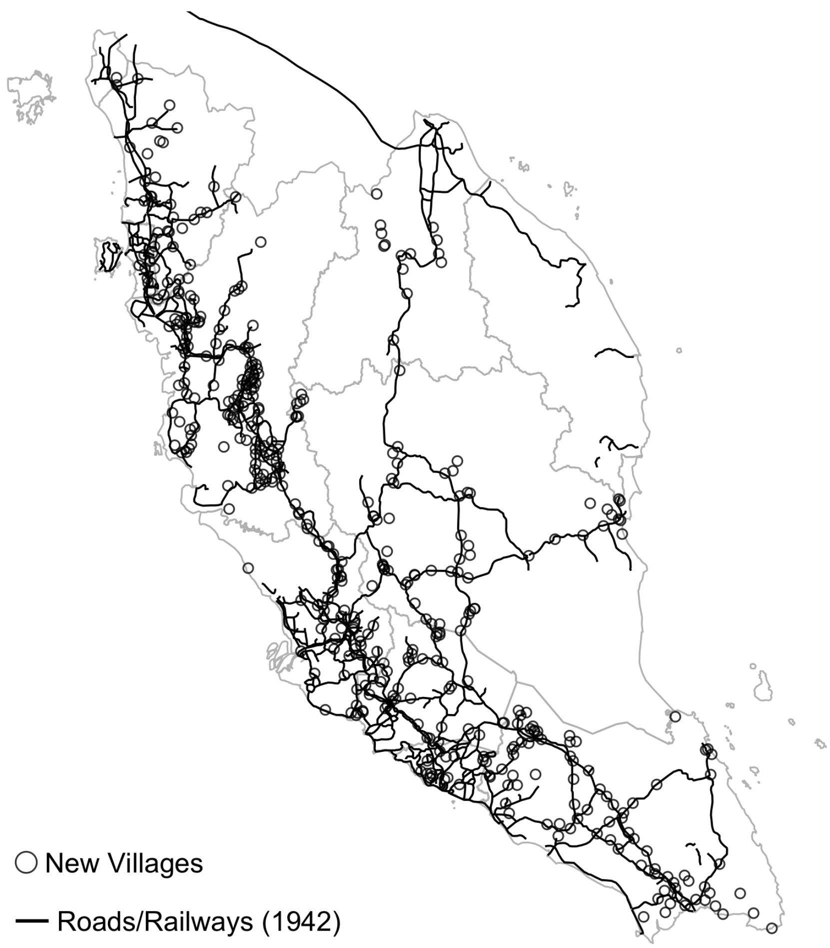
underscores a diversity trade-off. While higher diversity permit greater gains from idea exchange, excessive diversity can lead to tension and impede productive social interactions. This trade-off may also vary with development stage; in technologically backward economies like Malaysia during the studied period, lower diversity may be more crucial for productivity enhancement through channels of technology adoption. In contrast, closer to the technology frontier where innovation is presumably more important, diversity could prove more economically advantageous. Further empirical evidence along this line would be valuable.

While this paper focuses on ethnic composition as a source of heterogeneity in agglomeration forces, other demographic characteristics, such as religion and education, could also exhibit differential agglomeration spillovers within versus across groups. Considering the spatial variations in these demographic attributes, agglomeration elasticities are likely to vary spatially, implying that spatial or place-based policies could potentially enhance welfare.

The larger external economies of scale in the non-agricultural sector compared to agriculture suggest that reallocating resources from the latter to the former through industrial policies can increase aggregate output. Prominent examples include the growth trajectory of the East Asian Tigers and a few Southeast Asian countries, including Malaysia, during the postwar era. The heterogeneity in external economies across sectors may also apply to more detailed industry classifications. Further research into mechanisms through which size influences productivity, particularly why these mechanisms might differ by sector, is crucial for policy formulation. This includes understanding the relative contributions of firm-level increasing returns versus external economies at both the sector and location levels.

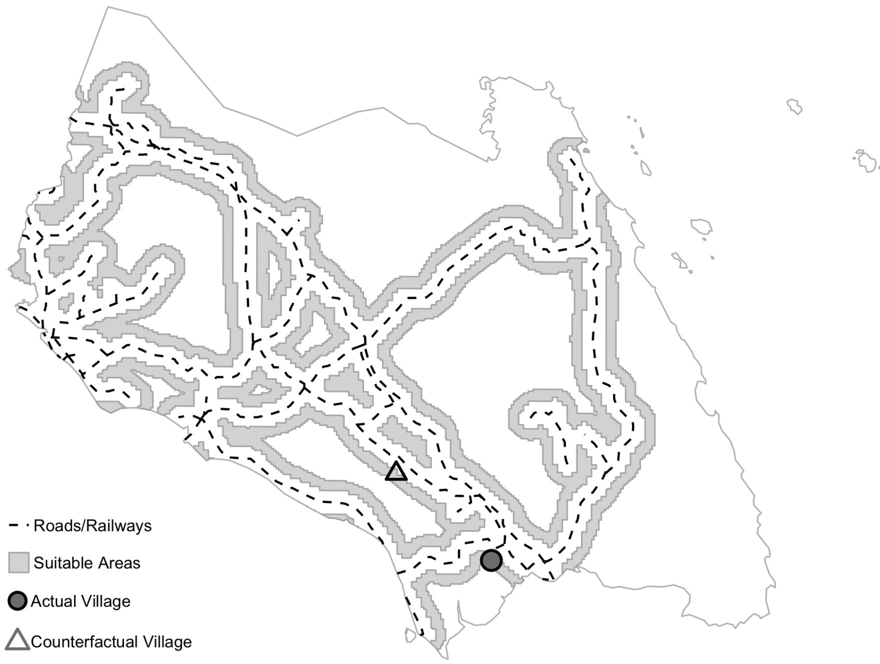
1.A Figures

Figure 1.1. The New Villages and Transportation Network



Notes: This figure shows the location of the New Villages (round circles) and the roads and railways in 1942 (line). The gray polygons indicate state boundaries. Data on the New Villages are from the Corry report. Data on roads and railways from U.S. Office of Strategic Services (1942).

Figure 1.2. Permutation Procedure: Counterfactual Site Selection



Notes: This figure illustrates the permutation procedure for counterfactual site selections. The black dot is one actual New Village in State Johor. Dashed lines are the roads and railroad network. Gray areas have equal distance to the depicted actual village and are equally suitable for resettlement. The triangle shows a counterfactual village location randomly and uniformly drawn from the suitable areas.

Figure 1.3. County Resettlement Density, Expected and Residualized

Panel A. County Resettlement Density, Expected

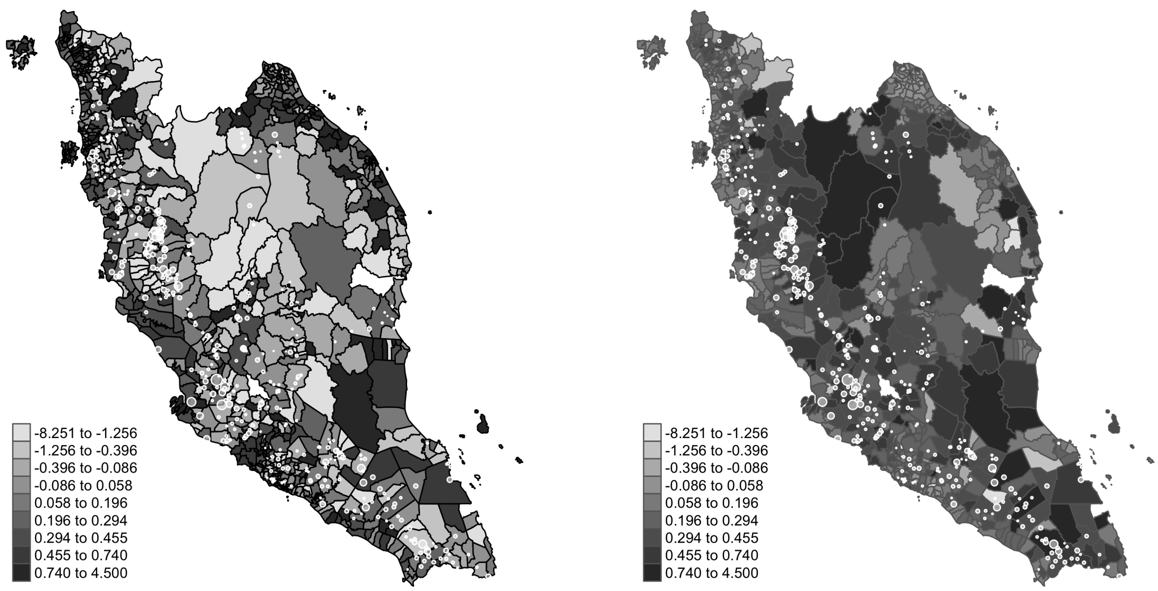
Panel B. County Resettlement Density, Residualized

Notes: This figure shows the expected county resettlement density and the residualized county resettlement density, with darker shades corresponding to greater resettlement density deciles. The white bubbles denote the New Villages, with their sizes proportional to the resettled population. The sample is restricted to the 249 counties with any New Village, which are where the identifying variations are drawn from. Panel A shows the calculated expected resettlement density, calculated with equation (1.4). Panel B shows the residualized resettlement density after partialling out state fixed effects, the expected resettlement density, and the baseline controls: an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944. Data on the resettlement density from the Corry report.

Figure 1.4. County Population Growth from 1947 to 1957, by Ethnic Group

Panel A. County Population Growth, Chinese

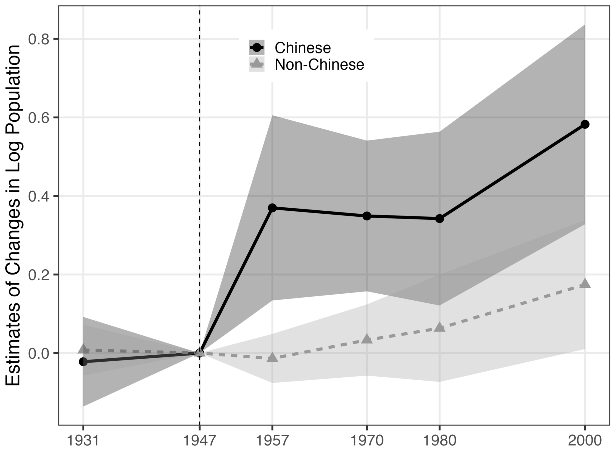
Panel B. County Population Growth, Non-Chinese



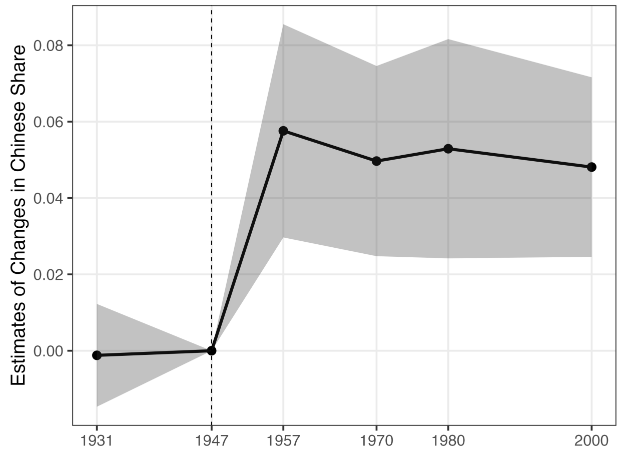
Notes: This figure shows county population growth from 1947 to 1957, by ethnic group. Panel A shows the log changes of Chinese population. Panel B shows the log changes of non-Chinese population. The white bubbles denote the New Villages, which are sized in proportion to the log resettled population in that village. Counties with missing population are shaded in white. Data from the tabulated Census of Population and the Corry report.

Figure 1.5. Changes in Population Distribution from 1931 to 2000, by County Resettlement Density

Panel A. Population Growth

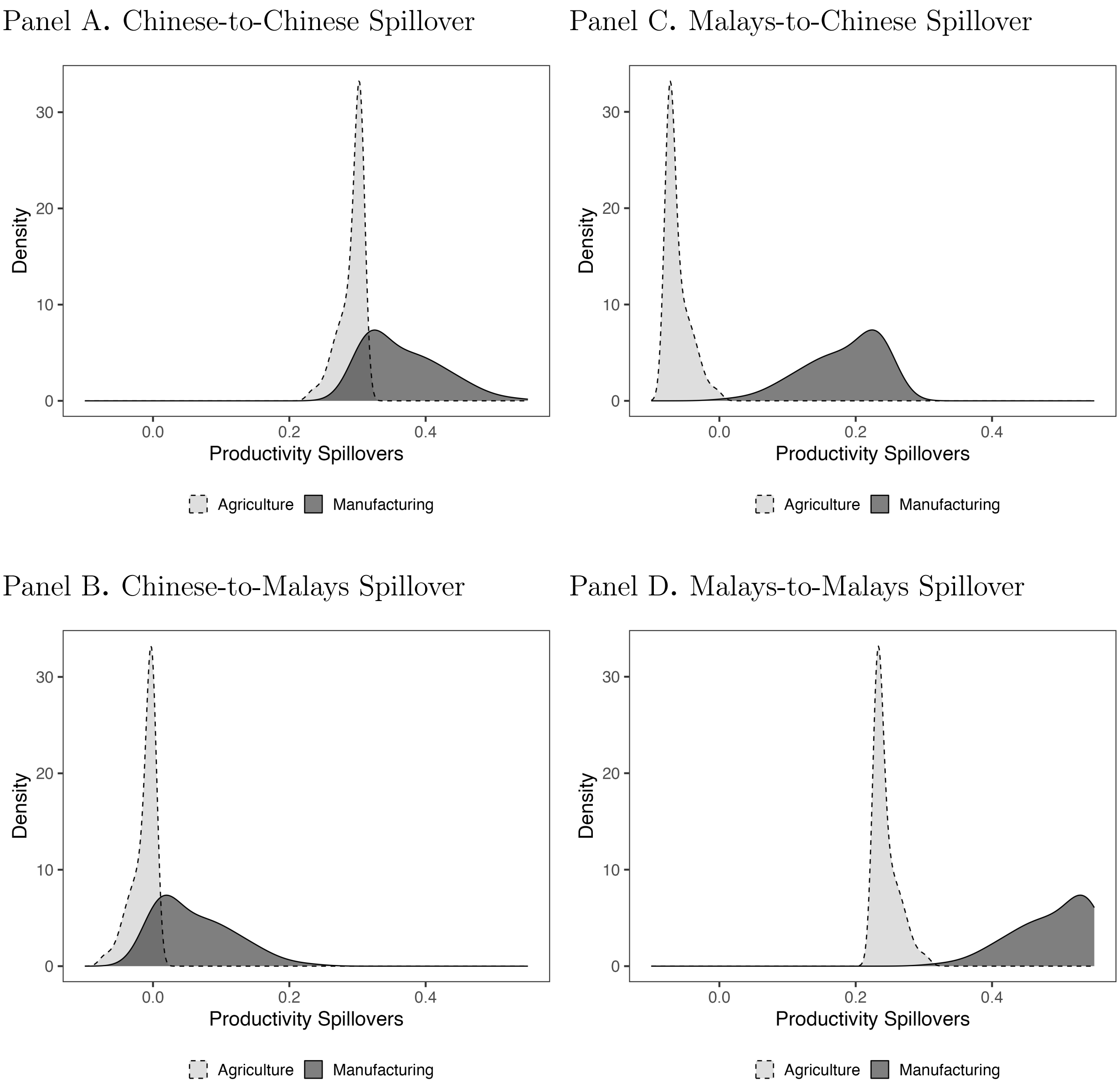


Panel B. Changes in Chinese Share



Notes: Regressions control for the expected resettlement density, whether a county has any resettlement, (log) county area, (log) distance to nearest road, road density, distance to nearest rail station, distance to coastline, 1947 Chinese population share and population density, and the land shares of rubber and mining. The shaded region reflects the 95% confidence interval under Conley standard errors, with a distance cutoff of 30 kilometers.

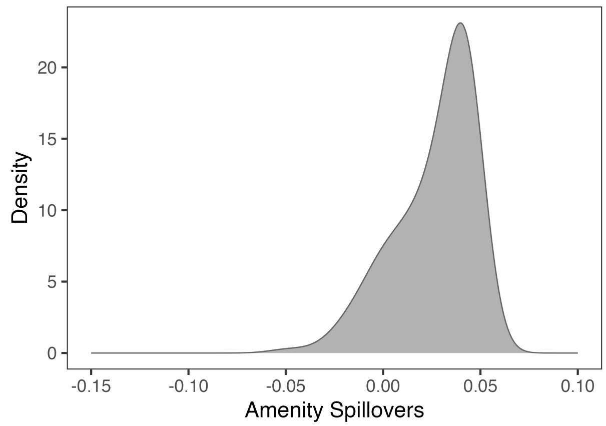
Figure 1.6. Distribution of Marginal Productivity Spillovers, by Ethnic Group



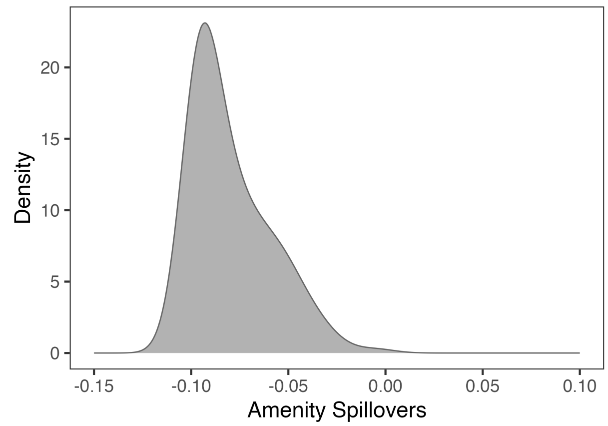
Notes: This figure shows the distribution of marginal productivity spillovers across counties, by pairwise combination of Chinese and Malays (non-Chinese). Panel A shows the elasticity of Chinese productivity with respect to local Chinese population; that is, the percent changes in Chinese productivity resulting from a one percent increase in the local Chinese population. Panel B shows the elasticity of Malays productivity with respect to local Chinese population. Panel C shows the elasticity of Chinese productivity with respect to local Malays population. Panel D shows the elasticity of Malays with respect to local Malays population. These elasticities are calculated from equations (1.7) and (1.8), holding fixed occupational shares.

Figure 1.7. Distribution of Marginal Amenity Spillovers, by Ethnic Group

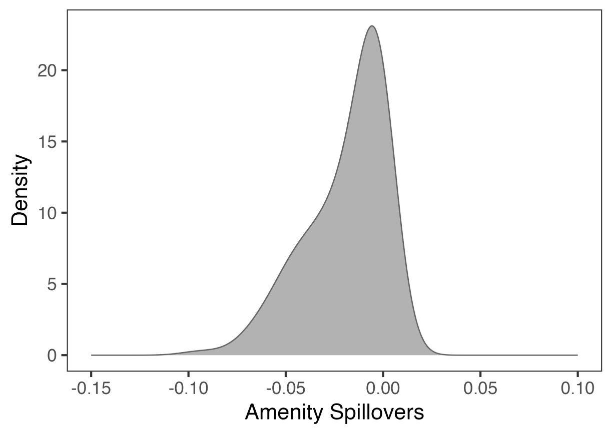
Panel A. Chinese-to-Chinese Spillover



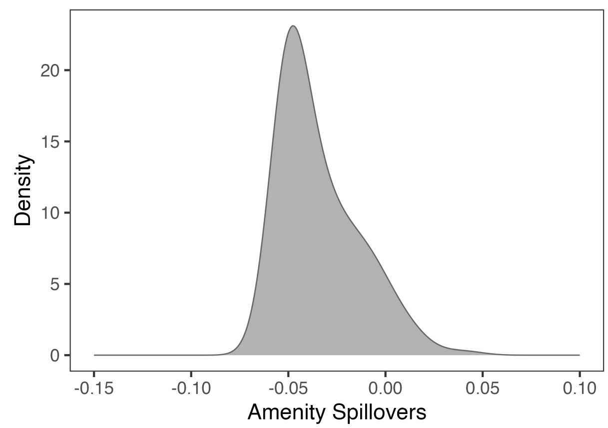
Panel C. Malays-to-Chinese Spillover



Panel B. Chinese-to-Malays Spillover



Panel D. Malays-to-Malays Spillover



Notes: This figure shows the distribution of marginal amenity spillovers across counties, by pairwise combination of Chinese and Malays (non-Chinese). Panel A shows the elasticity of Chinese utility with respect to local Chinese population; that is, the percent changes in Chinese utility resulting from a one percent increase in the local Chinese population. Panel B shows the elasticity of Malays utility with respect to local Chinese population. Panel C shows the elasticity of Chinese utility with respect to local Malays population. Panel D shows the elasticity of Malays with respect to local Malays population. These elasticities are calculated similarly as in equations (1.7) and (1.8) for the marginal productivity spillovers, holding fixed migration shares.

1.B Tables

Table 1.1. Balance of County Geography, Amenity, and Pre-resettlement Characteristics

	Geography				Amenity				Economic Activities				
	Elev. (1)	Rugged. (2)	Rice Suitab. (3)	Coconut Suitab. (4)	Dist. Police (5)	Dist. Post (6)	Dist. Hosp. (7)	Dist. Temple (8)	Log Pop. Density (9)	Land Use Rubber (10)	Land Use Mining (11)	Dist. Factory (12)	Dist. Cities (13)
Panel A. Within State													
ResettleDensity	-0.04	-9.84*	-0.04**	0.01	-0.71*	-0.94**	-1.75*	2.42	0.35***	0.08***	0.02	-2.60**	-4.67*
	(0.20)	(5.92)	(0.02)	(0.02)	(0.43)	(0.37)	(0.93)	(2.01)	(0.07)	(0.01)	(0.01)	(1.12)	(2.47)
Panel B: Resettlement Controls													
ResettleDensity	0.14	-8.55	-0.05*	-0.02	0.37	0.66	-0.04	2.72	0.10	0.02	0.02	-1.00	0.50
	(0.27)	(6.02)	(0.03)	(0.03)	(0.59)	(0.52)	(0.99)	(2.30)	(0.07)	(0.03)	(0.01)	(1.10)	(2.43)
Mean	0.94	62.77	1.21	1.12	9.35	11.33	23.50	66.13	4.06	0.24	0.01	26.23	87.69
SD	1.51	74.19	0.23	0.21	8.19	8.68	19.45	47.40	1.58	0.30	0.07	18.08	69.82
# Counties	777	777	777	777	777	777	777	777	777	777	777	777	777

Notes: This table shows the relationship between per-period county characteristics and county resettlement density. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. Columns 1-4 report the effect of resettlement density on measures of productive geography: elevation (Column 1); ruggedness (Column 2); suitability for padi rice cultivation (Column 3); and suitability for coconut cultivation (Column 4). Columns 5-8 report the effect on measures of amenity measured in 1945: distance to nearest police station (Column 5); distance to nearest post or telegraph office (Column 6); distance to nearest hospital (Column 7); and distance to nearest Chinese temple (Column 8). Columns 9-13 report the effect on economic activities prior to the resettlement: (log) population density in 1947 (Column 9); land use share of rubber in 1943 (Column 10); land use share of mining in 1943 (Column 11); distance to nearest industrial facility in 1945 (Column 12); and distance to nearest major cities, i.e., Singapore, George Town, Malacca, Ipoh, and Kuala Lumpur. Panel A reports regression estimates controlling only for state fixed effects, an indicator for any resettlement in the county, and (log) county area. Panel B additionally controls for the expected resettlement density; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; and Chinese population share of the county in 1947. The unit of observation is the county. Data on drawn from various sources: elevation from SRTM; ruggedness from Nunn and Puga (2012); rice and coconut suitability from FAO GAEZ v4; locations of police station, post/telegraph office, hospital, and temples from HIND 1076 topographical maps (Survey of India Offices (P.Z.O.), 1944); population density from the 1947 census of population; land use share of rubber and mining from GSGS 4474 land utilization maps; and locations of prewar industrial facilities from the US national archive, RG226 (U.S. Office of Strategic Services, 1944). Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: *p<0.1; **p<0.05; ***p<0.01.

Table 1.2. Post-Resettlement Population Distribution, by County Resettlement Density

	Log County Population			Chinese Share of County Population			Log Built-up Volume
	1957	1980	2000	1957	1980	2000	1990
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ResettleDensity	0.094***	0.108*	0.177**	0.048***	0.050***	0.041***	0.264***
	(0.034)	(0.062)	(0.075)	(0.012)	(0.011)	(0.011)	(0.087)
# Counties	777	777	777	777	777	777	777

Notes: This table shows the relationship between measures of population distribution and county resettlement density. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. Columns 1-3 report the effect of resettlement density on log county population in 1957 (Column 1), 1980 (Column 2), and 2000 (Column 3). Columns 4-6 report the effect of resettlement density on the Chinese share of county population in 1957 (Column 4), 1980 (Column 5), and 2000 (Column 6). Column 7 reports the effect on log built-up volume in 1990. All regressions are estimated using OLS and include state fixed effects, the expected resettlement density, and the baseline controls: an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944. The unit of observation is the county. Columns 1-6 use data from the tabulated Census of Population. Column 7 uses data from the Global Human Settlement Layer (GHSL) project. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: *p<0.1; **p<0.05; ***p<0.01.

Table 1.3. Sectoral Employment, by County Resettlement Density

	Total Employment by Industry:				
	Primary Sector (1)	Manuf. (2)	Utility and Construction (3)	Trade and Transport (4)	Services (5)
ResettleDensity	0.101***	0.258*	0.327**	0.216*	0.209
	(0.034)	(0.148)	(0.136)	(0.131)	(0.147)
# Counties	777	777	777	777	777

Notes: This table shows the relationship between sectoral employment and county resettlement density. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. Each column shows the effect of resettlement on total employment in a different industry sector: the primary sector, comprised of agriculture, hunting, forestry, fishing, mining, and quarrying (Column 1); manufacturing (Column 2); utility and construction (Column 3); wholesale/retail trade, transport, and communication (Column 4); finance, business and other services (Column 5). All regressions are estimated using the Poisson pseudo-maximum-likelihood (PPML) estimator and include state fixed effects, the expected resettlement density, and the baseline controls: an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944. The unit of observation is the county. Data from the tabulated Census of Population in 1980. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: *p<0.1; **p<0.05; ***p<0.01.

Table 1.4. Sectoral Employment Share, by County Resettlement Density

	Chinese Individuals (1)	Non-Chinese Individuals (2)	Difference (1) − (2) (3)
Panel A. Primary Sector			
ResettleDensity	0.031	-0.141**	0.172***
	(0.081)	(0.072)	(0.053)
Panel B. Manufacturing			
ResettleDensity	0.128*	-0.006	0.134*
	(0.070)	(0.114)	(0.081)
Panel C. Utility and Construction			
ResettleDensity	-0.103	0.255**	-0.357***
	(0.086)	(0.107)	(0.104)
Panel D. Trade and Transport			
ResettleDensity	0.016	0.160***	-0.145**
	(0.046)	(0.058)	(0.057)
Panel E. Services			
ResettleDensity	0.160*	0.024	0.136**
	(0.083)	(0.069)	(0.065)
# Individuals	21,086	38,819	
# Counties	515	744	

Notes: This table shows the relationship between sectoral employment share and county resettlement density, separately for Chinese and non-Chinese individuals. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. Each panel shows the effect of resettlement density on the probability that a person is employed in a different sector: the primary sector, comprised of agriculture, hunting, forestry, fishing, mining, and quarrying (Panel A); manufacturing (Panel B); utility and construction (Panel C); wholesale/retail trade, transport, and communication (Panel D); finance, business and other services (Panel E). Column 1 reports estimates restricting the sample to Chinese individuals only. Column 2 restricts the sample to non-Chinese individuals only. Column 3 reports the difference between the estimates in Columns 1 and 2. All regressions are estimated using the Poisson pseudo-maximum-likelihood (PPML) estimator and include state fixed effects, the expected resettlement density, and the baseline controls: an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944. The unit of observation is the individual. The sample is restricted to individuals aged 15 or above reporting employed in an industry. Data from the 2% individual-level Census of Population microdata in 1980. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: *p<0.1; **p<0.05; ***p<0.01.

Table 1.5. Manufacturing Activity, by County Resettlement Density

	All (1)	Chinese Owned (2)	Non-Chinese Owned (3)
Panel A. Number of Establishments			
ResettleDensity	13.319*	9.476*	3.843*
	(7.565)	(5.386)	(2.319)
Panel B. Log Establishments (PPML)			
ResettleDensity	0.138	0.139	0.123
	(0.121)	(0.103)	(0.181)
Panel C. Any Establishment			
ResettleDensity	0.010	0.018	0.083***
	(0.027)	(0.030)	(0.024)
# Counties	777	777	777

Notes: This table shows the relationship between measures of manufacturing activity and county resettlement density. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. Each cell corresponds to a regression. Panel A, Column 1 shows the effect of resettlement density on an indicator of whether the county has any manufacturing establishment. Column 2 (or 3) shows the effect on whether the county has any Chinese-owned (or non-Chinese owned) manufacturing establishment. Panel B reports the effect of resettlement density on the total number of establishments (Column 1) and the number of Chinese-owned establishment (Column 2) and non-Chinese owned establishments (Column 3). Panel C shows the effect of the same outcome as B but is estimated with the Poisson pseudo-maximum-likelihood (PPML) estimator, whereas Panels A and B are estimated with OLS. All regressions include state fixed effects, the expected resettlement density, and the baseline controls: an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944. The unit of observation is the county. Data from the Directory of Manufacturing in 1970. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: *p<0.1; **p<0.05; ***p<0.01.

Table 1.6. Human Capital, by County Resettlement Density

	All Individuals (1)	Chinese Individuals (2)	Non-Chinese Individuals (3)	Difference (2) − (3) (4)
Panel A. Years of Schooling				
ResettleDensity	0.180	0.416*	0.098	0.317**
	(0.157)	(0.224)	(0.123)	(0.150)
# Individuals	88,852	31,507	57,345	
# Counties	752	522	745	
Panel B. Primary Education				
ResettleDensity	0.044	0.069**	0.029	0.040**
	(0.027)	(0.032)	(0.024)	(0.016)
# Individuals	88,852	31,507	57,345	
# Counties	752	522	745	
Panel C. Secondary Education				
ResettleDensity	0.072	0.177**	0.024	0.153***
	(0.057)	(0.079)	(0.048)	(0.051)
# Individuals	88,852	31,507	57,345	
# Counties	752	522	745	
Panel D. Log Birth Weight of First Child				
ResettleDensity	0.051***	0.078***	0.038*	0.040
	(0.015)	(0.029)	(0.020)	(0.030)
# Individuals	1,451	396	1,055	
# Counties	164	79	152	

Notes: This table shows the relationship between measures of human capital and county resettlement density. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. Each panel shows the effect of resettlement density on a different outcome of human capital: years of schooling (Panel A); completion of primary education (Panel B); completion of secondary education (Panel C); log birth weight of of the first child. Column 1 reports pooled estimates for Chinese and non-Chinese households. Column 2 restricts the sample to Chinese households only. Column 3 restricts the sample to non-Chinese households only. Column 4 reports the difference between the estimates in columns 2 and 3. Panels A to C are estimated using the Poisson pseudo-maximum-likelihood (PPML) estimator, and Panel D is estimated using OLS. All regressions include state fixed effects, the expected resettlement density, and the baseline controls: an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944. The unit of observation is the individual. For panels A to C, the sample is restricted to individuals aged 20 or above from the 2% individual-level Census of Population microdata in 1980. For panel D, the sample is restricted to females who report having at least one child from the Malaysian Family Survey in 1989. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: *p<0.1; **p<0.05; ***p<0.01.

Table 1.7. Household Income, by County Resettlement Density

	All Households (1)	Chinese Households (2)	Non-Chinese Households (3)	Difference (2) − (3) (4)
Panel A. Log Earnings				
ResettleDensity	0.069*	0.111**	0.037	0.074**
	(0.038)	(0.052)	(0.031)	(0.038)
# Households	33,328	10,622	22,706	
# Counties	713	495	705	
Panel B. Log Earnings, Primary Sector				
ResettleDensity	0.029	0.073**	-0.009	0.082*
	(0.037)	(0.036)	(0.040)	(0.044)
# Households	9,726	1,660	8,066	
# Counties	679	349	649	
Panel C. Log Earnings, Non-Primary Sector				
ResettleDensity	0.077*	0.121**	0.044	0.078**
	(0.039)	(0.052)	(0.030)	(0.033)
# Households	23,602	8,962	14,640	
# Counties	698	445	689	

Notes: This table shows the relationship between household income and county resettlement density. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. Panel A, Columns 1-3 show the effect of resettlement density on log household earnings predicted from asset ownership for all households (Column 1); Chinese households (Column 2); and non-Chinese households (Column 3). Column 4 reports the difference between the estimates in columns 2 and 3. Panel B restricts the sample to households whose head employ in the primary sector, comprised of agriculture and mining. Panel C restricts the sample to households whose head employ outside the primary sector. All regressions are estimated by OLS and include state fixed effects, the expected resettlement density, and the baseline controls: an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944. The unit of observation is the household. Data from the 2% individual-level Census of Population microdata in 1980. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: *p<0.1; **p<0.05; ***p<0.01.

Table 1.8. Parameter Estimates

Parameter (1)	Description (2)	Value (3)	SE (4)
Panel A. Estimated Parameters			
κ	Distance elasticity of migration costs	0.520	(0.003)
α	Expenditure share on agriculture	0.316	(0.001)
θ	Skill dispersion	3.786	(0.050)
γ_A	Productivity spillover w.r.t. size, agric.	0.230	(0.166)
γ_M	Productivity spillover w.r.t. size, manuf.	0.549	(0.037)
γ^e	Productivity spillover w.r.t. ethnic share	0.305	(0.119)
β	Amenity spillover w.r.t. size	-0.053	(0.063)
β^e	Amenity spillover w.r.t. ethnic share	0.045	(0.132)
Panel B. External Parameters			
σ	Elasticity of substitution	8.00	
ν	Migration elasticity	3.00	
ξ	Distance elasticity of trade costs	0.18	

Notes: Bootstrap standard errors reported in parentheses.

1.C Theoretical Results

1.C.1 Properties of Fréchet Distribution

Let $\{x_i\}_{i=1}^n$ be i.i.d. random variables of the Fréchet distribution with scale ϕ_i and shape θ, such that the c.d.f. is given by

$$F_{x_i}(x) = e^{-\phi_i x^{-\theta}}, \ \forall i = 1, 2, \dots, n,$$

where $\mathbb{E}[x_i] = \Gamma(1 - 1/\theta)\phi_i^{1/\theta}$. Going forward, let $\Gamma_\theta \equiv \Gamma(1 - 1/\theta)$.

One property of Fréchet distribution is that the maximum of n i.i.d. Fréchet random variables is also Fréchet distributed, with the same shape parameter, and its scale parameter is the sum of scales across x_i, $i = 1, \dots, n$. Let $\tilde{x}$ denote the maximum of $\{x_i\}_{i=1}^n$, we have

$$F_{\tilde{x}}(x) = e^{-(\sum_{i=1}^n \phi_i) x^{-\theta}}, \tag{1.34}$$

where $\mathbb{E}[\tilde{x}] = \Gamma_\theta \left(\sum_{i=1}^n \phi_i\right)^{1/\theta}$.

Furthermore, a key property of Fréchet distribution is that the above, unconditional expectation of $\tilde{x}$ is also the conditional expectation, conditional on any x_i being the maximum; that is,

$$\mathbb{E}[\tilde{x}] = \mathbb{E}[x_i | x_i = \max_j \{x_j\}], \ \forall i = 1, \dots, n. \tag{1.35}$$

Any multiplication of a Fréchet random variable with a constant also has a Fréchet distribution. To see this, let $y_i = wx_i$ and we can write the c.d.f. of y_i as

$$\begin{aligned} F_{y_i}(y) &= \mathbb{P}(y_i \leq y) \\ &= \mathbb{P}(x_i \leq y/w) \\ &= e^{-\left(\phi_i w^\theta\right) y^{-\theta}}, \end{aligned} \tag{1.36}$$

where the scale parameter of y_i is $\phi_i w^\theta$.

Another useful property of Fréchet distribution is that the probability of x_i being the maximum among all the n i.i.d. random Fréchet distributed random variables is simply the ratio of the scale of x_i to the sum of scales across n:

$$\mathbb{P}(x_i = \max_{i=1,\ldots,n} x_i) = \frac{\phi_i}{\sum_{i=1}^{n} \phi_i}. \tag{1.37}$$

1.C.2 Sectoral Labor Supply

I now derive the key equations pertaining to the sectoral labor supply. Individuals draw their efficiency units independently across sectors of agriculture and manufacturing $\Lambda^e = \left(\Lambda_A^e, \Lambda_M^e\right)$ from the joint distribution:

$$F_n^e(\Lambda_A, \Lambda_M) = \prod_{k=A,M} F_{nk}^e(\Lambda_k),$$

where the marginal probability distribution is Fréchet:

$$F_{nk}^e(\Lambda_k) = \exp\left(-\phi_{nk}^e \Lambda_k^{-\theta}\right).$$

After knowing their efficiency units, they choose the sector that pays higher earnings. Let w_{nk} be the wage per efficiency unit for industry k in region n. The earnings of individual i

of ethnicity e in industry k, location n is thus

$$\begin{aligned} y^e_{ink} &= w_{nk}\lambda^e_{ink} \\ &= w_{nk}\Lambda^e_{ink} f(L^c_{nk}, L^m_{nk}) \\ &= w^e_{nk}\Lambda^e_{ink}, \end{aligned}$$

where

$$w^e_{nk} \equiv w_{nk} f(L^c_{nk}, L^m_{nk}).$$

Function $f(L^c_{nk}, L^m_{nk})$, which depends on local population distribution, captures human capital externalities.

Since y^e_{ink} equals a constant w^e_{nk} multiplied by a Fréchet random variable Λ^e_{ink}, it is also Fréchet distributed with shape θ and scale $\phi^e_{nk}(w^e_{nk})^\theta$. The expected earning for ethnicity e in industry k and region n is thus $\Gamma_\theta\left(\phi^e_{nk}(w^e_{nk})^\theta\right)^{1/\theta}$.

For an individual of ethnicity e in region n, the probability of choosing to work in industry k is

$$\pi^e_{nk} \equiv \mathbb{P}(y^e_{ink} = \max_s y^e_{ins}) = \frac{\phi^e_{nk}(w^e_{nk})^\theta}{\sum_s \phi^e_{ns}(w^e_{ns})^\theta} = \phi^e_{nk}\left(\frac{w^e_{nk}}{\bar{w}^e_n}\right)^\theta,$$

where

$$\bar{w}^e_n \equiv \left(\phi^e_{nA}\left(w^e_{nA}\right)^\theta + \phi^e_{nM}\left(w^e_{nM}\right)^\theta\right)^{1/\theta}.$$

Since people of ethnicity e choose the sector that pays more and this process continues until the (e-specific) earning equalize across the two sectors, in equilibrium, the average wage for ethnic group e in region n is given by

$$\mathbb{E}[\max_k y^e_{ink}] = \Gamma_\theta \left(\sum_k \phi^e_k (w^e_{nk})^\theta \right)^{1/\theta} = \Gamma_\theta \bar{w}^e_n.$$

Moreover, due to a Fréchet property shown in Equation (1.35), ethnic group e in region n attain, on average, the same earning across the two sectors.

It follows that the average skill of group-e in region n, sector k, is given by

$$\mathbb{E}[\underbrace{y^e_{ink}/w^e_{nk}}_{\Lambda^e_{ink}} | y^e_{ink} = \max_s y^e_{ins}] f(L^c_{nk}, L^m_{nk}) = \Gamma_\theta \bar{w}^e_n w^{-1}_{nk}.$$

Notice that it can also be written in terms of occupation share as

$$\Gamma_\theta \left(\phi^e_{nk}\right)^{1/\theta} \left(\pi^e_{nk}\right)^{-1/\theta} f(L^c_{nk}, L^m_{nk}),$$

where the neoclassical force $(\pi^e_{nk})^{-1/\theta}$ implies that a higher share of labor supply tends to lower the average skill in the sector due to selection. In contrast, the externality term $f(L^c_{nk}, L^m_{nk})$ tends to increase the average skills in the number of population.

The aggregate sectoral earnings from ethnicity e in industry k and region n is the local population of ethnicity e multiplied by the share working in industry k and by their average sectoral earning conditional on choosing k:

$$w_{nk} H^e_{nk} = L^e_n \pi^e_{nk} \Gamma_\theta \bar{w}^e_n.$$

This implies that the aggregate human capital supply in industry k, region n is

$$\begin{aligned} H_{nk} &= \Gamma_\theta \sum_e L_n^e \phi_k^e (w_{nk}^e)^\theta w_{nk}^{-1} (\bar{w}_n^e)^{1-\theta} \\ &= \Gamma_\theta \sum_e L_n^e \phi_k^e w_{nk}^{-1} w_{nk}^e (w_{nk}^e)^{1-\theta} (\bar{w}_n^e)^{1-\theta} \\ &= \Gamma_\theta \sum_e L_n^e \phi_k^e \left(L_{nk}\right)^{\gamma_k} \left(\frac{L_{nk}^e}{L_{nk}}\right)^{\gamma^e} \left(\frac{w_{nk}^e}{\bar{w}_n^e}\right)^{\theta-1} . \end{aligned}$$

1.C.3 Migration

Individuals of group e draw an idiosyncratic taste shock for each location and decide where to migrate before knowing their efficiency units. The taste shock u_n^e is assumed to drawn from the following location-specific Fréchet distribution

$$F_n^e(a) = \exp\left(-\bar{a}_n^e a^{-\nu}\right),$$

where the scale $\bar{a}_n^e$ captures the average attractiveness of location n for group e and the shape ν captures the dispersion of taste (which is assumed to be the same for all groups and locations).

The value of relocating from r to n for ethnicity e is

$$V_{rn}^e = \eta_{rn}^{-1} a_n^e \Gamma_\theta \bar{w}_n^e P_n^{-1}$$

where η_{rn} is the migration cost and the amenity term a_n^e depends on the local population:

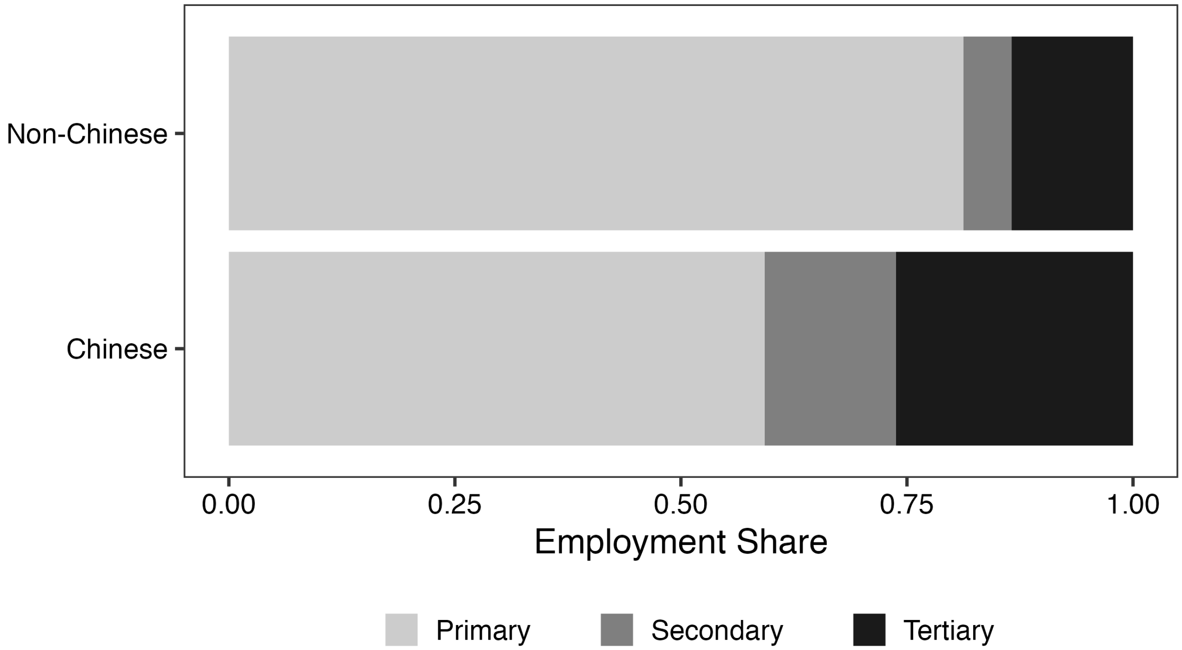
$$a_n^e = u_n^e \left(L_n\right)^\beta \left(\frac{L_n^e}{L_n}\right)^{\beta^e} .$$

As V_{rn}^e is a Fréchet random variable u_n^e multiplied by a constant $\eta_{rn}^{-1} L_n^\beta (L_n^e/L_n)^{\beta^e} \Gamma_\theta \bar{w}_n^e P_n^{-1}$, it is itself Fréchet distributed. The distribution of V_{rn}^e thus implies that the probability of relocating from r to n for ethnicity e is

$$m_{rn}^{e} \equiv \mathbb{P}\left(V_{rn}^{e} = \max_{l} V_{rl}^{e}\right) = \frac{\bar{a}_{n}^{e}\left(\eta_{rn}^{-1}\left(L_{n}\right)^{\beta}\left(L_{n}^{e}/L_{n}\right)^{\beta^{e}}\bar{w}_{n}^{e}P_{n}^{-1}\right)^{\nu}}{\sum_{l=1}^{N}\bar{a}_{l}^{e}\left(\eta_{rl}^{-1}\left(L_{l}\right)^{\beta}\left(L_{l}^{e}/L_{l}\right)^{\beta^{e}}\bar{w}_{l}^{e}P_{l}^{-1}\right)^{\nu}}.$$

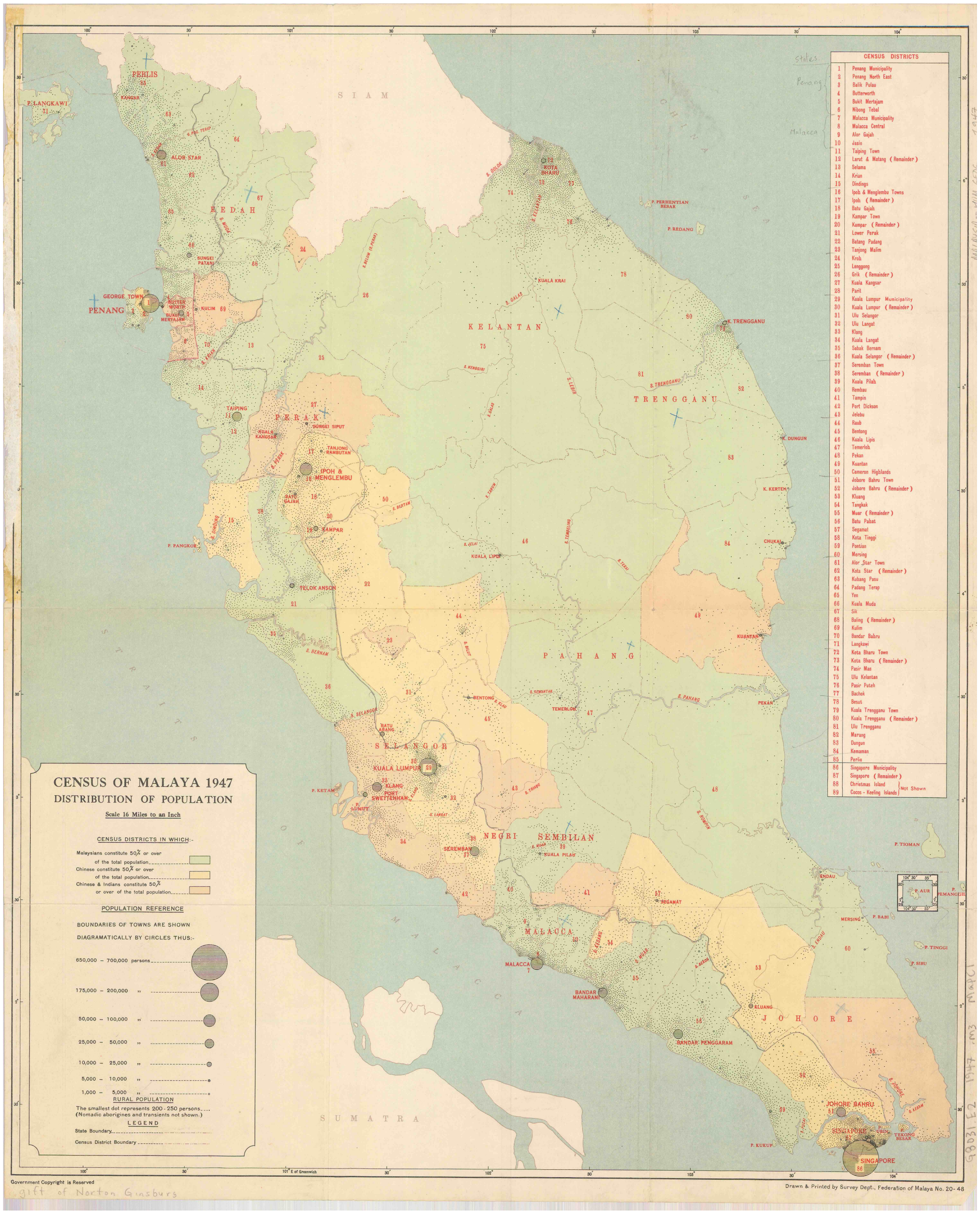
1.D Appendix Figures

Figure 1.8. Employment Share in 1947, by Ethnic Group



Notes: This figure shows the employment share across the primary, secondary, and tertiary industries for Chinese and non-Chinese, respectively. The primary sector includes agriculture and mining. The secondary sector includes manufacturing, utility, and construction. The tertiary sector includes transportation, communication, commerce, finance, business, and other services. Data from the 1947 Census of Population (Del Tufo, 1947).

Figure 1.9. Population Distribution in 1947

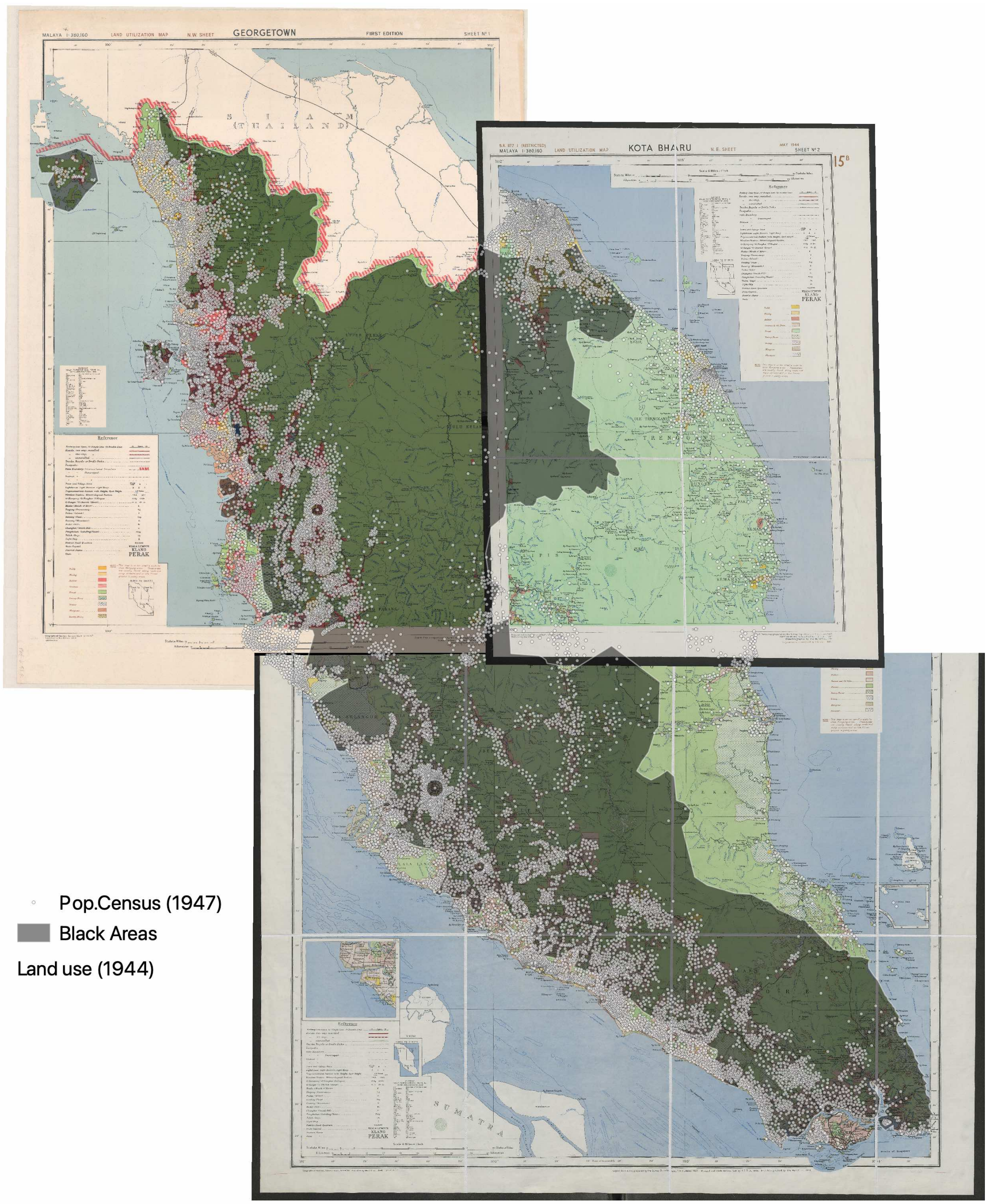


Notes: This figure shows the distribution of population in 1947. Data from the 1947 Census of Population (Del Tufo, 1947).

Figure 1.10. The Black Areas in 1957

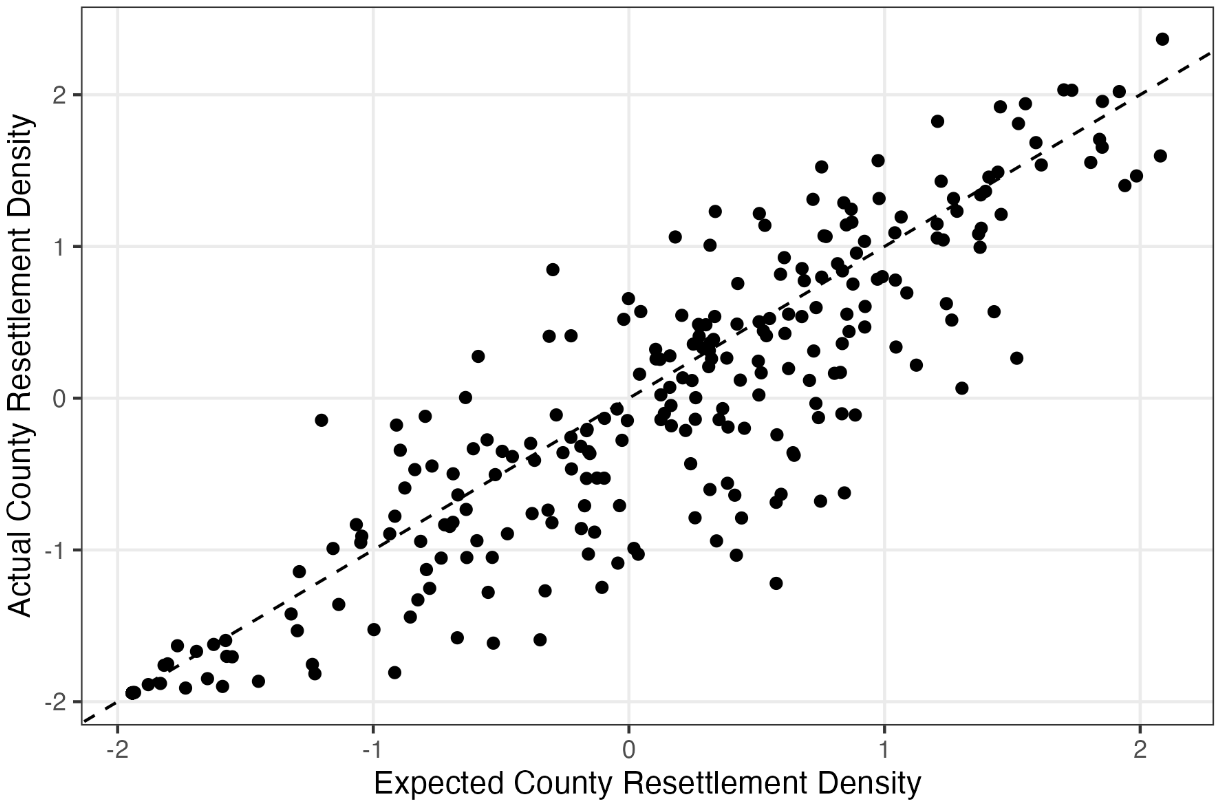
Notes: This figure shows the "Black areas" in 1957—areas considered to have substantial communists activities and were under various Emergency regulation. Data from the National Archives of the UK.

Figure 1.11. Population Distribution of the Squatters



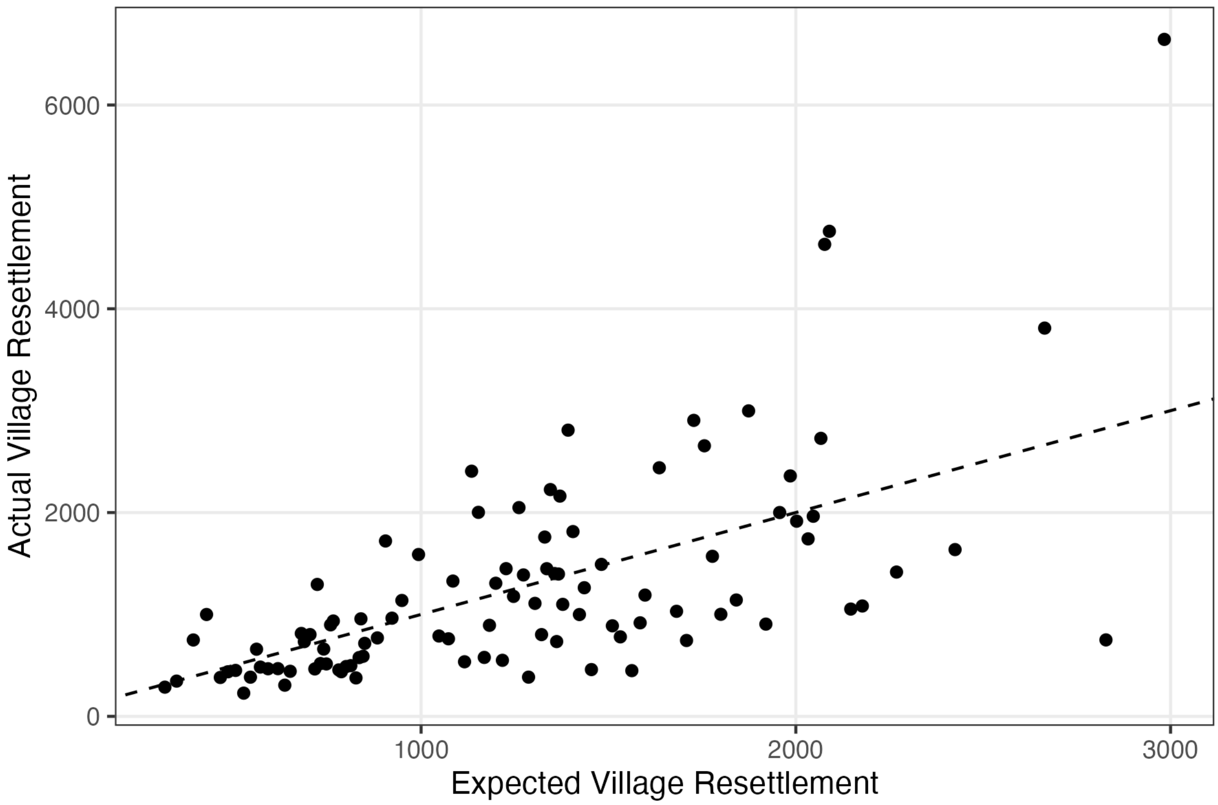
Notes: This figure shows the distribution of the squatters inferred from the intersection of three maps. The gray dots represent population clusters provided by the 1947 census. The areas shaded in dark are the "Black areas" with communist activities and were under various Emergency regulation. The areas shaded in green are areas classified as forest from land utilization maps in 1943 (War Office, 1943).

Figure 1.12. County Resettlement Density, compared to Expected Resettlement Density Predicted by Gravity Model



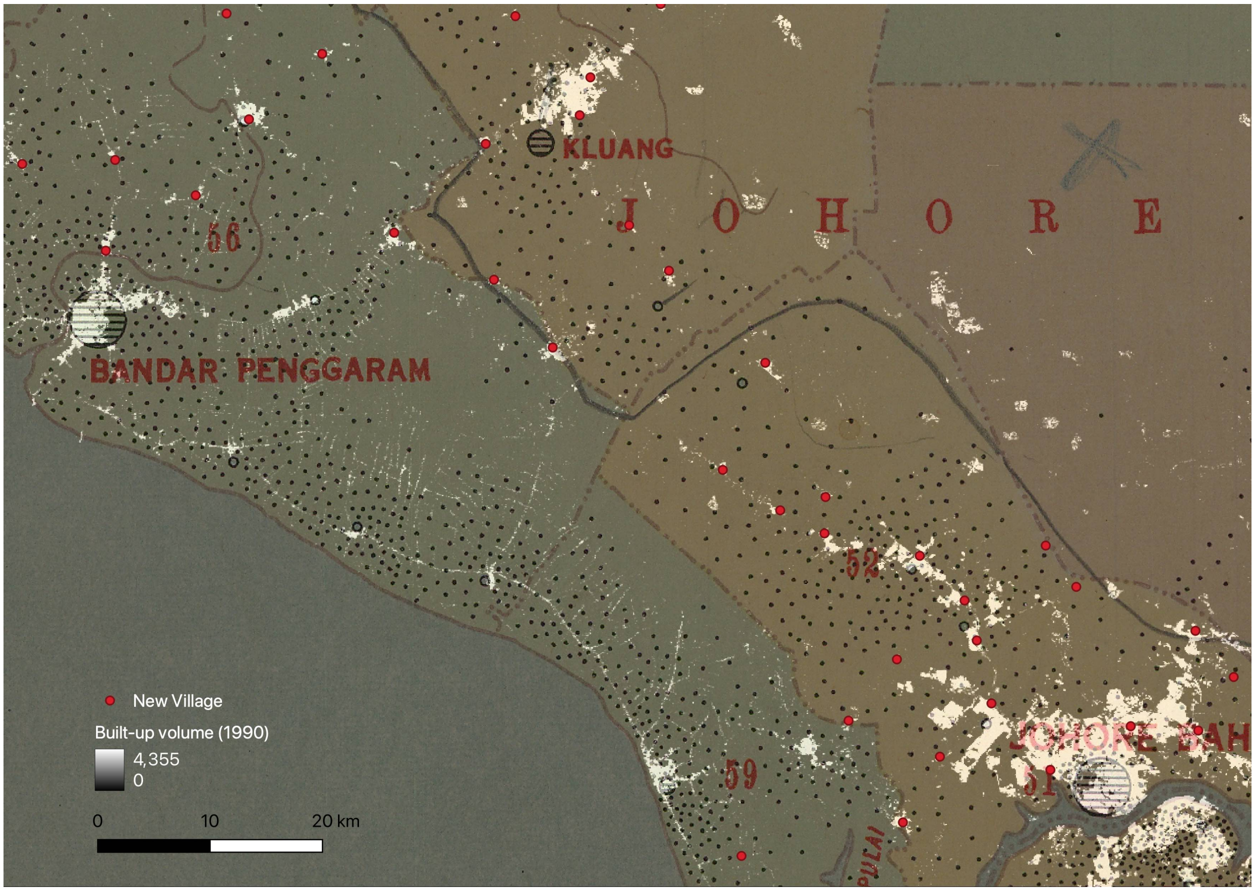
Notes: This figure compares the measured county resettlement density and the expected resettlement density calculated from equation (1.4). Expected resettlement density is calculated conditional on the actual locations of the New Villages. Data from the Corry report.

Figure 1.13. Village Resettled Population, compared to Expected Resettled Population Predicted by Gravity Model



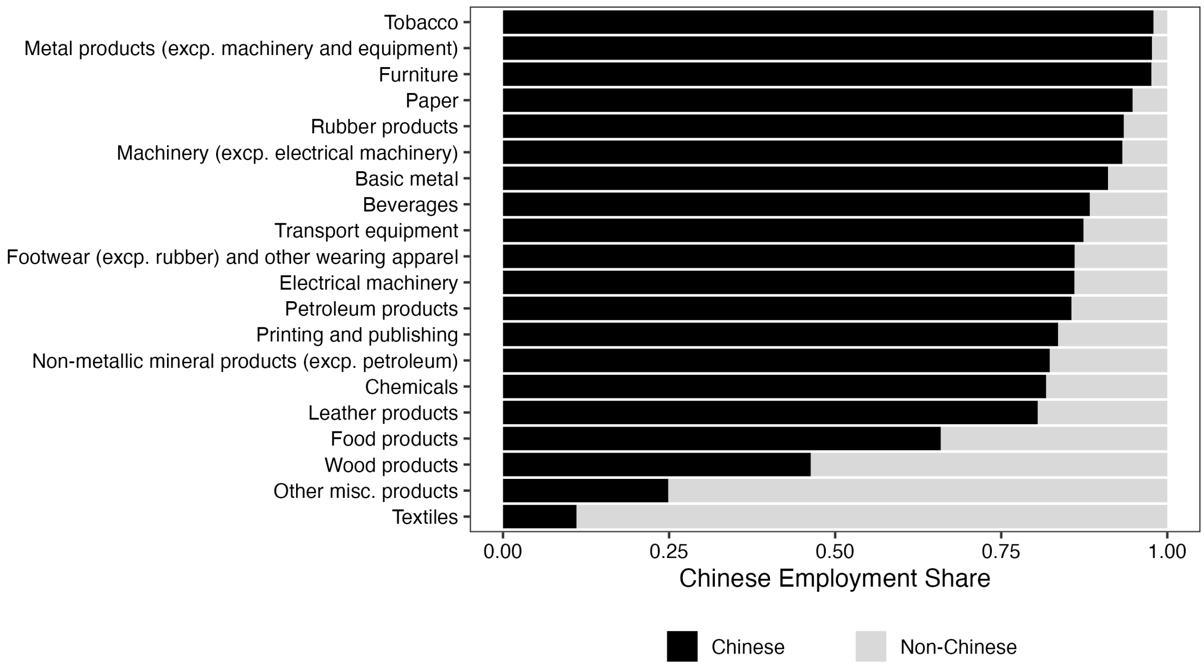
Notes: This figure compares the measured resettled population of each village and the counterfactual resettled population in the villages, conditional on their locations. The counterfactual village resettlement is calculated based on gravity equation (1.3), which models the dislocation-minimizing plan. Data from the Corry report.

Figure 1.14. Built-up Capital in 1990



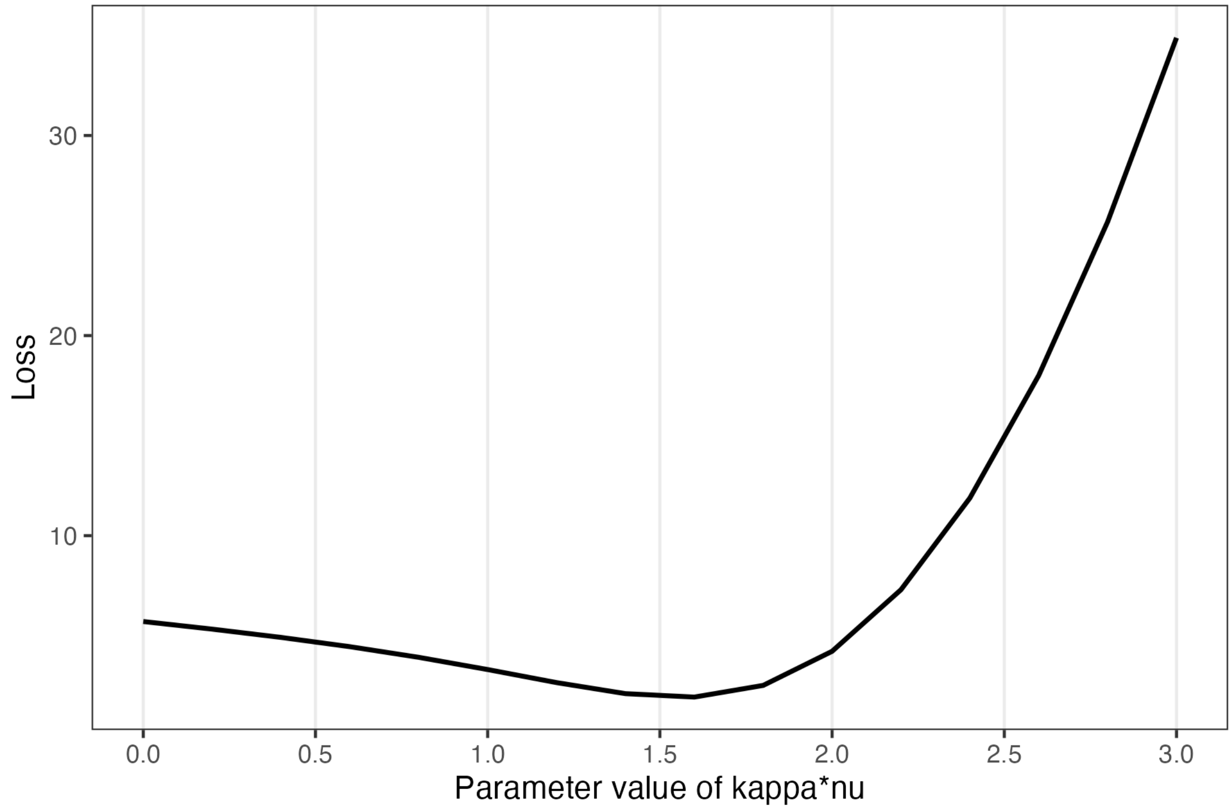
Notes: This figures shows the distribution of built-up capital in 1990 in a region within the state of Johor and the locations of the New Villages in that area. The New Villages are marked by red dots. Built-up volumes, shaded in white, are calculated using the surface and height data at a 100-meter resolution from the Sentinel-2 and Landsat satellite images. The base map shows the population clusters, marked by black dots, from the Census of Population in 1947. Data on built-up volume and New Villages from the GHSL project and the Corry report.

Figure 1.15. Chinese Employment Share Within Manufacturing in 1947, by Industry



Notes: This figure shows the Chinese employment share across industries within the manufacturing sector in 1947. Each bar shaded in dark shows the share of Chinese employment in total employment of a specific industry within manufacturing. Data from the 1947 Census of Population (Del Tufo, 1947).

Figure 1.16. Convexity of the Loss Function in Estimating Migration Costs



Notes: This figure shows convexity of the loss function for estimating migration cost. The y-axis plots the loss from equation (1.26), which is a function of observed bilateral migration flows and parameter value $\tilde{\kappa}$, shown in the x-axis. Data from the tabulated Census of Population in 1980.

1.E Appendix Tables

Table 1.9. Population in British Malaya from 1911 to 1957, by Ethnic Group

	Malays		Chinese		Indians and Others	
Year	Number (1)	Percent (2)	Number (3)	Percent (4)	Number (5)	Percent (6)
1911	1,367,245	59%	692,228	30%	239,169	12%
1921	1,568,588	54%	855,863	29%	439,172	17%
1931	1,863,723	49%	1,284,094	34%	572,205	17%
1947	2,395,686	49%	1,882,700	39%	529,594	12%
1957	3,126,773	50%	2,328,480	37%	695,923	13%

Notes: This table shows the population and share by ethnic group in British Malaya from 1911 to 1957. Columns 1 and 2 report the number of Malays and its share in total population of a given year. Columns 3 and 4 report the same figures for Chinese. Columns 5 and 6 report the same figures for Indians and other ethnic groups. Data from the Census of Population 1911-1957 (Vlieland, 1931; Del Tufo, 1947; Purcell, 1947; Fell, 1960).

Table 1.10. Migration Outcomes in 1980, by County Resettlement Density

	All Individuals (1)	Chinese Individuals (2)	Non-Chinese Individuals (3)	Difference (2) − (3) (4)
Panel A. Internal Migrant				
ResettleDensity	0.019	0.130*	0.012	0.118**
	(0.043)	(0.071)	(0.034)	(0.054)
# Individuals	173,375	57,512	115,863	
# Counties	752	526	744	
Panel B. Internal Migrant After 1960				
ResettleDensity	0.028	0.171*	0.025	0.146**
	(0.052)	(0.092)	(0.040)	(0.073)
# Individuals	172,514	57,225	115,289	
# Counties	752	524	744	

Notes: This table shows the relationship between migration outcomes and county resettlement density. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. Panel A shows the effect of resettlement density on whether a person is an internal migrant—those who moved to the current locality from another village or town within Malaysia. Panel B shows the effect of resettlement density on whether a person is an internal migrant who moved into the current locality within the last 20 years (or after 1960). Column 1 reports pooled estimates for Chinese and non-Chinese households. Column 2 restricts the sample to Chinese households only. Column 3 restricts the sample to non-Chinese households only. Column 4 reports the difference between the estimates in columns 2 and 3. Both panels are estimated using the Poisson pseudo-maximum-likelihood (PPML) estimator and include state fixed effects, the expected resettlement density, and the baseline controls: an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944. The unit of observation is the individual. The sample is restricted to individuals who are not external migrant—those moved to the current locality from outside Malaysia. Data from the 2% individual-level Census of Population microdata in 1980. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: *p<0.1; **p<0.05; ***p<0.01.

Table 1.11. Manufacturing Activity in 1970, by County Resettlement Density and Industry Employment Share of Chinese

	All	Chinese Owned	Non-Chinese Owned
	(1)	(2)	(3)
Panel A. Number of Establishments			
ResettleDensity	0.121	0.083	0.038
	(0.223)	(0.134)	(0.098)
ResettleDensity×ChiEmpShare	0.595	0.422	0.173
	(0.409)	(0.297)	(0.122)
Panel B. Log Establishments (PPML)			
ResettleDensity	-0.126	-0.016	-0.402
	(0.246)	(0.221)	(0.347)
ResettleDensity×ChiEmpShare	0.282	0.164	0.584
	(0.251)	(0.205)	(0.417)
Panel C. Any Establishment			
ResettleDensity	-0.009	0.000	-0.026*
	(0.025)	(0.024)	(0.016)
ResettleDensity×ChiEmpShare	0.053*	0.036	0.066***
	(0.031)	(0.029)	(0.025)
# Counties	777	777	777
# County-Industries	15,540	15,540	15,540

Notes: This table shows the relationship between measures of manufacturing activity and county resettlement density. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. "ChineseEmpShare" is the share of Chinese employment within the state-industry pair measured in 1947, with 20 categories of 2-digit industries within manufacturing (see Figure (1.15)). Each panel-column cell corresponds to a regression. Panel A, Column 1 shows the effect of resettlement density and its interaction with the pre-period Chinese employment share of the industry on an indicator of whether the county-industry has any establishment. Column 2 (or 3) shows the effect on whether the county-industry has any Chinese-owned (or non-Chinese owned) establishment. Panel B reports the effect of resettlement density on the total number of establishments (Column 1) and the number of Chinese-owned establishment (Column 2) and non-Chinese owned establishments (Column 3). Panel C shows the effect of the same outcome as B but is estimated with the Poisson pseudo-maximum-likelihood (PPML) estimator, whereas Panels A and B are estimated with OLS. All regressions include state fixed effects, the expected resettlement density, and the baseline controls: an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944. The unit of observation is the county. Data on manufacturing establishments are from the Directory of Manufacturing in 1970. Data on Chinese employment share are from the tabulated Population Census in 1947. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table 1.12. Educational Attainment in 1980 for Different Age Cohorts, by County Resettlement Density

	Years Schooling, By Cohort:			Primary Education, By Cohort:			Secondary Education, By Cohort:		
	20–35	36–50	>50	20–35	36–50	>50	20–35	36–50	>50
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. All Individuals									
ResettleDensity	0.218	0.158	-0.086	0.030**	0.062	-0.036	0.047	0.122	-0.040
	(0.144)	(0.154)	(0.096)	(0.015)	(0.041)	(0.081)	(0.038)	(0.116)	(0.164)
# Individuals	45,684	23,899	19,269	45,684	23,899	19,269	45,684	23,899	19,269
# Counties	746	740	739	746	740	739	746	740	739
Panel B. Chinese Individuals									
ResettleDensity	0.417*	0.315*	0.072	0.039**	0.070*	0.048	0.122*	0.254***	0.058
	(0.231)	(0.188)	(0.135)	(0.018)	(0.039)	(0.097)	(0.063)	(0.097)	(0.176)
# Individuals	15,597	8,843	7,067	15,597	8,843	7,067	15,597	8,843	7,067
# Counties	469	446	437	469	446	437	469	446	437
Panel C. Non-Chinese Individuals									
ResettleDensity	0.116	0.096	-0.102	0.020	0.052	-0.090	0.015	0.001	-0.115
	(0.115)	(0.130)	(0.097)	(0.015)	(0.039)	(0.088)	(0.034)	(0.119)	(0.165)
# Individuals	30,087	15,056	12,202	30,087	15,056	12,202	30,087	15,056	12,202
# Counties	738	731	730	738	731	730	738	731	730

Notes: This table shows the relationship between educational attainment and county resettlement density. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. Panel A shows the effect of resettlement density on individuals, pooling Chinese and non-Chinese together. Panels B and C show the effect only for Chinese individuals or non-Chinese individuals, respectively. Columns 1-3 report the effect of resettlement density on years of schooling for cohort aged 20-35 (column 1); 36-50 (column 2); and those above 50 (column 3), respectively. Columns 4-9 report the effect of resettlement density on the completion of primary education (columns 4-6) or secondary education (columns 6-9), by cohort. Columns 1-3 are estimated using OLS; and columns 4-9 are estimated using the Poisson pseudo-maximum-likelihood (PPML) estimator. All regressions include state fixed effects, the expected resettlement density, and the baseline controls: an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944. The unit of observation is the individual. Data from the 2% individual-level Census of Population microdata in 1980. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table 1.13. Household Asset Ownership, by County Resettlement Density

	All Households (1)	Chinese Households (2)	Non-Chinese Households (3)	Difference (2) − (3) (4)
Panel A. Owned the House				
ResettleDensity	0.043***	0.048***	0.030*	0.018
	(0.014)	(0.015)	(0.017)	(0.017)
Panel B. Have Vehicle				
ResettleDensity	0.038**	0.058**	0.020*	0.038*
	(0.018)	(0.028)	(0.012)	(0.022)
Panel C. Have Fridge				
ResettleDensity	0.038*	0.037	0.032	0.004
	(0.022)	(0.028)	(0.020)	(0.024)
Panel D. Have TV				
ResettleDensity	0.013	0.035***	0.003	0.032**
	(0.017)	(0.013)	(0.018)	(0.015)
Panel E. Have Phone				
ResettleDensity	0.023*	0.034	0.013	0.021
	(0.013)	(0.022)	(0.008)	(0.016)
# Households	37,124	11,604	25,520	
# Counties	759	521	751	

Notes: This table shows the relationship between household asset ownership and county resettlement density. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. Each panel shows the effect of resettlement density on a different indicator of asset ownership: the occupied house (Panel A); any motor car or van (Panel B); any refrigerator (Panel C); any black or color TV (Panel D); any phone (Panel E). Column 1 reports pooled estimates for Chinese and non-Chinese households. Column 2 restricts the sample to Chinese households only. Column 3 restricts the sample to non-Chinese households only. Column 4 reports the difference between the estimates in columns 2 and 3. All regressions are estimated by OLS and include state fixed effects, the expected resettlement density, the baseline county controls—an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944—and controls of household head's characteristics: age; years of schooling; and industry of employment. The unit of observation is the household. Data from the 2% individual-level Census of Population microdata in 1980 and Second Malaysian Family Life Survey 1988-1989. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: *p<0.1; **p<0.05; ***p<0.01.

Table 1.14. Household Income, Controlling for Household characteristics, by County Resettlement Density

	All Households (1)	Chinese Households (2)	Non-Chinese Households (3)	Difference (2) − (3) (4)
Panel A. Log Earnings				
ResettleDensity	0.063*	0.089**	0.038	0.051
	(0.033)	(0.042)	(0.029)	(0.034)
# Households	33,328	10,622	22,706	
# Counties	713	495	705	
Panel B. Log Earnings, Primary Sector				
ResettleDensity	0.032	0.066*	-0.005	0.071
	(0.038)	(0.035)	(0.041)	(0.045)
# Households	9,726	1,660	8,066	
# Counties	679	349	649	
Panel C. Log Earnings, Non-Primary Sector				
ResettleDensity	0.071**	0.096**	0.046*	0.050*
	(0.034)	(0.044)	(0.028)	(0.029)
# Households	23,602	8,962	14,640	
# Counties	698	445	689	

Notes: This table shows the relationship between household income and county resettlement density. "ResettleDensity" is the county resettlement density constructed according to equation (1.2), standardized such that it has a standard deviation of one. Panel A, Columns 1-3 show the effect of resettlement density on log household earnings predicted from asset ownership for all households (Column 1); Chinese households (Column 2); and non-Chinese households (Column 3). Column 4 reports the difference between the estimates in columns 2 and 3. Panel B restricts the sample to households whose head employ in the primary sector, comprised of agriculture and mining. Panel C restricts the sample to households whose head employ outside the primary sector. All regressions are estimated by OLS and include state fixed effects, the expected resettlement density, the baseline county controls—an indicator for any resettlement in the county; (log) county area; distance to the nearest road; road density of the county; distance to the nearest rail station; distance to coastline; Chinese population share of the county in 1947; (log) population density of the county in 1947; the share of lands used for rubber cultivation in 1944; and the share of lands used for mining in 1944—and controls of the household head's age and years of schooling. The unit of observation is the household. Data from the 2% individual-level Census of Population microdata in 1980. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses. Level of significance: *p<0.1; **p<0.05; ***p<0.01.

Table 1.15. Housing Elasticity in 1989

	Log Rents (1989)	
	OLS (1)	IV (2)
Panel A. Year 1980		
Log Population	0.274	0.326
	(0.051)	(0.144)
F-stat (1st Stage)		64.6
Panel B. Year 2000		
Log Population	0.267	0.270
	(0.051)	(0.110)
F-stat (1st Stage)		106.2
# Counties	103	103

Notes: This table shows the relationship between log housing rents in 1989 and log population in years 1980 (Panel A) and 2000 (Panel B). Column 1 reports the OLS estimates. Column 2 reports the IV estimates and the first-stage F statistics. The instrumental variable used is the residual resettlement density, as shown in Figure 1.3, Panel B. The unit of observation is the household. The sample is restricted to households reporting non-missing rent expenditure. Data from the Malaysian Family Life Survey in 1989. Conley standard errors with a distance cutoff of 30 kilometers are reported in parentheses.

CHAPTER 2

GAINING STEAM: INCUMBENT LOCK-IN AND ENTRANT LEAPFROGGING

2.1 Introduction

Technological innovation drives economic growth, but the widespread adoption of new technology can be slowed by firms continuing to use and invest in old technologies (Strassmann, 1959; David, 1990; Comin and Hobijn, 2010). We examine the adoption of steam power, an iconic general purpose technology (Bresnahan and Trajtenberg, 1995; Jovanovic and Rousseau, 2005). Steam power broke the dependence of mechanization on local geographic characteristics, particularly local waterpower availability, and steam was a central technological driver of widespread industrialization (Hunter, 1985; Atack et al., 2008).

Our primary goal is estimating the forces underpinning the slow transition from water to steam power in American flour mills and lumber mills, which were leading users of mechanical power. We estimate that steam created aggregate economic opportunities but hastened the exit of water powered incumbents, as switching barriers (particularly sunk costs) prevented many incumbents from upgrading technologies. As steam power improved, counties with less waterpower potential grew faster, both because steam power was relatively more useful in those places and because the *actual prior use* of waterwheels slowed steam's adoption.

It may seem surprising that switching barriers faced by incumbent establishments could be an important driver of the market-level spread of steam power, as there was substantial churn in establishments: only 2% of mills active in 1880 also existed in 1850. The fundamental force is that water power's relatively low fixed costs made it the optimal choice for relatively less productive entrants. One period's water powered entrants can become the next period's locked-in incumbents, so switching barriers can matter long after the initial incumbents have closed. We show that the *interaction* of switching barriers and high fixed costs slows aggregate technology adoption. If the new technology instead had a lower fixed cost (and higher marginal cost) *or* lower switching barriers, it would have diffused much faster.

To measure plant-level technology use and switching, we digitize the complete surviving establishment-level records from the US Census of Manufactures in 1850, 1860, 1870, and 1880.[1] These records include data on power use for every establishment. We create a panel by hand-linking mills over time based on their name, industry, and location, and we explore the influence of linkage error and unobserved resale of water power capital from incumbents to entrants. From the places with complete surviving establishment level records, we construct a balanced panel of 1199 county-industries (612 lumber-mill counties and 587 flour-mill counties), covering 690 unique counties and 80,000 establishment-year observations.

For causal identification, we use geographic variation in counties' access to waterpower. Local waterpower potential, measured in horsepower, is generated by the interaction of water flow and elevation changes. We use the geographic variation in waterpower potential, controlling for the main effects of water flow and terrain ruggedness that can otherwise influence local economic activity. We also control for other local characteristics that might impact local manufacturing activity, such as coal access and market access (Chandler, 1972; Hornbeck and Rotemberg, 2024). We measure local waterpower potential using modern hydrological models (McKay et al., 2012), which we validate with historical records.

The purchase cost of steam equipment declined from 1850 to 1880, leading to increases in aggregate steam-use in milling: in 1850, ten percent of mills were powered by steam, a share which increased to forty percent by 1880. We find that counties with higher potential for waterpower had more initial industrial activity. However, the decline in steam costs led to an "advantage of backwardness" (Gerschenkron, 1962). Counties with less waterpower potential

1. Samples of these manufacturing schedules were digitized by Bateman et al. (1971), Atack (1976), and Bateman and Weiss (1981), see also Sokoloff (1984) and **?**. Atack and Bateman (1999) provide detailed description of these samples. Recent efforts have digitized historical manufacturing microdata in a few contexts, including Japan, Russia, France, and Sweden (Braguinsky et al., 2015; Gregg, 2020; Juhász et al., 2023b; Berger and Ostermeyer, 2023).

adopted steam faster and experienced faster growth in their number of mills and mill output. Some incumbent mills switched from water to steam power, but county growth was driven by steam powered entrants. Incumbents were more likely to exit in counties with lower water power, despite more overall growth in these counties.

Lumber and flour mills were at the forefront of driving the adoption of steam power in the broader US economy, which brought mechanization to new industries and spurred productivity growth.[2] We find evidence of backwards linkages (Hirschman, 1958; Baldwin and Venables, 2015), as counties with less waterpower potential experienced disproportionate growth in makers of steam equipment such as engines and boilers, and shift to steam power outside of milling. Accelerating growth in upstream industries heightens the aggregate gains from early adoption of general purpose technologies like steam power, as it can encourage faster adoption in mechanizing industries.[3] This suggests that privately optimal technology adoption can be socially inefficient (Juhász et al., 2023a), and we evaluate potential counterfactual policies that might counteract the technological lock-in caused by historical advantages.

Because mills' technology adoption decisions depend on choices made by their competitors, as well as potential entrants, it is difficult to assess the equilibrium implications of the reduced-form estimates without some structure. A model also helps generalize lessons from steam power, isolating specific influences on technology adoption from other features of the technology itself and its economic environment. Further, given agglomeration spillovers in the adoption of this general purpose technology, a model allows us to consider the potential for welfare-enhancing policies whose effects depend on equilibrium responses.

To explore counterfactual technology adoption transitions, we develop and estimate a dynamic equilibrium model of firm entry and steam adoption. We build on Hopenhayn

2. For discussions of the role steam power played in the Industrial Revolution, see, for instance, Ashton (1948); North (1958); Kuznets (1967); Landes (1969); Rostow (1975); Atack et al. (2019), and Ridolfi et al. (2023).

3. A large literature studies technology adoption in the presence of network effects, including **??** and **?**.

(1992) and model firm dynamics with entry and investment in industry equilibrium, where firms make dynamic discrete choices in the tradition of Rust et al. (1987). In the model, heterogeneous firms make forward-looking decisions in each period about whether to enter, operate or exit, and which power source to use. Each power source is associated with costs and benefits that potentially vary over time and space, and incumbents face additional barriers to switching technologies.

Steam power was not a strictly dominant technology, and, even into the 20th century, water and steam power were both used by many millers. A key economic force in our model rationalizes that both water and steam power were used in equilibrium: one technology had lower marginal costs, and the other had lower fixed costs (where the fixed costs include both the purchase price and non-variable operating costs). We find that steam powered mills were larger than water powered mills (Atack et al., 2008; Ridolfi et al., 2023). Correspondingly, we estimate that steam had lower marginal costs and higher fixed costs (Melitz, 2003).

Steam power also attracted more-productive millers because it was easier to scale. While the direct marginal costs of water power were likely low in many places, our estimates reflect the difficulty of scaling up water power due to capacity constraints, and that many of the additional costs of steam power (such as skilled steam operatives) reflect fixed overhead costs. The size advantage of steam mills was not driven by steam's expansion of milling to new locations, as we find a similar pattern within counties.

Using variation across time and space, we estimate that the fixed costs of steam power declined over time and that higher local waterpower potential lowered the fixed costs of water power.[4] We also estimate the presence of agglomeration spillovers in steam-use.

A striking pattern in the data is that entrants were around four times more likely to use steam power than incumbent water powered mills, even though incumbents were typically larger and therefore predisposed to benefit more from steam. Nevertheless, the incumbents

4. Declining steam fixed costs are consistent with qualitative histories of steam use in rural US milling, in particular, which emphasize the development of practical low-cost engines (Hunter, 1985).

who did switch technologies grew faster than those who did not, consistent with costly switching barriers causing technological lock-in. We quantify that the barriers to switching from water to steam power were equal to about two months of revenue, and that sunk costs can account for around 90% of the switching barriers. Incumbents already had a functional power source, and most were reluctant to abandon it to switch to steam power.

We estimate the model using the Method of Simulated Moments, leveraging our reduced-form differences by county waterpower availability to identify key parameters in the model along with the patterns of establishment level usage of water and steam power.

We simulate the transition path from 1830 to 1900, as there was a secular decline in the price of steam power along with local agglomeration in steam-use and competition in local product markets. The estimated model closely matches the targeted moments. In addition, the model matches several non-targeted moments related to how waterpower potential leads to entrant-driven growth, as well as 19th-century accounts of the costs of power.

We use the model to estimate how technological lock-in from counties' waterpower potential delayed and reduced overall adoption of steam power in lumber and flour milling. Waterpower potential substantially slowed steam adoption: if the average county had one standard deviation lower waterpower potential, the share of mills using steam would have reached one-half 31 years earlier and been 18 percentage points greater in steady-state.

To quantify the role of barriers to switching, we evaluate a counterfactual economy where we remove all sources of lock-in. We estimate that without any switching barriers, the share of mills using steam would have reached 30% of US mills a decade earlier.[5] The delay was mainly caused by relatively low productivity entrants, initially attracted to the lower fixed costs of water, who then faced barriers in switching to steam power if their productivity grew. Though these barriers slowed adoption, switching was still an important mechanism for the

5. Even in the absence of barriers to switching, steam power would not have reached its steady state usage immediately, as the technology improved over time (David, 1969; Sandberg, 1969; Atack, 1979; Manuelli and Seshadri, 2014).

technological transition to steam power. For a counterfactual economy with infinite switching costs, we estimate that the steady state share of plants using steam would have been ten percent lower.

We estimate that switching barriers were sufficiently large that incumbent firms actually suffered overall from the introduction of steam power. While incumbents directly benefited from the option to switch to the new technology, this force is smaller than the increased competition from entrants.

The importance of switching barriers for steam use became relatively less important as steam reached maturity, and their removal would have resulted in a similar steady-state adoption rate. However, there would have been more entry in the absence of switching barriers, as entrepreneurs would have been attracted to the option value of seamlessly switching to steam power in the future. Therefore, switching barriers had persistently large effects on output.

In the presence of agglomeration spillovers, a natural policy intervention could mitigate switching barriers by purchasing the old sunk capital (i.e., "cash for clunkers"). We find that this type of subsidy would have generated positive social surplus, through raising steam adoption in the short run for the directly affected incumbents and through agglomeration spillovers on later entrants. However, the estimated agglomeration forces are weak enough that temporary policies do not have permanent effects on steam use (nor are there multiple equilibria).

Finally, to quantify the importance of the interaction of switching barriers and fixed costs, we estimate technology adoption rates in a counterfactual environment where the new technology has features of water (lower fixed costs and higher marginal costs) in comparison to an environment in which the new technology has features of steam (higher fixed costs and lower marginal costs). Even in the presence of switching barriers, the counterfactual lower fixed cost new technology would have rapidly reached its steady-state adoption. This is because relatively low productivity entrants would be attracted to the new technology, and so would not later become incumbents locked into the old technology.

The study of the transition from waterpower has a long intellectual history, for instance motivating Schumpeter (1942), and our establishment-level panel analysis complements a large literature studying long-run technology diffusion from a more aggregate perspective (Griliches, 1957; Jovanovic and Lach, 1989; Greenwood and Yorukoglu, 1997; **?**). The panel microdata allow us to measure directly plants changing their technologies over an extended period of time. We estimate large but not prohibitive barriers to switching, placing our results between common assumptions of either infinite switching costs (Chari and Hopenhayn, 1991; Atkeson and Kehoe, 2007; Collard-Wexler and De Loecker, 2015) or no lock-in (Basu and Weil, 1998; Acemoglu and Zilibotti, 2001; Greenwood et al., 2005; Benhabib et al., 2021; Miller et al., 2022). Our questions have similarities to those in a macroeconomic literature on "vintage capital," which considers the technology embedded in each successive generation of capital (Salter, 1960; Solow, 1962; Denison, 1964; Benhabib and Rustichini, 1991; Chari and Hopenhayn, 1991; Atkeson and Kehoe, 1999; Gilchrist and Williams, 2000; Jovanovic and Yatsenko, 2012; Caunedo and Keller, 2021), though we emphasize that an important driver of adoption speeds is if *entrants* use the old technology (and then become locked-in).

We focus on lumber mills and flour mills because they were heavy users of mechanical power that relied initially on local waterpower availability. Combined, they accounted for 20% of American manufacturing revenue at the start of our sample and 60% of mechanized establishments.[6] Lumber mills and flour mills sold primarily to local markets, and were classified by the Census as "neighborhood industries." Due to transportation costs and the high perishability of their finished products (Kuhlmann, 1929), these mills were broadly spread across the country and dependent on local geographic endowments for access to power.

6. Our setting is before the development of systematic industry codes, and for readability we combine the various names used to describe the industries into "flour" and "lumber." Similarly, we refer only to those two sectors when discussing "mills."

As a consequence, we model each county as having a distinct market, as in a recent literature that studies goods with prohibitive transport costs such as ready mix concrete (Syverson, 2004). By contrast, textile mills were geographically concentrated, as textiles were more broadly traded across domestic and international markets.

The importance of waterpower availability for power technology choices was understood contemporaneously (Montgomery, 1840), and occurs empirically across different contexts (Temin, 1966; Atack, 1979; Atack et al., 1980; Cooney, 1991; Bishop and Muñoz-Salinas, 2013; Chernoff, 2021; Gershman et al., 2022; Guilfoos, 2022).[7] Relative to this literature, our contribution is emphasizing the importance of establishment-level dynamics. This complements research on path dependence and inertia in other contexts, including railroad gauges (Veblen, 1915), prices (Rotemberg, 1982), keyboard layouts (David, 1985), consumer choice (Klemperer, 1995), migration (Kennan and Walker, 2011), city locations (Bleakley and Lin, 2012), health care (Handel, 2013), light bulbs (Armitage, 2023), telephone switchboards (Feigenbaum and Gross, 2023), and skills (Adão et al., ming). We also contribute to a literature studying why incumbents are slow to adopt new technologies (Chari and Hopenhayn, 1991; Jovanovic and MacDonald, 1994; Parente, 1994; Henderson, 1995; Jovanovic and Nyarko, 1996; Hall, 2004; Snow, 2004; Holmes et al., 2012; Verhoogen, 2023).[8] The most closely related model is from Humlum (2022), who studies robot adoption in modern firms but abstracts from entry decisions.

Because steam power reduced the dependence of manufacturing on local geography, it was adopted faster in places with less waterpower potential. This was caused by static forces, which raised the returns to adopting steam in those places, and dynamic forces, due to those

7. Duflo and Pande (2007), Lipscomb et al. (2013), Severnini (2023), and Brey (2023) leverage similar geographic characteristics to understand the effects of 20th-century dams. Arkolakis and Walsh (2023) use measures of solar insolation and wind speed to measure geographic variation in the potential for renewable energy production.

8. Frankel (1955) considers the importance of sunk costs for slow technological transitions, and Saxonhouse and Wright (1987) argue that sunk costs and durable capital led to a slow transition from spinning mules to ring-frame spinning in Lancashire, though both also abstract from the role of entry.

places also having fewer locked-in incumbents. Both static and dynamic forces were amplified by agglomeration spillovers, which encouraged further adoption of steam power in places where its adoption was already higher. Technology adoption was largely driven by entrants, but even entrants become stuck in prior technology when the new technology has higher fixed costs and lower marginal costs. Therefore, despite substantial firm entry, technological lock-in can dampen social gains from new technologies over a long time horizon.

2.2 Context and Data Construction

2.2.1 Water and Steam Power in US Mills

Water powered milling has a long history in the United States, as the Massachusetts Bay Colony built several watermills in the 1630s, some of which remained in use into the nineteenth century (Weeden, 1890). Mullin and Kotval (2021) note that Puritans believed every "town required four essential elements if it were to succeed: a meeting house with a pastor, a blacksmith, a sawmill and a grain mill." Flour and lumber mills were needed throughout the country, using the available local water power. They could use smaller rivers and did not typically require large installations. In contrast, textile mills could be agglomerated in major manufacturing centers in places with substantial waterpower capacity. Hunter (1979, 1985) provides an overview of water and steam power in the 19th century,[9] and we summarize a few key features of this context.

Most flour and lumber mills served their "local clientele" (Brown, 1923), though some "merchant mills" served cities and export markets (Kuhlmann, 1929). The nationalization of these industries occurred after our sample period. Flour milling began to concentrate in Minneapolis in the 1880s, after the development of less-perishable flours made possible by the middlings purifier and the roller mill (Kuhlmann, 1929; Perren, 1990). The rise of the

9. See Howes (2022) for a description of innovations in steam power before the 19th century.

milled lumber trade was facilitated by the emergence of manufacturers' associations to create and maintain standards (such as those regarding sapwood and knots). These associations did not exist in lumber until the 1880s, and did not reach prominence until the 1890s (Brown, 1923; National Industrial Conference Board, 1925).

The fundamental change from the arrival of steam power was a new source of mechanical power, less subject to natural constraints (Hunter, 1985): steam power was not as expensive to scale up, and it offered consistent year-round access to power. As a result, steam power was particularly useful in places with less local waterpower potential (Sharrer, 1982). These places had higher fixed costs for using water power, due to greater need for constructing dams, millponds, and riverwalls, which were generally more expensive to build than the wheels themselves (Monroe, 1825). Places with lower waterpower potential may have also required higher costs for securing water rights.[10] While water power technology improved over the 19th century, for instance with the development of the Jonval turbine in the 1840s and the Pelton wheel in the 1880s (Hunter, 1979), the more-substantial forces were that steam improved substantially over time and that waterpower availability varied substantially over space. For instance, a congressional report discussing options for a national armory on the "Western Waters" (Armistead et al., 1841) used, without updating, the estimated costs of water power from a previous Presidential report (Monroe, 1825).

While steam offered advantages, it was not a strictly dominant technology, as it required high non-variable costs: "the first cost of steam engines, and their annual expense, [did] not increase or diminish in proportion to the size of each engine" (Monroe, 1825). For instance, steam equipment required installation and continued maintenance oversight from trained engineers (Fisher, 1845).

Early steam engines were not widely adopted in the early United States.[11] With the

10. Swain (1888) reports the cost of water rights for 25 counties, which are negatively (though not significantly) correlated with our measure of waterpower potential.

11. Early Newcomen engines were coal-intensive and inefficient, wasting energy in the process of heating

introduction of the Corliss engine, patented in the US in 1849, manufacturing hubs in the US were increasingly using more-sophisticated and massive steam power systems. But these increasingly large and intricate systems were not particularly suitable for the small local mills throughout the US.

Local mills focused on relatively cheap "high-pressure" engines, patented and evangelized by Oliver Evans in the early 19th century, which did not use a condenser and instead used substantially higher pressure in the boiler. These engines were smaller and had substantially lower fixed costs, but were prone to explode (Burke, 1966; Mayr, 1975). Over the 19th century, many engineers adapted and improved on the standard designs (?), which allowed mill owners to purchase steam engines at steadily decreasing prices. Further, as local expertise in steam power spread geographically, increased local construction of steam machinery reduced shipping and installation costs (Greenberg, 1982).

In the second half of the 19th century, US mills began using "high-speed" engines that drew on earlier high-pressure boilers. High-speed engines were smaller and cheaper, though the parts needed to be made precisely to avoid the machine shaking dangerously and disintegrating.[12] New high-speed engine designs were introduced by Porter and Allen in 1862, and were described contemporaneously as a "revolution in engineering" (Scientific American, 1870). Porter (1868) argued that their design required efforts that machinists "were now thoroughly accustomed to," and that the "commercial benefits" to the engine included "the saving of space and the economy in first cost."

Many classic examples of switching barriers were likely relatively less important in this

---

and cooling water to drive a piston in a cylinder. In the late 18th century, James Watt introduced a separate condensing chamber so the primary cylinder never needed to be substantially cooled, which dramatically improved the efficiency and force of British engine designs. In the spirit of Arrow (1962), steam engine manufacturing was characterized by learning-by doing, as many subsequent improvements to Watt's design came as machinists gained experience and tinkered with the size and arrangement of the parts.

12. Although steam engines and boilers got safer over time, explosions are often described in histories of individual mills and, during the period, a plurality of steam engine explosions were in lumber mills (Scientific American, 1871, 1881).

context. Technological interrelatedness between components within the production process can rationalize lock-in in other contexts (David and Bunn, 1988; Bresnahan and Greenstein, 1996), but are unlikely to be relevant in our setting as the power remained rotational in nature, and the millstones or saws as well as the material inputs and outputs were the same regardless of the power source.[13] Similarly, catering to existing customers (?) or changing suppliers (Farrell and Klemperer, 2007) are unlikely to be important, as the milled products remained unchanged and the physical waterwheels were very durable. By contrast, the later introduction of electricity ushered in more wholesale changes in manufacturing operations, and it was more difficult for incumbents to change power sources (Devine, 1983; David, 1990; Damron, 2023).

In Appendix 2.F, we collect the histories of several mills who switched from water to steam power. The most common reason why mills switched to steam power we found was they outgrew the power availability of their local waterway, or they lost their local water rights (Emery, 1883). A few millers physically moved their operations to a new structure when switching power sources, but most retrofitted their existing mills in place even after losing the original motivation for their location. Many switches from water to steam power were associated with a change in ownership, often through sons taking over from their fathers, which suggests switching frictions on the part of operators and points to the importance of management (Bloom et al., 2013; Giorcelli, 2019).

2.2.2 County Waterpower Potential

We measure counties' waterpower potential, based on natural geographic characteristics, as a cost-shifter for local firms' use of water power. A key assumption for our analysis is that waterpower potential affected mills only through the costs of water power use. To support

13. The technologies are similar enough that some water powered mills used steam as an auxiliary power source (Hunter, 1985), and in our data around a third of establishments who switched from only-water to steam power continued using some water power.

this assumption, we focus on variation in local waterpower potential from the interaction of particular geographic characteristics, controlling for their main effects and other local characteristics.

For any river segment, its theoretical potential for generating waterpower (in units of horsepower) is given by multiplying: (1) the flow rate of water; (2) the change in elevation (fall height); and (3) a gravitational constant equal to roughly 0.1134:

$$\text{Theoretical Water Power} = \underbrace{\text{FlowRate}}_{\substack{\text{Cubic Feet}\\ \text{Per Second}}} \times \underbrace{\text{FallHeight}}_{\text{Feet}} \times \substack{\text{Gravitational}\\ \text{Constant}}.$$

For each river segment in the country, we use information from the National Hydrography Dataset Plus (NHDPlusV2), which is a national database of surface water from the US EPA and USGS. For measuring fall heights, we use the difference in elevation between the maximum and minimum elevation along each river segment. Given the absence of detailed and comprehensive direct measurements of historical water flow, and the potential influence of dams and other modern influences on modern rivers, we use monthly flow estimates from a USGS flow-balance model based primarily on natural and slowly changing climatic variables, such as rainfall, evaporation, and soil moisture. We use the average flow rate over the three lowest months of the year, which historical accounts argued was a key determinant of the feasibility of water power (Census Bureau, 1883).[14] Figure 2.1 shows flow rates and fall heights for each river segment across the US, whose interactions determine waterpower potential.

We calculate waterpower potential at the county level, summing over each river segment in the county. We exclude wide river segments (more than roughly 106 feet wide) because those segments were considered at the time to be too wide for use as a practical source of

14. We include in our calculations "seasonal" rivers with intermittent flows, though in practice many do not have any flows in the lowest three months of the year and therefore do not affect our measure of local waterpower potential.

water power, due to high dam costs, and were used instead for transportation.[15]

We validate the estimates of waterflow using historical records from the 1880 Census "Reports on the Water Power of the United States" (the "Water Census"). Consistent with the historical importance of water power, the US government spent resources to promote its use even in 1880: the stated purpose of the Water Census was to "describe the privileges actually in use and call attention to locations where power could be advantageously developed." For river segments covered in the historical Water Census, their flow rates are in close agreement with the modern data (Appendix Figure 2.11).[16]

Our measurement of county waterpower potential does not directly use the Water Census, however, because the Water Census has non-random incomplete coverage based on historical economic activity (Appendix Figure 2.12 Panel A). The Water Census was intended to focus on places with high waterpower potential or usage, systematically missing places that have lower waterpower potential and lower usage. Further, the Census data collection effort ran out of funds before getting to much of our sample area (Atack et al., 1980). In Section 2.3.1, we show how relying on only the Water Census would bias estimated impacts of county waterpower potential on water power usage.

Appendix 2.D describes in more detail our processing of the NHDPlusV2 data. We also compiled a variety of county-level information for supplementary analysis and controls, such as access to coal deposits, which we also describe in Appendix 2.D.

15. For example, the 1880 Water Census writes: "...the Mississippi as it flows past New Orleans gives an exhibition of tremendous force, and by damming it up to a head of 10 feet a power of nearly 700,000 horse-power would result, but the river would be flooded back for 300 miles, and the plan is therefore impracticable." Indeed, Appendix 2.D.3 shows that these wide rivers are not predictive of water powered mills.

16. There are some exceptions where the values diverge, which generally reflect segments where merging the two datasets is difficult (e.g., if a river splits into several sections and we are not sure how many segments to aggregate when comparing our smaller river segments to what the Census considered a river segment, or when distinct rivers in a county share a name).

2.2.3 Census of Manufactures, Establishment level Data

We collected and digitized all known establishment-level manuscripts from the Census of Manufactures in 1850, 1860, 1870, and 1880 (see Appendix Figure 2.13 for example images, and Appendix Table 2.13 for the coverage of manuscripts). We classify each establishment into one of 31 industries, following Hornbeck and Rotemberg (2024), using information on self-reported "name of business" and products the establishment produced.

We restrict our main analysis to county-industries with at least one active mill in 1850 and non-missing data in each decade from 1850 to 1880. Our sample covers lumber mills in 612 counties and flour mills in 587 counties. There are 690 unique counties with at least one of these industries in the sample. Our sample includes over 80,000 lumber or flour mills from 1850-1880, and cover 83% of all steam powered mills and 89% of reported steam-generated sales in the lumber and flour industries. Figure 2.2 Panel A shows the waterpower potential of the counties in our balanced sample.

Our data include the type of power used by each establishment, which was not geographically disaggregated in contemporaneous census tabulations (Hornbeck and Rotemberg, 2024). We also use the total annual revenue for each establishment, which inform distributions of establishment sizes that are unavailable in the previous more-aggregated data. We record establishment names, which were not entered in previous samples of the establishment-level manuscripts due to punchcard width limitations (Atack and Bateman, 1999), which allow us to link mills over time.

Not all manuscripts have survived, which we can assess using contemporaneously published Census tabulations at the county level for 1850-1880 (Haines, 2010) and county-by-industry level for 1860-1880 (Hornbeck and Rotemberg, 2024). Manuscripts for some entire states and decades were lost when the original manuscripts were returned to states. Manuscripts for some counties were lost for reasons such as being used as wrapping paper when returning other manuscripts (Atack and Bateman, 1999) and manuscripts for some industries (though neither lumber nor flour) were lost in 1880 (Delle Donne, 1973). To separate "missing" from "zero," we classify a county as having missing data if the county has no manuscripts but the tabulations report positive establishments; otherwise, we record the county as truly having no manufacturing activity.

For counties with surviving manuscripts, Appendix Figure 2.14 shows that our microdata generally align closely with the tabulated county-level data. However, we provide the first comprehensive information on lumber and flour mills in the period because the Census did not report county-industry statistics in 1870 and 1880 for small "local industries" (Appendix Figure 2.15 Panel A). For county-industry cells above the Census tabulation threshold, our data aligns closely (Panel B). Appendix 2.C discusses in detail our collection and processing of these data, data coverage issues, and how we group counties into time-consistent geographic units.

While mechanical power eventually spread throughout manufacturing (Atack et al., 2019, 2022), we focus on industries that had widely mechanized before steam arrived to study the transition of mechanical power from water to steam. Most water powered establishments in 1850 were either lumber or flour mills (Figure 2.3). Flour milling was the largest industrial sector in the economy during our period, by revenue, and lumber milling was the largest by number of establishments. Textile mills were also heavily-mechanized, though records for textiles in 1880 have been almost completely lost (see Appendix 2.C and Atack and Bateman 1999).

Among lumber and flour mills in 1850, 91% report using either water or steam power. Around 1% of mills used both water and steam power, which we classify as steam mills because they paid the fixed costs of steam and thereby benefited from the ability to scale relatively cheaply. Non-mechanized mills contributed little revenue share (Figure 2.3, Panel B), and our main analysis omits these non-mechanized mills.

Mills had substantial local competition. The median county-industry had 10 mills operating in a given year. Almost all county-industries had more than one mill (96%). Of these, 62% had at least one mill using each type of power and this share increased over time as steam power became more prevalent.

A useful feature of lumber and flour mills, for our analysis, is they primarily served local demand because cut lumber and ground flour were perishable and not economical to trade, especially to rural destinations (Hunter, 1979). Indeed, an important source of revenue for flour mills was "custom milling": grinding grain that customers brought themselves (Dondlinger, 1919; Le Bris et al., 2019). The Census asked specifically about this practice in 1880: 95% of mills did at least some custom milling in 1880, and it represented 41% of total flour milling output. While milling was dependent on local geographic endowments to generate power, the material inputs (logs and whole grains) for these mills were less perishable and could be transported long distances, so the local endowment of inputs was not as important for millers (Cronon, 2009).

Consistent with historical accounts that flour and lumber milling produced relatively non-tradable output, Appendix Figure 2.16 shows that the spatial concentration of lumber and flour mills was particularly low (in the spirit of Mian and Sufi 2014).[17] This contrasts with clothing and textile mills, whose output was more easily traded and so was much more concentrated geographically. Lumber milling remains diffused: in the 2021 County Business Patterns, 98% of commuting zones had a lumber mill and 25% had a flour mill.

Census schedules in 1870 and 1880 also asked mills for their installed horsepower, shown in Appendix Figure 2.17: steam powered mills typically used more horsepower than water powered mills, and most mills used between 10-60 horsepower with the mode around 25 horsepower.

17. The other least geographically concentrated sectors are leather and iron & steel (due to blacksmithing, as discussed by Atack and Margo 2019).

2.2.4 Data Linking

We create a linked panel of manufacturing establishments over time, which allows us to observe technology switching and entrant technology choices. The manuscripts do not have a time-consistent identifier for each establishment, just as in the Censuses of Population (Ferrie, 1996; Feigenbaum, 2016; Ruggles et al., 2018; Bailey et al., 2020; Abramitzky et al., 2021; Price et al., 2021), so we generate our own links.

We define a stable manufacturing establishment based on its owner name, industry, and place. If the owner shuts down an establishment and reopens an establishment in a different county, we consider that a new establishment. Similarly, if the owner changes their establishment to no longer be a mill, we consider the mill closed.[18] While we link establishments with partial ownership changes (such as a son taking over from his father), if the establishment's ownership changes entirely, with no clear link between previous and new owners, then we also consider that a new establishment. This is dictated by data availability, and also raises philosophical questions about what is a surviving establishment. Our view is that mill owners at the time were sufficiently involved in the operation of the establishment that entire ownership changes are akin to closing operations and selling capital assets to a new venture.[19]

We link establishments over time, within a county, using data on owner or company names, industry, product types, and (when available) nearest post office. Importantly, we do not use mills' type of power to make the panel identifiers. We hand-linked all lumber and flour mills,

18. These cross-county "migrations" appear unusual for millers, based on historical society records (Appendix 2.F), and when we hand-linked the establishments we allowed for cross-industry links and found very few outside of milling. Around 4% of surviving mills switched between lumber and flour.

19. We do find evidence of ownership transfers in historical accounts, though most business closures appear to be associated with the mill no longer being operated. The Census data do not allow us to directly observe the resale market (Lanteri, 2018), though measuring the importance of durable capital across in addition to within firms is an interesting direction for future research on technology transitions. We discuss the implications of unobserved reselling for our reduced-form estimates in Section 2.4. In Section 2.5, we model and estimate how local technology choices affect the relative purchase prices of steam and water power, which captures if the transition to steam power lowered the purchase price of water power.

across each decade. Two people searched for matches for each mill, and we reconciled any disagreements. We also trained a machine-learning ("ML") algorithm to predict the matches, described in Appendix 2.C.4, which allows us to analyze robustness to different confidence thresholds, and show that the distribution of predicted ML link probability for our actual matches is similar in counties with above and below median waterpower potential counties.

We also link establishment owners to the Census of Population, based on owner name, industry/occupation, and place, as described in Appendix 2.C.4. For our analysis, we use three owner characteristics from the Census of Population: their age; whether they were born outside the United States ("immigrant"); and if their listed occupation was a miller or manufacturer ("professional miller").[20]

2.3 Estimating Differences by County Waterpower Potential

Our analysis looks to estimate how local waterpower potential affected early water power usage and the growth of steam use. We contrast impacts on incumbents and entrants to explore how both the potential for waterpower and actual prior use of water power affected steam adoption.

To estimate cross-sectional effects of county waterpower potential on lumber and flour mill activity, we estimate the following regressions where each observation is a county-industry:

$$Y_{ic} = \beta LowerWaterpowerPotential_c + \gamma_i X_c + \lambda_i + \varepsilon_{ic}. \tag{2.1}$$

We define $LowerWaterpowerPotential_c$ as a negative standardized measure of (log) county waterpower potential per square mile, so the coefficient β can be interpreted as the effect of having one standard deviation lower waterpower potential. We focus on the estimated pooled β, across lumber and flour mills.

20. The modal listed occupation for a person we link to the Census of Manufactures is "farmer," and we explore whether self-reported "professional millers" are more likely to use the more modern technology.

The estimated effect of $LowerWaterpowerPotential_c$ is conditional on industry fixed effects λ_i and a set of county controls X_c, whose effects are allowed to vary by industry i. We include three types of baseline controls, within X_c. First, as waterpower potential comes from the interaction of water flow and elevation changes, we control for its components: total county water flow, summing over all river segments; and county ruggedness, defined as each county's average terrain ruggedness index (?).[21] Second, because access to markets also affected economic activity and some mills got access to their material inputs through waterways (Cronon, 2009), we also control for: whether the county has navigable waterways; distance to the nearest navigable waterway; and county market access in 1850 including the waterway and railroad network (Hornbeck and Rotemberg, 2024). Third, because an important source of fuel for steam mills was coal,[22] we control for: whether there are workable coal deposits in the county, the share of the county covered by coal deposits (Campbell, 1908), and access to coal via the transportation network.

We also estimate some pooled cross-sectional regressions, across 1850 to 1880. For this analysis we replace the industry fixed effects in Equation (2.1) with year-industry fixed effects (λ_{it}) and allow the effects of the control variables to vary jointly by year and industry ($\gamma_{it}X_c$).

The key identifying variation comes from the interaction of river flow rates and fall heights. For the baseline cross-sectional specification, the identification assumption is that counties with lower waterpower potential would have had similar mill activity in 1850 as counties with more waterpower potential, on average, aside from differences due to power use. In practice, the identification assumption is conditional on any other differences associated with the included control variables. The control variables look to adjust for direct effects of rivers, particularly through lower transportation costs and differential impacts from the railroad network, along with different economic outcomes associated with variable elevation, access to

21. County ruggedness is closely associated with the presence of changes in elevation, whereas fall height along river segments is not defined in the absence of rivers.

22. Some lumber mills used scrap wood for fuel (Cole, 1970).

markets, and access to coal. We discuss alternative controls in Section 2.3.3 and Appendix 2.G, including specifications without controls, with fewer controls, or with additional controls that adjust for other factors that might be associated with differential steam adoption and growth in mill activity across counties with different waterpower potential.

Our main sample is a balanced panel of county-industries, from 1850 to 1880, restricting our analysis to 690 counties with at least one lumber or flour mill in 1850 and surviving Census manuscripts in each decade. Figure 2.2 Panel B shows the residual waterpower potential of the counties in our sample after partialling out the baseline controls.[23]

To estimate changes over time in counties with lower waterpower potential, as steam technology improved, we estimate the following panel regressions where each observation is a county-industry-decade:

$$Y_{ict} = \beta_t LowerWaterpowerPotential_c + \gamma_{it} X_c + \lambda_{ic} + \lambda_{it} + \varepsilon_{ict}. \tag{2.2}$$

The estimated β coefficients report the relative change in counties with one standard deviation lower waterpower potential. We estimate the regressions separately by decade-pair, for instance estimating changes from 1850 to 1860 including only data from 1850 and 1860, which avoids interpretation issues associated with regression models that pool across many time periods (e.g., Roth et al., 2023). We include county-industry fixed effects (λ_{ic}), year-industry fixed effects ($\gamma_i t$), and interact our baseline control variables with year-industry dummies ($\gamma_{it} X_c$).

For the panel regressions, the identification assumption is that counties with lower waterpower potential would have *changed* similarly to counties with more waterpower potential, on average, aside from differences due to water power and steam. This assumption is conditional on differential changes associated with our baseline county controls (river flow, terrain ruggedness, navigable rivers and market access, coal deposits).

23. The Appalachia region generally has higher waterpower potential and in Appendix 2.G we show directly that our results are not driven by regional differences for Appalachia (with its own distinct topography and history).

For the cross-sectional and panel regressions, our main outcome variables relate to mill activity and their power source. We also examine outcomes separately for entrants and incumbents, which informs the role of switching barriers in the transition from water to steam power.

Some outcome variables are well-defined in levels, such as the share of mills using steam power, and for these outcomes we estimate Equation (2.2) using OLS. Shares are undefined when there are no mills, so we omit counties with no mills in one of the relevant decades. When estimating impacts on the share of mills using steam, we weight county-industries by their number of mills in the initial year to make our estimates comparable to a firm-level regression for an indicator of power adoption choice.

For outcomes such as total mills, we want to measure their elasticity with respect to waterpower potential. There are a few zeros in the sample, for county-decades where all incumbent mills closed after 1850 and there were no entrants. To estimate elasticities, and include growth on both extensive and intensive margins, we use Poisson Pseudo Maximum Likelihood (PPML) regressions (Silva and Tenreyro, 2006) rather than approaches such as $\log(1 + x)$ or inverse hyperbolic sine that are sensitive to units and therefore difficult to interpret (Chen and Roth, 2023).[24] Similarly, we use PPML to estimate the elasticity of the entry rate (entrants / previous mills) and the survival rate (incumbents / previous mills) with respect to waterpower potential.[25]

We focus on linear specifications, as Appendix Figure 2.12 Panels B and C show that the estimated impacts on mill activity from county waterpower potential are roughly linear. We report robust standard errors clustered by county. Mill activity serves largely local markets,

24. Formally, PPML estimates the average effect of county waterpower potential as a percentage of the baseline mean.

25. To estimate the elasticity of the entry rate with respect to waterpower potential, we use PPML regressions where the outcome in the current period is the number of entrants and the outcome in the previous period is the total number of establishments. This is equivalent to running a cross-sectional OLS regression for the log of entrants minus the log of total prior establishments, but does not require dropping counties without prior establishments or entrants. We use the same approach for the incumbent survival rate.

though waterpower potential is correlated across nearby counties, and we also estimate Conley (1999) standard errors that adjust for spatial correlation across counties assuming counties are independent beyond a distance cutoff. The Conley standard errors are similar to the clustered ones for distance cutoffs within 500 miles, and are 10-40% smaller for cutoffs up to 1000 miles.

The main outcomes that we are interested in are how milling was shaped by entrants vs. incumbents, and steam vs. water users. Table 2.1 shows the share of milling in each decade for each type of mill. In each Census year, most mills entered during the previous decade, and entrant establishments disproportionately used more steam power than incumbent establishments.

2.3.1 Waterpower Potential, Power Use, and Mill Growth

Table 2.2 reports that counties with one standard deviation lower waterpower potential had substantially fewer water powered mills in 1850 (Panel A) and substantially less revenue from water powered mills in 1850 (Panel B). Columns 2 and 3 report estimates separately for lumber mills and flour mills. The estimated coefficients of -1.06 and -1.13 imply 65% fewer water powered mills and and 68% less water powered revenue (Column 1).

By 1850, there had been faster adoption of steam power in counties with lower waterpower potential (Table 2.2, Panels C and D). The share of mills using steam power was 8.9 percentage points higher in these counties in 1850 (Panel C), and the share of revenue produced using steam power was 12 percentage points higher (Panel D).

Overall mill activity was still substantially lower in counties with lower waterpower potential (Panels E and F), though somewhat muted by the increased use of steam power. Particularly in lumber milling, where there was a more substantial early shift to steam power, there are more muted effects on total revenue in 1850.

Table 2.3 reports estimated changes in counties with lower waterpower potential. From

1850 to 1860, the share of mills using steam power grew 6.7 percentage points more in counties with lower waterpower potential (Column 1). Steam-use grew by 3.4 percentage points from 1860 to 1870 in lower waterpower counties. From 1870 to 1880, steam adoption began to catch up in counties with more waterpower potential by a statistically insignificant 0.9 percentage points. Figure 2.4 shows that steam use also increased from 1850 to 1880 in counties with average waterpower potential, but more so initially in counties with one standard deviation lower waterpower potential.

Counties with lower waterpower potential also experienced substantial relative growth in the total number of mills and total revenue (Table 2.3, Columns 2 and 3). The number of mills increased by 25% and revenue increased by 20% from 1850 to 1860. Growth continued at lower rates through 1880, suggesting continued benefits from lower waterpower availability and earlier steam adoption.

Table 2.4 shows this growth in counties with lower waterpower potential was driven by entrant firms. The entry rate was 38% higher, from 1850 to 1860, while the firm survival rate was 21% lower. In each period, entrants crowded-out local incumbent firms, which exited at higher rates in counties with lower waterpower potential despite the overall growth in these counties.

We can also separate incumbents by their prior-period power use. We refer to "water incumbents" and "steam incumbents" as surviving firms who used water and steam in the previous decade, regardless of their technology in the current period. Appendix Table 2.14 shows that waterpower potential had roughly similar effects on the exit probabilities of steam and water incumbents.

Table 2.5 shows that entrant firms mostly drove the greater adoption of steam power in counties with lower waterpower potential. In each decade, entrants were 16 – 19 percentage points more likely to be using steam power, relative to entrants in counties with higher waterpower potential (Column 1). Among incumbent firms that had been using water power ("water incumbents"), these firms were a more modest 3 – 5 percentage points more likely to adopt steam power in counties with lower waterpower potential (Column 2).

Steam adoption by entrant mills was substantially more responsive than switching to steam by water incumbents (Column 3, Table 2.5). Water incumbents' lower steam use, combined with the increased exit of incumbents from Table 2.4, suggest that incumbent mills were subject to switching barriers.

In Summary: The increase in steam use for lower water power counties was driven by more entrants in lower water power counties, as incumbents were crowded out (Table 2.4). Furthermore, in counties with less waterpower potential, entrants adopted steam more readily than water incumbents (Table 2.5). Section 2.5 quantifies this technological lock-in and its implications.

2.3.2 Non-Mill Manufacturing, Steam-Use, and Backward Linkages to Steam Production

This section shows differences by waterpower potential in broader manufacturing activity, outside lumber and flour mills. We also then narrow our focus to local steam engine production, which supported higher local steam-use across manufacturing. We restrict this analysis to 1850–1870 due to the missing Census manuscripts for some industries in 1880.

Table 2.6, Column 1, shows that counties with lower waterpower potential also had substantially less manufacturing activity in 1850 outside of lumber and flour mills. This is consistent with less local waterpower potential making locations less attractive, both due to lower water power use in other sectors and co-agglomeration of other sectors with milling that supported local economic activity generally. This difference declined slightly over time, as steam-use increased modestly (Column 2). In 1850, non-mills were already more likely to use steam power if located in counties with lower waterpower potential. Non-mills in these counties adopted steam power somewhat faster over the subsequent decades, though not as much as mills (shown in Table 2.3).

Differences in steam-use across the manufacturing sector can reflect both a direct effect, from restricted access to water power, and an indirect agglomeration effect from local complementarities in steam adoption. Lumber and flour milling were leading sectors for steam adoption, given their heavy initial reliance on mechanical power. Earlier steam-use by some agents could plausibly hasten steam adoption in the broader economy, given more-limited general knowledge of steam engine technology.[26] Installation and operation of steam power was not an off-the-shelf process; rather, steam was a more complicated and volatile technology, whose use might plausibly depend on the local knowledge base and, in turn, whose use might plausibly affect the local knowledge base. Delayed steam adoption by mills, in places with more waterpower availability, may have then held back steam adoption in local manufacturing more broadly.

One mechanism for these agglomeration effects is backward linkages in manufacturing of steam equipment: steam-use encouraging local manufacturing of steam equipment, which in turn encourages others to use steam power. Most manufacturing establishments purchased equipment from local manufacturers (Woodbury Report, 1838; Temin 1966), and a quarter of steam equipment manufacturers also report repair services in the Census of Manufactures, which highlights the importance of a local technical knowledge base.

Table 2.6, Column 3, shows that counties with lower waterpower potential had more manufacturers of steam engines, boilers, and related equipment (relative to all manufacturing establishments). The overall manufacturing sector was smaller in lower waterpower counties, but for manufacturing establishments in these counties there was a greater density of steam equipment makers to support steam adoption. This is consistent with the demand for steam power helping to create its own supply.

26. Indeed, Franck and Galor (2021, 2022) argue that an important driver of the spread of steam power in France was distance to Fresnes-sur-Escaut, the location of the first commercial steam engine in the country.

2.3.3 Potential Other Forces Driving Steam Adoption

Increases in local demand could have encouraged adoption of steam power in counties with lower waterpower potential, including steam power making these counties more attractive for a variety of activities that increase local demand for milling (Benhabib and Rustichini, 1993). Appendix Table 2.15, Column 1, shows that counties with lower waterpower potential experienced faster population growth during this period (7% to 10% per decade), but population is not driving our estimates on steam adoption. While counties with lower waterpower potential had a higher share of mills using steam power (Table 2.2) in 1850, but had lower population in 1850 (Appendix Table 2.15). Further, Appendix Table 2.15 shows that lower water power counties experienced increases in milling activity even in per capita terms. Our estimates from Table 2.4 are also inconsistent with population growth driving our results: if county growth were being driven by more customers, it would be difficult to rationalize the decreased survival of incumbents.

In Appendix Tables 2.16, 2.17, and 2.18, we show our results are similar when constraining the sample to only flour, which was less technologically less-tradable than lumber at the time due to its perishability. In Appendix 2.G, we explore the robustness of our results to controlling for a variety of other features of the economic environment that may have had direct effects on steam adoption or general effects on economic activity. We summarize our approach below.

Geographic variation in waterpower potential could be correlated with other factors affecting economic activity, in levels or in changes, and in Appendix Tables 2.19 and 2.20 we consider how our results change when controlling for alternative local factors. In Appendix Table 2.19, we show that our results are robust to including various characteristics that have been discussed as important drivers of steam power adoption across different contexts (Crafts, 1977; Floud and McCloskey, 1981; Allen, 2009; Mokyr, 2016): alternative measures of access to coal (Wrigley, 2010; Fernihough and O'Rourke, 2021; Reichardt, 2023); agricultural productivity and woodland that affect mills' material input availability (Ragnar, 1953); differences in labor availability reflected in manufacturing wages (Habakkuk, 1967; Allen, 2009) and mechanics and engineers (Hanlon, 2022), though also potentially outcomes of mills' steam adoption; capital availability through banks (Jaremski, 2014); and all of the above controls.

In Appendix Table 2.20, we show that our results are robust to other adjustments to our controlling for features of counties' economic environment. First, we show our results are robust to removing some or all of our controls for access to markets or coal. Our results are robust to controlling for time-varying market access and population, which are themselves potentially endogenous to steam adoption, or growth associated with counties' fixed 1850 population. Some estimates are smaller when controlling for population, but this also introduces bias because county population is endogenous to local waterpower potential (even in 1850). Our results are robust to controlling for alternative sources of potential growth: an indicator for being in Appalachia or on the frontier (Bazzi et al., 2020), the share of workers in agriculture (Eckert and Peters, 2023), having a portage site (Bleakley and Lin, 2012), exposure to the Civil War, and all of these time-invariant controls interacted with decade.

Our analysis focuses on county-level geographic variation in waterpower availability, though there could also potentially be within-county differences in location advantages for steam power. One salient locational characteristic could be the distance to the closest railroad, which was a source of fuel imported from other counties. We digitized historical maps of railroad station locations, and found locational variation within and between counties. Some counties had water power sites close to stations and in others they are far away, which could lead to differences across water incumbents in the feasibility of switching to steam power and therefore a potential source of technological lock-in. Nevertheless, Appendix Table 2.21 shows that distance to railroad station is not an additional substantive source of variation in steam suitability: it does not predict steam-use, water incumbents switching to steam, or a differential response of entrants versus incumbents.

2.3.4 Robustness to Linkage Error

A natural question is how much our estimates might be affected by measurement error, particularly errors in the construction of our panel links. For our main results, we invested in a resource-intensive approach that used hand-links, but there are inevitably false negatives and false positives in the links. The hand links are binary, such that mills are either linked or they are not. To create a measure of confidence for any given link, we train a supervised machine learning algorithm on the hand-made links (see Appendix 2.C.4 for details). We then use the estimated linking probabilities to explore the the quality of hand-links, and the sensitivity of our estimates to adding panel mills that were almost linked, or removing those for whom the links are less predictable.

Appendix Figure 2.18 Panel A shows the predicted match probability for the hand-links. For mills whose sector and ownership structure were unchanged from one decade to the next, the hand-links are very predictable: most match probabilities are above 0.8. For mills that changed milling sector (e.g. flour-to-lumber), and especially for mills that gained or lost some owners, the match probabilities are lower but still mostly above 0.5. For our regression analysis, a primary concern would be that linkage errors are correlated with county waterpower potential. Appendix Figure 2.18 Panel B shows that the distributions of predicted match probabilities are similar for mills in counties with low and high waterpower potential.

One advantage of the ML model for robustness analysis is that we can change the matching cutoff, which mechanically changes the firm survival rate along with the rate of false-negative and false-positive matches. Appendix Figure 2.19 shows how raising the cutoff lowers the share of ML links that are not hand-links (the "false match" rate, akin to a false discovery rate) but also lowers the share of hand-links that are made by the ML model (the "found match" rate, akin to the sensitivity). Our baseline machine-learning links use a predicted match probability of 0.6 as the benchmark cutoff for classifying a mill as surviving from one decade to the next, which is close to maximizing the "found match" rate while keeping the "false match" rate relatively low. Appendix Table 2.22 shows that with this cutoff, the survival rate is higher using the ML-links (compared to the hand-links), as many mills are only classified as surviving using the ML model. Most hand-links (67%) are also predicted by the ML model. Conditional on finding a match, it is rare that the ML-links and hand-links disagree on the identity of the match.

Appendix Tables 2.23 and 2.24 show that our results are not sensitive to changing the sample to include more- or less-confident matches based on the ML-link probabilities. Our results are similar if we restrict our panel sample to those mills linked by hand *and* the baseline ML model, rather than our main sample of hand-links, or use *only* the benchmark ML-links. Using the ML-links only, the results are also similar if we raise or lower the benchmark cutoff of 0.6 for classifying matches.

A useful feature of our approach is we classify whether mills have a "business name" (such as the "Rock Creek Mill") or whether mills are named after their proprietors (and might therefore be differentially subject to linkage error). Our estimates are similar when considering each type of mill separately.

We also explore potential measurement error in the type of power source recorded for mills, which is based on Census enumerator visits to the mills. The original manuscripts contain some corrections, with scratched out and re-written information by an occasional second enumerator, so the final recorded data could also differ in some cases from mills' actual operations. For instance, we searched in historical records for mills that reported power sources other than water and steam — in particular, some suspiciously large mills without reported mechanical power — and found that these mills often did actually use water or steam power. Some report "horse" as a power source, without further detail, which probably often represents water or steam power rather than horse-powered mills. We cannot systematically correct these mills' recorded power use, so our baseline estimates exclude these mills; but as there are few of these mills, our results are not sensitive to including them as non-steam powered mills.

Our main analysis restricts the sample to the panel of counties with at least one mill in 1850. In Appendix Table 2.25, we show that our results are similar when including different sets of counties: expanding the sample to include all counties that ever had a mill, or limiting the sample to counties with multiple mills in 1850. Our estimates are also not sensitive to dropping large county groupings, made in the construction of geographically-consistent counties, which potentially misclassify local waterpower availability, or the counties with extreme local waterpower potential. Our results are also similar if we exclude counties that were more involved with cross-county or international trade in mill output: the 20 largest cities at the time, or places that Kuhlmann (1929) describes as having "merchant mills" that exported their output.

It is important to use our geographically comprehensive measurement of waterpower potential. Because the 1880 Water Census effectively selected on the dependent variable (by omitting places with lower waterpower potential *and* lower water power use), we would expect estimates based on the 1880 Water Census to be biased toward zero, which we confirm when looking at the number of water powered mills in 1850 (Appendix Figure 2.12 Panel B) or 1850-1880 growth in mills (Panel C).

2.4 Key Empirical Patterns

Overall, there is a stable relationship between greater waterpower potential and slower adoption of steam power. The differences in steam adoption rates among entrants and incumbents is suggestive of technological lock-in, with barriers to steam adoption for those establishments that had been using water power. We use a quantitative model to estimate the magnitude of this lock-in and its implications for aggregate manufacturing outcomes given firm entry and exit. The model estimation draws on these estimated differences by county waterpower potential. The model also reflects other features of the economic environment, such as the costs and benefits of using steam power, which we describe further in Section 2.5. In this section, we describe several empirical patterns to motivate the model's structure, and which provide moments in the model's estimation.

Our view of the technological transition from water to steam is motivated by the following intuition. Each technology was associated with marginal costs and fixed costs (where fixed costs include both purchase and overhead costs). Because neither technology was clearly more attractive to millers, we model steam power as better on one cost dimension and water power as better on the other cost dimension. To distinguish which technology has which features, we use a logic in the spirit of Melitz (2003) (see also Olmstead and Rhode 2001; Cabral and Mata 2003, and Bustos 2011).

Millers have different productivities, for instance due to their ability to attract customers, manage suppliers, and operate the machinery (Huntington et al., 2023). Holding fixed productivity, firms will be larger if they use the lower marginal cost technology. For a given power technology, more-productive firms will have higher sales. More-productive firms are then more likely to prefer the high fixed cost and low marginal cost technology, because they can amortize the fixed costs over more units. Combined, this means that the technology associated with larger firms is the one with lower effective marginal costs. We compare a variety of firm size distributions and use a similar logic to study how the costs of steam and water power varied over space and time. Characterizing these size distributions relies on our digitization of the micro-level Census data, as these economic patterns were previously unknowable from aggregated tabulations or smaller samples of micro-data without firm names or panel links.

2.4.1 Cost Structures for Steam and Water

Figure 2.6 shows that steam powered mills were larger than water powered mills, on average. Given the Melitz (2003)-style logic discussed above, this implies steam power has higher fixed costs and lower marginal costs than water power.

This implication requires some further interpretation, though, as steam power, unlike water power, requires daily expenditure in order to access mechanical power. The empirical patterns reflect the realities of running steam engines and waterwheels. Even small steam mills employed full time engineers and firemen. To avoid ramping costs, mills used a relatively consistent amount of fuel to keep their engines on throughout the day (Fisher, 1845; Swain, 1888). As a result, many of these costs were fixed overhead costs, not marginal costs, which is in turn reflected in the firm-size distribution.

Furthermore, the *effective* marginal costs of water power were higher than their *inframarginal* variable costs. Waterwheels were limited by their local geography: the size, speed, seasonality, and reliability of their local waterway, as well as contractual water rights. The data reflect not only the actual monetary expenditure for the marginal power use for water mills, but also the shadow costs associated with expansion. Some water powered incumbents did grow (Appendix Figure 2.20), so water powered mills were not completely constrained, but expanding production further could require increasingly expensive modifications to their operations. On average, the water incumbents who stuck with waterpower expanded their horsepower capacity by 7%, and those who switched to steam power expanded their capacity by over 50%.

Finally, the relevant marginal costs are those of *production*, not of *power* alone. Appendix Figure 2.17 shows that steam mills had access to more power than water powered mills, lowering the non-power marginal costs of steam powered mills (for instance, because the mill could process more inputs per hour (Evans, 1795; Dedrick, 1931)).

To support the interpretation that steam powered mills had lower marginal costs, we analyze prices. Due to data constraints, we are only able to study prices for single-product lumber mills in 1880, but indeed find that steam use predicts lower output prices (by 6%).

Figure 2.6 shows that the size distributions for steam and water powered mills converged over time. This suggests a corresponding decline in the fixed cost of steam power, as less-productive firms started to find steam power more attractive, whereas a declining marginal cost of steam power would have increased the size premium of steam powered mills. This is consistent with the importance of the development of high-speed engines that reduced steam fixed costs for lumber and flour mills.[27]

One potential explanation for these results could be that steam power shifted activity to new locations that, for unrelated reasons, had mills of different sizes. This geographic shift is not driving our results, though: Appendix Figure 2.21 shows firm-size distribution patterns we find are similar within-counties (for counties with both types of mills).

For local waterpower potential to make water power use more attractive to firms (as in Figure 2.12 Panel B), it must have lowered the fixed costs or marginal costs of using water power. If waterpower potential lowered the marginal costs of water power, then counties with higher waterpower potential would have larger water powered mills (and, due to the resulting selection, also larger steam powered mills). Figure 2.7 shows this was not the case and, indeed, somewhat the opposite: in most decades, counties with higher waterpower potential have more small mills. Thus, we model county waterpower potential as lowering the time-invariant fixed costs of water power, such as the costs of water rights and constructing millponds.

Congestion was not an important force driving differences in steam power in the United States (Gordon, 1983). In our data, counties still had substantial available waterpower capacity.[28] Further, Table 2.5 shows that water incumbents are more likely to switch to

27. Figure 2.6 shows that the convergence of firm size distributions is partially driven by the left tail of low-productivity water mills disappearing over time. In our model, increasing competition (driven by the spread of steam power) crowded out the least productive water mills. Collard-Wexler and De Loecker (2015) document a similar pattern in US steel manufacturing during the spread of the minimill.

28. The median county used less than 10% of the available waterpower potential, and over 95% of counties

steam in places with lower waterpower potential. If the increased adoption of steam power was driven by difficulties finding available water power sites, water incumbents would be unaffected.

Figure 2.6 also shows there was substantial overlap in the size distributions of steam and water powered mills in every decade. This suggests a substantial idiosyncratic component to mills' technology adoption. One natural candidate for this heterogeneity is the preferences and talents of firm owners. Linking the Censuses of Manufactures and Population, Appendix Table 2.28 shows that owners who were immigrants or younger were more likely to use steam power, highlighting the role of owner characteristics for technology adoption.[29]

2.4.2 Operating Costs

We calculate that 19-24% of mills survived from one decade to the next (Appendix Table 2.26).[30] Firm exit implies that dynamic incentives are important, as only some firms successfully amortize their fixed costs of entry and technology adoption over a long time period.

Appendix Figure 2.22 shows that, on average, surviving firms are larger than exiting firms. This suggests a fixed cost of production in every period, with an additional idiosyncratic component, to rationalize the correlation between firm exit and initial size. Water incumbents were also more likely to survive than steam incumbents, consistent with explosions and the additional operating costs associated with steam power.

---

used less than half of the available waterpower potential. Hunter (1979) and Gordon (1983) report that standard estimates of waterwheel efficiency in the era were at least 50–70%.

29. McElheran et al. (2023) find that younger owners are more likely to adopt artificial intelligence technologies.

30. This implies an annual exit probability of around 15%, higher than modern annual exit probabilities of around 8% (Foster et al., 2016).

2.4.3 Barriers to Switching Technologies

Entrants' decisions to adopt steam is a useful contrast to incumbents' decisions, as entrant firms started with a clean slate. Figure 2.5 shows that entrants were four times more likely to use steam power than water incumbents.[31] The difference in steam adoption rates is not driven by differences in firm size and is slightly larger when conditioning on firm size (Appendix Table 2.27).

This difference in steam adoption rates, between entrants and incumbents, suggests there are barriers to switching from water to steam power. A barrier to switching technologies also causes only the highest-productivity water incumbents to adopt steam, while relatively lower-productivity entrants would use steam. Consistent with this logic, Appendix Figure 2.23 shows that incumbents are larger than entrants within each power technology: on average, incumbents are 20% larger when using water and 40% larger when using steam.

Switching barriers were not infinite, however, as both entrants and incumbents were more likely to adopt steam power over time. This is consistent with the technological improvements in steam power. Over the course of our sample, steam adoption rates increased by sixty percent for both entrants and water incumbents, from a base rate of thirty percentage points for entrants and 8 percentage points for water incumbents (Figure 2.5).

2.4.4 Alternative Reasons for Lower Steam Adoption among Incumbents

While the data patterns are consistent with fixed barriers to switch power technologies, we also consider several alternative explanations for the serial persistence in firm technologies.

Across different contexts, one leading alternative explanation for low technology switching by incumbents is learning-by-doing (Jovanovic and Nyarko, 1996). The idea would be that water powered incumbents could have freely adopted steam, but did not want to because

31. A few firms report switching from steam to water, which is rare enough that we do not report separate statistics for these firms, though we do include these firms when we estimate the model in Section 2.5.

they had learned to use water power and, for them, it continued to dominate steam. For this context, high rates of learning-by-doing for water power would be inconsistent with the longstanding use of water power in the US, but we can explore this further in the data.

Learning-by-doing would imply that water incumbents experience relatively fast growth, as they benefit both from learning and any other general economic changes that would increase firm size. To test for learning in the spirit of Bahk and Gort (1993), we compare the growth rate of water incumbents who keep using water power to the growth in the firm-size distribution for entrants over the same time period. Appendix Figure 2.24 shows that incumbents and successive generations of water powered entrants "grow" at a similar speed, consistent with no additional learning-by-doing boost for water incumbents.

We consider switching barriers as equivalent to an expenditure (a combination of the opportunity cost of scrapping a functional power source and other actual costs such as retrofitting). An alternative modeling approach could assume a *productivity* cost from switching technologies (see, e.g., Parente and Prescott 1994). For our context, productivity losses seem implausible, because most of the day-to-day operations of milling are the same with either power source. In the data, Appendix Figure 2.20 shows that switchers grow faster than stayers, which is not consistent with productivity losses from switching. Indeed, even though water incumbents were initially negatively selected (because only firms with relatively low initial productivity chose water power), those that switched to steam power were 2.6% larger than steam entrants.

Another potential reason why incumbents would not switch technologies is permanent unobserved heterogeneity (i.e., "steam types" and "water types"). Appendix Table 2.28 does show some specific examples of persistent firm heterogeneity (for instance the immigration status of the owner), but we do not include it in the model for a variety of reasons. First, the results from the owner-linking analysis are not quantitatively important on aggregate. While immigrant owners are much more likely to use steam power, they are a small share of overall millers. The effect of age is relatively small. Appendix Table 2.28 also shows that professional millers were more likely to use steam power, but this is not a permanent type.

Other features of the data also suggest that permanent idiosyncratic variation in costs and productivity is not driving the main data patterns. Appendix Figure 2.20 shows that firms' revenue grew more when they switched, which is not a general prediction of models with persistent types, but is a prediction of a model with switching barriers (as only the mills with productivity growth would choose to change technologies). Historical accounts of mills also discuss instances of mills switching technologies after a fire destroyed their original structure (Appendix 2.F), which suggests owners do not persistently prefer a particular technology, but instead face sunk fixed costs or other barriers to switching (Hornbeck and Keniston, 2017; Huesler and Strobl, 2023).

We can also use the timing of mills' water use and steam use to compare the implications of switching barriers that generate state-dependent technology choices against the implications of heterogeneous types. Methods of quantifying the importance of state dependence versus types require observing agents for many periods (Lancaster and Nickell, 1980; Chamberlain, 1985; Dano, 2023), whereas we observe mills for a maximum of four census rounds (and normally fewer). We provide two alternative tests, in the spirit of Chay et al. (1999), which are inconsistent with the presence of types driving relatively low switching rates.

One test of state dependence is to examine firms' technology choices, conditional on their prior use of water and steam power. Consider the sample of mills over four periods who start with water power, end with steam power, and use steam power exactly twice. These mills use steam power half of the time, and all have the same initial and final conditions (as in Hotz and Miller, 1993; Arcidiacono and Miller, 2011). Switching barriers would make it substantially more costly for these firms to alternate between technologies twice, as opposed to using water for two periods and then steam for two periods. By contrast, under heterogeneous types, switching is driven by period-specific idiosyncratic shocks, so each pattern would be equally likely. In our data, the vast majority of these mills switch technologies only once and then keep their new technology, which suggests switching barriers are driving technological choice.

The second test is based on the logic that under persistent heterogeneity, a Bayesian observer would update that a water powered incumbent who previously also used water power would be more likely to be a "water-type" than a water entrant, since the former chose water power multiple times. This would subsequently imply that the water incumbent stayers would be more likely to use water power than water entrants in subsequent decades, but this is not what we find in the data.

Another potential source of differences in technology use for entrants and incumbents could be differences across locations, if the (new) steam users locate in different places than the (pre-existing) water users. We find significant differences in adoption choices within counties, however, and we compare technologies choices within county-industries when estimating the relevant moments in Section 2.6.

Finally, we do not observe if entrants build their own mills, or if they purchase used mills. The resale of water infrastructure would generally attenuate the differences between entrants and incumbents: if persistent county-level infrastructure were important to the choices of entrants, then they too would face opportunity costs of using steam, and they would not be substantially more likely to use steam power than the water incumbents. Nevertheless, Table 2.5 shows that entrants are particularly more likely to use steam power in places with lower waterpower potential, which have relatively higher exit of waterpowered establishments.

In Summary: Steam power allowed firms to scale production at lower effective marginal costs, which required higher fixed costs. Those fixed costs declined over our sample period as steam technology improved. Both water and steam required fixed overhead costs, and millers faced some cost of switching power technologies. Counties with higher waterpower potential used relatively less steam power, due to their continued access to water power (direct effects of geography) and their previous use of water power (dynamic effects of geography, through technological lock-in). We now turn to a formal framework that fits these estimates and quantifies the influence of technological lock-in.

2.5 A Model of Steam Adoption

It is difficult to interpret all of the estimates jointly – the empirical patterns along with the estimated differences by waterpower potential – with only economic intuition. One main purpose of the model is to collect and synthesize the magnitudes of these different relationships. Further, the structural model allows us to evaluate how switching barriers – and policies aimed to alleviate them – matter for the aggregate spread of new technologies.

We develop a dynamic equilibrium model of technology adoption and firm entry. In the model, firms face a dynamic power source choice. The key tradeoff is that water power has a lower fixed adoption cost than steam, but a higher marginal cost that inhibits higher production levels. The only primitive that varies across counties is the cost of adopting water power. The only primitive that varies across time is the purchase price of steam. A falling price of steam power drives steam adoption but also incentivizes forward-looking firms to wait to adopt. The barriers to switching from water to steam power encourage firms to enter using steam. In this section, we describe the formal setup of the model.

2.5.1 Static Choices: Production and Demand

Each firm j in county c in year t maximizes its static profit by choosing its optimal levels of variable inputs x_{jct} and price p_{jct}, given its power source R, its baseline productivity φ, and the choices of other firms.

We assume all demand for mill products takes place locally and takes a nested CES form.[32] The price index P_{ct} equals $\left[\int p_{jct}^{1-\epsilon}\, dj\right]^{\frac{1}{1-\epsilon}}$, where ϵ is the elasticity of substitution

32. Appendix Table 2.14 shows that the competitive pressure from steam entrants has similar effects on the exit probabilities of steam and water incumbents. This result is consistent with entry raising competitive

across mills' products. Local demand for mill output Y_{ct} equals ${P_{ct}^{-\eta}}$, where η is the elasticity of demand for mill products. If firm j charges price p_{jct}, its quantity sold is: $y_{jct} = p_{jct}^{-\epsilon} P_{ct}^{\epsilon-\eta}$.

Firms produce using a constant-returns-to-scale technology in flexible inputs x (labor and materials), which are elastically supplied at a price w:

$$y_{jct} = \exp(\varphi_{jct} + \gamma_{R_{jct}} + \alpha_{R_{jct}} s_{ct}) x_{jct}. \tag{2.3}$$

Firms' overall productivity is determined by their baseline productivity φ_{jct} and an additional $\gamma_{R_{jct}}$ from their power choice R, which is either water (W) or steam (S). We normalize $\gamma_W = 0$ so $\gamma_S = \gamma$. The productivity boost from steam power is also a function of contemporaneous local steam usage ($\alpha_S s_{ct}$), where s_{ct} is the share of firms using steam and α_S is the strength of this agglomeration force.[33] Agglomeration effects (α_S) could reflect that increased local steam use generates greater local human capital in steam production.

Firms buy inputs x to maximize flow profits. Their price, output, and profit functions are:

$$p_{ct}(R,\varphi) = \frac{\epsilon}{\epsilon-1} \frac{w}{\exp(\varphi + \gamma_R + \alpha_R s_{ct})}, \tag{2.4}$$

$$y_{ct}(R,\varphi) = {P_{ct}}^{\epsilon-\eta} \left(\frac{\epsilon}{\epsilon-1} \frac{w}{\exp(\varphi + \gamma + \alpha_R s_{ct})} \right)^{1-\epsilon}, \tag{2.5}$$

$$\pi_{ct}(R,\varphi) = \frac{1}{\epsilon} {P_{ct}}^{\epsilon-\eta} \left(\frac{\epsilon}{\epsilon-1} \frac{w}{\exp(\varphi + \gamma_R + \alpha_R s_{ct})} \right)^{1-\epsilon}. \tag{2.6}$$

The next section describes how firms choose if they produce and with what power choice.

pressure by lowering the aggregate price index, and more consistent with monopolistic competition than a Bertrand model in which the initially water powered mills would be especially unable to match the low prices of the steam mills.

33. We normalize the agglomeration force in water power to zero, such that α_S captures the net agglomeration force in steam power.

2.5.2 Dynamic Choices: Firm Entry and Power Choice

We model a firm's dynamic choices in four stages (Hopenhayn, 1992; Melitz, 2003; Chernoff, 2021). In Stage 1, prospective entrants decide if they want to pay a fixed cost and enter the economy. In Stage 2, entrants draw their productivity φ_{jct} and incumbents update their productivity. In Stage 3, firms choose if they want to exit, given their revealed productivity and fixed operating cost. In Stage 4, surviving firms select their optimal power source and produce. After these four stages, the cycle starts over again. For the initial stages, we consider the possible power states to be E, W, or S (respectively for entrant, water, or steam). Entrants need to adopt water or steam power to produce in the final stage.

Stage 1: Entry. A prospective firm enters in county c in year t if its expected continuation value upon entry exceeds the fixed cost of entry:

$$\mathbb{E}_{\varphi}\left[V_{ct}(E, \varphi)\right] \geq f^{e}, \tag{2.7}$$

where $V_{ct}(E, \varphi)$ is the continuation value for an entrant.

Stage 2: Updating Baseline Productivity. The productivity of an incumbent mill j, φ_{jct}, follows an AR(1) process:

$$\varphi_{jct} = \pi\varphi_{jt-1} + \sigma\xi_{jt}, \tag{2.8}$$

where π and σ are parameters that represent the persistence and dispersion of latent productivity φ. Entrants draw their productivity from the stationary distribution of the same AR(1) process.

Stage 3: Sinking the Operating Cost. All firms pay a common deterministic operating cost f_o^R, given their power source $R \in \{E, W, S\}$. Furthermore, each firm j pays an idiosyncratic cost $\nu_{jct}^R(0)$ if it continues its operation, and $\nu_{jct}^R(1)$ if it chooses to exit. Each firm compares the expected value from paying the operating cost to the value from exit:

$$V_{ct}(R,\varphi) = \max\{\mathbb{E}_{\varepsilon}\left[V_{ct}^{o}(R,\varphi)\right] - f_{o}^{R} - \nu_{jct}^{R}(0), \Omega_{ct}^{R} - \nu_{jct}^{R}(1)\}, \tag{2.9}$$

where $V_{ct}^{o}(R,\varphi)$ is the continuation value after sinking the operating cost and Ω_{ct}^{R} is the resale value of technology R.

Stage 4: Choosing a Power Source and Producing. Having paid its fixed operating cost, each firm chooses its optimal power source as a function of adoption costs, switching barriers, and expectations over future productivity. The value function for an establishment with power source R and productivity φ is:

$$V_{ct}^{o}(R,\varphi) = \max_{R' \in \{W,S\}} \{\pi_{ct}(R',\varphi) - c_{ct}(R,R') - \varepsilon_{jct}(R') + \delta\mathbb{E}_{\varphi'}\left[V_{ct+1}(R',\varphi')\right]\}. \tag{2.10}$$

$\pi_{ct}(R,\varphi)$ is the firm's static profit from Equation (2.6), δ is the discount factor, and $\mathbb{E}_{\varphi'}\left[V_{ct+1}(R',\varphi')\right]$ is the expected continuation value given the law of motion for productivity in Equation (2.8). For each power source, the firm draws an idiosyncratic usage cost $\varepsilon_{jct}(R)$. To give some examples of idiosyncratic costs, Swain (1888) describes some millers preferring water power due to its "greater cleanliness, less annoyance, and less area required." If the firm chooses to change power sources, the firm pays $c_{ct}(R,R')$ to switch from power source R to power source R'. The firm then produces, charging the profit-maximizing price described in Equation (2.4).

2.5.3 Equilibrium

Firms make forward-looking decisions anticipating improvements in steam power and the competition from other firms in their local product market. For example, while lower steam costs create an option value for incumbents to switch to steam, these firms understand that cheaper steam may also induce other firms to enter, adopt steam, and compete for customers.

We study the local economies along their transition path as steam power becomes available at lower costs.

Definition 1 (Dynamic Equilibrium)**.** An equilibrium for county c is a time path for the mass of entrants M_{ct}, the mass of operating firms $F_{ct}(R, \varphi)$, and the policy functions for operation/exit $O_{ct}(R, \varphi)$ and power $R'_{ct}(R, \varphi)$, taking the time path of steam costs $c_{ct}(S)$ as given, such that:

(i). Firms enter, exit, and adopt power sources to maximize expected discounted profits (Equations (2.7), (2.9), and (2.10)).

(ii). Firms source inputs x to maximize flow profits period-by-period (Equation (2.6)).

(iii). Output markets clear:

$$P_{ct} Y_{ct} = w X_{ct} + \Pi_{ct}, \tag{2.11}$$

where $\Pi_{ct} = \int \pi_{ct}(R, \varphi) dF_{ct}(R, \varphi)$ are total local profits, and $X_{ct} = \int x_{ct}(R, \varphi) dF_{ct}(R, \varphi)$ is local demand for inputs.

(iv). The free entry condition holds:

$$\mathbb{E}_{\varphi} \left[V_{ct}(E, \varphi) \right] \leq f^e. \tag{2.12}$$

(v). The evolution of firm masses $\{F_{ct}\}_t$ is consistent with the policy functions $\{O_{ct}, R'_{ct}\}_t$.

2.5.4 The Arrival of Steam

We initiate the model in 1830, before steam power became broadly available to mills in the US. We assume the economy was in a steady state before steam, with differences across counties reflecting their different water costs.[34] In 1830, firms receive the news that steam will become increasingly available. After the surprise of steam power, firms have perfect foresight about the path of falling steam costs.[35] In particular, steam power first becomes purchasable at a high price in 1830, and its fixed adoption cost then monotonically declines until reaching its steady-state level in 1900.[36]

The falling steam cost is the only driving force along the transition path. In particular, we assume water technology is comparatively unchanged over this period, as it was a comparatively mature technology. Rosenberg and Trajtenberg (2004) estimate that horsepower per waterwheel was largely stable over time.

2.5.5 Parametric Assumptions

We make a series of parametric assumptions to solve and estimate our model.

Firm operating/exit costs are drawn from a Gumbel distribution with dispersion parameter ρ_O, and the adoption costs for each power source are drawn from Gumbel distributions with dispersion parameter ρ:

$$\nu^R_{jct}(\text{OPERATE/EXIT}) \overset{\text{iid}}{\sim} \mathit{GEV1}(\rho_O) \tag{2.13}$$

$$\varepsilon_{jct}(R) \overset{\text{iid}}{\sim} \mathit{GEV1}(\rho). \tag{2.14}$$

These distributional assumptions follow Rust et al. (1987).

34. The Census of Manufactures was professionalized and comprehensive beginning in 1850 (United States Census Bureau, 1900; Atack and Bateman, 1999), after the first introduction of steam power. Because some firms were already using steam, we cannot use the start of our data (1850) as the steady state before steam power. Instead, we initiate the model simulations in 1830, when very few steam engines were used in US milling (Woodbury Report, 1838), and estimate the model to match steam adoption from 1850 to 1880.

35. Humlum (2022) adopts a similar approach to modeling the arrival of robots in modern manufacturing. While we do not have measures of millers' expectations, contemporaneous accounts of steam technology are consistently optimistic about the potential for future improvements (e.g., **?**).

36. Steam power reached its peak adoption in US manufacturing around 1890-1900 (Jovanovic and Rousseau, 2005), prior to the large-scale arrival of electricity in milling (Fenichel, 1966).

The productivity innovations are drawn from a standard-normal distribution

$$\xi_{jt} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{2.15}$$

which implies that entrants draw their productivities from the normal distribution

$$\varphi_{jct} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{(1-\pi)^2}\right). \tag{2.16}$$

The resale value of each technology (Ω_{ct}^{R}) is a share of the current purchase price:

$$\Omega_{ct}^{R} = \omega^{R} c_{ct}(R). \tag{2.17}$$

The costs of switching power sources reflect buying prices, resale values, and other costs:

$$c_{ct}(R,R') = \begin{cases} 0 & \text{if } R = R' \\ c_{ct}(R') & \text{if } R = E \\ c_{ct}(R') + c(R,R') - \Omega_{ct}^{R} & \text{otherwise.} \end{cases} \tag{2.18}$$

Mills keeping their existing technology do not pay any further costs. Mills purchasing technology R' have to pay a fixed purchase price $c_{ct}(R')$. Switchers face two additional forces. First, incumbents face an additional switching cost to change power sources, $c(R,R')$, which captures all costs of changing technologies. Second, incumbents may sell their pre-existing technology (if $\omega^{R} > 0$), though the scrap value may not be equal to the purchase price of their old technology (Bertola and Caballero, 1994; Ramey and Shapiro, 2001).

We parameterize the fixed cost of steam adoption declining over time as follows:

$$c_t(S) = \kappa s_{ct} + c_S^{(initial)} + (c_S^{(terminal)} - c_S^{(initial)}) \exp\left(-c_S^{(slope)}(t - T_0)\right), \tag{2.19}$$

where the cost at period T_0 is $c_{T_0}(S) = c_S^{(initial)}$, and $\lim_{t\to\infty} c_t(S) = c_S^{(terminal)}$. This set-up implies that the price of steam varies over time but not space. Conversely, the price of water power varies over space due to local waterpower potential, but does not vary over time. Finally, we allow the price of steam power to be a function of local steam use (κ), capturing the potential for agglomeration (or congestion) in power adoption, such as information sharing and limited local access to the relevant capital.[37]

Given these distributional assumptions, the firm-level expected continuation value is:

$$\mathbb{E}_\nu[V_{ct}(R,\varphi)] = \rho_o \log\left[\exp\left(\frac{\Omega_{ct}^R}{\rho_o}\right) + \exp\left(\frac{\mathbb{E}_\varepsilon\left[V_{ct}^o(R,\varphi')\right] - f_o^R}{\rho_o}\right)\right], \tag{2.20}$$

while the expected continuation value after sinking the operating cost is:

$$\mathbb{E}_\varepsilon[V_{ct}^o(R,\varphi)] = \rho \log\left[\sum_{R'\in\{W,S\}} \exp\left(\frac{1}{\rho}(-c_{ct}(R,R') + \pi_{ct}(R',\varphi) + \delta\mathbb{E}_{\varphi'}[V_{ct+1}(R',\varphi')]\right)\right]. \tag{2.21}$$

The probability of exit, given the existing power source R and the baseline productivity φ, is:

$$\Pr_{ct}\left(\text{OPERATE/EXIT}|R,\varphi\right) = \frac{\exp\left(\frac{\Omega_{ct}^R}{\rho_o}\right)}{\exp\left(\frac{\Omega_{ct}^R}{\rho_o}\right) + \exp\left(\frac{\mathbb{E}_\varepsilon[V_{ct}^o(R,\varphi')] - f_o^R}{\rho_o}\right)}. \tag{2.22}$$

The conditional probability of choosing power source $R' \in \{W,S\}$, given a mill is starting

37. While we formally model κ as affecting the price of steam power, it also functionally serves as a local shifter for the *relative* price of steam. For instance, if the price of local water power falls in the local use of steam, due to a move along the supply curve for water power (in the spirit of Hansen and Prescott 2002), we would estimate a positive κ.

with power source R, is:

$$\Pr_{ct}\left(R'|R,\varphi\right) = \frac{\exp\left(\frac{1}{\rho}(-c_{ct}(R,R') + \pi_{ct}(R',\varphi) + \delta\mathbb{E}_{\varphi'}[V_{ct+1}(R',\varphi')])\right)}{\sum_{R''\in\{W,S\}}\exp\left(\frac{1}{\rho}(-c_{ct}(R,R'') + \pi_{ct}(R'',\varphi) + \delta\mathbb{E}_{\varphi'}[V_{ct+1}(R'',\varphi')])\right)}. \tag{2.23}$$

2.5.6 Solution Algorithms

The equilibrium for each economy is a complicated fixed point: heterogeneous firms make forward-looking decisions about entry, exit, and power adoption, and firms' decisions are interlinked through their competition in local product markets and agglomeration spillovers in steam power choices. We study the transition path of the economy, where falling steam costs drive the transition from water to steam power.

Appendix 2.H describes our solution algorithms. In brief, we solve firms' dynamic programs by combining value function iteration (in the steady states) with backward recursion (along the transition path). We solve the dynamic equilibrium using a fixed-point shooting algorithm in the aggregate state variables.

Existence and Uniqueness

Appendix 2.H discusses the properties of our solution algorithm, including the existence and uniqueness of the equilibrium. The convergence of our iterative algorithm is ensured by a congestion force due to competition in the product market, which in turn ensures the existence of an equilibrium. The congestion force behind the convergence property also tends to make the equilibrium unique. Strong steam agglomeration forces (κ and α_S) could, however, lead to multiple equilibria: a "low steam" equilibrium where few mills adopt steam (because the net agglomeration force is weak) and a "high steam" equilibrium where many mills use steam (because the net agglomeration force becomes strong). We verify that multiple equilibria are not present in our terminal steady state (when steam power is fully available and more firms are at the margin of steam use) by initiating our solution algorithm at different starting values for the equilibrium steam share.

2.6 Structural Estimation

In this section, we describe the quantification of the model developed in Section 2.5. We consider two counties: a baseline county with the average amount of water power in the United States, and a "lower waterpower" county with one standard deviation less waterpower potential. We assume that the only fundamental difference between the counties is the cost of water power $c_c(W)$. This structural modeling mirrors the identifying assumption in our reduced-form analysis in Section 2.3, using waterpower potential as a cost-shifter for local firms' use of water power (after controlling for county water flow, elevation changes, and other characteristics). In particular, the differences between the model counties correspond to our reduced-form regression coefficients β_t in Equations (2.1)-(2.2). One feature of our setting is that the transition to steam power had already started when comprehensive manufacturing census data started to be collected in 1850, as by then 10% of mills used steam power. We model the adoption curve directly, allowing us to interpret the reduced-form regressions as estimates of the effect of waterpower potential at different dates along the adoption curve.

2.6.1 Estimation Strategy

In this section, we describe the set of structural parameters and the target moments used for estimation. We estimate the structural model to match the empirical patterns and the reduced-form estimates from Section 2.3. In particular, we target a mix of estimates within county-industries and between counties. We estimate the parameters simultaneously using the Method of Simulated Moments (MSM). Appendix 2.I provides details on the MSM estimation procedure.

Within-County Moments

Most of the moments we match in the model come from predicting the value of a typical baseline county (denoted B). We have data on two sectors (flour and lumber), while in the model we consider one composite "milling" sector. To create this composite, we calculate the relevant moment Y_{ict} for each sector separately. We then predict Y_{ict} using our reduced-form specification in Equation (2.1).[38] We then take the average to generate Y_{Bt}, weighting by the number of mills. Specifically, the baseline moment we match is the predicted outcome for a county with average waterpower potential:

$$Y_{Bt} \equiv \mathbb{E}_i \left[\mathbb{E}_c[Y_{ict}] \right] = \mathbb{E}_i \left[{\gamma_{it}}' \mathbb{E}_c[X_{ic}] \right], \tag{2.24}$$

where X_{ic} consists of our baseline controls, our standardized measure of local waterpower potential (whose average is normalized to zero), and an industry fixed effect.

For some moments, we compare outcomes in the baseline county to those in a "lower water power" county (denoted L). The counterfactual moments for county L are identified under the assumption that local waterpower potential is a cost-shifter for local firms' use of water power (conditional on our included control variables). To calculate outcomes in county L, we follow Equation (2.24) but predict outcomes for a county with one standard deviation lower waterpower potential (while holding all of the other characteristics fixed at their average levels). The difference in moments between counties B and L corresponds to our estimated reduced-form impacts of lower water power, $\hat{\beta}_t$.

While the parameters are estimated jointly, many have an intuitive mapping to specific moments, which we discuss below. Appendix 2.I supports these intuitive explanations with a formal analysis of our sources of identification, using the local relationships between structural parameters and simulated moments, following Andrews et al. (2017).

38. We weight by the number of mills in each county-industry, for estimating these moments, because some of our moments relate to dispersion and we want these to reflect aggregate dispersion.

Steam productivity. A positive γ means that steam is relatively more productive, and consequently steam users will have higher sales. We therefore use the sales differential between steam and water users within each county, as in Figure 2.6, to help identify γ. Importantly, the observed difference in sales between steam and water users also reflects selection, as productive mills are more likely to use steam power. We model this selection directly and account for it when estimating γ jointly with the other parameters.

Baseline productivity process. We estimate the persistence of the baseline productivities π using the 10-year auto-correlation of log sales at the establishment level (0.4). To help estimate the dispersion of productivities σ, we use the standard deviation of log sales within each county (1.0).

Operating costs. Given the dynamics of productivity, higher operating costs f_o^R will make firms more likely to exit. We therefore use the share of water (or steam) users that subsequently exit the market, as in Table 2.26, to help estimate f_o^R.

Startup costs. Entrants have to pay $f_o^E + c(R)$ to start producing. A higher startup cost toughens the selection upon entry, increasing the relative sizes of entrant mills. We use the sales differential between incumbents and entrants (as in Figure 2.6) to help pin down $f_o^E + c(R)$.

Power adoption costs: Water power. We split the startup costs into general milling capital f_o^E and power-specific capital $c(R)$ by comparing water mills (who pay $f_o^E + c(W)$) to hand powered mills (who only pay f_o^E) in our data.[39] The capital premium for water users is 0.5 log points, implying $\frac{c(W)}{f^E + c(W)} = 0.4$.

39. We do not include hand powered mills in our broader analysis, as these mills only constitute 0.6% of total revenue in flour and lumber milling.

Power adoption costs: Steam power. A higher adoption cost of steam power $c_t(S)$ leads fewer firms to choose steam over water power. We use the share of establishments using steam power in 1850 and 1880, as in Figure 2.5, to help estimate $c_t(S)$.

Power switching barriers. Higher power-switching barriers lead incumbents to switch power technologies less often. To help estimate the barriers that incumbents face to switch technologies, we follow Equation (2.23) and use the (within-county) difference in adoption shares for entrants versus incumbents, as in Figure 2.5:

$$\log \frac{\Pr(R|R,\varphi)}{\Pr(R'|R,\varphi)} - \log \frac{\Pr(R|E,\varphi)}{\Pr(R'|E,\varphi)} = \frac{1}{\rho} \times \Big(c(R,R') + (1-\omega^R) c_{ct}(R) \Big) . \tag{2.25}$$

Entry costs. A higher entry cost will deter mills from entering the market. We use the share of producers who are entrants, as in Table 2.1, to inform our estimate of f^e.

Across-County Moments

The comparison across counties is crucial for identifying key model parameters, including the demand elasticity for milling and the strength of the steam agglomeration forces. We match four moments that are generated by comparing counties of different waterpower potentials.

Regional cost of water power. The additional fixed cost of water power in places with lower waterpower potential, $c_L(W) - c_B(W)$, lowers the attractiveness of using water power. Therefore, we estimate it using the relationship between waterpower potential and the share of mills using water power (as in Table 2.2).

Total demand elasticity. The total demand elasticity η determines how sensitive the demand for milling output is to milling prices. The primary moment used to identify η is the initial (1850) relationship between lower waterpower potential (which increases milling costs) and local milling activity.

Agglomeration in steam adoption. An agglomeration force in steam adoption costs (negative κ) will further boost the adoption of steam power in the low-water region. Hence, to identify the agglomeration in power costs, we use the impact of lower water power on the observed use of steam power from 1850 to 1880, as in Table 2.3.

Agglomeration in steam productivity. An agglomeration force in steam productivity (positive α_S) will further boost economic growth in the low-water region (where steam is diffusing faster). Hence, to identify the agglomeration in steam productivity, we use the impact of lower water power on revenue growth from 1850 to 1880, as in Table 2.3.

Calibrated Parameters

We calibrate the following parameters outside the estimation routine.

Firm demand elasticity. In our model, mills charge a constant sales-to-cost markup $\frac{1}{\epsilon-1}$ over variable costs (materials and labor). In Appendix 2.C, we calculate that the median sales-to-cost markup among flour and lumber mills is 20%, implying a firm demand elasticity of 6. In comparison, modern estimates range between 3 and 11 (Asker et al., 2014; Bloom, 2009; Sedláček and Sterk, 2017; Felbermayr et al., 2018; Acemoglu et al., 2018; Buera et al., 2021), and are relatively large in milling (Broda and Weinstein, 2006).

Time discounting. The discount factor (denoted as δ) is calibrated to reflect an annual interest rate of 6%. In Section 2.6.3, we support the forward-looking assumption by demonstrating that ignoring future returns (a scenario with $\delta = 0$) would imply an implausibly low estimate for the startup capital cost of milling.

Sunk costs. Our baseline setup assumes that water and steam capital is fully sunk and sets ω^R to zero. We explore the robustness of our estimates to these assumptions by allowing water to steam switchers to partially recover the value of their power assets, setting ω^W to 0.35 following Kermani and Ma (2023). We also explore counterfactuals where instead capital is fully recoverable.

Convergence rate for steam technology. The parameter $c_S^{(slope)}$ governs how fast steam adoption costs fall from their initial state $c_S^{(initial)}$ to their mature state $c_S^{(terminal)}$. We set the convergence rate to 4% per year, which implies that steam power matures by 1890. This assumption is consistent with the long-run diffusion patterns in Jovanovic and Rousseau (2005) and aligns with the power cost estimates presented in Atack (1979). We show that the estimated model can match the steam adoption patterns in all decades from 1850 to 1880, despite fixing the convergence rate to this literature-informed value.

Dispersion of cost shocks. We set the dispersion parameters ρ and ρ_O to 2, equivalent to about 6.5% of median 1850 sales. These values fall within the range of estimates in the literature (Chernoff, 2021; Humlum, 2022) and imply a limited amount of idiosyncratic variation in power and operation costs. As a validation of the amount of idiosyncrasies in power and exit choices, our estimated model can match the observed overlap between exiting and surviving firms (as in Figure 2.6) and the overlap in firm size distributions between steam and water users (as in Figure 2.22).

Estimation Procedure

We use an adapted Newton-Rhapson method to estimate our structural model. Appendix 2.I.1 details the algorithm and validates the method. In particular, we ensure that the estimated model satisfies the parameter-moment relationships predicted in Sections 2.6.1-2.6.1.

2.6.2 Estimation Results

Model Fit

Table 2.7 shows the targeted moments and how well the model does at matching the data. We estimate 15 parameters using 15 target moments. Due to the robust and monotone relationships between parameters and moments described in Sections 2.6.1-2.6.1, our estimation procedure matches the target moments exactly. In Section 2.6.3, we conduct overidentification tests of the model by comparing model simulations to the non-targeted regressions from Section 2.3.1.

Parameter Identification

Appendix 2.I.2 conducts a formal analysis of our sources of parameter identification, following the local sensitivity measures proposed by Andrews et al. (2017). In particular, we verify that the relationship between moments and parameters have the signs and magnitudes predicted in Sections 2.6.1-2.6.1. The analysis also highlights the importance of estimating the model parameters jointly, as many parameters affect multiple target moments simultaneously.

Parameter Estimates

Table 2.8 reports our estimated parameters. We discuss the estimated magnitudes below and, when possible, compare them to estimates in the literature and from contemporaneous sources.

Productivity. The steam power productivity premium, γ, lowers marginal production costs by about 9.3%. This structural estimate falls within the range of existing estimates of the efficiency of steam engines vs. waterwheels in the 19th century (Atack, 1979; Crafts, 2004; Chernoff, 2021). Our estimated parameters for the baseline productivity process (π, σ)

are within the standard range of estimates from modern data (Bachmann and Bayer, 2014; Coşar et al., 2016; Schaal, 2017; Ottonello and Winberry, 2020).[40]

Operating costs. The operating costs of steam power f_o^S are larger than those of water power f_o^W, constituting 30% and 10% of 1850 median sales, respectively. Large operating costs of steam are consistent with the qualitative evidence that steam engines required more upkeep and reflect that steam users exit at a higher rate, despite being larger and more productive (as in Table 2.26 and Figure 2.21). Swain (1888) estimates that the annual fixed costs of steam and water power, respectively, were around \$20 and \$10 dollars per horsepower, which applied to 1850 firm medians are around 16% and 8% of annual sales.

Startup costs. The startup cost of setting up a watermill $f_o^E + c_B(W)$ is around 44% of annual sales. These inferred costs are close to the capital stocks of water users directly observed in our data, as the value of the capital stock of the average water mill in 1850 was 51% of annual sales.

Power adoption costs. Figure 2.26 plots the estimated adoption costs of water and steam power over time. Water power in the baseline region $c_B(W)$ had an upfront cost of around 444 dollars, equivalent to about 18% of 1850 median sales. Steam initially had a higher upfront cost, and we estimate that in 1850 the additional upfront cost of steam power $c_{1850}(S)$ was about 611 dollars or 24% of median sales. By comparison, in our 1850 data, the typical water and steam mills had, respectively, around \$500 and \$2000 more capital installed than the hand-powered mills. Our estimated purchase prices are also somewhat smaller than contemporaneous accounts that 20 horsepower engines – including the boiler and other associated equipment – cost \$2,500 in the 1840s and \$2,000 in the 1880s (Armistead et al.,

40. For example, Bachmann and Bayer (2014) estimate (π, σ) to be $(0.9675, 0.0905)$, which falls close to our estimates of $(0.9663, 0.0875)$.

1841; Emery, 1883; Atack et al., 1980), though our estimated operating costs are slightly higher than those in contemporaneous accounts.

We estimate that as steam became more available and adaptable, the upfront cost of steam fell below water, converging to a level of around 8% of annual sales. Emery (1883) reports that the purchase prices of steam and water power were similar in 1880, which is consistent with our estimates. The continued use of water power in this later period reflects lower operating costs, idiosyncratic shocks, and switching costs.

Power switching barriers. The barrier to switching from water to steam includes sunk capital $(1-\omega^W)c(W)$ and other switching costs $c(W,S)$. This total switching barrier from water constitutes 19% of 1850 median annual sales or just above two months' worth of revenue. Notably, fully sunk water capital ($\omega^W = 0$) can account for the vast majority of these switching barriers (93%), and the switching costs $c(W,S)$ only represent 1.4% of annual sales. This implies that other forces that might make it difficult for enterprises to adopt new technologies (e.g., retrofitting, uncertainty about the costs and benefits of steam power, or some millers being stuck in their ways) are quantitatively less important for the transition to steam power. Sunk steam capital ($\omega^S = 0$) similarly accounts for the majority (81%) of switching barriers from steam to water power, though we estimate larger costs of switching from steam to water, perhaps due to the importance of location for water power.

Regional cost of water power. The additional water cost in the low-water region $c_L(W)$, 106 dollars, is around a quarter of the cost in the baseline region. By comparison, Atack et al. (1980) estimate that the average water-horsepower for all manufacturing in 1850 cost 67 percent more in the Midwest compared to New England.[41] One reason why our numbers might be smaller is that millers were relatively small power users, and therefore less affected

41. On average, counties in the Midwest have around 1.1 standard deviations less waterpower potential than counties in New England.

by more-limited local water power.

2.6.3 Model Validation

In this section, we examine the validity of our estimated model of steam adoption. First, we reproduce a series of non-targeted regressions from Section 2.3.1 on how waterpower potential shapes steam adoption and economic growth of incumbents and entrants. Second, we examine the validity of two key model features: the forward-looking behavior of establishments and agglomeration effects in steam power.

Testing the Model: Reproducing Regressions

In Table 2.9, we compare the data patterns in Tables 2.3, 2.4, and 2.5 to the patterns we find when we run equivalent regressions on simulated data from our model.

Table 2.3 shows that higher water costs cause faster steam adoption, and Table 2.5 shows that this is driven by entrants. However, over time the effect of local waterpower potential diminishes. Our estimated model demonstrates the same pattern. This is because higher costs of water affect steam adoption by making steam power a comparably cheaper technology (a *technology cost* effect), strengthening the selection of operating mills (a productivity *selection* effect), and weakening competition in local product markets (a *competition* effect). These effects are reinforced by an *agglomeration* effect in steam power. The *technology cost*, *selection*, *competition*, and *agglomeration* effects all lead to more steam use in places with higher water costs. Incumbents differ from entrants due to switching barriers, which make their steam adoption decisions less responsive to the cost of water power. Places with less waterpower potential approach their steady-state use of steam power earlier. As a result, along the adoption curve, the effect of waterpower potential on the *growth* in steam use diminishes and reverses over time, though in *levels* places with less waterpower potential are always more likely to use steam power.

Table 2.3 also shows that higher water costs cause faster revenue growth, and Table 2.5 again shows that this is driven by entrants. Our estimated model replicates this pattern. Higher costs of water increase the revenue growth from steam power through the *technology cost*, *selection*, and *agglomeration* channels described above. The *technology cost* and *agglomeration* benefits depend on mills' access to steam power, with diminished gains for water incumbents who face switching barriers. Incumbents are crowded out in places with higher water costs when the negative *competition* effect from new entrants is strong enough.

Our estimated model is also able to match the two potentially incongruous features of the data that incumbents in places with lower waterpower potential are both (1) more likely to invest and switch to steam power (Table 2.5) and (2) more likely to exit (Table 2.4). This reflects countervailing forces that dominate in different parts of the firm-productivity distribution: incumbents in places with lower waterpower potential places are relatively high productivity, and this selection means that (all else equal) they are more likely to choose to switch to steam power. However, the increased entry and greater steam-use in places with lower waterpower potential lowers the local price index, which lowers survival rates for the marginal incumbents (of which there are more in places with less waterpower potential).

Validating Model Features

Forward-looking behavior. Forward-looking expectations are at the heart of our adoption model: some establishments adopt steam power even though they anticipate that adoption costs will continue to fall, and other establishments choose water power, even knowing that they will face switching barriers if they later want to scale up production with steam power.

To illustrate the importance of allowing for expectations, we re-estimate the model assuming that establishments are fully myopic ($\delta = 0$) and compare our estimates to external benchmarks. We find that myopia would imply an implausibly low estimate for the startup capital cost of milling. With forward-looking millers, we estimate that the total startup costs, $f_o^E + c(W)$, are 44% of median firm sales, whereas we would estimate that the total startup costs are under 10% of median firm sales if millers were myopic. For comparison, the median 1850 water mill in our data has a capital stock worth 51% of annual sales (which is not a data feature used in the model estimation).

Agglomeration. Agglomeration effects in steam power are one prominent reason why adoption may be inefficiently slow, motivating a potential role for policy intervention. While Section 2.3.2 provides suggestive evidence of agglomeration spillovers through backward linkages, we can now use the estimated model to directly assess the quantitative importance of agglomeration effects in driving the economic impacts of steam power.

Increasing the local share of steam users from 0 to 100% further boosts the productivity of steam power by $\alpha_S = 2.5$ percentage points (over its baseline level of 9.3%). This agglomeration effect on marginal costs, potentially due to the increased local knowledge base, has a meaningful impact on the aggregate economic growth from steam power. In particular, in Table 2.30, we estimate the model while forcing $\alpha_S = 0$ and find that this constrained model can only account for around half of the differential growth we observe in the low-water region.

By contrast, we do not find economically significant agglomeration effects in steam purchase prices. Increasing the local share of steam users from 0 to 100% slightly increases the steam adoption cost by 1.8% of 1850 median sales (over a baseline level of 24%). In particular, in Table 2.30, we estimate the model while forcing $\kappa = 0$, and find that the constrained model can nevertheless still match the differential steam adoption and economic growth in the low-water region. One interpretation of this result is that it suggests that information about the existence of steam, and its broad costs and benefits, was not a barrier to adoption: having more steam-using neighbors did not make mills more likely to adopt, other than through the measured productivity spillover.

2.7 Counterfactual Experiments

In this section, we use our estimated model to assess the determinants of technology adoption and to evaluate policies aimed at alleviating barriers to adoption. In Section 2.7.1, we evaluate the importance of waterpower potential and switching barriers for the aggregate spread of steam power. In Section 2.7.2, we evaluate a "cash for clunkers"-style program that pays mills to switch from water to steam by buying the mills' sunk water capital. Finally, in Section 2.7.3, we show how the interaction of switching barriers and the new technology's high fixed costs leads to slow aggregate technological adoption, whereas technology adoption is much faster if there is only one of these. Even when there are substantial entry and exit of establishments, aggregate technology adoption is still slowed by sunk costs when the old technology's cost structure is relatively appealing to entrant firms – successive waves of entrants are still willing to become stuck in the old technology for future periods.

2.7.1 Local Waterpower Potential, Switching Barriers, and the Incidence of Steam Power

Waterpower Potential. In Figure 2.8, Panel A, we simulate the share of mills using steam power in the baseline region, and in a region with one standard deviation lower waterpower potential.[42] Higher costs of water power induce the use of steam: places with lower waterpower potential reach the baseline steady-state steam share 31 years faster and ultimately experience an 18% higher steady-state steam share.

Figure 2.8, Panel B shows the influence of water costs on total milling activity.[43] Initially,

42. Appendix Figure 2.27 shows that the simulated implications are similar if, instead of assuming power costs are fully sunk, we set $\omega^W = 0.35$ for water mills that switch to steam power. When sunk costs are lower, other estimated switching barriers are correspondingly higher to rationalize the relatively low switching rates in the data.

43. Note that the impacts on mill revenues (Figure 2.8, Panels B and D, and Table 2.10) capture consumer surplus, scaled by a factor of $\eta - 1 = 4.9$. This is because mill revenues and the price index (our theory-consistent measure of consumer surplus) are log-log linearly related: $\log(\text{Revenues}_{ct}) = (\eta - 1) \log P_{ct}$.

lower waterpower potential constrains milling, as mill revenues in 1830 are 75% lower than in the baseline region. With the arrival of steam power, places with lower waterpower potential catch up to the baseline region and shrink the gap in total milling activity to 11% by 1890. Limited access to water power created an "advantage of backwardness" in steam adoption, but this advantage was not strong enough for the lower waterpower region to overtake the baseline region in aggregate mill revenue. This is because the direct benefits from lower water costs in the baseline region continued to outweigh the benefits from higher steam power adoption in places with less waterpower potential.

Table 2.10 reports the impact of steam power on milling activity, separately for 1830 incumbents and all future entrants.[44] Entrants are the sole driver of higher economic activity from steam power in the baseline region (Column 1) and lower water power region (Column 2). Entrant revenue grows by 111% and 201% in these regions, respectively, whereas incumbent establishments earn only 0.2% and 0.3% more due to steam power. Lower incumbent revenue reflects increased competition from entrants, which entirely mitigates the direct benefits to incumbents from increased access to improved power technology. Quantitatively, the net effects on 1830 incumbents are small because steam power diffused relatively slowly.

These unequal gains from steam power are consistent with our findings in Section 2.3.1, in which incumbents have lower survival rates in regions with lower waterpower potential where steam is diffusing faster. These counterfactuals report the total impact of steam on milling, including "level effects" shared across regions, whereas the estimates from Section 2.3.1 identify only the relative impact of steam power across regions.

Table 2.11 shows the impact of steam power on incumbent firm values in 1830 (Equation (2.10)), decomposing the values into operating profits, the option value of exit, and the option value of steam power (see Appendix 2.J.1 for a formal definition of the components).[45] While

44. In this section, we evaluate the impacts on incumbents in 1830 because incumbency in later periods is endogenous to the arrival of steam power.

45. Because the free entry condition holds in equilibrium (Equation (2.12)), all of the value of steam power

steam power raised incumbent revenues (Table 2.10), Table 2.11 shows that steam lowered incumbent firm values by 0.1% in the baseline and lower waterpower regions (Columns 1 and 2). This is because impacts on firm values reflect both firm revenues and firms' costly adjustments. Some incumbents switch to steam power, which incurs costs along with increasing revenue. Much of the decline in incumbents' profits due to steam was counteracted by their value of exiting the market. The option value of steam power compensated 68%-73% of the losses in incumbent firm values after considering the effects on profits and exit.

Water Lock-in. Figure 2.8 also shows the importance of establishment-level switching barriers for *aggregate* steam power adoption and mill revenue. This extends our results in Section 2.3.1, which suggested a *relative* influence of switching barriers on the steam adoption of water incumbents compared to entrants. We simulate the arrival of steam power in two counterfactual scenarios: a "No Water Lock-In" scenario in which water mills face no switching barriers and choose power sources as freely as entrants ($\omega^W = 1$, $c(W, S) = 0$); and a "Full Water Lock-in" scenario in which water mills face insurmountably high costs of switching ($c(W, S) \to \infty$). Entrants are free to choose their power source in the baseline and both counterfactual scenarios.

Panel C shows that establishment-level switching barriers substantially delay aggregate steam adoption, despite substantial entry and exit in the economy. The economy reaches a 30% steam adoption rate 22 years faster when water mills face no switching barriers, compared to the scenario with full water lock-in (1855 vs. 1877). Switching barriers matter the most in the middle of the adoption curve, when steam technology is improving and more establishments are on the margin of choosing steam power. Switching barriers also continue to be important, however, and lower steam adoption by about eight percentage points even in the terminal steady state.

Steam adoption rates in our baseline economy fall roughly halfway between the "Full

---

generated by entrants is passed through to lower consumer prices.

Water Lock-In" and "No Water Lock-In" scenarios. Our baseline economy is closer to the "Full Water Lock-In" scenario early on the adoption curve and, over time, converges to the "No Water Lock-In" scenario. Technology switching is particularly important for the acceleration in steam adoption that we see in our data period, from 1850 to 1880.

Panel D of Figure 2.8 shows the impact of switching barriers on total mill revenue. Switching barriers of water mills continue to hamper the economic potential of steam power, even in the steady-state when steam technology is fully mature. Without water lock-in, the steady-state gains in total revenue from steam power would be 2.6 times larger. These substantial gains arise because there would be substantially more entry without switching barriers, as firms are attracted by the option of switching to steam in the future (Dixit and Pindyck, 1994). This substantially increases the number of active mills in the no lock-in scenario, whereas total revenue in the baseline economy is closer to the scenario with full water lock-in.

Columns 3 and 4 of Table 2.10 report how switching barriers shape the impact of steam power on incumbent and entrant establishments. When water mills do not face switching barriers, in Column 3, the introduction of steam power increases incumbent mill revenue by more than in the baseline, while Column 4 shows that with infinite switching barriers, incumbent mill revenue increases by less. Table 2.11 shows that effects on incumbent values are muted relative to those on revenue, as switching is also associated with increases in adoption and overhead costs. In total, while the option value of steam power is higher when water mills face no lock-in (1.9% vs. 0%), this benefit is counterbalanced by the increased competition from entrants, lowering the profitability of existing mills (by 3.8% vs. 0.6%).

This last result, in particular, highlights the importance of accounting for firm competition and forward-looking behavior. Removing lock-in effects would seemingly benefit the incumbent firms who are locked into water power, but that also benefits new firms who are more willing to enter when there are no future switching barriers. Quantitatively, removing switching barriers raises competition enough that, on net, incumbents do not benefit from the arrival of steam power.

2.7.2 "Cash for Clunkers" Policy Counterfactual

Agglomeration spillovers from technology use can make private adoption decisions inefficiently slow. Section 2.7.1 showed that establishment-level switching barriers cause substantial delays in aggregate technology adoption, which raises questions about the aggregate consequences of removing those barriers. In this section, we show that our estimated agglomeration effects are small enough that removing switching costs does not generate persistent long-run effects. However, the agglomeration effects are large enough that government subsidies to steam adoption would generate a 38% return.

We evaluate both temporary and permanent policies that counterfactually subsidize water incumbents switching to steam power by purchasing their old water power infrastructure, thereby eliminating the sunk costs. These policies are motivated by the 2009 "cash for clunkers" program (Blinder, 2008), which lasted for two months and incentivized drivers to trade-in old (fuel-inefficient) cars.

Figure 2.9 shows the effect of different counterfactual policies, along with their annual costs. Panel A shows the counterfactual effects of a one-year temporary policy, implemented in 1850. The share of mills using steam power instantaneously doubles, as many establishments take advantage of the subsidy. The share falls over time, however, and by 1865 there is no remaining impact of the program on the share of establishments using steam power. This response illustrates that our estimated agglomeration effects are too small to generate "big push" effects from a short-duration policy.

Even longer-duration policies would not have had permanent effects. Appendix Figure 2.28 shows the counterfactual effects of 5 and 20-year temporary policies, also starting in 1850. Compared to the one-year program, fewer establishments switch to steam immediately because some prefer to wait (knowing they can still take advantage of the policy later). There is a spike in steam adoption, and program cost, in the last year of these policies because of mills' forward-looking behavior.[46] Nevertheless, the policy effects fully dissipate within two decades of their termination.

Figure 2.9, Panel C shows that a permanent policy, paying firms' sunk costs in water, induces steam adoption that is broadly the same as our counterfactual with no switching barriers (shown in Figure 2.8). A permanent policy leads to a small steady-state increase in steam-use, but at very high costs because of low "additionality" (Russo and Aspelund, 2024): many subsidized steam switchers would have switched without the subsidy. Furthermore, the subsidy encourages many firms that would have entered using steam power to instead enter using water power and later switch to steam.

Table 2.12 conducts a cost-benefit analysis of the subsidies, comparing benefits (for producers and consumers) to the cost of the programs. Our baseline estimates are that for every \$1 in subsidies, the one-year temporary program generates \$1.38 in total benefits and the permanent program generates \$1.13 in total benefits.

Consumers receive the majority of benefits through lower goods prices from the subsidies (\$0.91 to \$1.13 per dollar spent in our baseline economy). Part of the total gains are also driven by agglomeration spillover effects on steam entrants: even though entrants do not qualify for the one-year subsidy, the agglomeration spillover from incumbents' steam adoption is strong enough to crowd in entry of steam users, further lowering the price index.

Incumbent firms directly benefit from switching subsidies, and the gains to 1850 incumbents from the one-year program are around \$0.47 per dollar spent. However, increased competition from future entrants attenuates the gains to 1850 incumbents under longer-duration subsidies.

The payoff from subsidizing steam switching depends crucially on the presence of agglomeration spillovers. Without steam agglomeration, we estimate that both the temporary

46. This is consistent with modern evidence on bunching at the expiration of subsidy policies (Chen, 2024).

and permanent programs would generate net deficits. Without agglomeration, subsidizing switching crowds out entrants, mitigating the direct effects on the price index.

This exercise shows how a model of the economic environment can be used to evaluate gains from potential subsidies to technology adoption. We show that a switching subsidy that compensates incumbents can generate a net surplus, but only when there are some agglomeration spillovers in the adoption of the new technology, and especially when implemented for a short time horizon. These substantial gains accrue even under our estimated "weak" agglomeration spillovers, which are insufficient to generate a permanent shift in steam adoption from temporary subsidies.

2.7.3 Fixed Costs and the Speed of Technology Adoption

The aggregate importance of switching barriers may seem surprising given the substantial amount of entry and exit in our data. From one decade to the next, about 80-85% of establishments exit. In this economic environment, how can switching barriers continue to matter even though most establishments are entrants? The answer is that water power continued to appeal to entrants far along the transition path to steam. Switching barriers influence aggregate technology adoption even with exit and entry when entrants often adopt the old technology, and then themselves have to face switching barriers.

In particular, water power appealed to less-productive entrants who did not yet have the scale to benefit from steam power: water had lower purchase prices (at the start of our sample period), and lower fixed costs of operation (throughout our sample period). When some of these entrants later had successful businesses with higher productivity, however, they faced switching barriers to scaling up with steam power. Importantly, these "entrant lock-ins" are not indicative of mistakes in adoption decisions. On the contrary, entrants in our model choose water power fully anticipating that they will have to pay switching costs in the future if their productivity increases and they later want to adopt steam.

To conclude our counterfactual analysis, we isolate how the interaction of fixed costs and switching costs slows technology adoption by holding constant the overall attractiveness of the new technology. To do this, we consider two hypothetical technologies that are equally attractive (so their steady-state adoption rates are 50%), but they have different purchase prices and different marginal costs. Technology 1 ("High FC & low MC") has a marginal cost advantage that is equal to our estimated marginal cost advantage of steam over water. Technology 2 ("Low FC & high MC") has a lower fixed adoption cost, chosen such that its steady-state adoption rate is 50%. Otherwise, for both technologies we hold all parameters fixed at those we estimate for water power (e.g., demand elasticities, overhead costs, and idiosyncratic shocks).

Figure 2.10 shows the adoption speed of new technologies in this environment, separately by whether all firms initially use technology 2 or technology 1. As a benchmark, the gray line shows the importance of switching costs: if the economy starts with technology 2 and we introduce an *identical* technology, it takes 5 years for the new technology to get close to steady-state adoption (47% adoption share).[47] The black line shows that higher fixed costs slow adoption: if technology 1 is introduced into an economy that only has technology 2, it takes 19 years for the new technology to reach 47% adoption. By contrast, the dashed line shows that lower fixed costs accelerate adoption: if technology 2 is introduced into an economy that only has technology 1, it reaches 47% adoption in its first year.[48] These estimates are all driven by the interaction of fixed costs and switching barriers: in the absence of switching barriers, adoption would immediately reach its steady-state level, regardless of the relative costs of the two technologies.

47. The steady-state share of firms using the new technology asymptotically trends to 50%, so we report the duration to reach an adoption rate of 47%.

48. When the new technology has relatively low fixed costs, it overshoots upon introduction and reaches over 50% adoption. This is because, initially, the price index is relatively high and so low-productivity establishments (who prefer technology 2) are initially able to profitably produce before getting crowded out in steady-state.

2.8 Conclusion

This paper studies the adoption of steam power in milling in the late 19th century. Steam power was a general purpose technology that alleviated the dependence of mechanized power on local geography. The adoption of steam power, and its impacts, depended on places' access to water power. Indeed, a general feature of new technologies is their impacts vary with differences in access to previously-available alternative technologies. Over time, steam technology improved both nationally (through technological change) and locally (through agglomeration). Nevertheless, as steam became increasingly more cost-effective than water power in more places and for more firms, many incumbents were resistant to changing technologies.

To understand the effect of improvements in steam power on milling, this paper draws on substantial data contributions. We compile a full panel dataset of manufacturing establishments in the United States during the Second Industrial Revolution. We link the data to the geographic distribution of waterpower potential, which allows insights into the adoption of steam power: places with less waterpower potential adopted more steam power, earlier, and steam adoption was driven predominantly by entrant mills.

We emphasize dynamic effects, through which prior use of water power (1) created lock-in effects discouraging steam adoption, (2) generated leapfrogging by entrants, and (3) made steam adoption inefficiently slow due to agglomeration spillovers.

We estimate a dynamic equilibrium model of entry and investment to characterize the forces that determine technology use across space and time. We estimate the importance of economic features for the slow spread of steam power, and evaluate policies that counteract the technological lock-in caused by historical advantages.

We find that the interaction of high fixed costs and switching barriers delays aggregate technology adoption. For technologies with both of these features, entry of new firms may not be a panacea against technological lock-in. High fixed costs made smaller entrants predisposed to use the old technology (the low-initial-cost and high-marginal-cost technology). Switching barriers then meant that these entrant firms became stuck with water power, even though the barriers were anticipated. Either feature on its own has little effect on adoption speeds.

Many recent quickly-embraced innovations, such as cloud computing (Lu et al., 2023), allow small firms to use new technologies without substantial fixed investments. Energy transitions have historically been protracted (Smil, 2014), and many modern environmentally friendly technologies, such as heat pumps and renewable energy sources, are associated with low marginal costs, but high fixed costs and switching barriers. Our results highlight how these characteristics can lead to slow adoption.

2.A Figures

Figure 2.1. Components of County Waterpower Potential

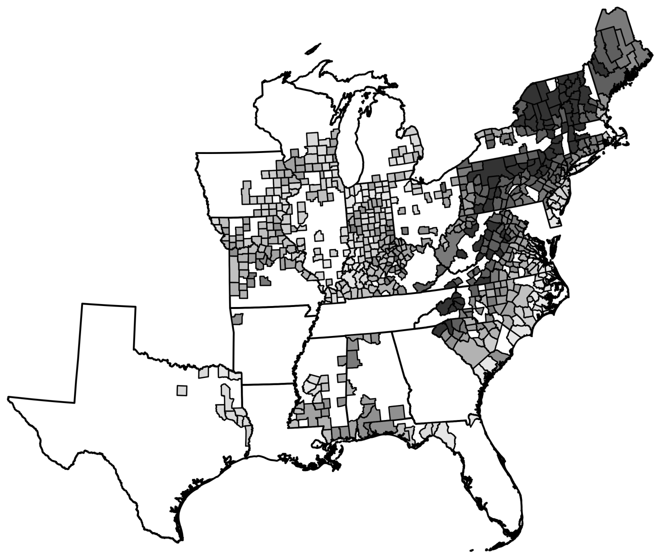
Panel A. Flow Rate of River Segments

Panel B. Fall Height of River Segments

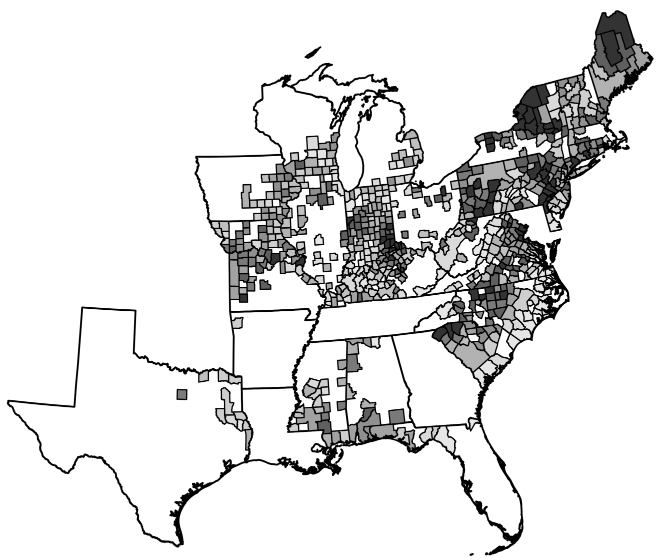
Notes: This figure plots the sources of waterpower potential in the United States, with darker shares corresponding to greater flow rates or fall heights. Panel A plots our estimated flow rates for each river segment, in cubic feet per second. Panel B plots the drop in elevation for each river segment, in feet per mile. Data from NHDPlusV2.

Figure 2.2. County Waterpower Potential, Measured and Residualized

Panel A. County Waterpower Potential, Measured



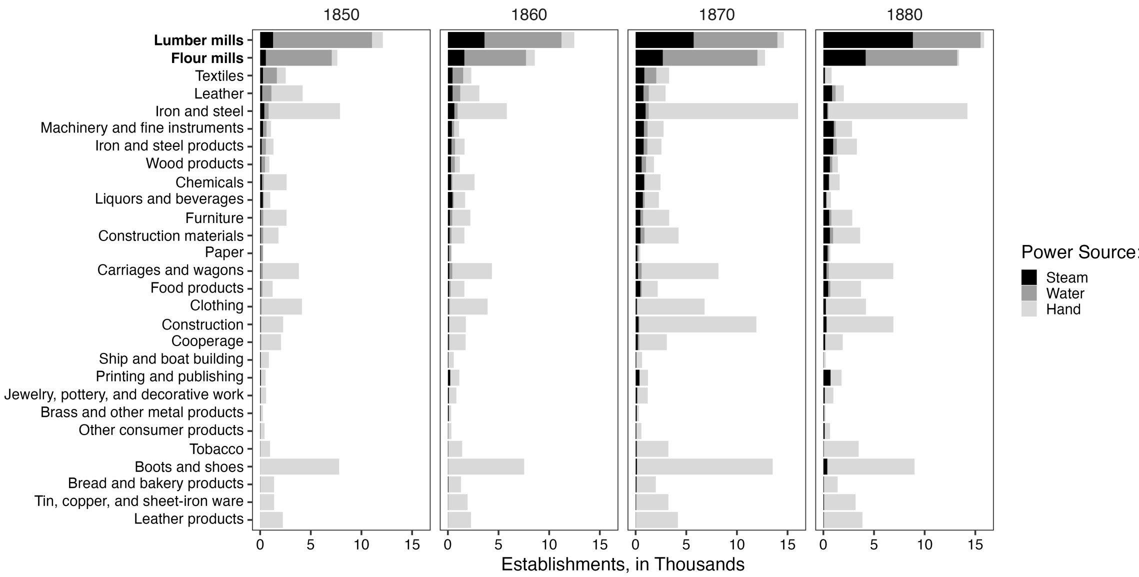
Panel B. County Waterpower Potential, Residualized



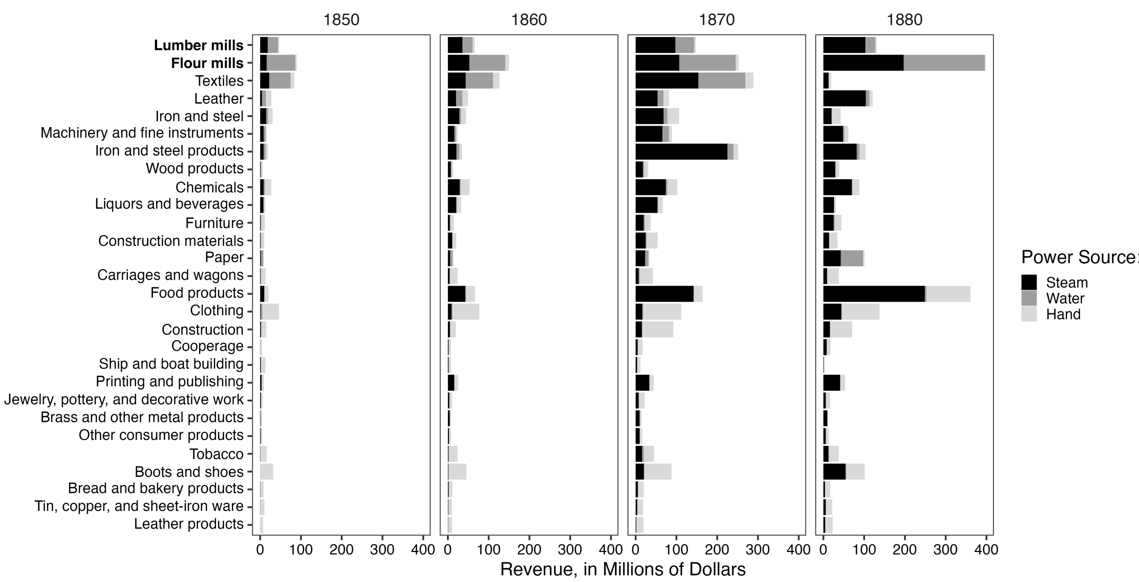
Notes: This figure shows our estimated county waterpower potential, with darker shares corresponding to greater waterpower potential deciles. The sample is restricted to our main balanced panel of 690 counties. Panel A shows our measure of county waterpower potential: summing across all river segments in the county the flow rate of the river segment times its fall height (and a gravitational constant), per square mile. Panel B shows the residual county waterpower potential, after controlling for our main baseline controls: total county water flow and terrain ruggedness; the presence of a navigable waterway, distance to the nearest navigable waterway, and county market access in 1850; the presence of coal in the county, the share of county area covered by coal deposits, and market access to coal deposits. Data from NHDPlusV2.

Figure 2.3. Power Source By Industry

Panel A. Number of Establishments, by Power Source

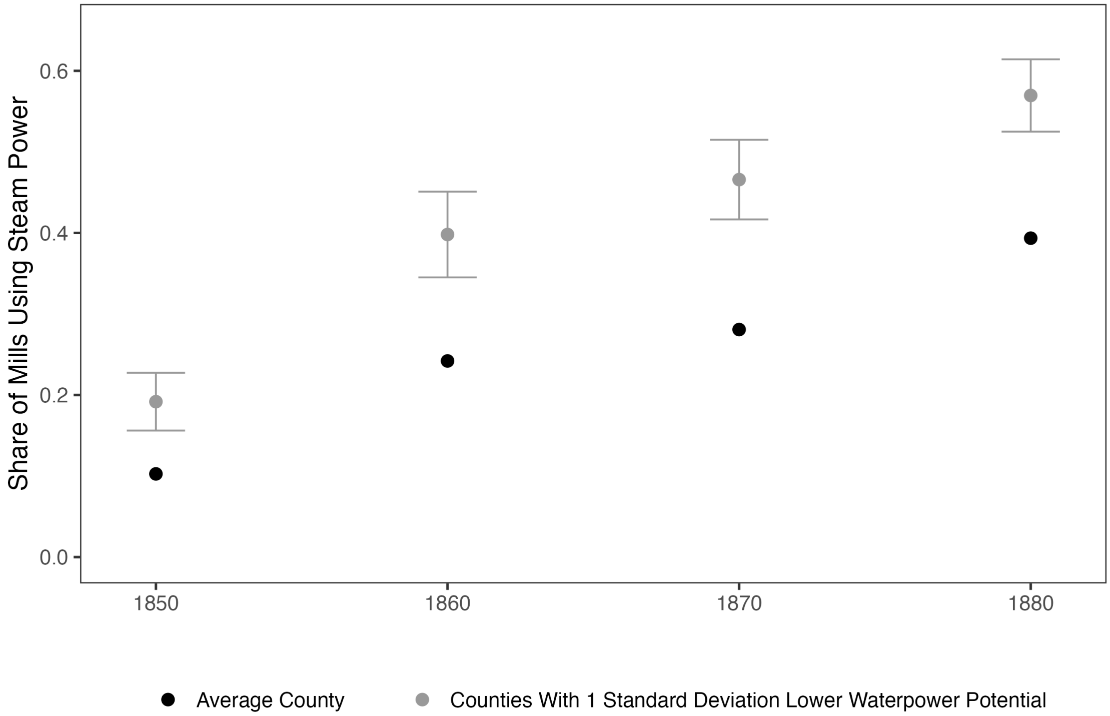


Panel B. Total Revenue, by Power Source



Notes: This figure plots power-use, by industry and decade. Industries are sorted by the number of establishments using either steam or water power in 1850 (in decreasing order). Panel A shows the number of establishments in each industry using steam, water, and hand power. Panel B shows the total revenue produced in establishments using steam, water, and hand power. We define "steam" to include all establishments using any steam power; "water" includes establishments using water power and no steam power; "hand" includes the remaining establishments that use neither steam nor water. Data from our digitized establishment-level Census of Manufactures (1850-1880).

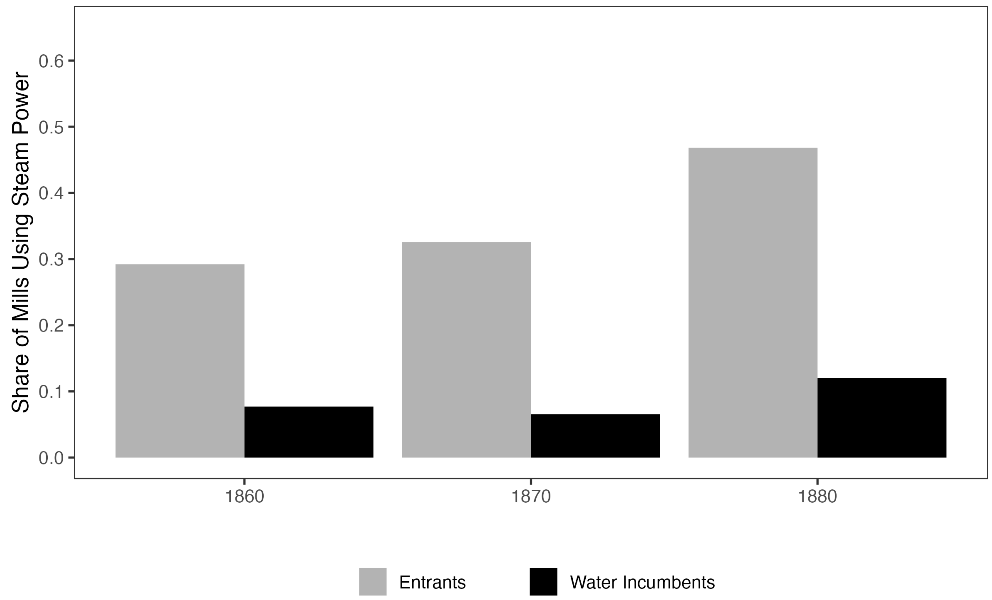
Figure 2.4. Share of Mills using Steam Power, by Decade and County Waterpower Potential



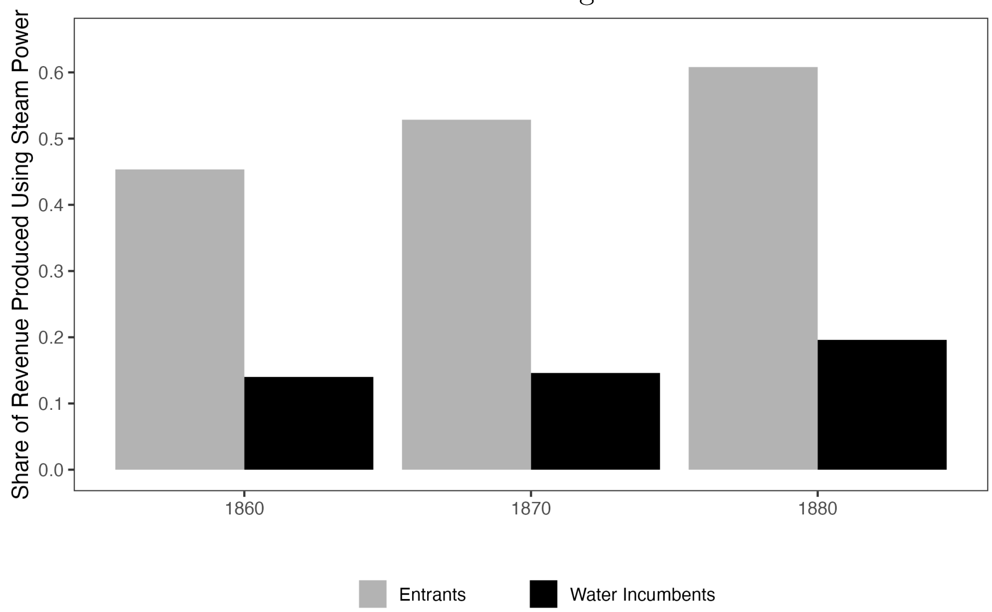
Notes: Darker circles represent the share of mills using steam power in the average county. For the lighter circles, we add the estimated increase in steam share from a one standard deviation decrease in county waterpower potential (conditional on our baseline controls as in Table 2), with an indicated 95% confidence interval. Standard errors are robust and clustered at the county level. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Figure 2.5. Steam-use Share, for Entrants and Water Incumbents

Panel A. Share of Mills Using Steam Power

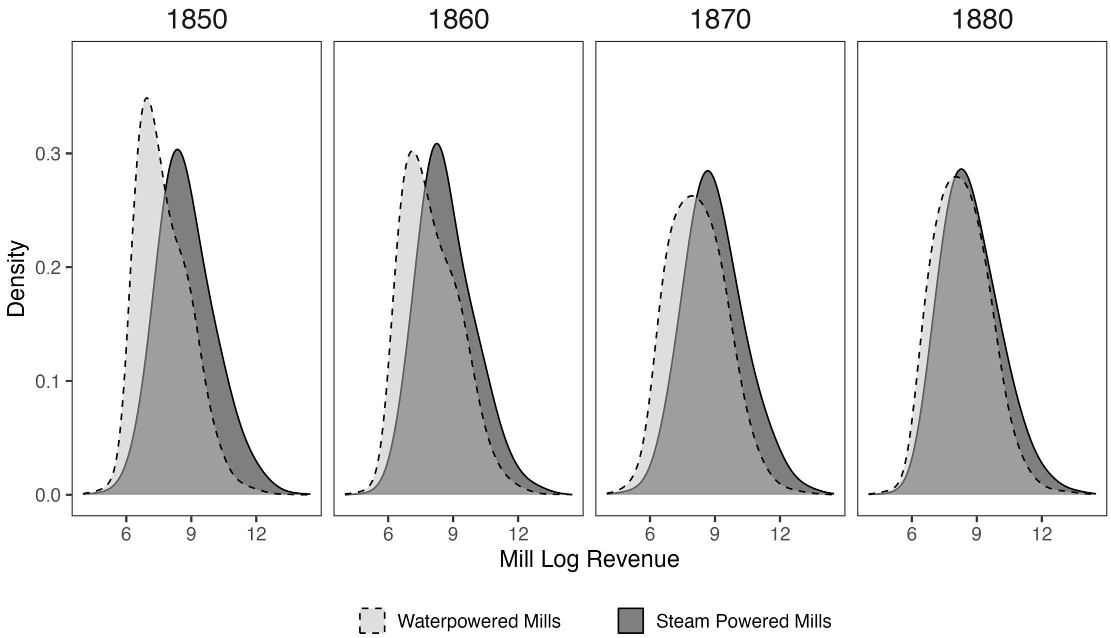


Panel B. Share of Revenue Produced Using Steam Power



Notes: This figure shows steam-use rates, by mill type ("Entrants" and "Water Incumbents"). Entrants began operations after the prior Census. Water Incumbents used water power in the prior Census. Panel A shows the share of mills using steam power, for each mill type. Panel B shows the share of revenue produced using steam power, for each mill type. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880).

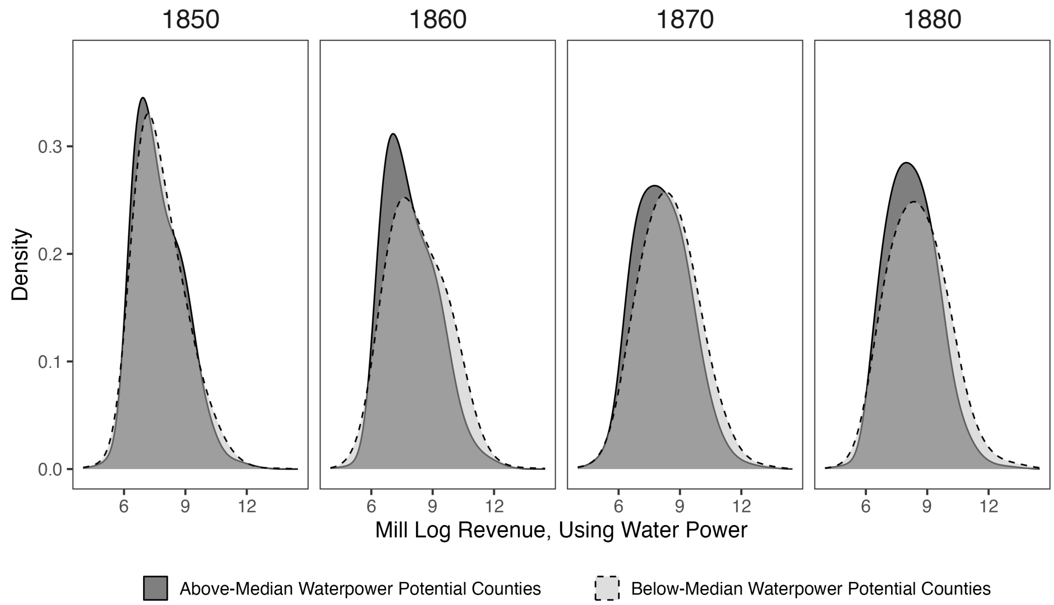
Figure 2.6. Mill Size Distribution, by Power Source



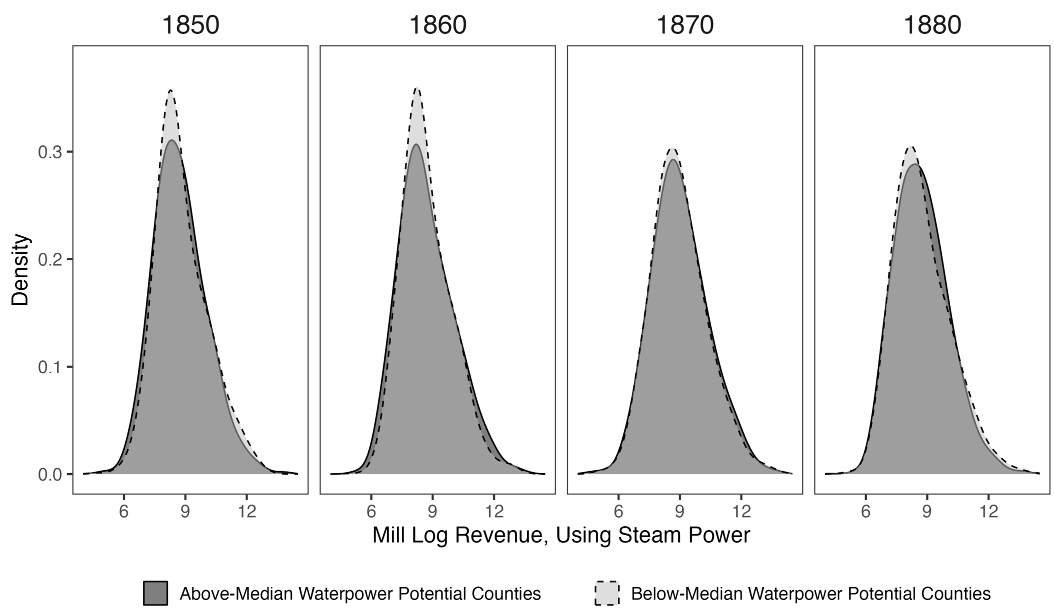
Notes: This figure shows the distribution of mill revenue, by power source, in each decade. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880).

Figure 2.7. Mill Size Distribution, by County Waterpower Potential

Panel A. Revenue Distribution of Water-Using Mills



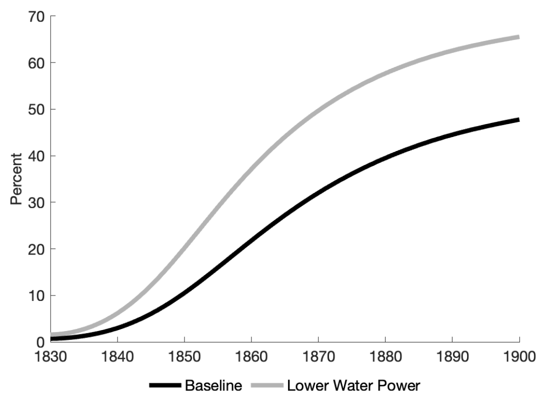
Panel B. Revenue Distribution of Steam-Using Mills



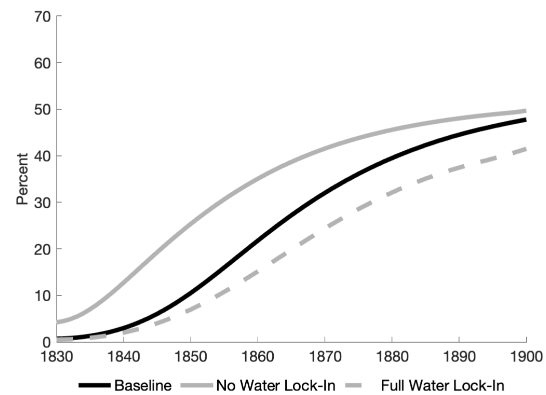
Notes: This figure shows the distribution of mill revenue in each decade, separately for counties with above-median and below-median waterpower potential. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Figure 2.8. Water Technology and the Impacts of Steam Power

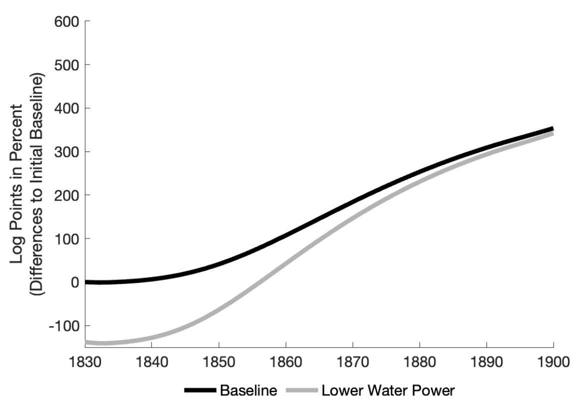
A. Water Costs and Steam Adoption



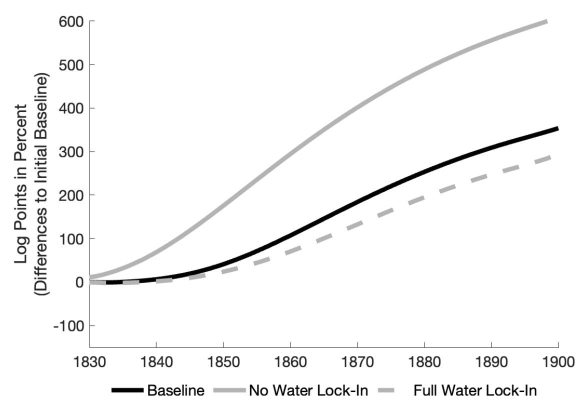
C. Switching Barriers and Steam Adoption



B. Water Costs and Mill Revenue



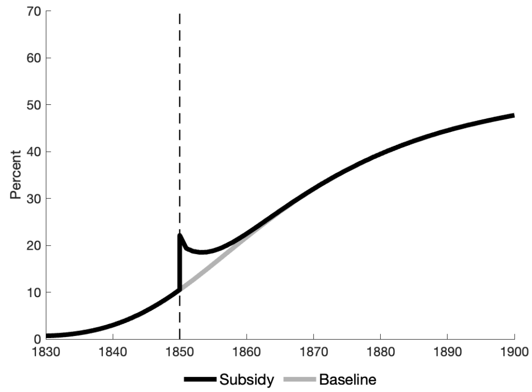
D. Switching Barriers and Mill Revenue



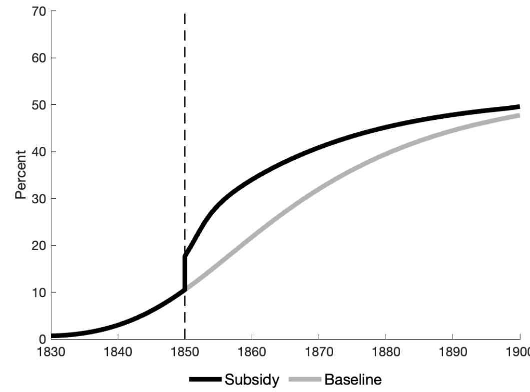
Notes: This figure shows the share of steam users and total mill revenue in model counties with different water technologies. Mill revenue is measured in log differences to the initial steady state of the baseline region. Panels A and B plot the impacts of steam power in the average county (black line) and a region with a standard deviation lower waterpower potential (gray line), where the only parameter difference between the regions is the fixed cost of water power adoption. Panels C and D plot the impacts of steam power as a function of switching barriers. The black line shows adoption for our baseline estimates, the gray line removes switching barriers ($\omega^W = 1, c(W,S) = 0$), and the dashed line represents prohibitive switching barriers ($c(W,S) \to \infty$).

Figure 2.9. Water-to-Steam Switching Subsidies: Steam Adoption and Annual Costs

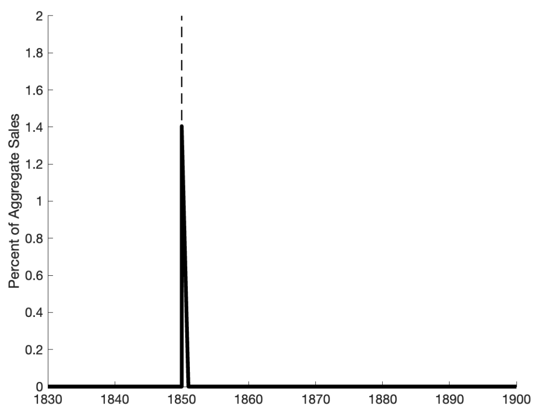
A. Temporary Subsidy: Steam Adoption



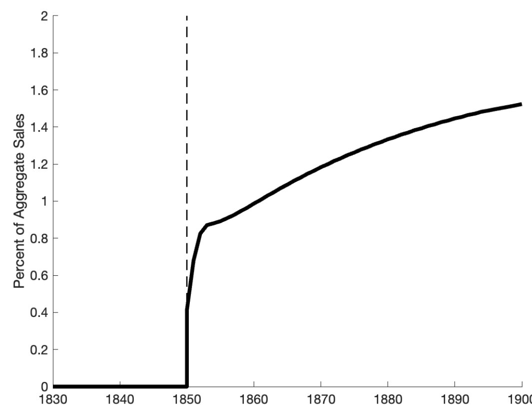
C. Permanent Subsidy: Steam Adoption



B. Temporary Subsidy: Annual Cost

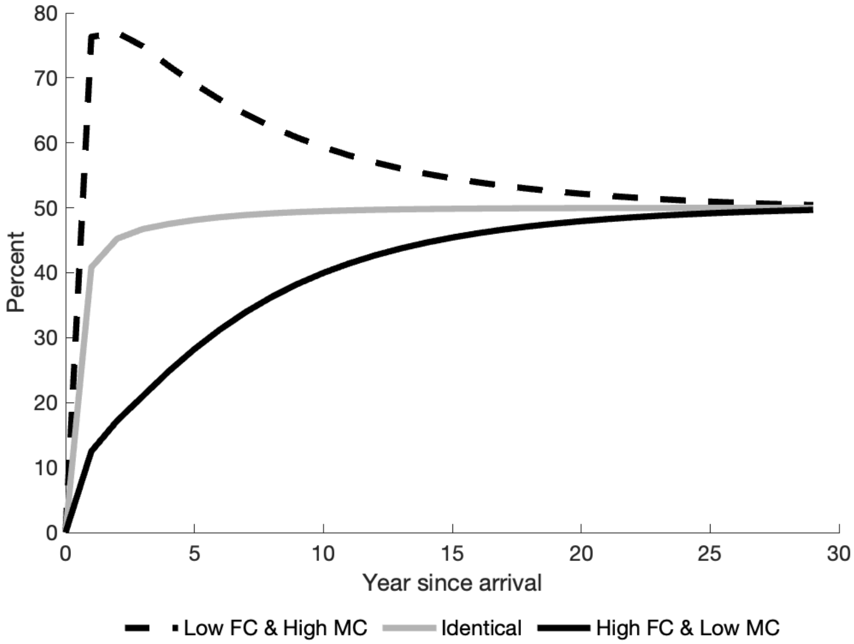


D. Permanent Subsidy: Annual Cost



Notes: This figure simulates counterfactual "cash-for-clunkers" policies that pay water incumbents $c_B(W)$ to switch to steam power, exactly offsetting the sunk cost of switching. Panel A shows the adoption of steam power with a one-year-only temporary policy in 1850, and Panel B shows its annual costs in percent of aggregate mill revenues. Panel C shows the adoption of steam power after a permanent policy introduced in 1850, and Panel D shows its annual costs. Panels A and C compare the counterfactual adoption of steam power (in black) to the factual adoption (in gray).

Figure 2.10. Technology Adoption and Fixed Costs



Notes: This figure simulates the adoption of new technologies under various scenarios. One technology ("High FC & low MC") has a marginal cost advantage equal to our estimated steam's marginal cost advantage over water in 1900, while the other technology ("Low FC & high MC") has a lower fixed cost, chosen such that in an economy with both, the steady-state adoption rate of each is 50%. Otherwise the technologies have the same parameters as those we estimate for water power. The gray line shows the adoption speed when introducing the latter technology in an environment that already has its equivalent (so the old and new technologies are identical other than through idiosyncratic shocks). The black line shows the adoption of the former technology in an environment that already has the latter. The dashed line shows the adoption of the latter technology in an environment that already has the former. The x-axis is years (the new technology is introduced in year 1), and the y-axis is the share of establishments using the new technology.

2.B Tables

Table 2.1. Composition of Milling

	Share of Total Milling					Share of Steam Milling		
	Steam Entrants (1)	Water Entrants (2)	Steam Incumbents (3)	Water Incumbents (Switchers) (4)	Water Incumbents (Stayers) (5)	Entrants (6)	Steam Incumbents (7)	Water Incumbents (8)
Panel A. Establishments								
1860	0.23	0.56	0.01	0.01	0.18	0.90	0.05	0.06
1870	0.28	0.58	0.03	0.01	0.11	0.90	0.07	0.03
1880	0.37	0.42	0.06	0.02	0.14	0.84	0.11	0.05
Panel B. Revenue								
1860	0.36	0.43	0.04	0.02	0.15	0.85	0.09	0.06
1870	0.43	0.38	0.08	0.02	0.09	0.83	0.14	0.03
1880	0.44	0.29	0.10	0.03	0.14	0.77	0.17	0.06

Notes: Columns 1–5, in Panel A, show the share of total mills that are steam entrants, water entrants, steam incumbents, or water incumbents (distinguishing between those who switched to steam and those who stayed with water power). Columns 6–8 show the share of steam mills in each decade that are steam entrants, steam incumbents, or water incumbents. Panel B reports corresponding numbers for the share of total revenue produced by each mill type. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880).

Table 2.2. Mill Activity in 1850, by County Waterpower Potential

	All Mills (1)	Only Lumber Mills (2)	Only Flour Mills (3)
Panel A. Number of Waterpowered Mills			
Lower Waterpower	-1.055	-1.246	-0.783
	(0.130)	(0.173)	(0.109)
Panel B. Revenue of Waterpowered Mills			
Lower Waterpower	-1.127	-0.974	-1.178
	(0.249)	(0.215)	(0.302)
Panel C. Steam Share of Mills			
Lower Waterpower	0.089	0.107	0.060
	(0.015)	(0.019)	(0.016)
Panel D. Steam Share of Revenue			
Lower Waterpower	0.123	0.160	0.060
	(0.022)	(0.031)	(0.021)
Panel E. Total Number of Mills			
Lower Waterpower	-0.956	-1.100	-0.738
	(0.119)	(0.156)	(0.105)
Panel F. Total Revenue of Mills			
Lower Waterpower	-0.876	-0.704	-0.973
	(0.215)	(0.173)	(0.291)
# County-Industries	1,199	612	587

Notes: This table shows the relationship between mill activity in 1850 and county waterpower potential. "Lower Waterpower" is a negative standardized measure of county water power potential, with standard deviation of one, so the estimates reflect differences in counties with one standard deviation lower waterpower potential.

Each panel shows the effect of water power potential on a different outcome in 1850: the total number of water powered mills (Panel A); the total revenue of water powered mills (Panel B); the share of mills using steam power (Panel C); the share of milling revenue from using steam power (Panel D); the total number of mills (Panel E); and total mill revenue (Panel F). Column 1 reports pooled estimates from county-industry regressions, for lumber and flour milling; Column 2 restricts the sample to lumber mills only; and Column 3 restricts the sample to flour mills only. Panels A, B, E, and F use PPML estimation, which approximates percent differences. Panels C and D are OLS regressions, weighting county-industries by their number of mills, which reflect percentage point differences in the shares.

All regressions include industry fixed effects and our baseline controls interacted with industry: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-industry in 1850. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850) and NHDPlusV2.

Table 2.3. Steam Adoption and Mill Growth, by County Waterpower Potential

	Steam Share of Mills (1)	Total Mills (2)	Total Mill Revenue (3)
Growth in Lower Waterpower Counties:			
From 1850 to 1860	0.067	0.220	0.183
	(0.016)	(0.062)	(0.081)
# County-Industries	1,084	1,199	1,199
From 1860 to 1870	0.034	0.113	0.203
	(0.013)	(0.052)	(0.097)
# County-Industries	1,061	1,199	1,199
From 1870 to 1880	-0.009	0.092	0.140
	(0.013)	(0.036)	(0.087)
# County-Industries	1,138	1,199	1,199

Notes: This table shows the relationship between growth in mill activity and county waterpower potential. "Lower Waterpower" is a negative standardized measure of county water power potential, with standard deviation of one, so the estimates reflect differences in counties with one standard deviation lower waterpower potential.

The outcomes are the share of mills using steam power (column 1), the total number of mills (column 2), and total mill revenue (column 3). Each row corresponds to growth over the indicated decade, using only data from the indicated years.

Column 1 reports OLS estimates, restricting the sample to county-industries with at least one mill in both decades (for the steam share to be defined) and weighting by the number of mills in that county-industry in 1850. These estimates reflect percentage point differences in the shares. Columns 2 and 3 report PPML estimates for a balanced panel of county-industries (including zeros), which approximate percent differences.

All regressions include county-industry fixed effects, industry-year fixed effects, and our baseline controls interacted with industry and year: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.4. Entry Rates and Survival Rates, by County Waterpower Potential

	Entry Rate (1)	Survival Rate (2)	Difference (1) − (2) (3)
Elasticity with Respect to Lower Waterpower:			
In 1860	0.323	-0.230	0.554
	(0.074)	(0.065)	(0.089)
# County-Industries	1,199	1,199	
In 1870	0.168	-0.266	0.434
	(0.058)	(0.057)	(0.072)
# County-Industries	1,199	1,199	
In 1880	0.158	-0.158	0.316
	(0.045)	(0.040)	(0.061)
# County-Industries	1,199	1,199	

Notes: This table shows the elasticity of mill entry and mill survival, over the previous decade, with respect to county waterpower potential. "Lower Waterpower" is a negative standardized measure of county waterpower potential, with standard deviation of one, so the estimates reflect differences in counties with one standard deviation lower waterpower potential.

Column 1 reports results for entry, column 2 reports results for incumbent survival, and column 3 reports the difference in these estimates. Each row corresponds to a different PPML regression, using data from the indicated Census year and previous Census year, which approximates percent differences in the rates.

All regressions include county-industry fixed effects, industry-year fixed effects, and our baseline controls interacted with industry and year: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.5. Steam Adoption Shares for Entrants and Water Incumbents, by County Waterpower Potential

	Entrants (1)	Water Incumbents (2)	Difference (1) − (2) (3)
Adoption in Lower Waterpower Counties:			
In 1860	0.169	0.034	0.135
	(0.024)	(0.021)	(0.023)
# County-Industries	1,076	607	
In 1870	0.188	0.049	0.139
	(0.022)	(0.018)	(0.025)
# County-Industries	1,151	560	
In 1880	0.172	0.051	0.121
	(0.022)	(0.024)	(0.025)
# County-Industries	1,169	685	

Notes: This table shows the relationship between county waterpower potential and the steam use of entrant mills and water incumbent mills. "Lower Waterpower" is a negative standardized measure of county waterpower potential, with standard deviation of one, so the estimates reflect differences in counties with one standard deviation lower waterpower potential.

The outcome in column 1 is the share of entrants using steam power, restricted to county-industries with at least one entrant in that year. Column 2 reports the share of "water incumbents" (mills that used water power in the previous Census year) who switched to steam power. For column 2, the sample is restricted to county-industries with at least one surviving water incumbent. Column 3 reports the difference between the estimates in columns 1 and 2. Each row corresponds to a different OLS regression, using data from the indicated Census year only, which reports percentage point differences in the shares.

All regressions include industry fixed effects and our baseline controls interacted with industry: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

For each row, each observation is a county-industry, weighted by the number of mills in 1850. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.6. Non-Mill Manufacturing Establishments, Non-Mill Steam-Use, and Steam Manufacturing, by County Waterpower Potential

	Total Non-Mill Establishments (1)	Steam User Share of Non-Mill Establishments (2)	Steam Makers, Relative to All Establishments (3)
Differences in Lower Waterpower Counties:			
In 1850	-0.584	0.017	0.444
	(0.224)	(0.005)	(0.164)
# Counties	690	674	690
In 1860	-0.443	0.024	0.340
	(0.338)	(0.008)	(0.207)
# Counties	690	661	690
In 1870	-0.529	0.034	0.504
	(0.236)	(0.009)	(0.240)
# Counties	690	678	690

Notes: This table shows the relationship between county waterpower potential and local non-mill manufacturing activity (i.e., outside the flour mill and lumber mill industries). "Lower Waterpower" is a negative standardized measure of county waterpower potential, with standard deviation of one, so the estimates reflect differences in counties with one standard deviation lower waterpower potential.

The outcome in column 1 is the total number of non-mill manufacturing establishments. The outcome in column 2 is the share of non-mill establishments using steam power. The outcome in column 3 is the number of steam makers (establishments reporting making engines or boilers) relative to the number of all manufacturing establishments. Each row corresponds to a different regression, using data from the indicated year only. Columns 1 and 3 report PPML estimates, including zeros, which approximate percent differences. Column 2 reports OLS estimates, weighting by the number of non-mills in that county in 1850, which reflect percentage point differences in the shares.

All regressions include our baseline controls: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

For each row, each observation is a county. We exclude 1880 because data for several non-mill industries are mostly lost for 1880. Robust standard errors are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.7. Model Fit to Target Moments

Parameter (1)	Moment (2)	Years (3)	Model (4)	Data (5)
Panel A. Baseline County				
$c(W,S)$	Water Choice Differential:	1850–1880	0.552	0.553
	Water Incumbents vs. Entrants			(0.062)
$c(S,W)$	Steam Choice Differential:	1850–1880	0.977	0.977
	Steam Incumbents vs. Entrants			(0.123)
$c_S^{(initial)}$	Steam Adoption Rate	1850	0.102	0.103
				(0.006)
$c_S^{(terminal)}$	Steam Adoption Rate	1880	0.393	0.393
				(0.011)
f_e	Entry Rate	1850–1860	0.737	0.750
				(0.006)
f_o^E	Log Sales Differential:	1850–1880	0.132	0.131
	Incumbents vs. Entrants			(0.015)
f_o^W	Water Exit Rate	1850–1880	0.789	0.789
				(0.003)
f_o^S	Steam Exit Rate	1850–1880	0.835	0.835
				(0.006)
γ	Log Sales Differential:	1850–1880	0.855	0.855
	Steam vs. Water Users			(0.029)
π	Log Sales Autocorrelation	1850–1860	0.412	0.412
				(0.019)
σ	Log Sales Standard Deviation	1850–1860	1.019	1.019
				(0.011)
Panel B. Differences in Lower Waterpower Counties				
$c_L(W)$	Steam Adoption Rate	1850	0.089	0.089
				(0.016)
η	Log Total Output	1850	-0.876	-0.876
				(0.215)
κ	Change in Steam Adoption Rate	1850, 1880	0.092	0.092
				(0.019)
α_S	Growth of Output	1850, 1880	0.525	0.525
				(0.118)

Notes: This table shows the empirical fit of our estimated model. The table shows each parameter of the model (Column 1) and the moment (in time period) that most closely targets it (Columns 2 and 3). Column 4 reports the model-simulated moments, and Column 5 contains the empirical estimates with robust standard errors in parentheses. Panel A includes the within-county moments described in Section 2.6.1, and Panel B includes the across-county moments described in 2.6.1. Our estimation procedure, described in Section 2.6.1, matches these target moments exactly, up to a preset numerical tolerance of 1%.

Table 2.8. Parameter Estimates

Parameter (1)	Description (2)	Value (3)	Dollars (4)	Source (5)
Panel A. Power Costs				
$c(W,S)$	Switching costs from water	0.014	36	Table 2.7
$c(S,W)$	Switching costs from steam	0.058	145	Table 2.7
$c_S^{(initial)}$	Steam cost (initial)	0.441	1102	Table 2.7
$c_S^{(terminal)}$	Steam cost (terminal)	0.084	211	Table 2.7
$c_S^{(slope)}$	Steam cost (time-slope)	0.040		Section 2.6.1
$c_B(W)$	Water cost in baseline county	0.178	444	Table 2.7
$c_L(W)$	Water cost in lower water power county	0.220	551	Table 2.7
κ	Agglomeration in steam adoption	0.018	44	Table 2.7
ρ	Dispersion in power costs	0.064	159	Table 2.7
Panel B. Entry and Operating Costs				
f_e	Entry costs	0.004	10	Table 2.7
f_o^E	Startup cost	0.266	666	Table 2.7
f_o^W	Operating cost of water user	0.103	257	Table 2.7
f_o^S	Operating cost of steam user	0.299	748	Table 2.7
ρ_o	Dispersion in operating costs	0.064	159	Table 2.7
Panel C. Productivity				
γ	Steam productivity premium	0.093		Table 2.7
π	Autocorrelation in baseline productivities	0.966		Table 2.7
σ	Dispersion in baseline productivities	0.088		Table 2.7
α_S	Agglomeration in steam production	0.025		Table 2.7
Panel D. Demand				
ϵ	Elasticity of firm demand	6.000		Section 2.6.1
η	Elasticity of local demand	5.877		Table 2.7
Panel E. Other Parameters				
β	Water share in startup cost	0.400		Section 2.6.1
ω	Power resale value	0.000		Section 2.6.1
δ	Discount factor	0.940		Section 2.6.1

Notes: This table shows the estimated values of our model parameters and their sources of identification. Columns 1 and 2 list each parameter and its description. Column 3 reports the parameter values. Panel A includes the parameters of power adoption costs, Panel B includes the parameters of entry and operating costs, Panel C includes the production technology parameters, Panel D includes the parameters of product demand, and Panel E includes other calibrated parameters. Parameter values in Panels A and B are in units of 1850 median firm sales, while Panels C, D, and E are unit-free elasticities unless otherwise noted. Parameters with Table 2.7 as their sources are directly estimated, with the other parameters calibrated in Section 2.6.1.

Table 2.9. Non-Targeted Differences between Lower Waterpower and Baseline Regions

Moment (1)	Years (2)	Model (3)	Data (4)
Panel A. Steam Adoption and Mill Growth (Table 2.3)			
Change in Steam Share of Mills	1850–1860	0.054	0.067 (0.016)
Change in Steam Share of Mills	1860–1870	0.030	0.034 (0.013)
Change in Steam Share of Mills	1870–1880	0.008	-0.009 (0.013)
Total Mills	1850–1860	0.184	0.220 (0.062)
Total Mills	1860–1870	0.157	0.113 (0.052)
Total Mills	1870–1880	0.142	0.092 (0.036)
Total Revenue	1850–1860	0.217	0.183 (0.081)
Total Revenue	1860–1870	0.169	0.203 (0.097)
Total Revenue	1870–1880	0.139	0.140 (0.087)
Panel B. Entry Rates and Survival Rates (Table 2.4)			
Entry rate	1850–1860	0.218	0.323 (0.074)
Entry rate	1860–1870	0.189	0.168 (0.058)
Entry rate	1870–1880	0.170	0.158 (0.045)
Survival rate	1850–1860	-0.047	-0.230 (0.065)
Survival rate	1860–1870	-0.102	-0.266 (0.057)
Survival rate	1870–1880	-0.127	-0.158 (0.040)
Panel C. Steam Adoption of Entrants and Water Incumbents (Table 2.5)			
From Entrants	1850–1860	0.145	0.169 (0.024)
From Entrants	1860–1870	0.173	0.188 (0.022)
From Entrants	1870–1880	0.181	0.172 (0.022)
From Water Incumbents	1850–1860	0.068	0.034 (0.021)
From Water Incumbents	1860–1870	0.088	0.049 (0.018)
From Water Incumbents	1870–1880	0.089	0.051 (0.024)

Notes: This table replicates non-targeted regressions from Section 2.3.1 on our model-simulated data. Each panel reports the regression estimates from a different table. Columns 1 and 2 describe each regression moment, Column 3 reports the model-simulated values, and Column 4 repeats the empirical values from the relevant table in Section 2.3.1 with standard errors in parentheses.

Table 2.10. The Impact of Steam on Mill Revenue 1830-1900 (PDV in %)

	Baseline (1)	Lower Waterpower (2)	No Water Lock-In (3)	Full Water Lock-In (4)
Total	85.21	165.77	267.01	53.28
Incumbents	0.16	0.28	1.08	-0.10
Entrants	110.86	200.79	305.55	72.44

Notes: This table reports the impact of steam on the present discounted values of mill revenues of incumbent and entrant establishments. Incumbents refer to establishments that have been active since 1829 or earlier. Entrants refer to the establishments that entered the region in 1830 or later. Incumbents represent 34% of revenues in the initial steady state without steam power. Columns (1)-(4) report the impact of steam power (measured in percent log points) relative to this initial steady state. Column 1 considers our baseline region, while Column 2 considers an economy with one standard deviation lower waterpower potential. Column 3 considers a counterfactual without switching barriers ($\omega^W = 1, c(W,S) = 0$). Column 4 considers a counterfactual with prohibitive switching barriers ($c(W,S) \to \infty$).

Table 2.11. The Impact of Steam Power on Firm Values in 1830 (in Percentage Points)

	Baseline (1)	Lower Waterpower (2)	No Water Lock-In (3)	Full Water Lock-In (4)
Total	-0.05	-0.08	0.02	-0.05
Operating Profits	-0.98	-1.50	-3.76	-0.61
Option Value of Exit	0.80	1.24	1.91	0.56
Option Value of Steam	0.13	0.18	1.86	0.00

Notes: This table decomposes the percent impact of steam power on firm values in 1830. "Option Value of Steam" reflects the difference in firm value relative to a mill that cannot access steam power. "Option Value of Exit" reflects the additional difference in firm value relative to a water mill that is forced to stay in business indefinitely (labeled "Operating Profits"). Appendix 2.J.1 provides formal definitions of these components. Column 1 considers our baseline region, and Column 2 considers an economy with one standard deviation lower waterpower potential. Column 3 considers a counterfactual without switching barriers ($\omega^W = 1, c(W, S) = 0$). Column 4 considers a counterfactual with prohibitive switching barriers ($c(W, S) \to \infty$).

Table 2.12. Costs and Benefits of Steam Switching Subsidies

	Baseline		No Agglomeration	
	Temporary Subsidy (1)	Permanent Subsidy (2)	Temporary Subsidy (3)	Permanent Subsidy (4)
Incumbent Firms	0.47	0.00	0.40	0.00
Consumers	0.91	1.13	0.01	0.83
Government	-1.00	-1.00	-1.00	-1.00
Total	0.38	0.13	-0.59	-0.16

Notes: This table shows the costs and benefits per dollar of steam switching subsidies, measured in present-discounted values in 1850. Columns 1-2 evaluate the subsidy programs in the baseline economy, and Columns 3-4 consider a counterfactual economy without agglomeration in steam power ($\alpha_S = \kappa = 0$). Columns 1 and 3 evaluate a 1-year temporary program, and Columns 2 and 4 evaluate a permanent program, where both programs are enacted in 1850. "Incumbent Firms" refer to producer surplus, measured by the impact on firm values in 1850. "Consumers" refer to consumer surplus, measured by the equivalent-variation impact on consumer prices; see Appendix 2.J.2 for details. "Government" refers to the direct cost of the switching subsidies.

2.C Establishment-Level Manuscripts from the Census of Manufactures

We have digitized establishment-level data from the original published manuscripts of the Census of Manufactures for 1850, 1860, 1870, and 1880 . We are grateful to Jeremy Atack for providing us many manuscripts; the rest we located in a variety of state, non-profit, and university archives. Most manuscripts were already microfilmed, and the rest we photographed or acquired photos of from archive staff. Our data include some manuscripts that had not been found during the construction of previously-digitized samples described in Atack and Bateman (1999), including Rhode Island and Nevada.

The Census of Manufactures was professionalized and comprehensive beginning in 1850 (Atack and Bateman, 1999). Before 1880, Census enumeration was done in person by U.S. Marshals and all establishments received the same questionnaire, though it changed slightly over time. In 1880, the Census of Manufactures was split into three broad parts: (1) a "general" schedule; (2) a "special agent" schedule; and (3) a "special" schedule. First, many industries received a "general" schedule, similar to that used in 1850, 1860, and 1870. Second, some important sectors were instead given "special agent" schedules, which involved sector-specific questions and specially trained enumerators. These "special agent" manuscripts for 1880 are all believed to be lost (Delle Donne, 1973), which include most manufactures of: cotton, wool, and worsted goods; silk and silk goods; iron and steel; the coke industry; the glass industry; the mining of metals, coal, and petroleum; distilleries and breweries; shipbuilding; and fisheries.[49] Some establishments in these industries were surveyed in the "general" schedule (Atack et al., 2004).

A third category of sectors were enumerated in "special schedules" with sector-specific

49. In 1880, cities with over 8,000 inhabitants were surveyed separately from their counties, also by special agents. While Delle Donne (1973) reports that the special agent city records were lost, we found the city manuscripts and they are included in our samples (the city manuscripts were with the other records, so we are not sure why they were considered lost).

questions, but these were administered by the regular enumerators and these manuscripts were *not* lost along with the "special agent" schedules. For 1880, these special schedules include "Lumber and Saw Mills" and "Flouring and Grist Mills," along with other manufacturing sectors: agricultural implements; paper mills; boots and shoes; leather; brick and tile; cheese and butter; and slaughtering and meat packing. For example, the additional sector-specific questions include: the extent of custom milling for flour mills; and whether a lumber mill does its own logging.

2.C.1 Variable Coverage

The 1860 Census instructions to enumerators discuss the data collection guidelines in useful detail. In addition to establishment count, our main variables of interest are:

Manufacturing Revenue. Products were valued at the factory gate, excluding transportation costs to customers: "In stating the value of the products, the value of the articles *at the place of manufacture* is to be given, exclusive of the cost of transportation to any market" (emphasis original, United States Census Bureau 1860a). We consider a mill active if it reports positive revenue, and include only active mills in our analysis.

From 1850 to 1870, establishments were asked about the quantities and values for each product, but both units and types were not consistently recorded and so we were unable to create a reliable measure of prices. In 1880, the quantities of common products were more consistently defined in special schedules (e.g., "number of thousands of feet of lumber") but the value of sales was recorded at the establishment level, not the product level, for the lumber and flour milling special schedules. In the general schedule, and for less-common products in special schedules, the only recorded output was total value of sales at the establishment level, with no disaggregation by product or reported quantities at any aggregation level. When using price data, we therefore use data from single-product lumber mills in 1880 (both because flour mills are more likely to produce multiple products, and flour prices were often regulated and therefore less informative about marginal costs).

Input Expenditure. To estimate the demand elasticity ϵ, we need a measure of variable input expenditure. We calculate variable input expenditure as the sum of reported labor costs and materials. Total wages paid are reported directly in 1870 and 1880. In 1850 and 1860, we calculate labor costs as the sum (for men and women) of the monthly wage bill times twelve. Materials expenditures are reported directly in the data. For estimating the demand elasticity, we need the input expenditure, so for this calculation we only include mills that report all inputs (94% of the sample). Equation 2.5 shows that prices are a multiple $\frac{\epsilon}{\epsilon-1}$ of marginal costs, so $\epsilon = \frac{\frac{y_{jct}}{x_{jct}}}{\frac{y_{jct}}{x_{jct}}-1}$. We find that for the median mill, revenues are around 20% higher than expenditures, which implies $\epsilon = 6$.

For the custom milling of flour, millers were paid in wheat, keeping a fraction of what their customers brought. The "millers toll" (the price that could be charged for custom flour milling) was regulated, ranging across regions from a quarter to a sixteenth. The markup for wheat sold on the market was higher (Dondlinger, 1919). Consistent with these regulations, we estimate lower markups in flour (10%) than lumber (33%).

Power Source. The Census also asked all establishments for their number of horsepower used in 1870 and 1880. The kind of power source was asked in every year. Across manufacturing, the most common responses were variations on "steam," "water," "horse," and "hand," which we processed to make those broad categories (as well as "other" and "nothing"). Wind power was relatively rare, and by the time of our sample most American enterprises using tides for power had closed (Charlier and Menanteau, 1997). In milling, "steam" and "water" were by far the most common power sources. For our main analysis, we exclude mills who report other categories, mostly because there are very few and therefore are difficult to quantitatively model, but also due to concerns about measurement error for the larger ones. We found historical records for steam or water power use for several suspiciously-large self-reported "non-mechanized" mills. Since we cannot systematically correct these non-mechanized mills' recorded power-use, we drop them from the main analysis. The one exception is that some mills use "steam" or "water" in their industry name (e.g., "steam mill"), but do not also directly report steam or water as their power source, and for those mills we assume they used the named power source. We do use the reported capital stock of "hand" and "manual" mills in order to estimate the share of the capital for water powered mills that was due to water power (as opposed to other milling equipment or structures)).

Industry. In all years, the general schedule Census asked establishments to report the type of business that they were in. Before 1880, the general schedule Census also asked for the types of products they made. In 1880, most flour mills and lumber mills were surveyed on their own special schedules. Two percent of the flour and lumber mills in 1880 were recorded in the general schedule, and we include those mills in our analysis unless the same mill was already also recorded in the special schedule. Below, we describe our processing of the industry strings.

The Census of Manufactures included some establishments outside of manufacturing, including mining, fisheries, and liquor packaging. We do not include those establishments in our analysis. In Appendix Figure 2.14, we compare totals from our sample of only manufacturing establishments to the published totals compiled by the Census. If the Census included non-manufacturing establishments in their totals (which we can observe that they did in 1850 and 1860), then that might lead to differences. On the whole, non-manufacturing made up less than 2% of the establishments in the data.

For Table 2.6, we define "steam makers" as follows. First, we search for establishments whose products are variations on "steam", "engine", "heat", or "boiler". We then constrained the set to establishments who self-reported being in a potentially relevant industry: "iron and steel", "iron and steel products", "brass and other metal products", "machinery and fine instruments", or have industry unclassified/unknown. Finally, we manually verified that the product strings plausibly related to steam products and were not false positives. For instance, we found several establishments that passed these criteria but also produced baked goods, which we did not classify as steam makers. Because product names are not available in the 1880 general schedule, we only classified steam makers in 1850, 1860, and 1870.

Location. The manuscripts record county and state in each decade, based on contemporaneous county names and boundaries. In addition, the name of the closest post office is available for 90% of establishments in 1860, 1870, and the 1880 general schedule. Post office is rarely recorded on the 1850 manuscripts and 1880 special schedules.

2.C.2 Digitization and Processing of the Census Manuscripts

We worked with Digital Divide Data to double-enter and reconcile data from the manuscript images. In total, there were 99,198 manuscript images with manufacturing establishments, including 49,547 pages from 1880. The average page had 7 establishments. Appendix Table 2.13 shows the coverage for which states and decades we were able to find and digitize. When we have records for a state and decade, the records are normally complete for the entire state. For some states and decades, there are some entire counties missing or parts of counties from comparing our establishment totals to the published county-level tabulations.[50] We track each establishment's decade, state, county, page, and row.

To help clean the data, we received assistance from many UChicago undergraduates, graduate students, and full-time research professionals. The team randomly checked many entries, finding a very low error rate. We also used a useful feature of the manuscripts to

50. There are 7 counties that, in the manuscripts and tabulated data, have more than 10 firms in an initial decade, have no firms in the subsequent decade, and then have more than 10 firms. We drop these counties, given our concerns for enumeration error (or the manuscripts being lost contemporaneously). This is in the spirit of Allcott et al. (2016), who similarly drop firms with observations in a given year that are very different from both adjacent observations.

verify numeric entries on many sheets: many 19th-century enumerators entered *totals*, such as writing the total production value for the entire page or for a given firm. We also digitized these row totals and page totals, and compared the entered total with the sum of the relevant responses. Consistent with our general verification of the data, the most common sources for discrepancies were that the total was calculated incorrectly by the enumerator or the total reflected a sum of values that were later crossed out and replaced with other values. In these cases, we made no changes. We also manually checked entries when a ratio seemed highly unusual, such as the output to employment ratio, which was inspired by the data cleaning processes at the current U.S. Census (Fellegi and Holt, 1976; Thompson and Sigman, 1999; Rotemberg and White, 2021). We manually changed any cells where we found a difference between entered values and the manuscripts themselves, but did not otherwise "correct" the original written entries.

We manually processed the entered strings for product names, material inputs, and self-reported industry, along with categorizing the entered power strings based on relevant information such as "water" and "steam." The overall goal was to standardize misspellings and British spellings, expand abbreviations, and assign strings to broader categories. To clean industries, we also used the product strings.

The data include many self-reported industries in each decade, which we group together for our analysis. Following Hornbeck and Rotemberg (2024), we homogenized industry names into 31 categories, using additional information on products when needed. Our analysis focuses on flour and lumber milling, which were relatively straightforward to classify since they had unique outputs. To give a sense of the raw data, there were over 4000 distinct industry strings in the original manuscripts that we associate with the flour and lumber industries, including: "grist," "flower mill," "wood & lumber," "steam saw mill," and "mill" (for the last, we could only identify the industry by the products).

In Appendix Figure 2.15, we compare total lumber and flour milling from the establishment data to contemporaneous tabulations, described and digitized by Hornbeck and Rotemberg (2024). Note that although the tabulation data is useful for detecting missing data, it should not be considered as the ground truth. Some counties may have manuscripts that were not tabulated in the census reports or were mistabulated, for instance because of difficulties defining the industry for each establishment.

Some values for string variables were entered in the "wrong place," when the surveyor had run out of room, which we manually corrected. Similarly, we corrected when numeric variables were entered in a string column. Some entries were marked with a question mark, when the data processing team could not read part or all of a cell. We looked at those entries, and were rarely able read them either.

The Census recorded an enterprise as one establishment even if it contained multiple locations within the same Census subdivision, if these activities across sites were for the "same concern, and all engaged in the same manufacture" (United States Census Bureau, 1860a). There were also some entries in the Census that were associated with one owner but represent multiple industries (for instance, below we discuss the case of E. E. Locke & Co, which operated a distillery and a mill). We split each establishment into multiple industries, so as to consider only the output of each industry. For instance, when we consider the revenue of E. E. Locke & Co, we only consider the revenue of the mill and not that of the distillery. This is particularly relevant for the mills in the period that produced both cut lumber and flour, which we classify as separate mills in our analysis. This approach follows historical Census practice to, for multi-industry establishments, "[separate] the two parts of the business and [assign] each to its appropriate place in the Statistics of Industries" (United States Census Bureau, 1870a). We often refer to "firms" for convenience, though note that the Census enumeration is at the establishment level (unless there were multiple buildings within the same enumeration area) and activity is recorded where it takes place, not at headquarters, so we are then referring to single-establishment "firms."

2.C.3 Adjustment for County Border Changes

Some county borders change over our sample period, and we group together counties with overlapping geographies to create time-consistent borders. This approach is preferable for our analysis of individual mills and establishment-level panel-linking. This differs from an alternative approach of splitting aggregate county activity based on geographic area and aggregating to baseline county borders (Hornbeck, 2010), which would make it difficult to interpret split shares of individual establishments in establishment-level data.

Our baseline county boundaries start with 1850 borders. Issues arise when county polygons from 1860, 1870, or 1880 overlap with multiple 1850 county borders. We group together 1850 counties so that every county from 1860 to 1880 corresponds to a unique grouped 1850 cell. The first step is to group together all of the 1850 counties that overlap with at least 5% of the area of the same 1860, 1870, or 1880 county.

The second step is then grouping together all of the 1850 counties that were linked in the previous step. As an example, suppose 1860 county a overlaps with 1850 counties i and j, and 1870 county b overlaps with 1850 counties j and k. In the first step, we would group i and j and j with k. In the second step, we create a time-consistent boundary that covers i, j, and k.

We use conservatively large county groupings because we do not want to split individual establishments across counties and we want to find the same establishments in subsequent decades. Two grouped counties have an area larger than a circle with a radius of 50 miles, which is too large to be considered a single market, so we drop them from our analysis. We focus our analysis on counties east of the 98th meridian, where county borders are more stable and settlement patterns are less irregular (Webb, 1931). For simplicity, we continue to call the grouped geographies "counties." Our baseline sample covers 750 counties using the actual 1850 borders, which we group into 690 consistent geographies. This covers 83951 flour and lumber mills, and around 90% of all steam-generated sales in those industries.

2.C.4 Creating a Linked Panel of Mills

We link mills by hand, from one decade to the next, in combination with a machine-learning linkage model. We employed a team of data associates to compare a mill in one decade to plausible matches in the subsequent decade. We matched mills on name and location, but did not force establishments to be in the same industry in every decade. Because mills rarely switched between lumber and flour, and we consider working in a different manufacturing sector to be part of the outside option in our model, and so we consider industry switches to be "exits." We never make links using information on power source.

To guide the large-scale hand-links, we first matched a few counties ourselves and compared every mill to every manufacturing establishment in the subsequent decade. We then trained a machine learning algorithm on those matches. For the large-scale hand-linking, we then only considered potential matches with a relatively high linking probability. For the possible matches, we mostly included all candidates with over a 9% linking probability. For mills with many potential links, we only sent the top twenty; for mills with few potential links, we sent the top five as long as their linking probabilities were above 5%. In practice, the potential links with a low match probability were rarely hand-chosen as an actual match. For the analysis in the paper, we then retrained the machine-learning model on the full set of matches. Below, we describe our approach in more detail.

Hand-Linking Procedure

Our first step was to create some panel links by hand, linking establishments in 1860 to their 1870 counterparts in 97 counties. We chose relatively small counties, to start, so it was feasible to compare all possible matches in the same county. We matched 2,709 establishments in 1860 to 5,518 establishments in 1870, adding up to 282,341 comparisons.

To make the links, we considered each establishment's name, industry classification (including the self-reported string and our own cleaned industry measures), and the nearest post office. We also had access to the original CMF manuscript images for each establishment to double-check mistakes, either in the original handwriting or its transcription. Each hand-linking sheet was completed by two UChicago students, and assigned to a third person to reconcile any discrepancies. For each 1860 establishment, we sorted all 1870 candidates by Jaro-Winkler (JW) name similarity, and by whether or not their broad industries matched, to increase the likelihood that links were at the top of each block of names.

Broadly, we made two types of matches in the data. "Direct" matches are when the establishment names in both periods are close matches. This is similar to common practice in literature linking men across decades in the Census of Population (Ferrie, 1996; Feigenbaum, 2016; Ruggles et al., 2018; Bailey et al., 2020; **?**; **?**). However, an important difference between linking men and linking establishments is that many mills *actually* changed their names, especially when adding owners. While additional data would be needed to link women who change their last names, our Census of Manufactures data can tolerate moderate changes in ownership. For instance, Appendix Figure 2.13 shows the manuscript images for a mill that was initially owned by Alson Rogers, which later passed to his son Lucian. To account for "ownership transfers," we also match establishments where part of the name is very similar but another part is different in a manner consistent with a partial change in ownership. In practice, this second category includes partnership formation or newer members taking on the family business.[51]

Model Specification

From hand-linking establishments, we noticed there were broadly four categories for how the establishment's name was reported (consistent with guidance from Jeremy Atack). These were not formal rules, and the way names were written down varied across time and space, but we list the categories below along with our interpretation of their meaning.

51. In our replication files, we denote direct matches as "y", ownership transfer matches as "o", and non-matches as "n". We denote direct matches where the industry changed within milling as "s."

(i). Establishments with sole proprietorship contain a single owner's name. Names were sometimes initialized, and the names did not consistently follow a first/last name order.

(ii). Establishments owned by families normally appeared as a person's name followed by *& sons* or *& brothers.* Others appeared with two first names separated by an ampersand, followed by a last name.

(iii). Establishment that were a partnership or expanded partnership reported two or more names of the proprietors; limited partnerships reported one or more people's names followed by *& co.*

(iv). Establishments that reported names that were impersonal, and often included tokens related to the business and location.

For our mills, in particular, there were two broad types of naming patterns: those with general company names, sometimes including the name of the water power source; and those named after people. Across Census decades, the order of people's names can change. Even for establishments with a single owner, the order of first and last names can change, along with changes in the use of initials.

These features motivate us to build two separate linking models: one matching the whole establishment name, and one matching owners' names with flexibility in their ordering.[52] We use two random forest models to predict establishment pairs, either tracking the company as a whole or tracking individual owners.[53] Both linking models predict establishment pairs to be: a same-owner match, an ownership transfer match, or not a match. We describe this approach in more detail below.

52. We are grateful to Jeremy Atack for suggesting this approach.

53. We generated linking models based on several classifier families, including logistic regression, random forests, and extreme gradient boosting (Chen and Guestrin, 2016). After evaluating their performance on the validation data, we settled on a random forest trained using the R library `ranger`. The random forest model provided the most reliable output, with respect to false positive and negative rates, and the empirical distribution of predicted probability does not concentrate on the two ends which leaves room for setting the probability threshold and varying the false positive and false negative errors.

Name Classifier. We built a name classifier to categorize establishments by their naming pattern type, extract the name of the owners, and identify the name order. While owner names are embedded in establishments owned by sole proprietors, families, partners, or expanded partnerships, the names were often initialized and would switch first-last name orders.

We first use a list of company tokens to identify establishments with impersonal names, which includes: names of locations, such as state and county names; and tokens related to their product or business, such as tanning, manufacturing, lumber, etc.

For establishments without those company tokens, we implement the following steps to extract and format the owner names. First, we remove the non-name tokens, such as "& co" or "& sons," and split the establishment names into owners' names. For a family-owned establishment with two first names and one last name, we assign the last name to both owners (e.g., turn "J & D. Taflinger" into "J Taflinger" and "D. Taflinger.") We then standardize common nicknames and abbreviations to their original names (e.g., Wm to William and Geo to George.) We determine the name order using the first and last name frequency in the 1880 Census of Population. When both names can be first or last names, we keep both orders and look for both of them in the next Census decade.

Owner Linking Model. Our owner-linking model predicts links based on three sets of information: establishment name, industry, and post office. We define several sets of variables for each of the first, middle, and last names: Jaro-Winkler string distance, whether the name is initialized, and whether the initial matches exactly. When there are missing values, which are incompatible with the random forest model, we assign the median value and define an indicator flag for missing. For industry, we use our industry classification based on the raw industry string to create matching indicators for broad and detailed industries. We also create a measure of industry distance based on the industry classification and similarity in their reported kinds of products. For post office, we use the Jaro-Winkler string distance between post office names and an indicator for missing values.

For establishments with multiple owners, the model predicts matches at the establishment-owner level. At the predicting stage, we take the maximum of the predicted probability for each establishment pair (from all owner pairs) to let the output be at the establishment-pair level. This process allows a firm to match when one owner is the same, even if other owners are different, which mimics how humans generally make links.

Company Linking Model The company-linking model also predicts links based on establishment name, industry, and post office. However, instead of extracting the owner information from the establishment names, this model uses the full string of establishment names and looks for establishments with similar whole names. We use the Jaro-Winkler string distance for the full names, in addition to string distance after removing business and location tokens and the minimum string distance between those remaining tokens among all token pairs. The remaining name distances measure the name similarity unrelated to the business itself, which removes false matches that only have closer string distances on the full name because of common tokens (e.g., "Eagle Mill" and "James Mill").

Model Prediction Reconciliation and Hand-Linking

We use both models to predict matches, separately, and then take the maximum of the predicted probabilities. For the set of potential matches that we consider when making hand-links, we select the top 20 pairs with a linking probability above 9%. If there are 5 or fewer pairs to send, we send the top 5 pairs with a linking probability above 5%.

We worked with Digital Divide Data (DDD) in Kenya to hand-link the matches, at scale. Our team helped train the DDD associates in person, who also had experience linking individuals across decades in the Census of Population. We then continued to work closely with them remotely, handling the data process ourselves while their managers handled HR.

We sent DDD lists of all potential matches with identifying information: establishment name, industry, post office, and product kinds produced. We did not include the estimated linking probabilities. Two randomly-assigned separate members of the DDD team found the best match for each establishment, or indicated no close match, and a third random member reconciled any disagreements between the original two members.

We then iterated on these hand-links using the machine-learning model, asking them to manually check "unlikely" matches or "likely" non-matches. We used the same protocol as for the original data, sending DDD the information about the firm but not the estimated link probability. First, we flagged the following three sets of potential matches for review: (1) links that were made for which the algorithm predicted link probability was below 40%, (2) mills with no links, but for which the algorithm predicted at least one link probability above 40%, and (3) if DDD and the highest-predicted link were different (and the predicted link probability of the actual match was at least 0.1 lower than the best predicted match). For all mills that met one of these three criteria, we resent all of the candidate matches back to DDD for hand-linking. After iteration, the "unlikely" hand-linked matches were generally found to be reasonable matches (and missed by the machine-learning model) and the predicted "likely" matches were also generally decided to be matches after a second look. The automated linking model performed relatively worse in identifying ownership transfers, compared to the hand-links (Figure 2.18 Panel A).

Using this final hand-linked data, after iteration with the original model, we re-estimate the model to create final model-predicted links for our analysis. We consider two mills linked in the baseline ML linking specification if the predicted match probability is above 0.6. To eliminate a small number of multiple links from handlinking (3% of all links), we keep the mostly likely period 2 link for every period 1 establishment and then keep the most likely period 1 link for every period 2 establishment. There are a few tied matches (0.8% of all links), in cases where adjacent establishments in the same industry have the same owners; in these cases, we randomly select one of the establishments.

Linking Mill Owners to the Census of Population

We link mill owners to the complete Census of Population, using a similar procedure to our panel links. We construct an owner-name dataset with each probable person name ordering in the establishment name. For each owner-name, we keep up to 20 most likely matches in the Census from the same year and county who: were over 18 years old; had a matching first initial or first name Jaro-Winkler distance less than 0.3; and had a last name Jaro-Winkler distance less than 0.3. In rare cases when more than 20 individuals meet these criteria, we keep people with milling-adjacent occupations and those with the lowest string distances.

We sent the list of potential matches to Digital Divide Data, where two team members selected the best match (or no match) and a third team member reconciled all disagreements. Team members matched on the basis of: mill owner name and Census name; mill industry and Census person occupation.

Using the final match list, we first collapse between multiple matches, where for every owner name, we take the top match, sorting by milling status, last name distance, first name distance, and, for very rare cases, a seeded random variable. The same is done to collapse between multiple name orderings of the same owner, such that there is a list of unique owners paired to a single census person.

For mills with multiple owners who match to the Census of Population, we use all matches to characterize firm-level ownership characteristics: average owner age, whether any owner was born outside the United States (immigrant), and whether any owner has a self-reported occupation associated with being a "professional miller."[54] In most cases, only one owner

54. These occupations, listed in decreasing prevalence among the owners, are: Millers; Lumbermen and raftsmen; Sawyers; Manufacturers; Saw and planing mill operatives; Carpenters and joiners; Traders and dealers in lumber; Machinists; Mill and factory operatives (not specified); Mechanics (not specified); Traders and dealers in produce and provisions; Woolen mill operatives; Paper mill operatives; Cotton-mill operatives; Employees in manufacturing estabs. (not specified); and Traders and dealers in coal and wood.

name is linked to the Census of Population.

2.D Measuring County Waterpower Potential

This section describes how we measure county waterpower potential. We start with data on rivers in the United States (Section 2.D.1); define theoretical waterpower potential (Section 2.D.2); discuss our exclusion of rivers that were impractical for water power (Section 2.D.3); and aggregate flowline-level water power to the county-level, including adjustment for river segments that cross county boundaries (Section 2.D.4).

2.D.1 NHDPlusV2 Data

National Hydrography Dataset Plus is a national geospatial surface water framework for water resource analysis, developed and maintained by the U.S. EPA in partnership with the U.S. Geological Survey (USGS).

We use NHDPlus Version 2 (NHDPlusV2), released in 2012 (McKay et al., 2012).[55] NHDPlusV2 is built from multiple data sources, including: the medium-resolution (1:100,000) National Hydrography Dataset (NHD), 30 meter National Elevation Dataset (NED), and the National Watershed Boundary Dataset (WBD).

We generate waterpower potential for each "flowline" or "river segment," which is the basic unit in the NHD linear surface-water network. We use the two types of flowlines that represent natural rivers: "Stream Rivers" and "Artificial Paths." A Stream River (SR) is a river segment, often extending between tributary confluences. An Artificial Path (AP) represents a flow-path through a waterbody in the surface water network: for particularly wide rivers, normally those wider than 50 feet and longer than 2640 feet, an "artificial path"

55. Another version is NHDPlus High Resolution (NHDPlus HR), which is at a higher resolution (1:24,000-scale or better) (Moore et al., 2019), but does not currently include monthly streamflow estimates. The resolution of NHDPlusV2 is sufficient for us, particularly given that we later aggregate data to the county level.

is drawn to represent the flow-path within the waterbody.

2.D.2 Theoretical Water Power

For each river segment r, the theoretical water power generated from the flow of water along this segment can be derived using the following formula (assuming no friction):

$$\text{Theoretical Water Power}_r = \underbrace{\text{FlowRate}_r}_{\substack{\text{Cubic Feet}\\ \text{Per Second}}} \times \underbrace{\text{FallHeight}_r}_{\text{Feet}} \times \substack{\text{Gravitational}\\ \text{Constant}}, \tag{2.26}$$

where the gravitational constant roughly equals 0.1134 when the theoretical water-power is measured in horsepower. This formula closely approximates horsepower calculations in the 1880 Water Census.

Intuitively, the theoretical water power available is proportional to the flow rate of water (volume per second) and its falling height.

Flow Rate. Our data from NHDPlusV2 are based on the Enhanced Unit Runoff Method (EROM), a five-step procedure, to estimate mean monthly flow rates of rivers under natural conditions:

Step 1. Unit runoff based on a flow-balance model, taking into account: precipitation, potential evapotranspiration, evapotranspiration, and soil moisture.

Step 2. Adjustment for excessive evapotranspiration.

Step 3. Adjustment in a log-log regression estimated using reference gauge.

Step 4. Adjustment for flow transfers, withdrawals, and augmentations.

Step 5. Gage-adjustment based on actual observed flow at the gauge.

Step 4 is significant for our purposes, because the model predicts waterpower potential in the absence of the hydrological infrastructure built in the United States since the 19th century. The modeled water volume reflects natural waterflows, close to those observed in the 19th century (verified in Appendix Figure 2.11).

Fall Height. NHDPlusV2 data also provide the maximum and minimum elevation values for each river segment. Following the hydrology literature, we approximate the fall height (or hydraulic head) using the difference in elevation along each river segment.

2.D.3 Practical Water Power.

As discussed in the 1880 Water Census: "There is a sharp distinction to be made between *theoretical* and *actually available* water power" (emphasis original). Some sources of water power were infeasible (e.g., the Mississippi River). We discuss two reasons why theoretical water power was not usable in practice – river width and seasonality – and how this enters into our calculations of county waterpower potential.

River Width

We exclude wide rivers, such as the lower Mississippi River, that were impractical to dam for the purposes of generating water power. These rivers were also used for water transportation, which crowded out water power for manufacturing because millers had to provide rights of way. We use the maximum "top" (surface) width of rivers for NHD segments from the National Water Model (NWM), developed by NOAA (2016).[56]

For each county, we calculate local waterpower potential excluding rivers with maximum widths above a cutoff. Appendix Figure 2.25 plots the coefficient on Lower Waterpower against each cutoff, where the outcome is the number of water mills in 1850 (as in Table 2.2).

56. For more details of the National Water Model, see `https://water.noaa.gov/about/nwm`.

There is a sharp attenuation in the relationship for very wide rivers. Our main measure of county waterpower potential therefore excludes rivers that are wider than the 96th percentile (106.3 feet). This cutoff mostly excludes "Artificial Paths" in the database, including most of the lower Mississippi River network, which were impractical for water power use. We also exclude Niagara Falls from our analysis, as water-wheels during our sample period were "inadequate" for the magnitude of the falls (Adams, 1927): there was only one nearby water-mill in our sample, that opened in the late 1870s.

Seasonality

The seasonality of water flow rates is also important for the practical use of water power, in addition to average flow rates, because it determines whether watermills can be active throughout the year. Some mills were more seasonal, using water power when available, but the strong tendency was for mills to focus on year-round water power availability.

For many rivers, water flow rates varied over the year. We use the average flow rate over the three lowest months of the year, as historical accounts viewed this as a key determinant of feasible water power (Census Bureau, 1883). Consistent with these accounts, while we include "intermittent" rivers in our analysis, they do not on their own predict water power-use (Appendix Table 2.29, Column 2). Similarly, the average flow rates across all 12 months are less predictive of county water power-use in 1850 than our baseline approach (Appendix Table 2.29, Column 3).

2.D.4 Aggregating to County Waterpower Potential

The above procedure constructs river segment waterpower potential, which we aggregate to the county level for our analysis of US Census data. For flowlines that intersect county boundaries, we split flowlines into multiple segments that are contained entirely within county boundaries. We allocate the total river segment waterpower potential in proportion to the share of its length inside each county. We then sum across all river segments in a county.

2.E Other County-Level Data

This section provides additional detail on some of our supplementary data sources.

Market Access, Navigable Waterways, and Railroad Stations. We use measures of county "market access" in 1850, and decadal changes from 1850 to 1880 (Donaldson and Hornbeck, 2016; Hornbeck and Rotemberg, 2024). Market access is approximated as:

$$MA_c = \sum_{d \neq c} (\tau_{cd})^{-\theta} L_d. \tag{2.27}$$

The market access of county c is the trade-cost-weighted sum of population L in other counties d, where the iceberg trade cost τ is raised to the power of the trade elasticity. We set $\theta = 3.05$, following Hornbeck and Rotemberg (2024), and control for the log of county market access in 1850 and decadal changes in log county market access.

Measured transportation costs are based on least-cost routes using railroads, navigable waterways, and wagon transportation. We also control directly for whether the county is on a navigable river (as defined by Fogel 1964) or other navigable waterway (canal, lake, or ocean), and log distance to the nearest navigable waterway (based on average distance from 200 random points in the county to the nearest navigable waterway). Using maps of the railroad network in Colton (1882), we also collect detailed locations of railroad stations.

Coal Access. We digitized maps of workable coal deposit locations from Campbell (1908), a survey run by the United States Geological Survey. The map shows workable deposits for each type of coal (lignite, subbituminous, bituminous, and anthracite), and we calculate both if the deposits overlap with a county and the share of the county with a deposit. In addition to using measures of coal in the county, we also calculate the lowest-cost "iceberg" transportation cost from any workable deposit to each county along the transportation network.

Specifically, we assume that if there is coal in a county, there is no transportation cost to access coal. If there is no coal in a county, we calculate (a) the cheapest cost to a county with coal, using the iceberg transportation costs calculated by Hornbeck and Rotemberg (2024). We also calculate (b) the minimum wagon cost (again using the Hornbeck and Rotemberg 2024 costs) from the border of the county to the nearest coalfield. We then calculate the relative cost of shipping as the transportation cost divided by the price of coal, using the minimum of (a) and (b).

The actual price we use does not affect our regressions (because we take logs and use a national commodity price), but to be consistent we followed Cole (1938) and calculated the weighted average price of coal in 1880 (40% anthracite and 60% bituminous), using commodity prices from the Statistical Abstract of the United States[57]

Local Milling Material Availability. We define counties' wheat suitability using crop suitability data from the Global Agro-Ecological Zones project of the Food and Agriculture Organization (GAEZ-FAO), from Rusanov (2021). We also use counties' acreage share in woodland in 1870 (Haines, 2010).

Portage Site Locations. Following Bleakley and Lin (2012), we use data from Semple (1903) and Fenneman (1946) to measure whether counties contain actual or potential portage sites based on the fall line. We also included the historic location of portage sites along the Ohio, Missouri, and Mississippi rivers described by Bleakley and Lin (2012).

The Water Census We digitized the "detailed tables" of the water census, which gives us information of waterpower potential at the level of the site, which we then aggregate to the county level as we do with the NHDPlusV2.

57. Available at:
`https://fraser.stlouisfed.org/files/docs/publications/stat_abstract/pages/18654_1915-1919.pdf`.

2.F Switching Case Studies from Historical Society Records

For some cases in which incumbent water mills adopted steam power, we looked through historical society records (and other documents, when possible) for guidance on why these mills adopted steam and what impediments to steam adoption may have confronted incumbent water mills. This qualitative history of switching helps motivate assumptions of our model for why water incumbents faced higher costs of steam power than entrants.[58] The available historical detail was limited in most cases, or we were unable to find records for the mills, though we could generally see that most millers did not change locations and verify Census data on when mills switched to steampower.

Below, we provide some examples of millers (in alphabetical order) for whom we were able to find more-detailed information. These case studies suggest some of the push and pull factors behind mills switching from water to steam power:[59]

The Blanchards Brick Mill was built in 1842 in Watertown, Wisconsin (Watertown Historical Society, 2022). Due to concerns about low flow from the Rock River, the proprietors started construction of a steam mill (next door to their original mill) in the 1840s, though in our data the mill did not switch to steam until the 1860s.

The Canal Mill in Erie, Pennsylvania was sold by Jehiel Towner to Oliver & Bacon in 1865, who immediately converted it to a steam mill (Bates, 1884). Oliver & Bacon had previously operated a mill called Hopedale, located in the same county but outside the city, but left it to purchase the Canal Mill.

The Ellis Mill was built around 1838 by Moses Ellis, in Fayette County, Indiana (Barrows, 1917). After Moses' death in the 1840s, his son Lewis operated the mill for a few years, until he abandoned the watermill in the 1850s and built a steam mill in nearby Bentonville.

58. We are particularly grateful to David Kirchenbauer and Tony Li for outstanding research assistance in finding these historical sources. We also include examples of switching that we found in secondary sources.

59. We provide an additional example in Appendix Figure 2.13.

Elhanan Garland owned a water powered mill on the East bank of a stream in Kenduskeag, Maine, and Moses Hodson owned a water powered mill on the West bank of that same stream (Hubbard, 1861). After a lawsuit, it was determined that Garland had the senior water rights for using two stones of grist mill, but Hodson's rights were prior to Garland's for other purposes (such as a saw mill). Garland subsequently switched to steam power, but did not change locations.

Charles Gwinn, who was already a prominent miller exploiting high water power availability in Baltimore, built a steam powered mill there in 1813. He did not use steam power for very long, though, as it became clear that steam was "too costly to operate for milling flour" relative to water, in Baltimore at that time (Scharf, 1874; Sharrer, 1982).

The Graue Mill in Oak Brook, Illinois (which is now a museum, conveniently close to Chicago) was a gristmill that opened in 1852 (York Township Historical Society, 2023). The ground was relatively flat, so the immigrant owner (Frederick Graue) had to construct a dam to create a three foot fall. In order to expand, Graue spent three years retrofitting his mill for steam use (including the help of a visiting millwright). Graue had also made his own bricks on site, for the building, and seemed quite entrepreneurial and adventurous in further modifications prior to the steam engine's explosion.

The Hardesty Brothers inherited a profitable grist mill in Canal Dover, Ohio after their father died in 1869 (Hardesty, 2019). Within a decade, they borrowed money to buy a steam engine (without changing the location of their mill). The mill dissolved a few years later, and Hardesty (2019) speculates that one possible reason was due to the heavy financing needs.

Chauncey B. Knight inherited a water powered flour and grist mill built by his grandfather Nicholas Knight in Monroe, New York (Flour and Feed, 1945). Close to what is now Harriman State Park, the location has excellent access to water power. Knight converted the mill to run on steam power, which was the first steam mill in the county. Knight recounted that "it was freely predicted that it would be a failure," as many thought steam "could not compete with water power which was so much cheaper." Knight's mill was large enough to process corn meal, wheat bran and middlings, and malt sprouts by the "carload," with the bulk discounts allowing his mill to sell meal much more cheaply than his competitors.

E. E. Locke & Co operated a distillery along with a mill in Mifflin, Pennsylvania (Ellis and Hungerford, 1886). The mill only used water power in 1850 and only used steam power in 1860. The distillery and mills of E. E. Locke were destroyed by a fire in 1857. The rebuilding and the restoration was finished by 1858. We suspect that the mill switched from water to steam after and because of the fire, and otherwise the broad site of the mill stayed the same.

David and Andrew Luckenbach purchased a grist mill from their father in 1861 in Bethlehem, Pennsylvania (Jackson, 1975). As the business expanded, "the water power provided by Monocacy Creek was found unsatisfactory," and they installed steam engines in 1877 after a fire destroyed the original mill.

J.S. Manning owned a mill in Columbus, Wisconsin that used only water power in 1870 and used only steam power in 1880 (Jones, 1914). He purchased the mill in 1849, which was already the busiest mill in Central Wisconsin. It is described that the wait for grist work was often weeks. Manning is described as switching to steam power to keep up with demand. When the mill switched from water to steam power, the location of the mill did not change, though new machinery was added to the pre-existing mill.

John Orf purchased a mill in Allen County, Indiana in 1856 (Bates, 1945). Water from the Wabash and Erie Canal was taken into a mill pond just east of the St. Mary's aqueduct and run across an overshot wheel. Anticipating the canal's closure, Orff retrofitted the mill to be able to run on either steam or water power in the 1870s. The canal closed in the 1880s, at which point Orf's mill used steam power exclusively.

The Phoenix Mill in Millwakee, Wiscosin was built by brothers William and Edward Sanderson in 1847 (Andreas, 1881). William died in 1868, and Edward added Isaac van

Schnaick as a partner. They expanded the business, and switched to steam power.

The Shoemaker Mill was built in 1746 on a mill race off Tookany Creek in Montgomery County, Pennsylvania Rothschild (1976). The family operated the mill for 100 years before it was purchased by Charles Bosler, an employee. After Charles died, his son Joseph enlarged the mill and converted to steam power.

Williams & Lufbury owned a water powered lumber mill in Rahway, NJ (International Publishing Co, 1887). The mill used water power in the 1850 Census and steam power in the 1860 Census, without changing location. During that time, dams were abolished within the city limits.

Emery (1883) describes an (unnamed) water mill forced to switch to steam power because it lost its water rights. Emery (1883)'s goal was to describe the cost of switching to steam power, as testimony for a hearing to determine how much the mill should be compensated.

2.G Alternative Specifications

In this section, we discuss in more detail the robustness specifications described in Section 2.3.3. In each table, the first row corresponds to our main specification for comparison.

Appendix Tables 2.19, and 2.20 consider other county-level characteristics that could affect the relative adoption of steam power across counties with different waterpower potential. Correspondingly, the outcomes in these tables are our main county-level outcomes: the number of water establishments (column 1) and the steam share (column 2) in 1850, the growth in total establishments over each decade (columns 3-5), and the change in the share of mills using steam power (columns 6-8).

Appendix Tables 2.19 and 2.20 include additional controls for potential drivers of county-level steam adoption and economic growth. Appendix Table 2.19, rows 2 and 3, include additional controls for county access to coal (in addition to our baseline controls that include an indicator for any workable coal in the county, the share of the county covered by workable coal deposits, and access to workable coal deposits via the transportation network). Row 2 includes separate controls for each type of coal (lignite, subbituminous, bituminous, and anthracite). Row 3 controls for a cubic polynomial in the share of the county covered by workable coal deposits. Because different access to material inputs may have influenced flour and lumber mills' steam adoption (Ragnar, 1953), row 4 controls for county wheat suitability (from FAO-GAEZ data provided by Rusanov 2021) and row 5 controls for share of the county covered by woodland (as in Hornbeck 2010). Rows 6–8 control for county access to labor and capital inputs: row 6 controls for local wages in manufacturing in the Census data (Allen, 2009); row 7 controls for the share of county population who report being engineers or mechanics (Hanlon, 2022); row 8 controls for the number and total capital of local banks (Jaremski, 2014). Row 9 includes all of the above controls. Our results are broadly robust across these specifications, though the point estimates fall in row 9.

Appendix Table 2.20 adjusts our baseline controls for different influences on county growth. Rows 2–4 use subsets of our baseline controls: row 2 excludes our baseline controls for market access and navigable rivers; row 3 excludes our baseline controls for coal; and row 4 excludes both sets of controls. Row 5 controls for contemporaneous market access. Row 6 controls for contemporaneous population. This is itself an endogenous outcome to water power availability and the arrival of steam power, so this is not our preferred specification, but rather gives a sense of how much the evolution of overall economic activity matters as a control. Rows 7–12 alternatively control for time-invariant county characteristics (interacted with year), which adjust for potentially differential growth patterns across counties with different waterpower potential, though even 1850 county outcomes are influenced by county waterpower potential. Rows 7–10 control for variation in counties' initial settlement, which may have been associated with differential growth subsequently: row 7 controls for 1850 population; row 8 controls for being in Appalachia; row 9 controls for being on the frontier (Bazzi et al., 2020); and, given the historical pattern of spatial convergence in structural transformation, row 10 controls for the 1850 population share working in agriculture (Eckert and Peters, 2023). Row 11 controls for whether counties had historical portage sites, which less directly relevant by our sample period but had persistent path-dependent effects on economic activity (Bleakley and Lin, 2012).[60] Exposure to the Civil War had direct effects on economic activity (Margo, 2002; Feigenbaum et al., 2022),[61] and so row 12 includes controls for differential exposure to the Civil War, following Hornbeck and Rotemberg (2024): whether there was a battle in the county; the number of battles; the total number of casualties; an indicator for if the number of casualties was over 500; if the county was on the Union/Confederacy border; if the state had legal slavery in 1864; if the state seceded from the union; and the share of industrial activity in broadly war-related industries.[62] Row 13 controls for all of the time-invariant controls listed in rows 8–12, and row 14 controls for all of the time-invariant controls listed in rows 7–12.

The estimates are broadly robust across these specifications in Appendix Tables 2.19 and 2.20, though the estimated initial differences in 1850 are more sensitive to controls for population.[63] We view time-varying population as an example of "bad controls" that introduce

60. Conceptually, there are two differences between waterpower potential and portage sites, which create independent variation in the two. First, portage sites were on navigable rivers, whereas local waterpower potential can also come from non-navigable rivers. Second, portage sites reflect any discrete changes in elevation, whereas waterpower potential varies more continuously in terrain ruggedness. For example, the St. Anthony Falls in Minneapolis has a elevation change of 49 feet, almost double the height of the Falls of Ohio by Louisville. Both were portage sites, but the former was more useful for water power.

61. Feigenbaum et al. 2022 argue that during his 1864 March, Sherman's troops explicitly targeted lumber mills. Following their identification strategy, we confirm in our data that counties affected by Sherman's March experienced a decline in lumber mills. We also find that the survival rate fell. We do not find an effect on switching for the water incumbents, but we only have data on only nine affected counties with surviving mills (since the microdata for Georgia was lost), so the test is underpowered.

62. These broad war-related industries include: artificial limbs and surgical appliances; awnings and tents; coffins; cutlery, edge tools, and axes; drugs; chemicals and medicines; explosives and fireworks; flags and banners; gun- and lock-smithing; gunpowder; lead; military goods; ship and boat building; bronze; canning and preserving; carriage and wagon materials; carriages and wagons; clothing (general); cooperage; gloves and mittens; and hats and caps.

63. The controls related to the Civil War affect the point estimates in 1850, though by little.

bias (Angrist and Pischke, 2009), as county population is endogenous to our mechanism: milling in lower waterpower potential places benefited more from the adoption of steam power, lowering the local price index and drawing population to those places. Indeed, Appendix Table 2.15 shows that population grew more in counties with lower waterpower potential, so controls for population potentially capture the direct effects of steam power. Row 7, columns 1 and 2, suffers from the same issue: population in 1850 is also endogenous to county waterpower potential and the existing steam power, which makes it difficult to interpret effects conditional on counties' contemporaneous population. For this reason, we only include the time-invariant controls in our omnibus regressions (rows 13 and 14). Row 13, which does not control for 1850 population, is our preferred omnibus regression.

Appendix Tables 2.23 and 2.24 explore the influence of linkage error for our results. These tables compare entrant and incumbent outcomes, which are the estimates most likely affected by linkage errors. Appendix Table 2.23 shows how the entry rate (columns 1–3) and incumbent survival rate (columns 4–6) vary with county waterpower potential, in each decade. Appendix Table 2.24 shows results for steam use by entrants (columns 1–3) and water incumbents (columns 4–6). The rows correspond to the same alternative specifications across the two tables. For rows 2–5, we use the machine-learning (ML) links described in Appendix 2.C.4. Our benchmark ML model considers mills linked across decades if they have a match probability of at least 0.6. In row 2, we limit the panel links to only mills that are matched *both* by hand and by the benchmark ML model. In row 3, we use only the benchmark ML links. Row 4 restricts the matches to those with a ML-link probability of 0.8, and row 5 expands the matches to those with a ML-link probability of at least 0.4. Rows 2–5 change the survival and entry rates, mechanically, but do not qualitatively change the relationship between waterpower potential and entry or survival. In rows 6 and 7, our estimates are similar for mills with a predicted "business name," often based on a local geographic feature, or other mills named after their proprietors. Our baseline regression sample includes mills who report positive sales, regardless of their input costs, though we further limit the sample to mills who report all inputs to calculate the elasticity of substitution. Rows 8 and 9 show that our regression results are robust to these sample choices: row 8 restricts the sample to mills who report all inputs, and row 9 expands the sample to include the mills with unreported output (who were likely inactive at the time). Finally, row 10 includes mills that do not explicitly report using water or steam power, where we consider a mill as steam powered only if it explicitly mentions steam.

Appendix Table 2.25 shows the robustness of our results to changes in the county sample. Rows 2–5 consider the role of zeros in the data. Row 2 expands the sample to an unbalanced panel of all counties that ever had a mill in our sample period. Rows 3 and 4 constrain the sample to counties that had at least 3 or 5 mills in 1850, which are counties that are substantially less likely to report no mills in subsequent decades. When we limit the sample to at least 3 mills or 5 mills in 1850, we exclude 94 and 175 counties, respectively. Our baseline sample drops the two grouped counties with areas larger than a circle with a radius of 50 miles, and row 5 shows that our similar when we include them. Rows 6 and 7 exclude counties with extreme values of measured waterpower potential: row 6 drops the 1% largest and smallest values, and row 7 drops the 5% largest and smallest values. Rows 8 and 9 exclude counties that were more involved in trading mill output: row 8 drops the 20 largest cities in our sample, and row 9 drops cities that Kuhlmann (1929) describes as having export-oriented "merchant mills" (Baltimore, Buffalo, Chicago, Cincinnati, Cleveland, Milwaukee, Minneapolis, Oswego, Philadelphia, Richmond, Rochester, St. Louis, and Washington DC).

2.H Solution Algorithms

2.H.1 Dynamic Programming

The expected operating values $\mathbb{E}_\varepsilon[V^o_{ct}(R,\varphi)]$ are the key determinant of firms' forward-looking decisions. Once firms know the operating values, their optimal decisions about entry, exit, and power adoption in Equations (2.7)-(2.10) are only determined by contemporaneous features of the economy.

The expected operating values satisfy the Bellman equation:

$$\mathbb{E}_\varepsilon[V^o_{ct}(R,\varphi)] = \mathbb{E}_\varepsilon \max_{R'} \left\{ \begin{aligned} &\pi_{ct}(R,\varphi) - c_{ct}(R,R') - \varepsilon_{jct}(R) \\ &+ \delta \mathbb{E}_{(\varphi'|\varphi)} \mathbb{E}_\nu \max \left\{ \mathbb{E}_\varepsilon \left[V^o_{ct+1}(R',\varphi') \right] - f^{R'}_o - \nu^{R'}_{jct}(0), \Omega^{R'}_{ct} - \nu^{R'}_{jct}(1) \right\} \end{aligned} \right\}. \quad (2.28)$$

Equation (2.28) involves two maximization steps over distributions of idiosyncratic cost shocks (for adoption ϵ and operation/exit ν, respectively). The parametric assumptions in Section 2.5.5 simplify these steps. In particular, when the cost shocks follow Gumbel distributions, Equation (2.28) simplifies to a log-sum expression for the expected maximum (EMAX) (Train, 2009; Keane et al., 2011):

$$\mathbb{E}_\varepsilon[V^o_{ct}(R,\varphi)] = \rho \log \left[\sum_{R'} \exp \left\{ \frac{1}{\rho} \left(\pi_{ct}(R,\varphi) - c_{ct}(R,R') + \delta \mathbb{E}_{(\varphi'|\varphi)} \rho_o \log \left[\exp \left(\frac{\mathbb{E}_\varepsilon \left[V^o_{ct+1}(R',\varphi') \right] - f^{R'}_o}{\rho_o} \right) + \exp \left(\frac{\Omega^{R'}_{ct}}{\rho_o} \right) \right] \right) \right\} \right]. \quad (2.29)$$

We use the recursive scheme in Equation (2.29) to solve for the expected operating values in the steady states and along the transition path between the steady states. To do so, we discretize the productivity process using the **?** method on 100 grid points. We assume that firms have perfect foresight about the price index and steam share (our two aggregate state variables) up to unanticipated aggregate shocks to the economy (e.g., the first arrival of steam power or policy announcements).

Steady State

Equation (2.29) is a contraction mapping when operating values are stationary, $\mathbb{E}_\varepsilon[V^o_{ct+1}(R,\varphi)] = \mathbb{E}_\varepsilon[V^o_{ct}(R,\varphi)]$, so we can solve for the unique fixed point $\mathbb{E}_\varepsilon[V^o_{ct}(R,\varphi)]$ by iterating on Equation (2.29) until convergence. Convergence of the value function iteration procedure is ensured by Blackwell's sufficient conditions for contraction mappings (Stokey et al., 1989, Theorem 4.6).

Transition Path

Starting from the terminal steady-state values $\mathbb{E}_\varepsilon[V^o_{cT_1}(R,\varphi)]$, we may solve for the operating values along the transition path $\{\mathbb{E}_\varepsilon[V^o_{ct}(R,\varphi)\}_{t=T_0}^{T_1-1}$ using backward recursion on Equation (2.29) from $T_1 - 1$ to the initial period T_0.

2.H.2 Dynamic Equilibrium

This section discusses how we solve for the dynamic equilibrium of our economy.

We first describe our algorithms for solving the equilibrium in steady states and along a transition path. In brief, we use a shooting algorithm that iterates on the time paths for the mass of operating firms and entrants to find a fixed point of the equilibrium policy functions.

We then discuss the properties of our solution algorithm, including the existence and uniqueness of equilibrium. The convergence of our iterative algorithm is ensured by a congestion force in the product market. The convergence property also ensures that an equilibrium exists and tends to be unique, although strong agglomeration effects in steam adoption can lead to multiple equilibria, which we consider directly.

Steady State

This section describes how we solve for the steady state equilibrium. We use a nested algorithm, where the outer loop searches for the mass of entrants M_c that closes the free entry condition, and the inner loop iterates over the mass of operating firms $F_c(R, \varphi)$ to find a fixed point of the equilibrium policy functions for exit and power adoption. Our solution algorithm reads as follows.

(i). Set an initial grid for the mass of entrants $\{M_c^{(0)}, M_c^{(1)}, M_c^{(2)}, ...\}$. For each grid point $(i) = 0, 1, 2, ..., :$

(ii). Solve for the equilibrium mass of operating firms $F_c^{(i)}(R, \varphi)$:

(a) Set an initial guess for the mass of firms $F_{ct}^{(i,0)}(R, \varphi)$. For each iteration $(j) = 0, 1, 2, ..., :$

(b) Solve for the expected operating values $\mathbb{E}_\varepsilon[V_{ct}^{o(i,j)}(R, \varphi)]$ by iterating on the contraction mapping in Equation (2.29).

(c) Simulate the mass of operating firms: given $F_{ct}^{(i,j)}$ and $M_{ct}^{(i)}$, use the policy functions for exit and power adoption (Equations (2.9)-(2.10)) to simulate the firm mass $F_{ct}^{(i,NEW)}(R, \varphi)$.

(d) Update the mass of operating firms:

$$F_{ct}^{(i,j+1)}(R, \varphi) = \lambda F_{ct}^{(i,NEW)}(R, \varphi) + (1 - \lambda) F_{ct}^{(i,j)}(R, \varphi), \tag{2.30}$$

where $\lambda = 0.5$ is the relaxation parameter in the Gauss-Seidel update.

(e) Repeat Steps (ii)b-(ii)d until $\sum_{R,\varphi} |F_{ct}^{(i,j+1)}(R, \varphi) - F_{ct}^{(i,j)}(R, \varphi)| \leq tol_F$ for a small tolerance level tol_F.

(iii). Evaluate the free entry condition:

(a) Compute entry values $EV_{ct}^{(i)} = \mathbb{E}_\varphi \left[V_c^{(i)}(E, \varphi) \right]$ by plugging $\mathbb{E}_\varepsilon[V_c^{o(i)}(R, \varphi)]$ into Equation (2.9) and integrating over the stationary distribution for φ.

(b) Compute the deviation from the free entry condition:

$$\mathcal{F}_c^{(i)} = EV_c^{(i)} - f^e \tag{2.31}$$

(iv). Update the mass of entrants $M_c^{(i+1)}$ to set the predicted free entry condition to zero. We use a linear interpolation based on the previous iterations $\{M_c^{(k)}, \mathcal{F}_c^{(k)}\}_{k=0}^{i}$.

(v). Repeat Steps (ii)-(iv) until $|\mathcal{F}_c^{(i)}| \leq tol_M$ for a small tolerance level tol_M.

We solve for the initial steady state ($T_0 = 1830$) and the terminal steady state (after $T_1 = 1900$). In the initial equilibrium, water power is the only available power source, which we model with a prohibitively high cost of steam adoption $c_{T_0}(S)$. In the terminal equilibrium, the cost of steam power has reached its new steady-state level.

Transition Path

This section describes how we solve for the transition path between the initial steady state ($T_0 = 1830$) and the terminal steady state ($T_1 = 1900$).

The dynamic equilibrium along the transition path is a technically challenging fixed point: We simulate a 70-year transition path, where heterogeneous firms make forward-looking decisions about entry, exit, and power adoption, as steam costs are falling over time, and decisions are interlinked through competition in product markets and agglomeration spillovers in steam power.

We use a nested shooting algorithm, where the outer loop searches for a time path for the mass of entrants that closes the free entry condition, and the inner loop iterates over the mass of operating firms to find a fixed point of the equilibrium policy functions for exit and power adoption. Our solution algorithm reads as follows.

(i). Set an initial guess for the mass of entrants $M_{ct}^{(0)}$. For each iteration $(i) = 0, 1, 2, ...,$:

(ii). Solve for the equilibrium mass of operating firms $F_{ct}^{(i)}(R,\varphi)$:

(a) Set an initial guess for the mass of operating firms $F_{ct}^{(i,0)}(R,\varphi)$. For each iteration $(j) = 0, 1, 2, ...,$:

(b) Solve for the expected operating values $\mathbb{E}_{\varepsilon}[V_{ct}^{o(i,j)}(R,\varphi)]$ by iterating on the contraction mapping in Equation (2.29).

(c) Simulate the mass of operating firms: given $F_{ct}^{(i,j)}$ and $M_{ct}^{(i)}$, use the policy functions for exit and power adoption (Equations (2.9)-(2.10)) to simulate the firm mass $F_{ct}^{(i,NEW)}(R,\varphi)$.

(d) Update the mass of operating firms:

$$F_{ct}^{(i,j+1)}(R,\varphi) = \lambda F_{ct}^{(i,NEW)}(R,\varphi) + (1-\lambda)F_{ct}^{(i,j)}(R,\varphi), \tag{2.32}$$

where $\lambda = 0.5$ is the relaxation parameter in the Gauss-Seidel update.

(e) Repeat Steps (ii)b-(ii)d until $\sum_{R,\varphi,t} |F_{ct}^{(i,j+1)}(R,\varphi) - F_{ct}^{(i,j)}(R,\varphi)| \leq tol_F$.

(iii). Evaluate the free entry condition:

(a) Compute entry values $EV_{ct}^{(i)} = \mathbb{E}_{\varphi}\left[V_{ct}^{(i)}(E,\varphi)\right]$ by plugging $\mathbb{E}_{\varepsilon}[V_{ct}^{o(i)}(R,\varphi)]$ into Equation (2.9) and integrating over the stationary distribution for φ.

(b) Compute the deviations from the free entry condition:

$$\mathcal{F}_{ct}^{(i)} = EV_{ct}^{(i)} - f^e \tag{2.33}$$

(iv). Update the path of entrants $M_c^{(i+1)}$ to set the predicted free entry condition to zero. We use a Newton-Rhapson method to update the mass of entrants:

$$M_c^{(i+1)} = M_c^{(i)} - \lambda J_{\mathcal{F}}\left(M_c^{(i)}\right)^{-1}\mathcal{F}_{ct}^{(i)}, \tag{2.34}$$

where $J_{\mathcal{F}}\left(M_c^{(i)}\right)$ is the Jacobian of the free entry condition $\mathcal{F}_c^{(i)}$, evaluated numerically around $M_c^{(i)}$, and $\lambda = 0.5$ is a dampening parameter that mitigates overshooting and ensures stable convergence toward clearing the free entry condition.

The Newton-Rhapson method is versatile but also potentially unstable, as Equation (2.34) is a system of 70 free entry conditions in 70 unknown masses of entrants. To mitigate erratic fluctuations in $M_c^{(i+1)}$, we apply a lowess smoother (that allows for breakpoints at shocks) to the path of entrants after each Newton update.

(v). Repeat Steps (ii)-(iv) until $||\mathcal{F}_c^{(i)}|| \leq tol_M$.

To ensure smooth convergence at the end of our transition path, we extrapolate the final years of M_c^* before running the inner loop for the mass of operating firms one final time.

As a consistency check, we verify that the mass of operating firms has reached its terminal steady-state values by T_1. Otherwise, the time horizon T_1 has to be expanded.

Approximate Path of Entrants. The algorithm for finding the path of entrants in Section 2.H.2 is versatile and exact but also computationally expensive. We aid our algorithm with an approximate method that works well when the economy is transitioning smoothly between two known steady states (as in our baseline simulations). As we describe below, we use the approximation as starting values in Step (i) of Section 2.H.2, and to ease the computational burden of the structural estimation in Section 2.6.

Our approximation to the path of entrants M_{ct} is based on the knowledge that: (i) the economy transitions between the steady states found in Section 2.H.2, and (ii) the only driving force along the transition path is a steadily falling steam cost. In particular, we know that lower steam costs induce more entry, more steam adoption, and a lower price index. Hence, we search for a transition path where the mass of entrants evolves smoothly between the steady states:

$$M_{ct}(\xi) = \exp\left(\log M_{cT_0} + \left(\frac{t - T_0}{T_1 - T_0}\right)^{\xi}\left(\log M_{cT_1} - \log M_{cT_0}\right)\right) \quad t \in [T_0, T_1], \tag{2.35}$$

where $\xi > 0$ governs the speed of convergence to the terminal steady state. Our goal is to find the value ξ^* that satisfies free entry and the other equilibrium conditions.

(i) Set an initial grid for the mass of entrants $\{\xi_c^{(0)}, \xi_c^{(1)}, \xi_c^{(2)}, ...\}$. For each grid point $(j) = 0, 1, 2, ...,:$

(ii) Perform Steps (ii)-(iii) of Section 2.H.2 for each value of $\xi^{(j)}$.

(iii) Update the parameter $\xi^{(j)}$ to set the predicted free entry condition to zero. We use a linear interpolation based on the previous iterations $\{\xi^{(k)}, \mathcal{F}_c^{(k)}\}_{k=0}^{j}$.

(iv) Repeat Steps (ii)-(iii) until $|\sum_t \mathcal{F}_{ct}^{(j)}| \leq tol_M$.

The approximate path of entrants $M_{ct}(\xi^*)$ performs well in our baseline simulations: The mean absolute deviation of the free entry condition $\mathcal{F}_{ct}^*$ is less than 0.005% of average firm sales. The approximation has the advantage of greater computational efficiency compared to the exact method in Section 2.H.2. In particular, the approximate and exact algorithms take, respectively, 4 seconds and 8.5 minutes to solve the baseline equilibrium. This difference in computational time is valuable when estimating the model, where the equilibrium needs to be solved and simulated repeatedly at various parameter values. Hence, to ease the computational burden of the estimation procedure, we use the approximate path of entrants when estimating the model in Section 2.6. The versatility of the exact algorithm is useful when evaluating the counterfactual experiments in Section 2.7.

Existence of Equilibrium

The convergence of our iterative algorithm (and thus the existence of an equilibrium) is ensured by the competition between firms in product markets, creating a congestion force (as summarized by the price index P_{ct}). For intuition, we describe a few practical examples of the congestion force.

First, suppose entry values exceed the fixed entry cost (such that the free entry condition in Equation (2.12) is not met) at our initial guess. More firms will then enter the market. The additional entrants strengthen the competition (i.e., lower the price index P_{ct}), which lowers profits ($\frac{\partial \pi_{ct}(R,\varphi)}{\partial P_{ct}} > 0$ in Equation (2.6)) and the value of entry.

Similarly, suppose the optimal survival rates exceed our initial guess. More firms will then stay in business. The additional operating firms lower the price index P_{ct}, which decreases operating values and, thus, optimal survival rates.

Finally, suppose the optimal steam adoption rates exceed our initial guess. More firms will then adopt steam power. The additional steam users lower the price index P_{ct} (when steam has lower marginal costs, $\gamma > 0$), which decreases optimal steam adoption (because of the profit complementarities between steam power and the price index, $\frac{\partial \pi_{ct}(S,\varphi)}{\partial P_{ct}} > \frac{\partial \pi_{ct}(W,\varphi)}{\partial P_{ct}}$ when $\gamma > 0$).

Uniqueness of Equilibrium

As Section 2.H.2 describes, the convergence of our solution algorithm relies on a monotone relationship between the mass of firms (steam users) and the price index: a higher price index induces more entry/survival (steam use), which in turn lowers the price index. This monotone relationship also tends to ensure the equilibrium of the economy is unique.

To see this, suppose – for the sake of contradiction – that the economy could sustain two equilibria with different masses of entrants. The price index in the "low entry" equilibrium would then be higher, all else equal. However, that higher price index would induce more entry, contradicting its "low entry" nature.

A strong steam agglomeration force (i.e., a very positive α_S or very negative κ) could, however, lead to multiple equilibria. For example, suppose that the agglomeration force is so strong that a higher steam share s_{ct} makes even more mills want to adopt steam (i.e., $\frac{d\pi_{ct}(S,\varphi)}{ds_{ct}} \geq \frac{d\pi_{ct}(W,\varphi)}{ds_{ct}}$). In this case, the economy could sustain multiple equilibria: a "low steam" equilibrium where few mills adopt steam (because the agglomeration force is weak) and a "high steam" equilibrium where many mills use steam (because the agglomeration force becomes strong).

The potential for multiple equilibria is larger when steam is more available, so that more firms are at the margin of steam adoption. We check for multiple equilibria in our terminal steady state (when steam power is fully available) by initiating our solution algorithm at different starting values for the equilibrium steam share (from 0% to 100%). The solution algorithm converges to our baseline equilibrium for all initial values. We also do not find persistent effects of "cash for clunkers" style programs described in Section 2.7.2, even those that temporarily raise steam adoption to well above its steady-state usage.

2.I Structural Estimation

2.I.1 Estimation Procedure

We estimate the structural model using a Newton-Rhapson algorithm that leverages the relationships between parameters and moments discussed in Sections 2.6.1-2.6.1. The method iteratively adjusts the parameter values $\theta \in \mathbb{R}^K$ to match model-simulated moments $f(\theta) \in \mathbb{R}^K$ to their target values $y^* \in \mathbb{R}^K$.

Starting from an initial value θ_0, the Newton method updates the parameter estimates as follows:

$$\theta_{n+1} = \theta_n - \lambda J_f(\theta_n)^{-1}(f(\theta_n) - y^*), \tag{2.36}$$

where $J_f(\theta_n)$ is the Jacobian of the moment function f, evaluated numerically around θ_n, and $\lambda = 0.5$ is a dampening parameter that mitigates overshooting and ensures stable convergence to the target values.

The theoretical relationships between parameters and moments described in Sections 2.6.1-2.6.1 are critical for the performance of the Newton method. In particular, the method works well when parameters and moments have smooth (especially linear) relationships (such that J_f does not change too rapidly) and the parameters have distinct (especially one-to-one) mappings to each target moment (such that J_f is well-conditioned and non-singular).

We make three adjustments to the estimation procedure to ensure these regularity conditions are robustly met.

First, we estimate the baseline productivity process (π, σ) and entry costs f^e in an initial step to match their target moments before the arrival of steam power. Second, we implement an adaptive grid search in the steam production parameters (γ, f_o^S), executing the Newton method on each grid point. Third, we adopt a dimensional continuation strategy for our Newton method, gradually incorporating more parameter-moment pairs into the estimation problem:

(a) *Steam adoption within regions:* estimate $c_S^{(initial)}, c_S^{(terminal)}, c(R, R')$ to match their target moments.

(b) *Steam adoption between regions:* add $(c_L(W), \kappa)$ and their target moments to the estimation problem.

(c) *Output between regions:* add (α_S, η) and their target moment to the estimation problem.

(d) *Startup and fixed costs*: add (f_o^E, f_o^W) and their target moments to the estimation problem.

Our estimation algorithm only proceeds to the next step once the incorporated moments are sufficiently close to their target values. These adjustments ensure that our estimation algorithm is well-behaved. We validate that J_f, at all iterations n, has the signs and magnitudes predicted in Sections 2.6.1-2.6.1.

2.I.2 Identification of Structural Parameters

We now further analyze the local relationships between parameters and moments around the best-fit values θ^*. Appendix Tables 2.31 and 2.32 report two standard measures of parameter identification: the Jacobian of the moment function, which captures how simulated moments change with parameter values,[64] and the sensitivity measure of Andrews et al. (2017), which captures how estimated parameters change with target moments.[65]

We show these relationships for our Newton-based estimation, which relies directly on the Jacobian for the estimation (see Section 2.I.1). We order the table rows and columns such that the diagonal elements capture the relationship between parameters and their target moments, as discussed in Sections 2.6.1-2.6.1. The tables yield several insights into the identification of our structural model.

First, the simulated moments are highly sensitive to our parameters, suggesting that our parameter estimates are tightly identified. For example, increasing the water-to-steam switching costs by 1% of firm sales brings the incumbent-to-entrant steam switching rate 6.4 percentage points away from its perfectly fitted target values, cf. the first element of the Jacobian matrix.

64. The Jacobian is a commonly used diagnostic to assess the empirical properties of structural models (see, e.g., Berger and Vavra (2015); Ottonello and Winberry (2020); Balke and Lamadon (2022)).

65. The sensitivity matrix M is related to the Jacobian J as follows: $M = (J'WJ)^{-1}J'W$, where W is a weighing matrix that does not matter in our exactly-identified case.

Second, there is a particularly strong link between model parameters and each of their target values, as the Jacobian and sensitivity matrices have pronounced excess mass along their diagonals. This suggests that the selected target moments are particularly important for identifying each of the parameters.

Third, and reassuringly, all the diagonal elements have the theory-predicted signs, as the relationship between moments and parameters have the directions predicted in Sections 2.6.1-2.6.1.

Finally, the Jacobian and sensitivity matrices also have important off-diagonal elements, which highlight the importance of estimating the model parameters jointly. For example, Appendix Table 2.32 shows that a higher water exit rate implies that steam costs must be higher to rationalize the observed level of steam adoption.

2.J Counterfactual Experiments

2.J.1 Option Value Decomposition

In this section, we describe how to decompose firm values into operating profits, the option value of exit, and the option value of steam power, as discussed in Section 2.7.1.

The value of a water mill (Equations (2.9)-(2.10)) is determined by its productivity (its idiosyncratic state variable, φ), the steam adoption cost path (the exogenous aggregate state variable, $\mathbf{c_t^S} = \{c_\tau^S\}_{\tau=t}^{\infty}$), as well as the paths for the price index and steam adoption rate (the endogenous state variables, $\mathbf{P_t}$ and $\mathbf{s_t}$):

$$\mathbb{E}_\varepsilon[V_{ct}^o(\varphi, W)] = V(\varphi, \mathbf{c_{Bt}^S}, \mathbf{P_{Bt}}, \mathbf{s_{Bt}}), \tag{2.37}$$

where subscript B denotes the baseline values.

The option value of steam power reflects the differences in firm value if the water mill cannot access steam power, keeping all other state variables fixed at their baseline values:

$$\text{OVS}_t(\varphi, W) = V(\varphi, \mathbf{c}^{\mathbf{S}}_{\mathbf{Bt}}, \mathbf{P_{Bt}}, \mathbf{s_{Bt}}) - V(\varphi, \infty, \mathbf{P_{Bt}}, \mathbf{s_{Bt}}) \tag{2.38}$$

The option value of exit reflects the additional difference in firm value relative to a water mill that is forced to stay in business indefinitely:

$$\text{OVE}_t(\varphi, W) = V(\varphi_t, \infty, \mathbf{P_{Bt}}, \mathbf{s_{Bt}}) - \text{OP}_t(\varphi, W). \tag{2.39}$$

The value of staying in business with water is the present-discounted value of operating profits:

$$\text{OP}_t(\varphi, W) = \sum_{\tau=0}^{\infty} \delta^{\tau} \mathbb{E}\left[\pi(\varphi_{t+\tau}, W, P_{Bt+\tau}) - f_o^W | \varphi_t = \varphi\right], \tag{2.40}$$

where the flow profit π_t is determined by the mill's productivity φ_t and the price index P_t.

Finally, combining Equations (2.37)-(2.40), we can decompose the value of a water mill into operating profits, the option value of exit, and the option value of steam power:

$$\mathbb{E}_{\varepsilon}[V^o_{ct}(\varphi, W)] = \text{OP}_t(\varphi, W) + \text{OVE}_t(\varphi, W) + \text{OVS}_t(\varphi, W). \tag{2.41}$$

Table 2.11 reports the effect of steam power on each of the terms of Equation (2.41).

2.J.2 Consumer Surplus

We measure the consumer surplus from a policy using equivalent-variation impacts on consumer prices. That is, we calculate the transfer that would deliver the change in real consumption that is equivalent to the one caused by the policy's impact on consumer prices.

As specified in Section 2.5.1, consumers' utility from mills' products is CES with elasticity

ϵ, such that $P_{ct} = \left[\int p_{jct}^{1-\epsilon}\, dj\right]^{\frac{1}{1-\epsilon}}$ is the utility-consistent price index.

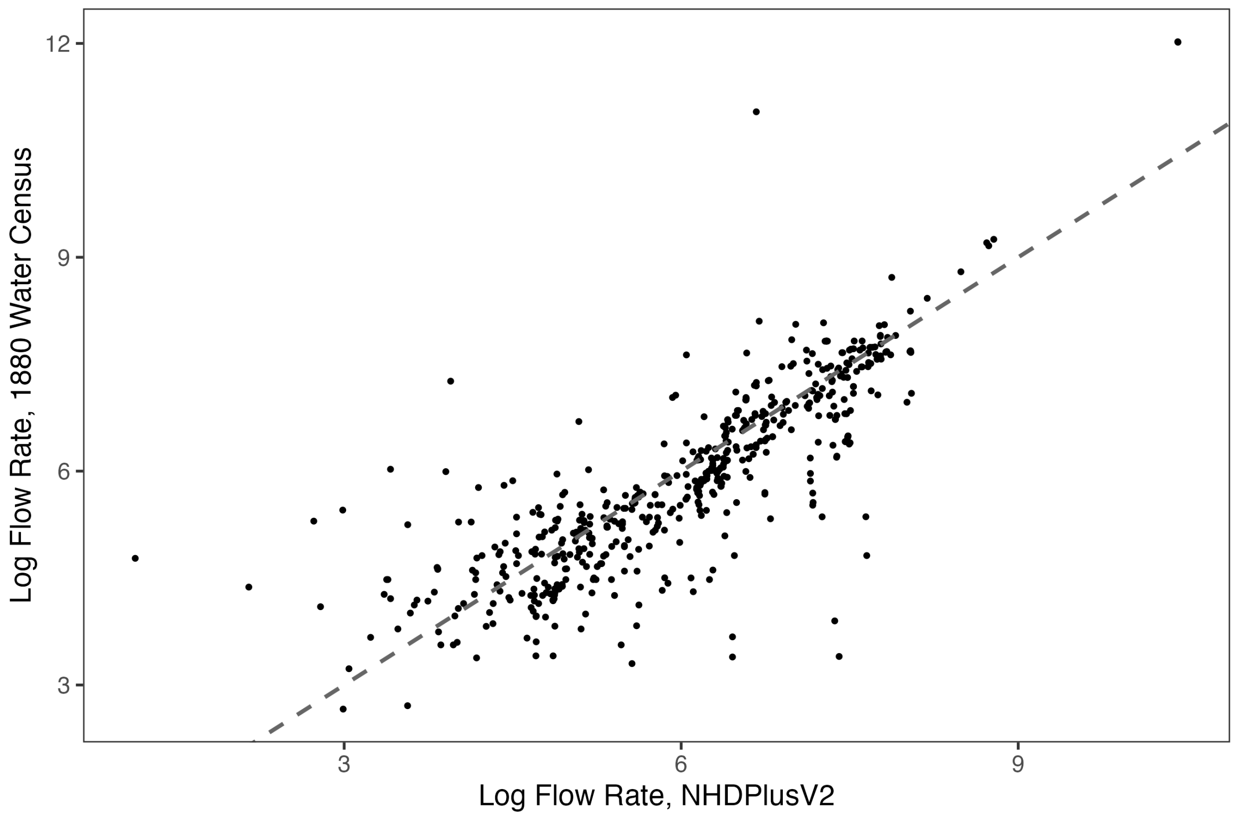
The consumer surplus (CS) of a policy enacted in year t_0 is

$$\mathrm{CS}(P_1, P_0 | C_{0t}) = \sum_{t=t_0}^{\infty} \delta^{t-t_0} C_{0t} \times \left(\frac{1}{P_{1t}} - \frac{1}{P_{0t}} \right), \tag{2.42}$$

where P_{1t} and P_{0t} are the consumer prices in year t of the policy and baseline equilibria, and C_{0t} is the baseline path of nominal consumption.

2.K Appendix Figures

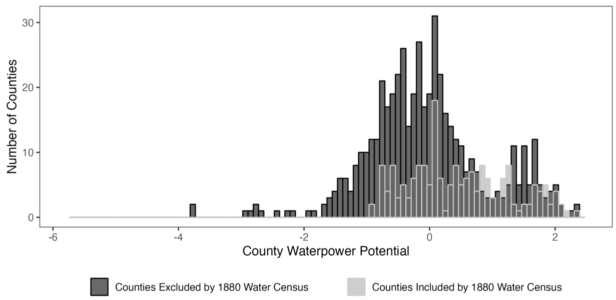
Figure 2.11. River Segment Flow Rates, in the 1880 Water Census Compared to NHDPlusV2



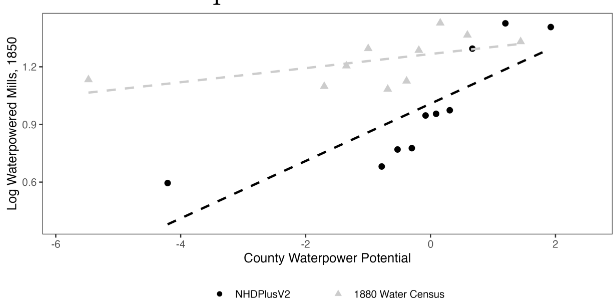
Notes: This figure compares the log water flow rates of river segments that we linked by name from the 1880 Water Census to the National Hydrography Dataset Plus Version 2.0 (NHDPlusV2). Each point represents one linked river segment.

Figure 2.12. Selected Coverage in the 1880 Water Census, Compared to Comprehensive NHDPlusV2 Data

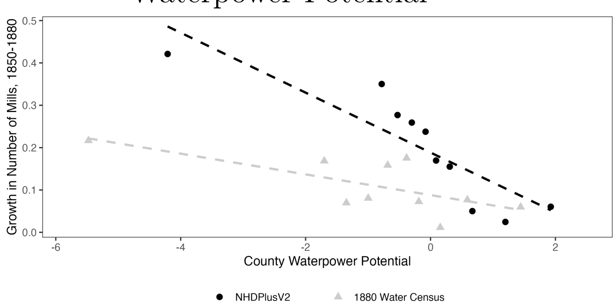
Panel A. Distribution of County Waterpower Potential, for Counties Included and Excluded by 1880 Water Census



Panel B. Measured Relationship between 1850 Water Powered Mills and County Waterpower Potential



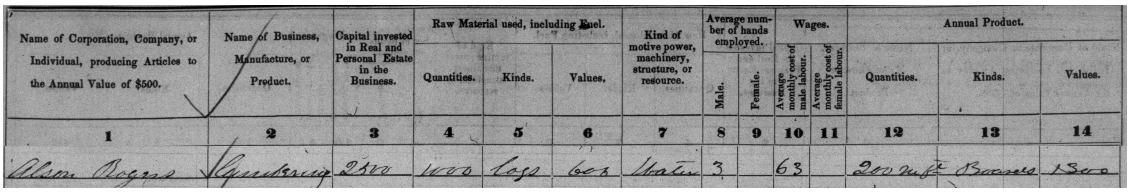
Panel C. Measured Relationship between 1850-1880 Mill Growth and County Waterpower Potential



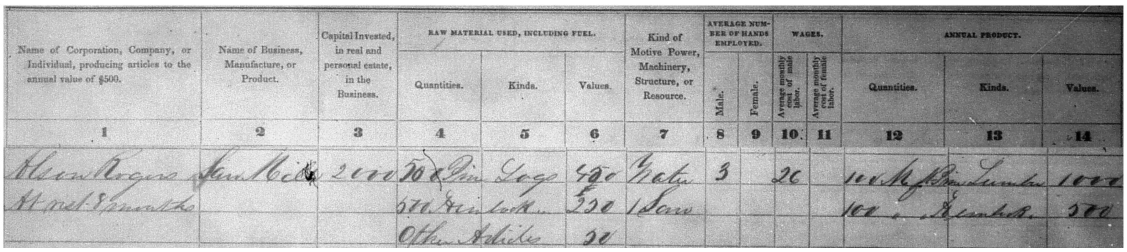
Notes: Panel A shows the distribution of country waterpower potential, measured using NHDPlusV2 data, for counties included by the 1880 Water Census (light gray) and counties excluded by the 1880 Water Census (dark gray). Panel B shows a binscatter of the unadjusted relationship between the number of water powered mills in 1850 and county waterpower potential, using the full NHDPlusV2 data and the Water Census data. Panel C shows a binscatter of the unadjusted relationship between the growth in the number of mills between 1850 and 1880 and county waterpower potential, using the full NHDPlusV2 data and the Water Census data. Panels B and C use PPML estimation, which approximates percent differences in the rates. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880), NHDPlusV2, and Census Bureau (1883).

Figure 2.13. Example Census Images: The Rogers' Lumber Mill

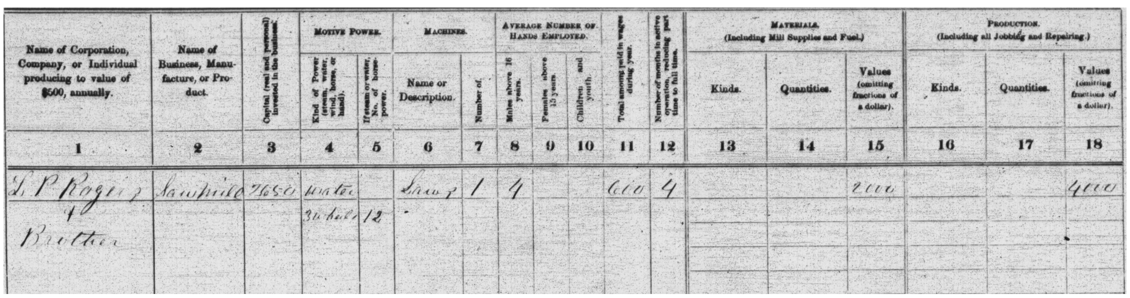
Panel A. 1850



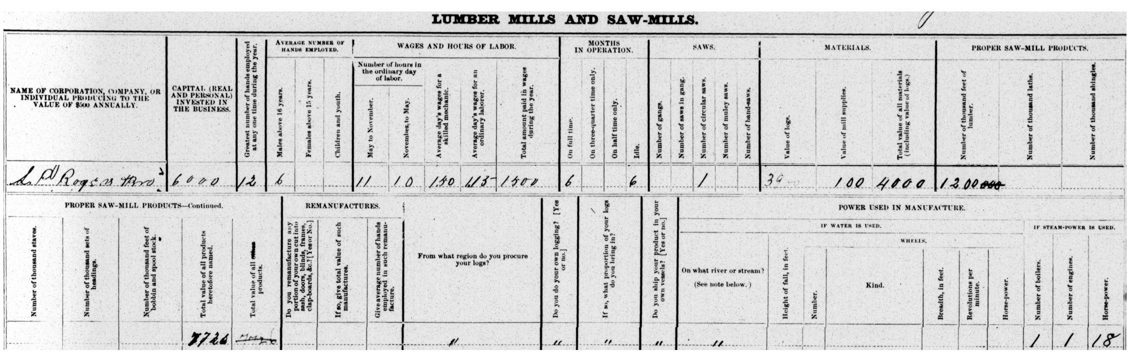
Panel B. 1860



Panel C. 1870

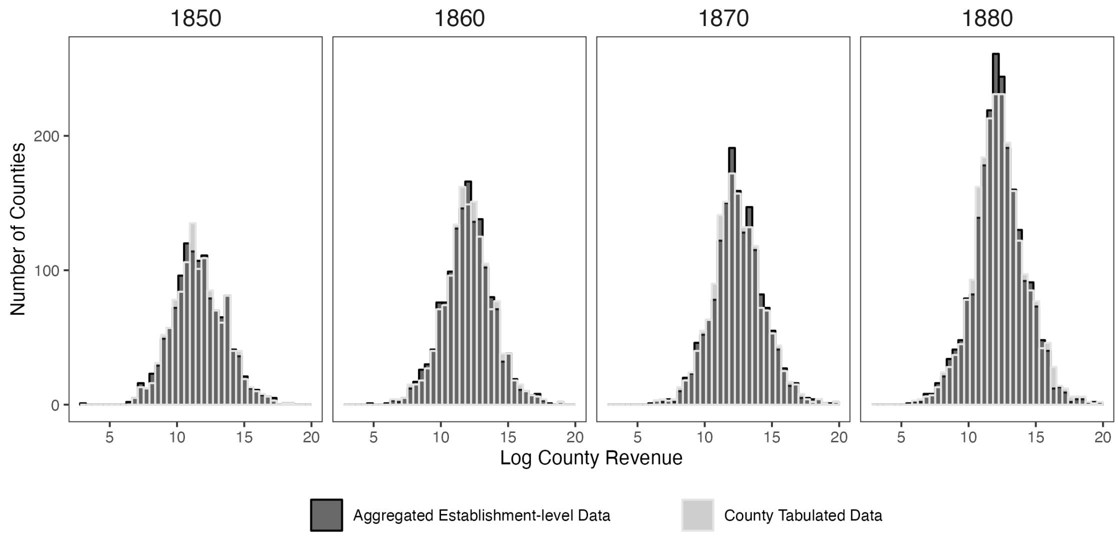


Panel D. 1880



Notes: This figure shows example images for the Census of Manufactures in each decade, and follows the Rogers' Mill across each decade. Alson Rogers settled in Warren, Pennsylvania and started in the lumber business after marrying in 1835. After he passed away in 1867, his sons Lucian (the "L.P." seen in the 1870 and 1880 Census images) and Burton took over the business, and built a steam engine. Sources: Schenck and Rann (1887), Census of Manufacturers (1850-1880), Census of Population (1850-1880).

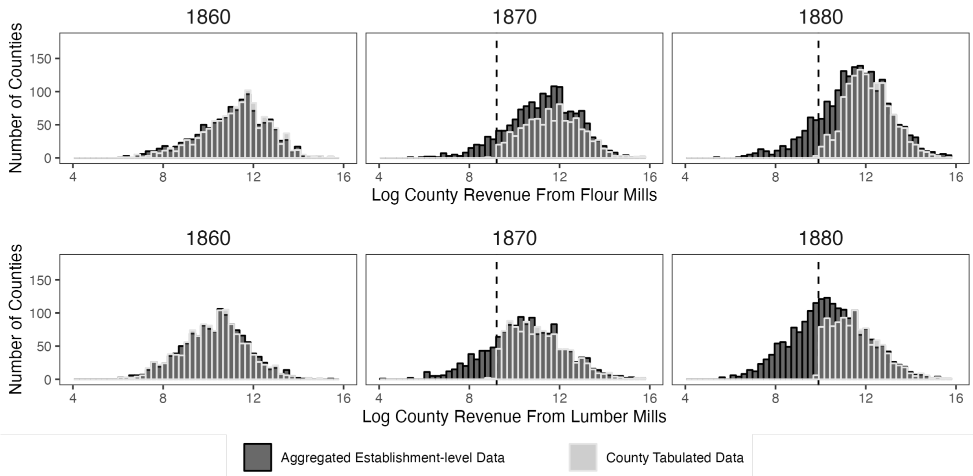
Figure 2.14. Distribution of County-Level Manufacturing Revenue, in County Tabulations and Aggregated Establishment-level Data



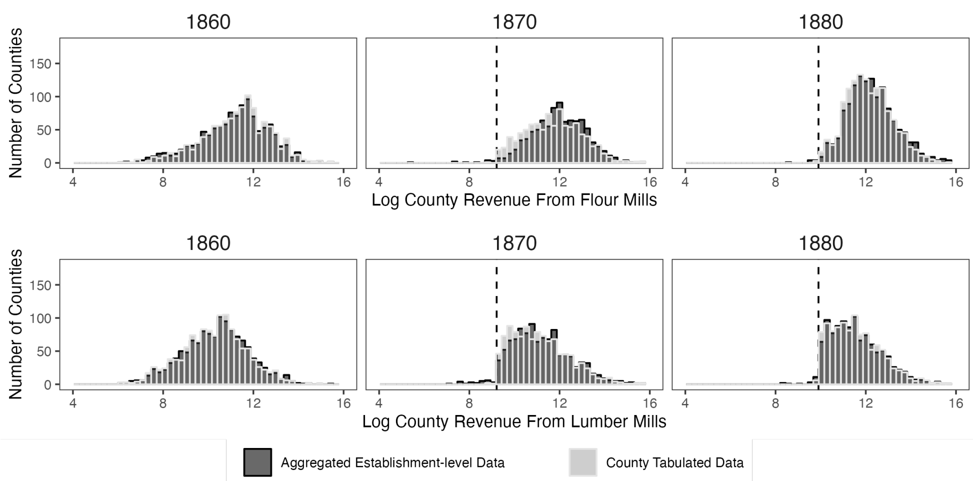
Notes: This figure shows the distribution of total recorded manufacturing revenue by county, comparing county-level tabulations made contemporaneously by the Census against the county-level sums of our digitized establishment-level data from Census manuscripts. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880), county-level tabulations (Haines, 2010).

Figure 2.15. Unreported Data in County-Industry Tabulations, for Flour and Lumber Mills, Compared to Aggregated Establishment-Level Data

Panel A. Distribution of County Revenue for Flour Mills and Lumber Mills, in County-Industry Tabulations or Aggregated Establishment-level Data

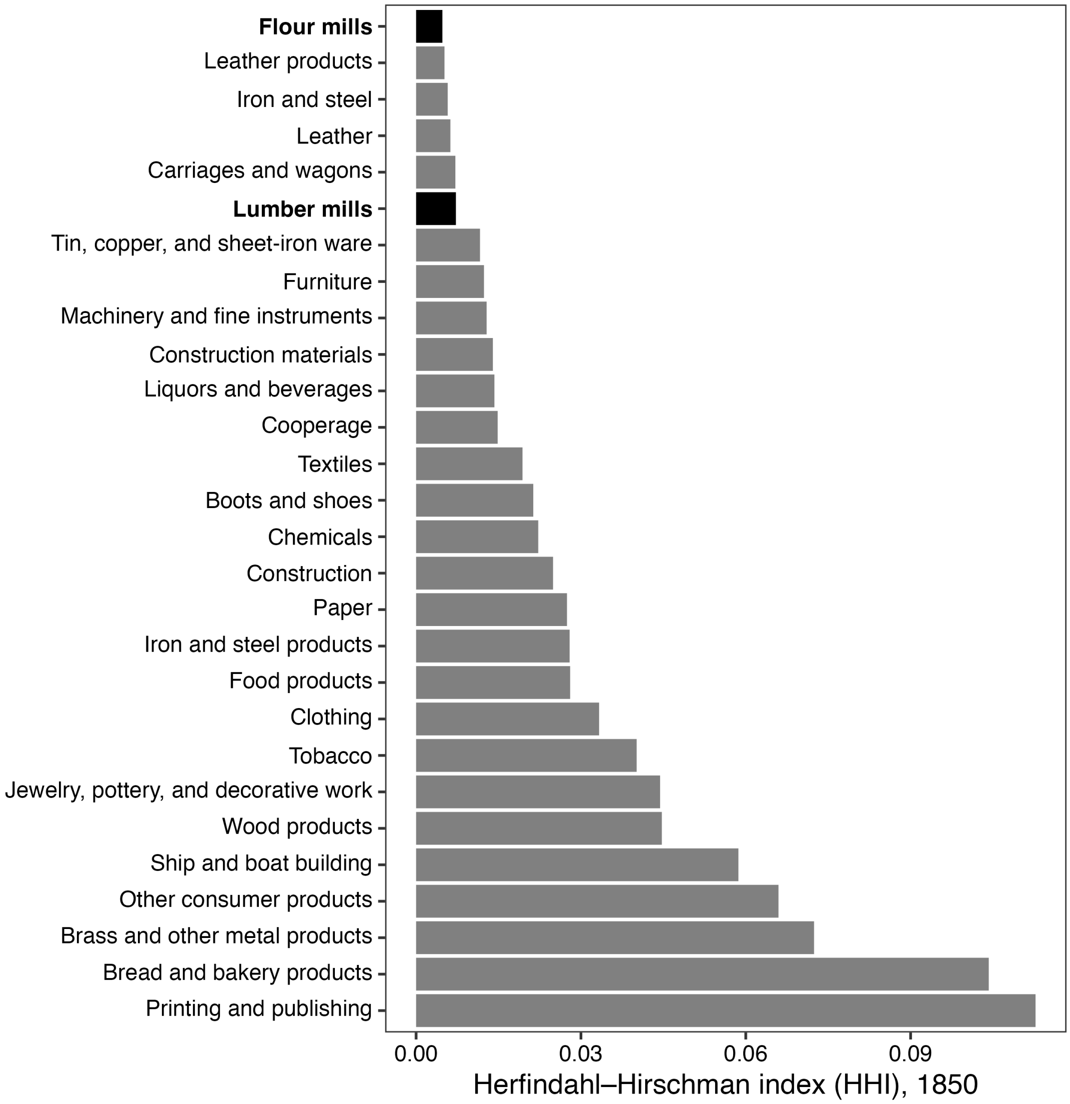


Panel B. Restricted to the Same Counties: Distribution of County Revenue for Flour Mills and Lumber Mills, by Data Source



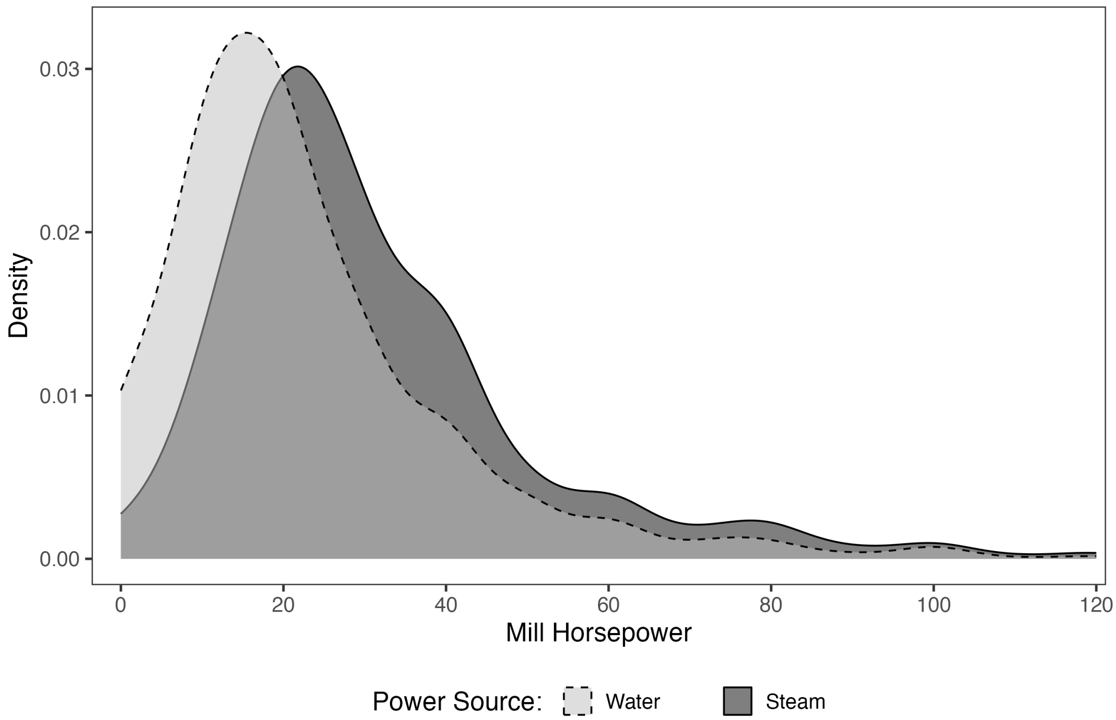
Notes: This figure shows the distribution of total flour mill revenue and total lumber mill revenue, by county, comparing county-industry tabulations for 1860-1880 made contemporaneously by the Census against the county-industry-level sums of our digitized establishment-level data from Census manuscripts (the Census did not publish county by industry tabulations in 1850). Panel A reports the distribution of values for county-industries with data in either source. Panel B reports the distribution of values for only those county-industries for which we have data from both sources. The Census had a *de jure* minimum value of total revenue for reporting county-industry values in 1870 and 1880, which corresponds to the vertical lines, and the Census also omitted tabulations for some other county-industry cells. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1860-1880) and county-industry-level tabulations digitized by Hornbeck and Rotemberg (2024).

Figure 2.16. Geographic Concentration of Production in 1850, by Industry



Notes: For each sector, this figure shows the Herfindahl–Hirschman index of revenue across counties in 1850 (sorted in increasing order). Data restricted to counties in our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850).

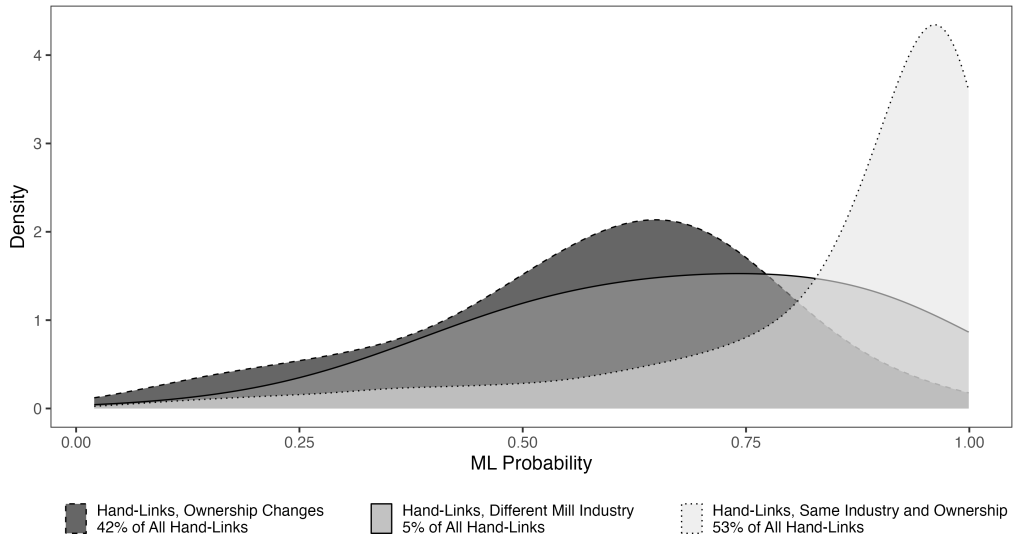
Figure 2.17. Distribution of Total Horsepower Installed, by Power Source



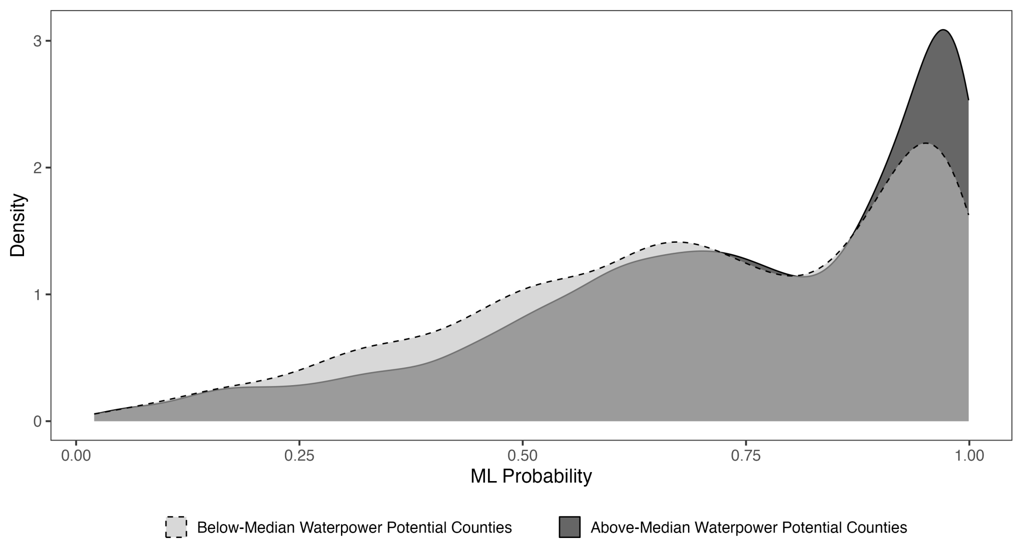
Notes: This figure shows the distribution of horsepower installed for flour mills and lumber mills in 1870 and 1880, pooled across both industries and decades. For this figure, we truncated the data at 120 horsepower. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1870 and 1880).

Figure 2.18. Distribution of Hand-Links' ML-Model Probability, by Type and Waterpower Potential

Panel A. Distribution of Hand-Links' ML-Model Probability, by Hand-Link Type

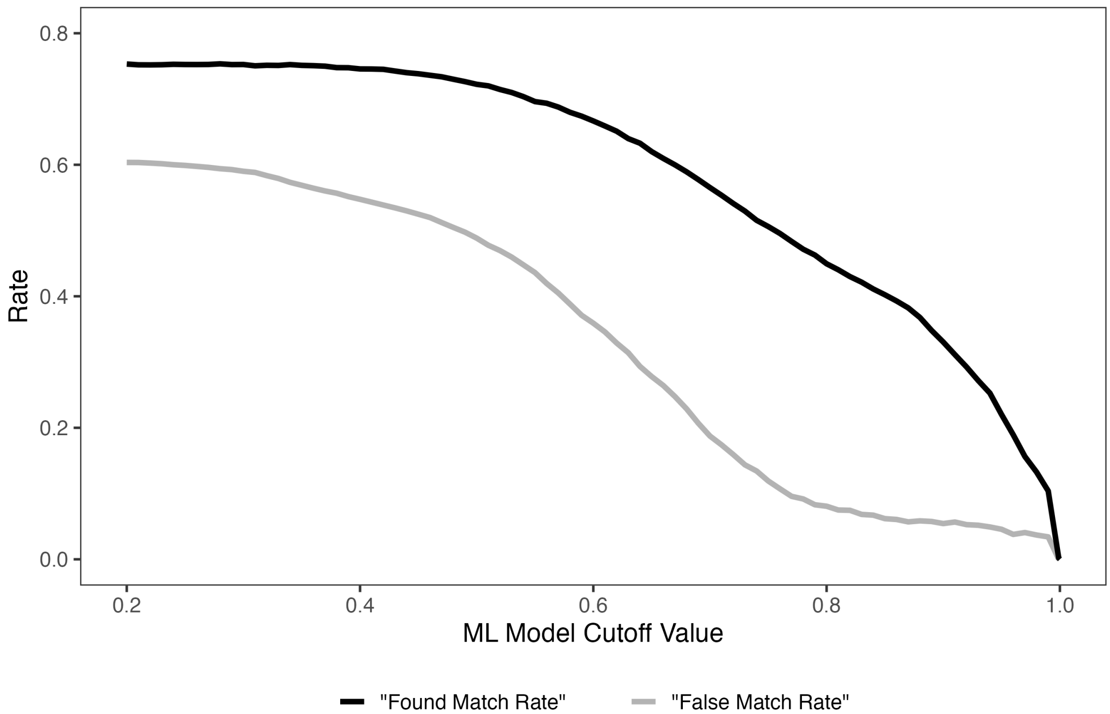


Panel B. Distribution of Hand-Links' ML-Model Probability, by County Waterpower Potential



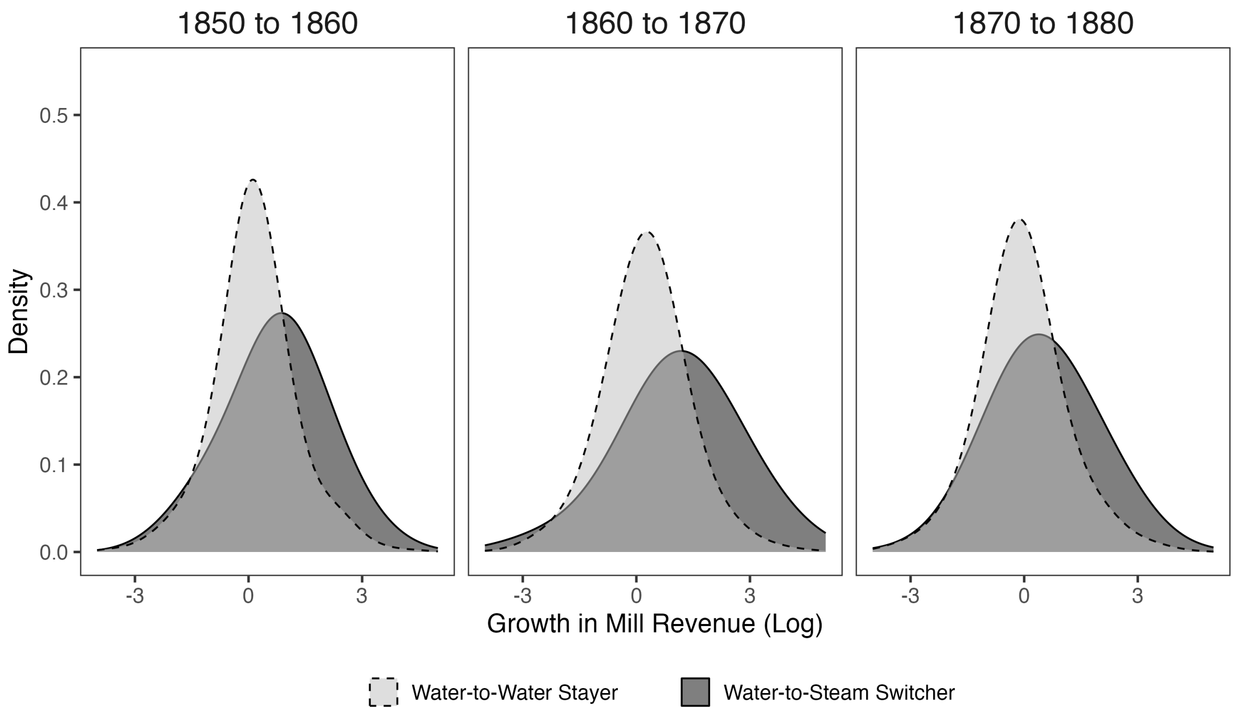
Notes: Panel A shows the distribution of hand-links by machine-learning probability, separately by the type of hand-link: those in the same industry and same ownership structure; those in a different mill industry (i.e., switched from flour to lumber milling); and those with ownership changes (i.e., added/removed some owners or changes to first names/initials). Panel B shows the distribution of machine-learning probabilities assigned to hand-links, separately for counties with above-median waterpower potential and below-median waterpower potential. The ML-Linking model is described in Appendix 2.C.4. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Figure 2.19. "False Match Rate" and "Found Match Rate" of Machine-Learning Model, Compared to Hand-Links, by ML-Model Cutoff Value



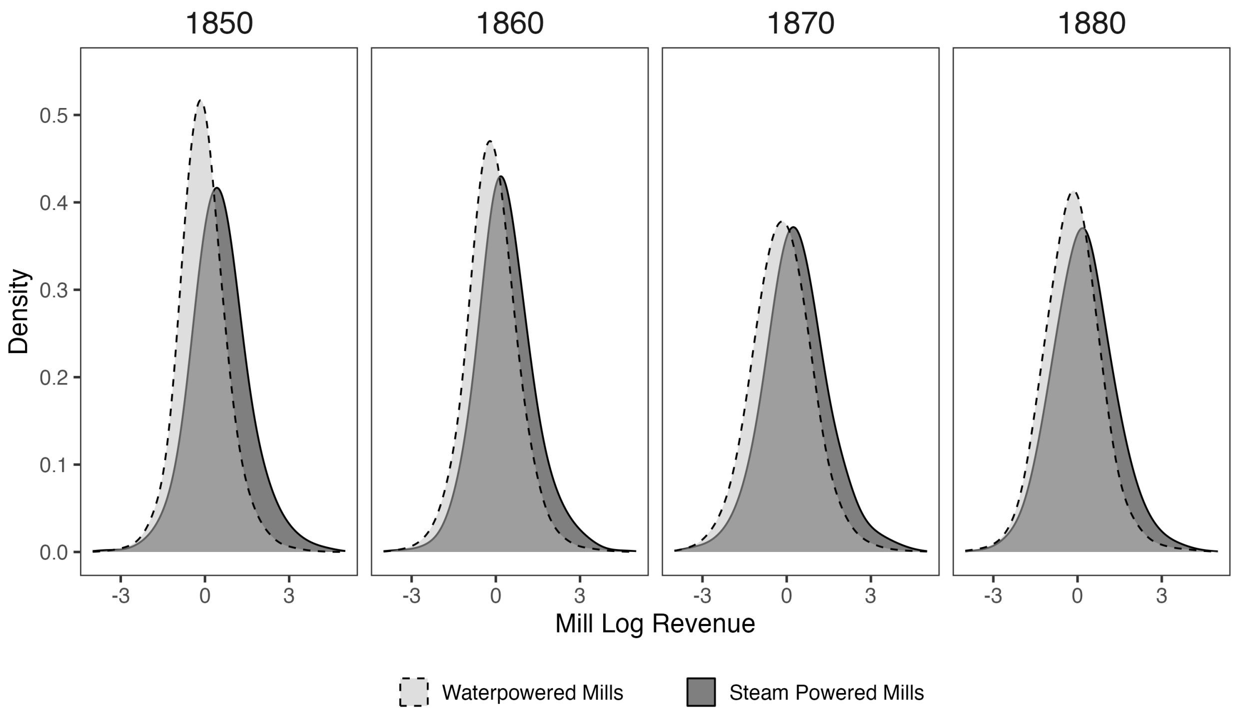
Notes: For different cutoff values on the machine-learning model predictions, the light gray line shows the share of links made by the machine-learning model that are not hand-links ("False Match Rate," if hand-links are assumed correct). The black line shows the share of hand-links made by the machine-learning model ("Found Match Rate"). The ML model reports a probability that mills in adjacent decades are the same, and the chosen ML-model cutoff value is the lowest probability that we would classify as a match. If there are multiple mills above the cutoff, we match only the highest probability mill. The ML-Linking model is described in Appendix 2.C.4. Data are for all lumber and flour mills in our digitized establishment-level Census of Manufactures (1850-1880).

Figure 2.20. Growth in Mill Revenue, by Steam Switching Choice



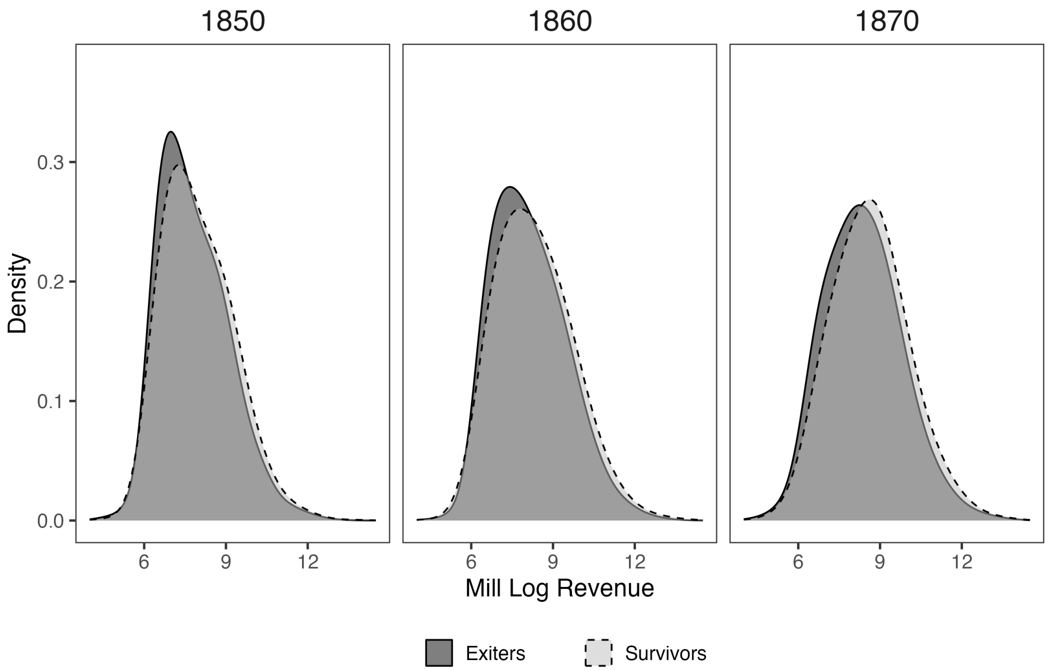
Notes: This figure shows the growth in mill revenue, by decade, for water incumbents who (1) kept using water power or (2) switched from water to steam power. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880).

Figure 2.21. Mill Size by Power Source, Within-County



Notes: This figure shows the distribution of mill revenue, in each decade, for each type of power source (steam or water). For each mill, we subtract mean log revenue in their county-industry (flour or lumber). Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880).

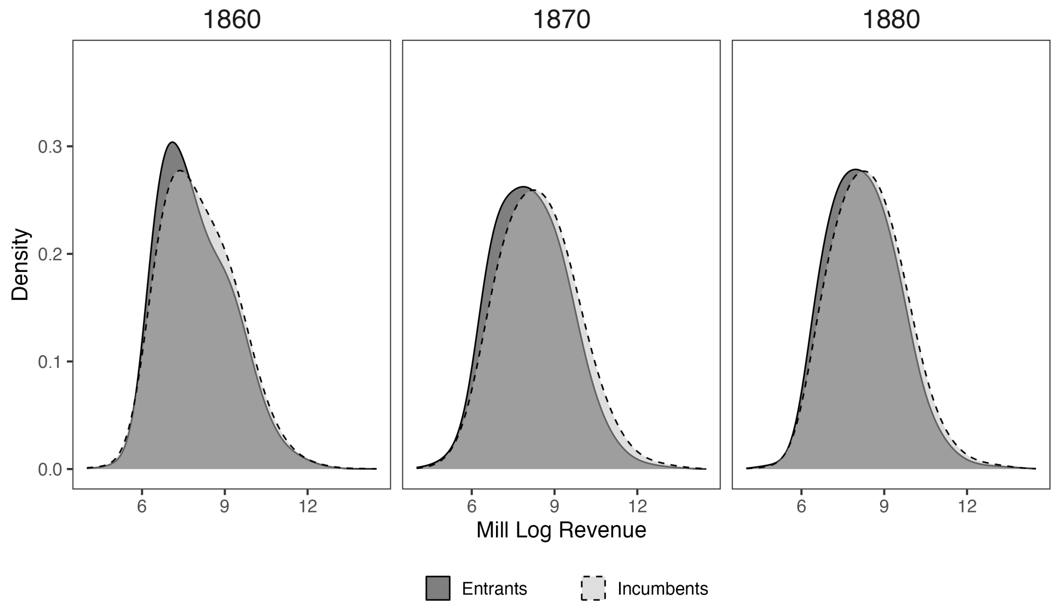
Figure 2.22. Initial Mill Size, for Exiters and Survivors



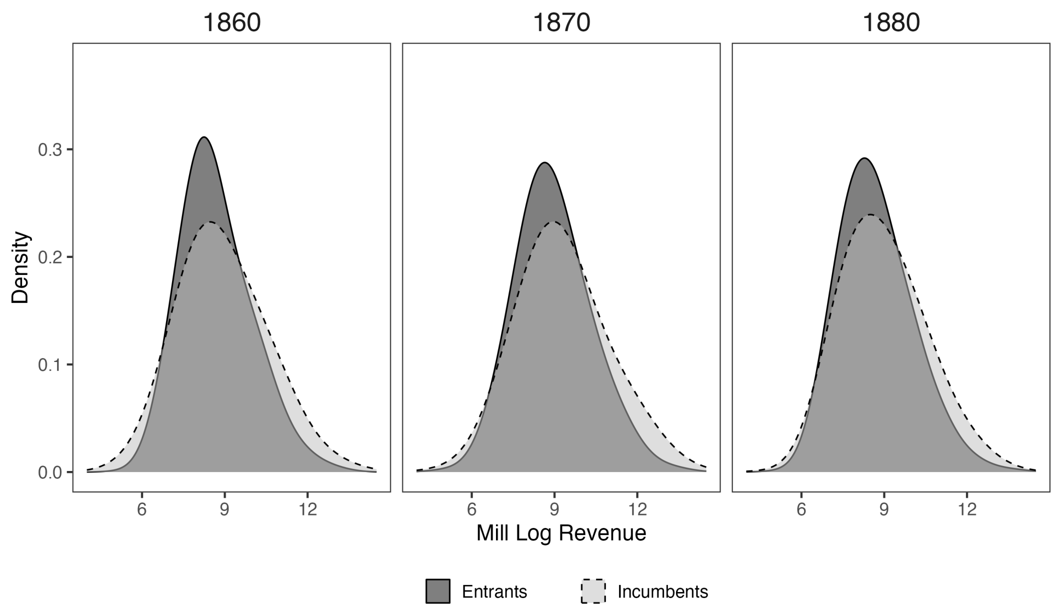
Notes: This figure shows the distribution of mill revenue in each baseline decade, separately for “Exiters” who close in the subsequent decade and “Survivors” who remain in operation by the next Census. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880).

Figure 2.23. Mill Size for Entrants and Incumbents, within Power Source

Panel A. Log Revenue of Entrants and Incumbents Using Water Power



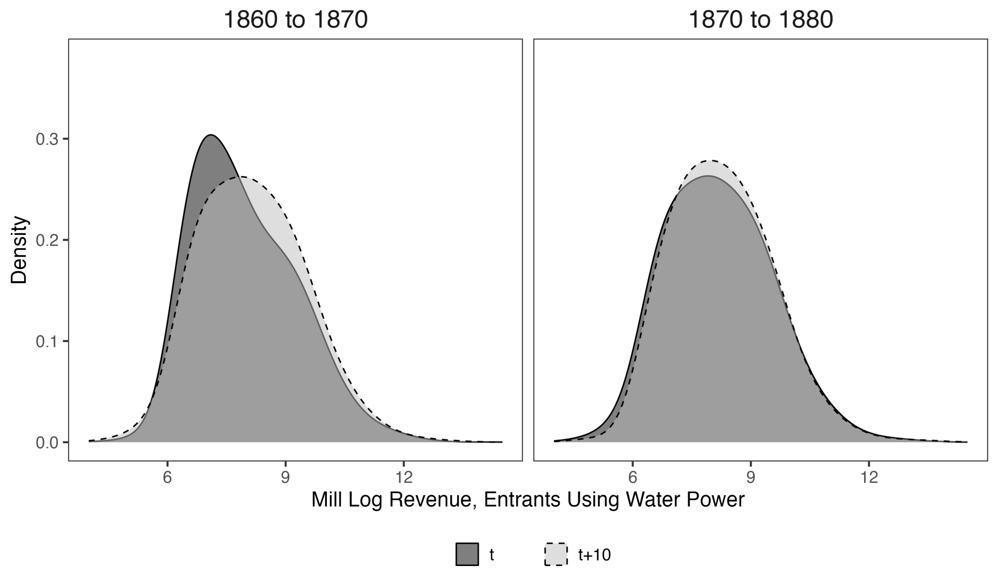
Panel B. Log Revenue for Entrants and Incumbents Using Steam Power



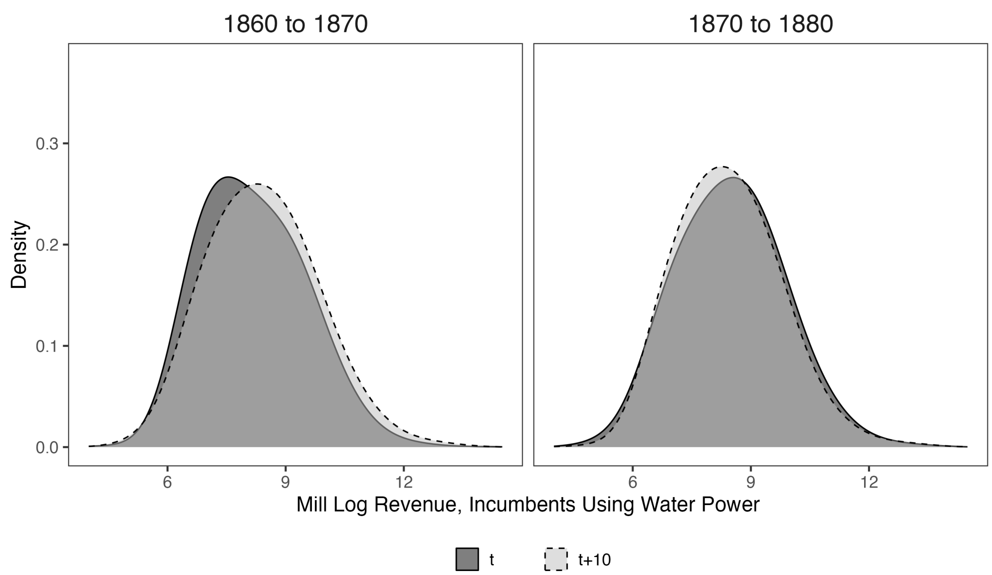
Notes: This figure shows the distribution of mill revenue, in each decade, comparing entrant mills and incumbent mills using the same power source (water power in Panel A, steam power in Panel B). Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880).

Figure 2.24. Mill Growth for Incumbents and Successive Generations of Entrants

Panel A. Log Revenue of Entrants Using Water Power

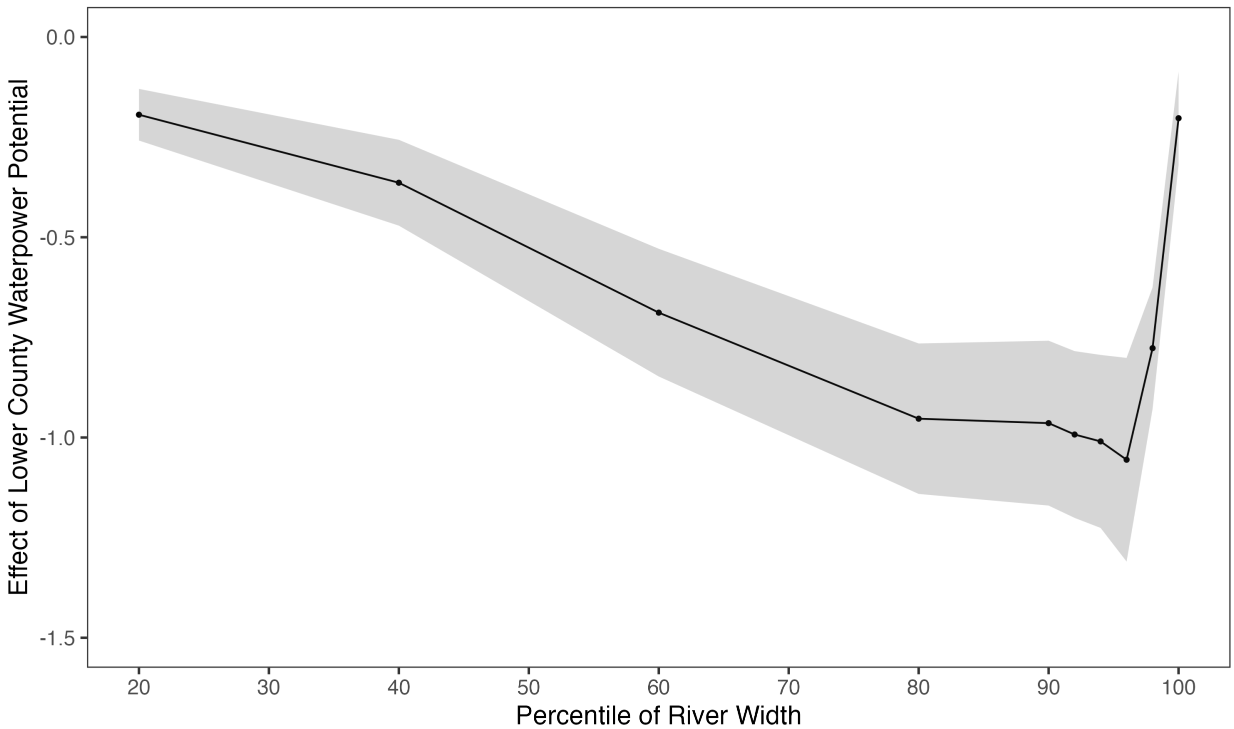


Panel B. Log Revenue of Incumbents Using Water Power



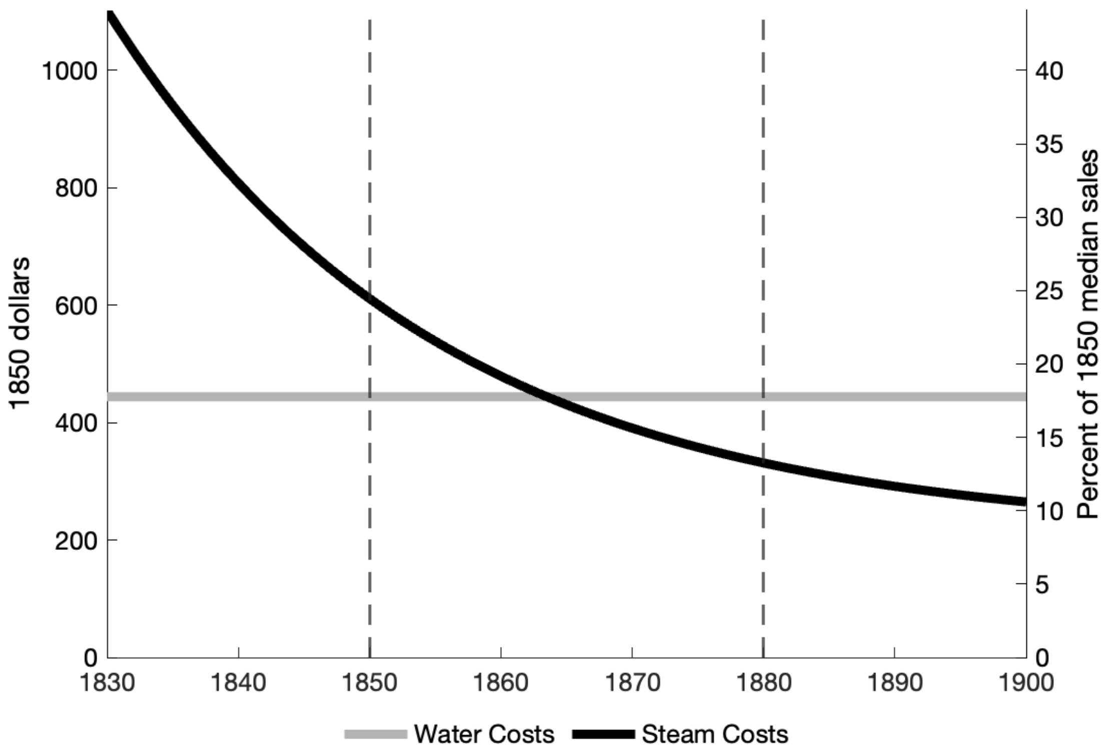
Notes: This figure plots the distribution of mill revenues for water mills, by decade. The top panel shows the size distributions of water entrants in t and $t+10$. The bottom panel shows the size distributions of the water incumbents (who do not subsequently switch to steam power) in t and $t+10$. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Figure 2.25. Estimated Relationship between Water Powered Mills in 1850 and County Waterpower Potential, Excluding Rivers with Widths Above Different Cutoffs



Notes: This figure shows the estimated relationship between a county's number of water powered mills in 1850 and a one standard deviation decrease in county waterpower potential, where county waterpower potential is measured excluding rivers that are wider than the indicated cutoff percentile of river widths. We sort rivers into percentile bins, based on their width, estimate our main specification from Panel A of Table 2.2, and plot the estimated coefficient on Lower Water power along with its 95% confidence interval. All regressions include our baseline controls interacted with industry: an indicator for the presence of navigable waterways in the county, distance to the nearest navigable waterway, county market access in 1850, an indicator for workable coal deposits in the county, the share of the county covered by coal deposits, and access to coal via the transportation network. Robust standard errors are clustered by county. Data from our main sample (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850) and NHDPlusV2.

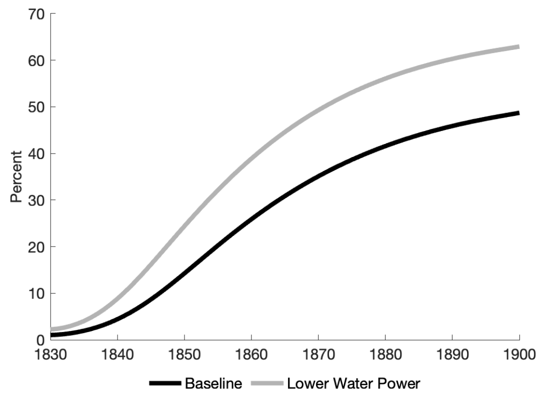
Figure 2.26. Water and Steam Adoption Costs: Structural Estimates



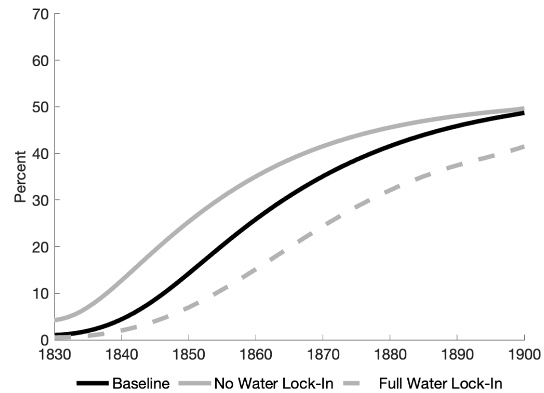
Notes: This figure plots our structural estimates of the adoption costs of water $c_B(W)$ and steam power $c_t(S)$, estimated in Table 2.8. The right axis is in percent of 1850 median firm sales, which the left axis converts to 1850 dollars using median firm sales in our 1850 data.

Figure 2.27. Water Technology and the Impacts of Steam Power (with Partial Reversibility of Water Power)

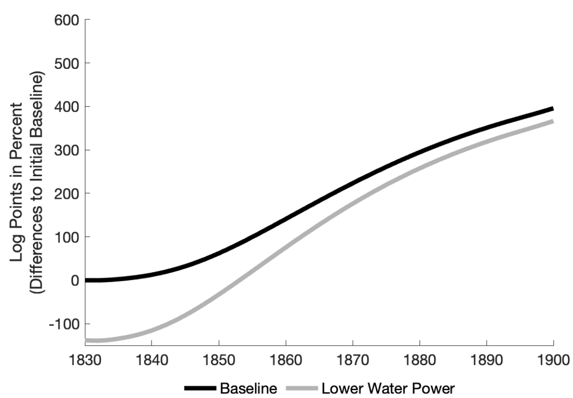
A. Water Costs and Steam Adoption



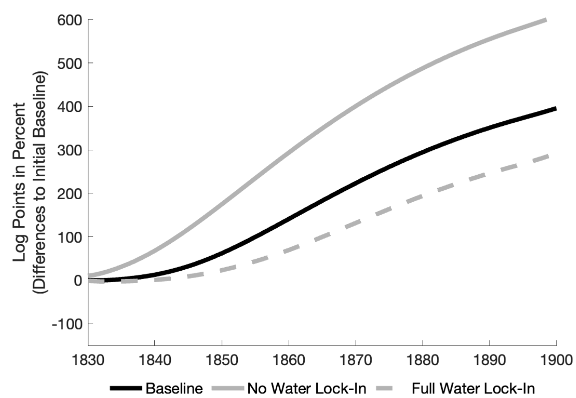
C. Switching Barriers and Steam Adoption



B. Water Costs and Mill Revenue



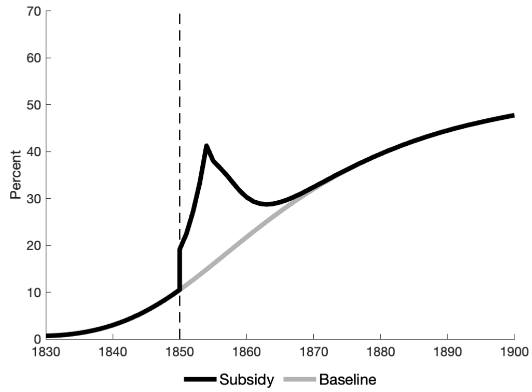
D. Switching Barriers and Mill Revenue



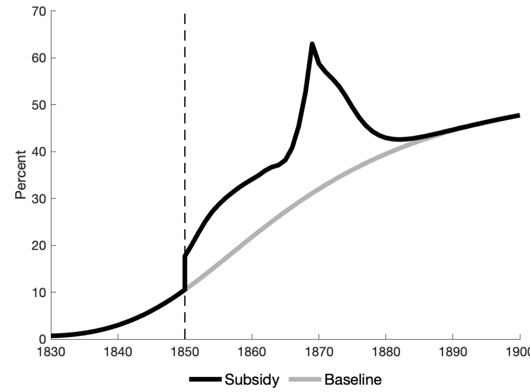
Notes: This figure shows the share of steam users and total mill revenue in model counties with different water technologies. The figure is based on a re-estimation of the structural model that assumes partial reversibility of water power. We set $\omega^W = 0.35$ for water mills that switch to steam power, reflecting the average liquidation rate estimated by Kermani and Ma (2023). Mill revenue is measured in log differences to the initial steady state of the baseline region. Panels A and B plot the impacts of steam power in the average county (black line) and a region with a standard deviation lower waterpower potential (gray line), where the only parameter difference between the regions is the fixed cost of water power adoption. Panels C and D plot the impacts of steam power as functions of switching barriers. The black line shows adoption for our baseline estimates, the gray line removes switching barriers ($\omega^W = 1, c(W, S) = 0$), and the dashed line represents prohibitive switching barriers ($c(W, S) \to \infty$).

Figure 2.28. Water-to-Steam Switching Subsidies: Steam Adoption and Annual Costs

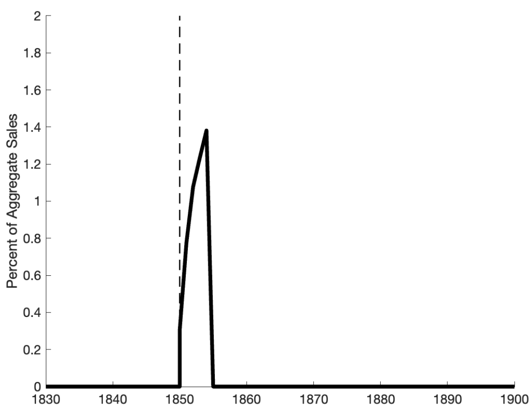
A. 5-Year Expiration: Steam Adoption



C. 20-Year Expiration: Steam Adoption



B. 5-Year Expiration: Annual Cost



D. 20-Year Expiration: Annual Cost

Notes: This figure simulates counterfactual "cash-for-clunkers" policies that pay water incumbents $c_B(W)$ to switch to steam power, exactly offsetting the sunk cost of switching. Panel A shows the adoption of steam power with a 5-year policy in 1850, and Panel B shows its annual costs. Panel C shows the adoption of steam power with a 20-year policy introduced in 1850, and Panel D shows its annual costs. Panels A and C compare the counterfactual adoption of steam power (in black) to its factual adoption (in gray).

2.L Appendix Tables

Table 2.13. Coverage Rates

State	1850	1860	1870	1880	State	1850	1860	1870	1880
AL	✓	✓	✓	✓	MT	-	-	✓	✓
AR	✓	✓	✓	✓	NE	-	✓	✓	✓
CA	✓	✓	✓	✓	NV	-	-	✓	✓
CO	-	-	✓	✓	NH	✓	✓	✓	✓
CT	✓	✓	✓	✓	NJ	✓	✓	✓	✓
DE	✓	✓	✓	✓	NY	✓	✓	82%	99%
DC	✓	✓	✓	✓	NC	✓	84%	✓	✓
FL	✓	✓	✓	✓	ND & SD	-	-	0%	18%
GA	0%	0%	0%	✓	OH	✓	26%	74%	68%
IL	✓	✓	46%	✓	OR	✓	✓	✓	✓
IN	✓	✓	✓	✓	PA	✓	✓	✓	✓
IA	✓	✓	✓	✓	RI	✓	✓	✓	✓
KS	-	✓	✓	✓	SC	✓	✓	✓	✓
KY	✓	✓	✓	✓	TN	✓	30%	35%	✓
LA	0%	0%	0%	✓	TX	✓	✓	85%	✓
ME	✓	✓	✓	✓	UT	-	✓	✓	✓
MD	✓	✓	0%	✓	VT	✓	✓	✓	✓
MA	✓	✓	32%	✓	VA	✓	✓	✓	✓
MI	✓	✓	49%	✓	WA	-	✓	✓	✓
MN	✓	✓	✓	✓	WV	-	-	✓	✓
MS	✓	✓	✓	✓	WI	✓	✓	✓	✓
MO	✓	✓	✓	✓					

Notes: This table shows our coverage of counties. Percents indicate estimates of the share of establishments that we digitized, given the published county-level tabulations. In 1850, the Census records for three counties in California (Contra Costa, San Francisco, and Santa Clara) were lost and never tabulated, we have complete coverage of the remaining counties in California. Dashes indicate that no survey was conducted, checkmarks indicate that we have complete coverage.

Table 2.14. Survival Rates, by County Waterpower Potential and Initial Power Source

	Water Survival Rate (1)	Steam Survival Rate (2)	Difference (1) − (2) (3)
Elasticity with Respect to Lower Waterpower:			
In 1860	-0.173	-0.490	0.317
	(0.068)	(0.210)	(0.217)
# County-Industries	1,199	1,199	
In 1870	-0.237	-0.188	-0.049
	(0.064)	(0.116)	(0.126)
# County-Industries	1,199	1,199	
In 1880	-0.180	-0.002	-0.179
	(0.048)	(0.070)	(0.079)
# County-Industries	1,199	1,199	

Notes: This table shows the elasticity of survival in both water and steam mills, over the previous decade, with respect to county waterpower potential from 1860-1880. "Lower Waterpower" is a negative standardized measure of county waterpower potential, with standard deviation of one, so the estimates reflect differences in counties with one standard deviation lower waterpower potential.

Column 1 reports results for water powered incumbents, column 2 reports results for steam powered ones, and column 3 reports the differences. Each row corresponds to a different PPML regression, using data from the indicated Census year and previous Census year, which approximates percent differences in the rates.

All regressions include county-industry fixed effects, industry-year fixed effects, and our baseline controls interacted with industry and year: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1860-1880) and NHDPlusV2.

Table 2.15. Per Capita Manufacturing Growth and Steam Adoption, by Waterpower Potential

	Population (1)	Mills Per Capita (2)	Mill Revenue Per Capita (3)
Panel A. Differences in Lower Waterpower Counties:			
In 1850	-0.284	-0.672	-0.592
	(0.226)	(0.233)	(0.232)
Panel B. Growth in Lower Waterpower Counties:			
From 1850 to 1860	0.094	0.126	0.088
	(0.029)	(0.065)	(0.082)
From 1860 to 1870	0.067	0.046	0.136
	(0.040)	(0.060)	(0.066)
From 1870 to 1880	0.075	0.017	0.065
	(0.024)	(0.044)	(0.101)
# County-Industries		1,199	1,199

Notes: This table shows the relationship between per capita growth in mill activity and county waterpower potential. "Lower Waterpower" is a negative standardized measure of county waterpower potential, with standard deviation of one, so the estimates reflect differences in counties with one standard deviation lower waterpower potential.

The outcome in column 1 is (log) population, the outcome in column 2 is mills per capita, and the outcome in column 3 is milling revenue per capita. Panel A reports cross-sectional differences in 1850. Panel B reports growth rates over the following decades. Each row corresponds to a different regression, using only data from the indicated years. Column 1 reports OLS estimates, and columns 2-3 report PPML estimates, which approximate percent differences.

All regressions industry fixed effects and our baseline controls interacted with industry: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network. Panel B regressions also include county-industry fixed effects, industry-year fixed effects, and our baseline controls interacted with industry and year.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.16. Steam Adoption and *Flour* Mill Growth, by County Waterpower Potential

	Steam Share of Mills (1)	Total Mills (2)	Total Mill Revenue (3)
Growth in Lower Waterpower Counties:			
From 1850 to 1860	0.018	0.114	0.154
	(0.020)	(0.069)	(0.110)
# Counties	535	587	587
From 1860 to 1870	0.038	0.163	0.194
	(0.019)	(0.072)	(0.088)
# Counties	531	587	587
From 1870 to 1880	0.013	0.053	0.160
	(0.015)	(0.041)	(0.120)
# Counties	574	587	587

Notes: This table shows the relationship between growth in mill activity and county waterpower potential, limiting the sample to flour mills. "Lower Waterpower" is a negative standardized measure of county waterpower potential, with standard deviation of one, so the estimates reflect differences in counties with one standard deviation lower waterpower potential.

The outcomes are the share of flour mills using steam power (column 1), the total number of mills (column 2), and total mill revenue (column 3). Each row corresponds to growth over the indicated decade, using only data from the indicated years.

Column 1 reports OLS estimates, restricting the sample to counties with at least one flour mill in both decades (for the steam share to be defined) and weighting by the number of flour mills in that county in 1850. These estimates reflect percentage point differences in the shares. Columns 2 and 3 report PPML estimates for a balanced panel of counties (including zeros), which approximate percent differences.

All regressions include county fixed effects, year fixed effects, and our baseline controls interacted with year: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.17. *Flour* Mill Entry Rates and Survival Rates, by County Waterpower Potential

	Entry Rate (1)	Survival Rate (2)	Difference (1) − (2) (3)
Elasticity with Respect to Lower Waterpower:			
In 1860	0.183	-0.153	0.336
	(0.084)	(0.093)	(0.120)
# Counties	587	587	
In 1870	0.203	-0.117	0.320
	(0.082)	(0.071)	(0.101)
# Counties	587	587	
In 1880	0.129	-0.223	0.352
	(0.050)	(0.057)	(0.079)
# Counties	587	587	

Notes: This table shows the elasticity of mill entry and mill survival, over the previous decade, with respect to county waterpower potential, limiting the sample to flour mills. "Lower Waterpower" is a negative standardized measure of county waterpower potential, with standard deviation of one, so the estimates reflect differences in counties with one standard deviation lower waterpower potential.

Column 1 reports results for entry, column 2 reports results for incumbent survival, and column 3 reports the difference in these estimates. Each row corresponds to a different PPML regression, using data from the indicated Census year and previous Census year, which approximates percent differences in the rates.

All regressions include county fixed effects, year fixed effects, and our baseline controls interacted with year: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.18. Steam Adoption of Entrants and Water *Flour* Mills, by County Waterpower Potential

	From Entrants (1)	Water Incumbents (2)	Difference (1) − (2) (3)
Adoption in Lower Waterpower Counties:			
In 1860	0.091	0.033	0.059
	(0.029)	(0.033)	(0.037)
# Counties	530	333	
In 1870	0.103	0.063	0.040
	(0.022)	(0.027)	(0.035)
# Counties	575	326	
In 1880	0.126	0.047	0.079
	(0.026)	(0.023)	(0.027)
# Counties	577	416	

Notes: This table shows the relationship between county waterpower potential and the steam use of entrant mills and water incumbent mills, limiting the sample to flour mills. "Lower Waterpower" is a negative standardized measure of county waterpower potential, with standard deviation of one, so the estimates reflect differences in counties with one standard deviation lower waterpower potential.

The outcome in column 1 is the share of entrants using steam power, restricted to county-industries with at least one entrant in that year. Column 2 reports the share of "water incumbents" (mills that used water power in the previous Census year) who switched to steam power. For column 2, the sample is restricted to county-industries with at least one surviving water incumbent. Column 3 reports the difference between the estimates in columns 1 and 2. Each row corresponds to a different OLS regression, which report percentage point differences in the shares.

All regressions include our baseline controls: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

For each row, each observation is a county, weighted by the number of flour mills in 1850. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.19. Robustness to Alternative Drivers of Steam Use

	Water Mills	Steam Share	Growth in Total Mills			Steam Diffusion of Mills		
	1850	1850	1850 to 1860	1860 to 1870	1870 to 1880	1850 to 1860	1860 to 1870	1870 to 1880
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. Baseline	-1.055	0.089	0.220	0.113	0.092	0.067	0.034	-0.009
	(0.130)	(0.015)	(0.062)	(0.052)	(0.036)	(0.016)	(0.013)	(0.013)
2. Each type of coal separately	-1.056	0.089	0.221	0.113	0.091	0.067	0.034	-0.009
	(0.130)	(0.015)	(0.062)	(0.052)	(0.036)	(0.016)	(0.013)	(0.013)
3. Square and cubic in county coal shares	-1.046	0.086	0.202	0.131	0.097	0.069	0.030	-0.007
	(0.129)	(0.015)	(0.061)	(0.053)	(0.035)	(0.016)	(0.012)	(0.013)
4. FAO suitability for wheat	-1.065	0.088	0.203	0.127	0.100	0.065	0.033	-0.011
	(0.130)	(0.015)	(0.061)	(0.052)	(0.037)	(0.017)	(0.013)	(0.013)
5. Woodland share in county	-1.004	0.093	0.182	0.097	0.086	0.059	0.030	-0.017
	(0.131)	(0.015)	(0.065)	(0.054)	(0.037)	(0.017)	(0.013)	(0.014)
6. 1850 local MFG wages	-1.045	0.094	0.230	0.116	0.080	0.059	0.035	-0.010
	(0.133)	(0.015)	(0.063)	(0.055)	(0.037)	(0.017)	(0.013)	(0.014)
7. 1850 engineers and mechanics	-1.063	0.091	0.216	0.115	0.093	0.069	0.031	-0.007
	(0.129)	(0.015)	(0.062)	(0.052)	(0.036)	(0.017)	(0.013)	(0.013)
8. 1850 access to banks	-1.044	0.086	0.219	0.115	0.090	0.067	0.034	-0.008
	(0.129)	(0.015)	(0.062)	(0.052)	(0.036)	(0.016)	(0.013)	(0.013)
9. All above	-0.983	0.082	0.138	0.140	0.090	0.050	0.023	-0.011
	(0.123)	(0.015)	(0.066)	(0.058)	(0.037)	(0.018)	(0.013)	(0.014)

Notes: This table shows the robustness of the relationship between waterpower potential and the number of 1850 water establishments, the 1850 steam share, and mill growth from 1850-1880. This table focuses on additional controls for alternative factors which may have driven steam adoption.

All regressions include industry-year fixed effects, and our baseline controls interacted with industry and year. All regressions other than those reported in Columns (1) and (2) additionally include county-industry fixed effects. Our baseline controls are an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.20. Robustness to Alternative Drivers of County Growth

	Water Mills	Steam Share	Growth in Total Mills			Steam Diffusion of Mills		
	1850	1850	1850 to 1860	1860 to 1870	1870 to 1880	1850 to 1860	1860 to 1870	1870 to 1880
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. Baseline	-1.055	0.089	0.220	0.113	0.092	0.067	0.034	-0.009
	(0.130)	(0.015)	(0.062)	(0.052)	(0.036)	(0.016)	(0.013)	(0.013)
2. No controls for MA/navigable rivers	-1.253	0.078	0.288	0.120	0.104	0.083	0.036	-0.011
	(0.136)	(0.015)	(0.063)	(0.051)	(0.036)	(0.017)	(0.012)	(0.013)
3. No controls for coal	-1.060	0.096	0.238	0.110	0.098	0.080	0.038	-0.005
	(0.125)	(0.016)	(0.064)	(0.052)	(0.036)	(0.019)	(0.012)	(0.014)
4. No extra controls	-1.270	0.080	0.306	0.125	0.114	0.087	0.036	-0.010
	(0.131)	(0.015)	(0.063)	(0.050)	(0.035)	(0.018)	(0.012)	(0.015)
5. Time-varying market access	-1.049	0.088	0.211	0.114	0.103	0.066	0.035	-0.008
	(0.126)	(0.015)	(0.059)	(0.052)	(0.034)	(0.016)	(0.013)	(0.013)
6. Time-varying population	-0.764	0.090	0.149	0.101	0.072	0.053	0.024	-0.013
	(0.108)	(0.015)	(0.064)	(0.057)	(0.037)	(0.017)	(0.013)	(0.014)
7. 1850 population	-0.815	0.094	0.179	0.112	0.078	0.056	0.032	-0.011
	(0.115)	(0.016)	(0.061)	(0.054)	(0.036)	(0.016)	(0.013)	(0.014)
8. Appalachia	-1.039	0.088	0.220	0.114	0.092	0.066	0.034	-0.009
	(0.130)	(0.015)	(0.062)	(0.052)	(0.036)	(0.016)	(0.013)	(0.013)
9. Frontier	-1.050	0.089	0.215	0.110	0.095	0.066	0.035	-0.009
	(0.130)	(0.015)	(0.062)	(0.052)	(0.036)	(0.017)	(0.013)	(0.013)
10. 1850 agricultural share	-1.041	0.093	0.207	0.103	0.084	0.065	0.032	-0.012
	(0.129)	(0.015)	(0.062)	(0.052)	(0.035)	(0.016)	(0.012)	(0.013)
11. Portage sites	-1.063	0.091	0.219	0.114	0.091	0.069	0.032	-0.008
	(0.129)	(0.015)	(0.062)	(0.052)	(0.036)	(0.017)	(0.012)	(0.014)
12. Civil war controls	-0.920	0.087	0.225	0.127	0.051	0.061	0.033	-0.009
	(0.122)	(0.015)	(0.063)	(0.055)	(0.037)	(0.017)	(0.013)	(0.013)
13. Time-invariant controls from rows 8-12	-0.914	0.091	0.217	0.110	0.052	0.062	0.033	-0.008
	(0.121)	(0.015)	(0.063)	(0.055)	(0.036)	(0.017)	(0.012)	(0.013)
14. All time-invariant controls (rows 7-12)	-0.667	0.092	0.184	0.123	0.045	0.055	0.032	-0.006
	(0.100)	(0.016)	(0.063)	(0.056)	(0.036)	(0.017)	(0.012)	(0.013)

Notes: This table shows the robustness of the relationship between waterpower potential and the number of 1850 water establishments, the 1850 steam share, and growth 1850-1880. This table focuses on additional controls for alternative factors which may have driven county growth.

Unless otherwise specified (in rows 2-5), all regressions include industry-year fixed effects, and our baseline controls interacted with industry and year. All regressions other than those reported in Columns (1) and (2) additionally include county-industry fixed effects. Our baseline controls are an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.21. Steam use, by Distance to Railroad Station

	From Entrants (1)	From Water Incumbents (2)	Difference (1) − (2) (3)
Lower Waterpower	0.177	0.045	0.132
	(0.021)	(0.014)	(0.017)
Log Distance, WPP-to-RR Station	0.017	-0.033	0.050
	(0.042)	(0.028)	(0.041)
Log Distance, to RR Station	-0.017	0.035	-0.052
	(0.047)	(0.032)	(0.045)
# County-Industries	1,190	841	

Notes: This table shows the relationship between waterpower potential, railroad station placement, and the steam use of entrant and incumbent mills from 1860-1880. “Lower Waterpower” is a negative standardized measure of county waterpower potential, with standard deviation of one, so the estimates reflect differences in counties with one standard deviation lower waterpower potential. “Log Distance, WPP-to-RR Station” is the log of the average distance from water segments to the closest railroad stations, weighting by potential horsepower. “Log Distance, to RR Station” is the log of the average distance from railroad stations from all points in the county.

The outcome in column 1 is the share of entrants using steam power, the outcome in column 2 is the share of water incumbents (incumbents who used water power in the previous decade) who switched to steam power, and column 3 reports the difference. Each row corresponds to different OLS regressions, using data pooled across all 1860-1880. The sample is restricted to all county-industry-years at least one current entrant (in column 1) or incumbent (in column 2).

All regressions include our baseline controls interacted with year and industry: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network. Regressions are weighted by the number of mills in the county in 1850.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.22. Confusion Matrix: Hand Links vs. Predicted Links

	Machine Learning Links			
	Linked (Same)	Linked (Different)	Not Linked	Total
Hand Links	(1)	(2)	(3)	(4)
Panel A. 1850 to 1860				
Linked	2,590	69	942	3,601
Not Linked	-	217	14,114	14,331
Panel B. 1860 to 1870				
Linked	2,313	256	816	3,385
Not Linked	-	2,237	11,885	14,122
Panel C. 1870 to 1880				
Linked	3,486	187	1,849	5,522
Not Linked	-	1,096	16,697	17,793

Notes: This table shows the confusion matrix for the panel links. The rows report matches made by the hand-linking procedure, and the columns correspond to matches made by the machine-learning model, both of which are described in Appendix 2.C.4. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880).

Table 2.23. Robustness to Measurement and Linking Error: Entry and Survival

	Entry Rate			Survival Rate		
	1850 to 1860	1860 to 1870	1870 to 1880	1850 to 1860	1860 to 1870	1870 to 1880
	(1)	(2)	(3)	(4)	(5)	(6)
1. Baseline	0.323	0.168	0.158	-0.230	-0.266	-0.158
	(0.074)	(0.058)	(0.045)	(0.065)	(0.057)	(0.040)
2. Links that are both ML and hand-linked	0.274	0.142	0.134	-0.165	-0.229	-0.193
	(0.068)	(0.055)	(0.041)	(0.079)	(0.069)	(0.048)
3. Only ML links	0.287	0.143	0.135	-0.211	-0.212	-0.179
	(0.069)	(0.056)	(0.042)	(0.078)	(0.068)	(0.046)
4. Raising ML linking threshold to 0.8	0.256	0.139	0.126	-0.179	-0.349	-0.258
	(0.065)	(0.054)	(0.040)	(0.089)	(0.073)	(0.057)
5. Lowering ML linking threshold to 0.4	0.320	0.143	0.130	-0.242	-0.192	-0.110
	(0.073)	(0.056)	(0.042)	(0.068)	(0.066)	(0.043)
6. Only business-name mills	0.308	0.244	0.124	-0.241	-0.258	-0.172
	(0.070)	(0.058)	(0.049)	(0.083)	(0.078)	(0.058)
7. Only non-business name mills	0.270	0.043	0.173	-0.205	-0.290	-0.229
	(0.083)	(0.072)	(0.052)	(0.097)	(0.084)	(0.064)
8. Only mills with all positive inputs	0.341	0.198	0.142	-0.207	-0.253	-0.124
	(0.079)	(0.061)	(0.047)	(0.068)	(0.061)	(0.043)
9. Include inactive mills with zero output	0.313	0.174	0.148	-0.240	-0.282	-0.157
	(0.072)	(0.057)	(0.045)	(0.067)	(0.058)	(0.039)
10. Include mills using manual/other power	0.314	0.164	0.150	-0.212	-0.244	-0.152
	(0.073)	(0.057)	(0.045)	(0.066)	(0.057)	(0.040)

Notes: This table shows the robustness of the relationship between waterpower potential and the share of entrants and water incumbents using steam. This table focuses on linking and measurement error.

All regressions include county-industry fixed effects and our baseline controls interacted with industry: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.24. Robustness to Measurement and Linking Error: Steam Use

	Entrant Steam Share			Incumbent Steam Share		
	1860	1870	1880	1860	1870	1880
	(1)	(2)	(3)	(4)	(5)	(6)
1. Baseline	0.169	0.188	0.172	0.034	0.049	0.051
	(0.024)	(0.022)	(0.022)	(0.021)	(0.018)	(0.024)
2. Links that are both ML and hand-linked	0.167	0.188	0.168	0.022	0.042	0.062
	(0.023)	(0.022)	(0.021)	(0.021)	(0.019)	(0.027)
3. Only ML links	0.166	0.188	0.168	0.029	0.045	0.064
	(0.023)	(0.022)	(0.021)	(0.022)	(0.019)	(0.027)
4. Raising ML linking threshold to 0.8	0.162	0.185	0.169	0.001	0.019	0.049
	(0.023)	(0.022)	(0.021)	(0.024)	(0.020)	(0.027)
5. Lowering ML linking threshold to 0.4	0.168	0.188	0.171	0.027	0.045	0.056
	(0.024)	(0.022)	(0.021)	(0.021)	(0.018)	(0.024)
6. Only business-name mills	0.161	0.175	0.159	0.031	0.082	0.070
	(0.027)	(0.025)	(0.024)	(0.029)	(0.035)	(0.033)
7. Only non-business name mills	0.149	0.177	0.164	0.034	0.005	0.050
	(0.030)	(0.023)	(0.024)	(0.022)	(0.021)	(0.031)
8. Only mills with all positive inputs	0.164	0.188	0.164	0.035	0.045	0.065
	(0.023)	(0.023)	(0.022)	(0.023)	(0.018)	(0.026)
9. Include inactive mills with zero output	0.167	0.189	0.171	0.035	0.052	0.052
	(0.024)	(0.021)	(0.021)	(0.021)	(0.017)	(0.023)
10. Include mills using manual/other power	0.166	0.187	0.169	0.035	0.047	0.059
	(0.024)	(0.022)	(0.021)	(0.021)	(0.017)	(0.024)

Notes: This table shows the robustness of the relationship between waterpower potential and the share of entrants and water incumbents using steam. This table focuses on linking and measurement error.

All regressions include county-industry fixed effects and our baseline controls interacted with industry: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.25. Robustness to Sample

	Water Mills	Steam Share	Growth in Total Mills			Steam Diffusion of Mills		
	1850	1850	1850 to 1860	1860 to 1870	1870 to 1880	1850 to 1860	1860 to 1870	1870 to 1880
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. Baseline	-1.055	0.089	0.220	0.113	0.092	0.067	0.034	-0.009
	(0.130)	(0.015)	(0.062)	(0.052)	(0.036)	(0.016)	(0.013)	(0.013)
2. Include extensive margin of counties	-1.152	0.088	0.298	0.092	0.118	0.070	0.034	-0.009
	(0.132)	(0.015)	(0.062)	(0.050)	(0.035)	(0.017)	(0.013)	(0.013)
3. At least 3 mills in 1850	-0.957	0.081	0.170	0.112	0.070	0.069	0.038	-0.010
	(0.133)	(0.016)	(0.064)	(0.057)	(0.042)	(0.017)	(0.013)	(0.014)
4. At least 5 mills in 1850	-0.859	0.073	0.135	0.113	0.057	0.067	0.048	-0.013
	(0.136)	(0.017)	(0.069)	(0.063)	(0.047)	(0.018)	(0.014)	(0.015)
5. Exclude large grouped counties	-1.105	0.099	0.227	0.111	0.092	0.071	0.031	-0.009
	(0.129)	(0.015)	(0.062)	(0.053)	(0.037)	(0.017)	(0.012)	(0.014)
6. Exclude top and bottom 1% WPP counties	-1.161	0.088	0.238	0.121	0.113	0.072	0.036	-0.010
	(0.127)	(0.016)	(0.064)	(0.056)	(0.039)	(0.017)	(0.013)	(0.014)
7. Exclude top and bottom 5% WPP counties	-1.131	0.085	0.238	0.117	0.120	0.061	0.039	-0.014
	(0.147)	(0.019)	(0.075)	(0.063)	(0.042)	(0.020)	(0.015)	(0.016)
8. Exclude largest 20 cities in 1850-1880	-1.048	0.088	0.225	0.100	0.098	0.072	0.038	-0.012
	(0.130)	(0.015)	(0.064)	(0.055)	(0.039)	(0.016)	(0.013)	(0.014)
9. Exclude merchant mill cities	-1.011	0.091	0.211	0.101	0.095	0.064	0.034	-0.007
	(0.126)	(0.015)	(0.064)	(0.055)	(0.038)	(0.017)	(0.014)	(0.014)

Notes: This table shows the robustness of the relationship between waterpower potential and the number of 1850 water establishments, the 1850 steam share, and growth 1850-1880. This table focuses on alternative choices for the sample of counties in the analysis.

All regressions include industry-year fixed effects, and our baseline controls interacted with industry and year. All regressions other than those reported in Columns (1) and (2) additionally include county-industry fixed effects. Our baseline controls are an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Except for the stated modifications in each row, data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.26. Survival Rates

	Survival Rate By Initial Power Source		
	All (1)	Water (2)	Steam (3)
From 1850 to 1860	0.201	0.208	0.138
From 1860 to 1870	0.194	0.214	0.136
From 1870 to 1880	0.237	0.257	0.194

Notes: This table shows the measured survival rate of mills, by decade. Column 1 reports the share of all mills that survive in each decade, column 2 reports survival for water powered mills, and column 3 reports survival for steam powered mills. We denote a mill as surviving if we can find a record for it in the subsequent Census.

Each observation is a county-industry-year. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880).

Table 2.27. Incumbency, Size, and Steam Use

	Steam Adoption		
	(1)	(2)	(3)
Water Incumbent	-0.175		-0.177
	(0.009)		(0.009)
Mill Log Revenue		0.091	0.091
		(0.004)	(0.004)
# Mill-Years	63,755	63,755	63,755

Notes: This table shows how incumbency and size predict steam use. Column 1 shows the bivariate relationship of (water) incumbent status and steam use, Column 2 the bivariate relationship between revenue and steam use, and Column 3 includes both as independent variables.

All regressions include industry fixed effects and our baseline controls interacted with industry: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850, an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.28. Steam Use and Characteristics of Owners

	Mean Value	Uses Steam			
	(1)	(2)	(3)	(4)	(5)
Immigrant	0.069	0.076			0.075
	[0.253]	(0.015)			(0.015)
Age, in years	44.7		-0.0018		-0.0016
	[13.3]		(0.0002)		(0.0002)
Professional Miller	0.395			0.041	0.035
	[0.489]			(0.006)	(0.006)
# Mills	30,777	30,777	30,777	30,777	30,777
Mean of Dependent Variable		0.203	0.203	0.203	0.203

Notes: This table shows the relationship between owner characteristics and steam use. We link (when possible) Census of Manufacturers establishments to the Census of Population, as described in the text.

Column 1 shows the mean value for each characteristic of the linked millers in the sample. Column 2 shows the relationship between steam use and immigrant status, column 3 the relationship with age, and column 4 the relationship with the owner self-reporting their occupation as a miller (or milling-related). Column 5 includes all covariates jointly.

All regressions include our baseline controls interacted with year and industry: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a mill-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.29. Lumber and Flour Mill Activity in 1850, by County Waterpower Potential, by Different River Classifications

	Baseline (1)	Intermittent River (2)	12-Month Average (3)
Panel A. Number of Waterpowered Mills			
Lower Waterpower	-1.055	0.023	-0.553
	(0.130)	(0.036)	(0.106)
Panel B. Revenue of Waterpowered Mills			
Lower Waterpower	-1.127	0.017	-0.678
	(0.249)	(0.059)	(0.170)
Panel C. Steam Share of Mills			
Lower Waterpower	0.089	-0.005	0.048
	(0.015)	(0.003)	(0.015)
Panel D. Steam Share of Revenue			
Lower Waterpower	0.123	-0.007	0.052
	(0.022)	(0.005)	(0.023)
Panel E. Total Number of Mills			
Lower Waterpower	-0.956	0.018	-0.496
	(0.119)	(0.035)	(0.095)
Panel F. Total Revenue of Mills			
Lower Waterpower	-0.876	-0.003	-0.474
	(0.215)	(0.051)	(0.151)
# County-Industries	1,199	1,191	1,199

Notes: This table shows the relationship between 1850 milling activity and waterpower potential. "Lower Waterpower" is a negative standardized measure of county waterpower potential (as described in the text) with standard deviation of one.

Column uses the benchmark measure of waterpower potential from the main text, as in Table 2.2 column 1 (where waterpower potential is proportional to the fall height times the average flow rate in the three lowest months in the year). Column 2 instead calculates waterpower potential only from intermittent rivers, and Column 3 uses the 12-month average flow rate. "Artificial Path" rivers are not formally labeled as intermittent or not, and so we predict their classification as a function of their observables, such as their monthly flows. Each panel shows the effect of waterpower potential on a different outcome. Panel A shows total number of water powered mills and Panel B shows the total revenue of water powered mills. Panel C shows the share of mills using steam power, and Panel D shows the share of milling revenue from steam power. Panel E shows the total number of mills, and Panel F shows total milling revenue. Panels A, B, E, and F use PPML estimation. Panels C and D weight counties by their number of mills.

All regressions include industry fixed effects and our baseline controls interacted with industry: an indicator for the presence of navigable waterways in the county; distance to the nearest navigable waterway; county market access in 1850; an indicator for workable coal deposits in the county; the share of the county covered by coal deposits; and access to coal via the transportation network.

Each observation is a county-industry-year. Robust standard errors clustered by county are reported in parentheses. Data from our main sample counties (Figure 2.2), using our digitized establishment-level Census of Manufactures (1850-1880) and NHDPlusV2.

Table 2.30. Model Fit without Agglomeration

Parameter (1)	Moment (2)	Years (3)	Model: $\alpha_S = 0$ (4)	Model: $\kappa = 0$ (5)	Data (6)
Panel A. Baseline County					
$c(W,S)$	Water Choice Differential: Water Incumbents vs. Entrants	1850–1880	0.546	0.557	0.553 (0.062)
$c(S,W)$	Steam Choice Differential: Steam Incumbents vs. Entrants	1850–1880	0.983	0.972	0.977 (0.123)
$c_S^{(initial)}$	Steam Adoption Rate	1850	0.100	0.100	0.103 (0.006)
$c_S^{(terminal)}$	Steam Adoption Rate	1880	0.393	0.390	0.393 (0.011)
f_e	Entry Rate	1850–1860	0.750	0.750	0.750 (0.006)
f_o^E	Log Sales Differential: Incumbents vs. Entrants	1850–1880	0.134	0.131	0.131 (0.015)
f_o^W	Water Exit Rate	1850–1880	0.789	0.789	0.789 (0.003)
f_o^S	Steam Exit Rate	1850–1880	0.834	0.834	0.835 (0.006)
γ	Log Sales Differential: Steam vs. Water Users	1850–1880	0.853	0.864	0.855 (0.029)
π	Log Sales Autocorrelation	1850–1860	0.412	0.412	0.412 (0.019)
σ	Log Sales Standard Deviation	1850–1860	1.019	1.019	1.019 (0.011)
Panel B. Differences in Lower Waterpower Counties					
$c_L(W)$	Steam Adoption Rate	1850	0.089	0.088	0.089 (0.016)
η	Log Total Output	1850	-0.882	-0.886	-0.876 (0.215)
κ	Change in Steam Adoption Rate	1850, 1880	0.093	0.098	0.092 (0.019)
α_S	Growth of Output	1850, 1880	0.250	0.529	0.525 (0.118)

Notes: This table shows the empirical fit of our estimated model, without agglomeration in steam power. The table shows each estimated parameter of the model (Column 1) and the moment that most closely targets it (Columns 2 and 3). Columns 4 and 5 show the model-simulated moments without agglomeration in steam productivity ($\alpha_S = 0$) and steam adoption costs ($\kappa = 0$), respectively. The columns restrict each parameter to zero and exclude the corresponding target moment from the estimation. Column 6 presents the empirical estimates with robust standard errors, clustered by county, in parentheses.

Table 2.31. Jacobian: Effect of Parameter on Moments, $\frac{dM}{d\theta_k}$

	$c(W,S)$ (1)	$c(S,W)$ (2)	$c_S^{(initial)}$ (3)	$c_S^{(terminal)}$ (4)	f_o^E (5)	f_o^W (6)	$c_L(W)$ (7)	η (8)	κ (9)	α_S (10)
Water Use: W-Inc – Ent	**6.40**	0.05	0.38	-0.05	0.62	-27.38	0.00	0.11	-0.22	-10.14
Steam Use: S-Inc – Ent	0.57	**10.00**	1.32	4.75	4.38	30.95	0.00	-0.10	1.82	9.45
Steam Share 1850	-0.33	-0.03	**-0.53**	-0.57	0.27	9.30	0.00	0.07	-0.10	1.60
Steam Share 1880	-0.30	0.53	-0.35	**-2.10**	0.50	18.82	0.00	0.07	-0.92	9.38
Firm Size: Inc – Ent	-0.80	0.33	-0.43	-1.55	**-3.00**	11.28	0.00	0.21	-0.62	3.82
Water Exit	-0.22	0.03	-0.12	-0.33	-0.27	**4.52**	0.00	-0.00	-0.12	1.41
Steam Share: L – B	0.00	0.07	-0.35	-0.40	0.25	6.73	**2.23**	0.13	-0.22	3.26
Output: L – B	0.38	0.68	-1.82	-2.70	5.15	46.92	-10.93	**-6.94**	-2.27	33.47
Steam Share Growth: L – B	0.43	-0.07	0.35	0.38	-0.12	-7.20	1.27	-0.12	**-0.22**	-0.09
Output Growth: L – B	1.35	0.27	0.80	-0.27	-0.07	-6.93	9.38	4.07	-3.10	**17.47**

Notes: This table shows the Jacobian of the moment function, capturing how simulated moments (in the rows) change with parameter values (in the columns). We order the table rows and columns such that the diagonal elements (in bold font) capture the relationship between parameters and their target moments, as discussed in Sections 2.6.1-2.6.1. The table includes the moment-parameter pairs of our Newton-based estimation in Table 2.7. The Jacobian matrix contains the local derivatives of simulated moments with respect to parameter values, evaluated numerically around our baseline parameter estimates. All parameters except η and α_S are measured in percent of 1850 median firm sales.

Table 2.32. Sensitivity: Effect of Moment on Parameters, $\frac{d\theta}{dM_k}$

	Water Use: WI – E	Steam Use: SI – E	Steam Share 1850	Steam Share 1880	Firm Size: I – E	Water Exit	Steam Share: L – B	Output: L – B	Steam Share Growth: L – B	Output Growth: L – B
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$c(W,S)$	**0.21**	-0.00	-0.85	-0.27	-0.24	2.92	0.35	0.03	-0.69	0.04
$c(S,W)$	-0.05	**0.09**	0.88	0.35	0.63	-4.68	-0.12	-0.01	0.52	-0.06
$c_S^{(initial)}$	0.04	0.01	**-1.31**	0.54	-1.14	8.33	-1.00	-0.14	1.26	-0.10
$c_S^{(terminal)}$	0.02	0.05	0.90	**-0.84**	-0.30	3.61	-0.64	0.06	0.52	0.15
f_o^E	-0.04	0.00	0.62	0.29	**-0.18**	-1.24	-0.22	-0.02	0.37	-0.02
f_o^W	0.01	0.00	0.10	-0.03	-0.09	**0.87**	-0.11	-0.01	0.14	-0.00
$c_L(W)$	-0.01	0.00	0.03	0.05	-0.02	-0.00	**0.30**	-0.02	0.35	-0.04
η	0.03	-0.01	-0.46	-0.08	-0.05	-0.04	-0.04	**-0.03**	-1.43	0.17
κ	0.03	-0.08	-5.54	-0.01	0.03	-0.96	3.27	0.04	**-4.80**	-0.08
α_S	-0.01	-0.01	-0.72	-0.04	0.05	-0.29	0.39	0.03	-0.65	**0.03**

Notes: This table shows the sensitivity measure of Andrews et al. (2017), capturing how parameter estimates (in the rows) change with moment values (in the columns). We order the table rows and columns such that the diagonal elements (in bold font) capture the relationships between parameters and target moments discussed in Sections 2.6.1-2.6.1. The table includes the moment-parameter pairs of our Newton-based estimation in Table 2.7. The sensitivity matrix M is related to the Jacobian J in Table 2.31 as follows: $M = (J'IJ)^{-1}J'I$, where I is the identity matrix. All parameters except η and α_S are measured in percent of 1850 median firm sales.

REFERENCES

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

Abel, M. (2019). Long-run effects of forced resettlement: evidence from apartheid south africa. *The Journal of Economic History*, 79(4):915–953.

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., and Pérez, S. (2021). Automated linking of historical data. *Journal of Economic Literature*, 59(3):865–918.

Acemoglu, D., Akcigit, U., Alp, H., Bloom, N., and Kerr, W. (2018). Innovation, reallocation, and growth. *American Economic Review*, 108(11):3450–91.

Acemoglu, D. and Zilibotti, F. (2001). Productivity differences. *The Quarterly Journal of Economics*, 116(2):563–606.

Adams, E. D. (1927). *Niagara Power: History of the Niagara Falls Power Company, 1886-1918*, volume 1. The Niagara Falls Power Company.

Adão, R., Beraja, M., and Pandalai-Nayar, N. (Forthcoming). Fast and slow technological transitions. *Journal of Political Economy: Macroeconomics*.

Ahlfeldt, G. M. and Pietrostefani, E. (2019). The economic effects of density: A synthesis. *Journal of Urban Economics*, 111:93–107.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The economics of density: Evidence from the berlin wall. *Econometrica*, 83(6):2127–2189.

Alesina, A. and La Ferrara, E. (2005). Ethnic diversity and economic performance. *Journal of economic literature*, 43(3):762–800.

Allcott, H., Collard-Wexler, A., and O'Connell, S. D. (2016). How do electricity shortages affect industry? evidence from india. *American Economic Review*, 106(3):587–624.

Allen, R. C. (2009). *The British industrial revolution in global perspective*. Cambridge University Press.

Allen, T. and Donaldson, D. (2022). Persistence and path dependence in the spatial economy. Technical report, National Bureau of Economic Research.

Amsden, A. H. (1992). *Asia's next giant: South Korea and late industrialization*. Oxford University Press, USA.

Ananat, E., Fu, S., and Ross, S. L. (2013). Race-specific agglomeration economies: Social distance and the black-white wage gap. Technical report, National Bureau of Economic Research.

Ananat, E., Shihe, F., and Ross, S. L. (2018). Race-specific urban wage premia and the black-white wage gap. *Journal of Urban Economics*, 108:141–153.

Anderson, J. E. and Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American economic review*, 93(1):170–192.

Andreas, A. (1881). *History of Milwaukee, Wisconsin*. The Western Historical Company.

Andrews, I., Gentzkow, M., and Shapiro, J. M. (2017). Measuring the Sensitivity of Parameter Estimates to Estimation Moments. *The Quarterly Journal of Economics*, 132(4):1553–1592.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Arcidiacono, P. and Miller, R. A. (2011). Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica*, 79(6):1823–1867.

Arkolakis, C. and Walsh, C. (2023). Clean growth. *Working Paper*.

Armington, P. S. (1969). A theory of demand for products distinguished by place of production (une théorie de la demande de produits différenciés d'après leur origine)(una teoría de la demanda de productos distinguiéndolos según el lugar de producción). *Staff Papers-International Monetary Fund*, pages 159–178.

Armistead, W., Lawson, T., and Long, S. (1841). *House Documents, Otherwise Publ. as Executive Documents: 13th Congress, 2d Session-49th Congress, 1st Session, Volume 4*, chapter Report of the Board of Officers Appointed to Select a Suitable Site of the Establishment of a National Armory on the Western Waters. Washington, D.C.

Armitage, S. (2023). Technology adoption and the timing of environmental policy: Evidence from efficient lighting. *Working Paper*.

Arrow, K. J. (1962). The economic implications of learning by doing. *The Review of Economic Studies*, 29(3):155–173.

Artuç, E., Chaudhuri, S., and McLaren, J. (2010). Trade shocks and labor adjustment: A structural empirical approach. *American Economic Review*, 100(3):1008–1045.

Ashraf, Q. and Galor, O. (2013). The "out of africa" hypothesis, human genetic diversity, and comparative economic development. *American Economic Review*, 103(1):1–46.

Ashton, T. S. (1948). *The Industrial Revolution, 1760-1830*. Oxford University Press.

Asker, J., Collard-Wexler, A., and De Loecker, J. (2014). Dynamic inputs and resource (mis) allocation. *Journal of Political Economy*, 122(5):1013–1063.

Atack, J. (1976). Estimation of economies of scale in nineteenth cenury united states manufacturing and the form of the production function. *PhD Dissertation. Indiana University, Bloomington*.

Atack, J. (1979). Fact in fiction? the relative costs of steam and water power: a simulation approach. *Explorations in Economic History*, 16(4):409–437.

Atack, J. and Bateman, F. (1999). Nineteenth-century us industrial development through the eyes of the census of manufactures a new resource for historical research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 32(4):177–188.

Atack, J., Bateman, F., and Margo, R. A. (2004). Skill intensity and rising wage dispersion in nineteenth-century american manufacturing. *The Journal of Economic History*, 64(1):172–192.

Atack, J., Bateman, F., and Margo, R. A. (2008). Steam power, establishment size, and labor productivity growth in nineteenth century american manufacturing. *Explorations in Economic History*, 45(2):185–198.

Atack, J., Bateman, F., and Weiss, T. (1980). The regional diffusion and adoption of the steam engine in american manufacturing. *The Journal of Economic History*, 40(2):281–308.

Atack, J. and Margo, R. A. (2019). Gallman revisited: blacksmithing and american manufacturing, 1850–1870. *Cliometrica*, 13(1):1–23.

Atack, J., Margo, R. A., and Rhode, P. W. (2019). 'automation' of manufacturing in the late nineteenth century: The hand and machine labor study. *Journal of Economic Perspectives*, 33(2):51–70.

Atack, J., Margo, R. A., and Rhode, P. W. (2022). 'mechanization takes command?': Powered machinery and production times in late nineteenth-century american manufacturing. *The Journal of Economic History*, 82(3):663–689.

Atkeson, A. and Kehoe, P. J. (1999). Models of energy use: Putty-putty versus putty-clay. *American Economic Review*, 89(4):1028–1043.

Atkeson, A. and Kehoe, P. J. (2007). Modeling the transition to a new economy: lessons from two technological revolutions. *American Economic Review*, 97(1):64–88.

Bachmann, R. and Bayer, C. (2014). Investment dispersion and the business cycle. *American Economic Review*, 104(4):1392–1416.

Bahk, B.-H. and Gort, M. (1993). Decomposing learning by doing in new plants. *Journal of Political Economy*, 101(4):561–583.

Bailey, M. J., Cole, C., Henderson, M., and Massey, C. (2020). How well do automated linking methods perform? lessons from us historical data. *Journal of Economic Literature*, 58(4):997–1044.

Baillargeon, D. (2021). Spaces of occupation: Colonial enclosure and confinement in british malaya. *Journal of Historical Geography*, 73:24–35.

Balboni, C. (2024). In harm's way? infrastructure investments and the persistence of coastal cities. Working paper.

Baldwin, R. and Venables, A. J. (2015). Trade policy and industrialisation when backward and forward linkages matter. *Research in Economics*, 69(2):123–131.

Balke, N. and Lamadon, T. (2022). Productivity shocks, long-term contracts, and earnings dynamics. *American Economic Review*, 112(7):2139–2177.

Barrows, F. I. (1917). *History of Fayette County, Indiana: Her People, Industries and Institutions*. BF Bowen.

Basu, S. and Weil, D. N. (1998). Appropriate technology and growth. *The Quarterly Journal of Economics*, 113(4):1025–1054.

Bateman, F., Foust, J. D., and Weiss, T. J. (1971). Large-scale manufacturing in the south and west, 1850–1860. *Business History Review*, 45(1):1–17.

Bateman, F. and Weiss, T. (1981). *A deplorable scarcity: the failure of industrialization in the slave economy*. UNC Press Books.

Bates, R. M. (1945). *The water-powered mills of Allen County, Indiana*. Fort Wayne, Indiana.

Bates, S. P. (1884). *History of Erie county, Pennsylvania*. Warner, Beers & Co, New York.

Bayer, P., Ferreira, F., and McMillan, R. (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of political economy*, 115(4):588–638.

Bayer, P., McMillan, R., and Rueben, K. (2004). An equilibrium model of sorting in an urban housing market.

Bazzi, S., Fiszbein, M., and Gebresilasse, M. (2020). Frontier culture: The roots and persistence of 'rugged individualism' in the united states. *Econometrica*, 88(6):2329–2368.

Bazzi, S., Gaduh, A., Rothenberg, A. D., and Wong, M. (2016). Skill transferability, migration, and development: Evidence from population resettlement in indonesia. *American Economic Review*, 106(9):2658–2698.

Becker, S. O., Grosfeld, I., Grosjean, P., Voigtländer, N., and Zhuravskaya, E. (2020). Forced migration and human capital: Evidence from post-wwii population transfers. *American Economic Review*, 110(5):1430–1463.

Benhabib, J., Perla, J., and Tonetti, C. (2021). Reconciling models of diffusion and innovation: A theory of the productivity distribution and technology frontier. *Econometrica*, 89(5):2261–2301.

Benhabib, J. and Rustichini, A. (1991). Vintage capital, investment, and growth. *Journal of Economic Theory*, 55(2):323–339.

Benhabib, J. and Rustichini, A. (1993). A vintage capital model of growth and investment: Theory and evidence. In *General Equilibrium, Growth and Trade II*, pages 248–300. Academic Press.

Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers*, (13).

Berger, D. and Vavra, J. (2015). Consumption dynamics during recessions. *Econometrica*, 83(1):101–154.

Berger, T. and Ostermeyer, V. (2023). Institutional change and the adoption of new technologies: The case of steam. *Working Paper*.

Bertola, G. and Caballero, R. J. (1994). Irreversibility and aggregate investment. *The Review of Economic Studies*, 61(2):223–246.

Besley, T. and Mueller, H. (2012). Estimating the peace dividend: The impact of violence on house prices in northern ireland. *American Economic Review*, 102(2):810–833.

Bishop, P. and Muñoz-Salinas, E. (2013). Tectonics, geomorphology and water mill location in scotland, and the potential impacts of mill dam failure. *Applied Geography*, 42:195–205.

Bleakley, H. and Lin, J. (2012). Portage and path dependence. *The Quarterly Journal of Economics*, 127(2):587–644.

Blinder, A. S. (2008). A modest proposal: Eco-friendly stimulus. *New York Times*, page B5.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica*, 77(3):623–685.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013). Does management matter? evidence from india. *The Quarterly Journal of Economics*, 128(1):1–51.

Borusyak, K. and Hull, P. (2023). Nonrandom exposure to exogenous shocks. *Econometrica*, 91(6):2155–2185.

Braguinsky, S., Ohyama, A., Okazaki, T., and Syverson, C. (2015). Acquisitions, productivity, and profitability: evidence from the japanese cotton spinning industry. *American Economic Review*, 105(7):2086–2119.

Bresnahan, T. and Greenstein, S. (1996). Technical progress and co-invention in computing and in the uses of computers. *Brookings Papers on Economic Activity. Microeconomics*, 1996:1–83.

Bresnahan, T. F. and Trajtenberg, M. (1995). General purpose technologies 'engines of growth'? *Journal of Econometrics*, 65(1):83–108.

Brey, B. (2023). Sparking knowledge: Early technology adoption, innovation ability and long-run growth. *GEP Research Paper*.

Broda, C. and Weinstein, D. E. (2006). Globalization and the gains from variety. *The Quarterly Journal of Economics*, 121(2):541–585.

Brown, N. C. (1923). *The American Lumber Industry: Embracing the Principal Features of the Resources, Production, Distribution, and Utilization of Lumber in the United States*. J. Wiley.

Bryan, G. and Morten, M. (2019). The aggregate productivity effects of internal migration: Evidence from indonesia. *Journal of Political Economy*, 127(5):2229–2268.

Buera, F. J., Hopenhayn, H., Shin, Y., and Trachter, N. (2021). Big push in distorted economies. Technical report, National Bureau of Economic Research.

Burke, J. G. (1966). Bursting boilers and the federal power. *Technology and Culture*, 7(1):1–23.

Bustos, P. (2011). Trade liberalization, exports, and technology upgrading: Evidence on the impact of mercosur on argentinian firms. *American Economic Review*, 101(1):304–40.

Cabral, L. M. B. and Mata, J. (2003). On the evolution of the firm size distribution: Facts and theory. *American economic review*, 93(4):1075–1090.

Caliendo, L., Dvorkin, M., and Parro, F. (2019). Trade and labor market dynamics: General equilibrium analysis of the china trade shock. *Econometrica*, 87(3):741–835.

Campbell, M. R. (1908). Coal fields of the united states. *United States Geological Survey, Washington, D.C.*

Carlitz, R., Morjaria, A., and Mueller, J. M. (2022). State-building in a diverse society. Technical report, National Bureau of Economic Research.

Caunedo, J. and Keller, E. (2021). Capital obsolescence and agricultural productivity. *The Quarterly Journal of Economics*, 136(1):505–561.

Census Bureau (1883). Tenth Census. Volumes 16 & 17. Reports on the Water Power of the United States. Technical report.

Chamberlain, G. (1985). Heterogeneity, omitted variable bias and duration dependencen. *Longitudinal Analysis of Labor Market Data, in eds JJ Heckman and B. Singer, Cambridge UP: Cambridge.*

Chandler, A. D. (1972). Anthracite coal and the beginnings of the industrial revolution in the united states. *Business History Review*, 46(2):141–181.

Chari, V. V. and Hopenhayn, H. (1991). Vintage human capital, growth, and the diffusion of new technology. *Journal of Political Economy*, 99(6):1142–1165.

Charlier, R. H. and Menanteau, L. (1997). The saga of tide mills. *Renewable and Sustainable Energy Reviews*, 1(3):171–207.

Chay, K. Y., Hoynes, H. W., and Hyslop, D. (1999). A non-experimental analysis of true state dependence in monthly welfare participation sequences. *American Statistical Association*, pages 9–17.

Chen, J. and Roth, J. (2023). Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics*, 139(2):891–936.

Chen, L. (2024). The dynamic efficiency of policy uncertainty: Evidence from the wind industry. *Working Paper*.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chernoff, A. W. (2021). Firm heterogeneity, technology adoption and the spatial distribution of population: Theory and measurement. *Canadian Journal of Economics/Revue canadienne d'économique*, 54(2):475–521.

Ciccone, A. and Peri, G. (2006). Identifying human-capital externalities: Theory with applications. *The Review of Economic Studies*, 73(2):381–412.

Cole, A. H. (1938). *Wholesale commodity prices in the United States, 1700-1861*. Harvard University Press.

Cole, A. H. (1970). The mystery of fuel wood marketing in the united states. *Business History Review*, 44(3):339–359.

Collard-Wexler, A. and De Loecker, J. (2015). Reallocation and technology: Evidence from the us steel industry. *American Economic Review*, 105(1):131–71.

Colton, G. (1882). Colton's railroad map (intermediate size) of the united states. *New York: GW & CB Colton & Co.*

Combes, P.-P. and Gobillon, L. (2015). The empirics of agglomeration economies. In *Handbook of regional and urban economics*, volume 5, pages 247–348. Elsevier.

Comin, D. and Hobijn, B. (2010). An exploration of technology diffusion. *American Economic Review*, 100(5):2031–2059.

Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of econometrics*, 92(1):1–45.

Cooney, E. (1991). Eighteenth century britain's missing sawmills: A blessing in disguise? *Construction History*, 7:29–46.

Corry, W. (1954). A general survey of new villages. *Kuala Lumpur: Government Printers (being a report to His Excellency Sir Donald MacGullivray)*, page 24.

Coşar, A. K., Guner, N., and Tybout, J. (2016). Firm dynamics, job turnover, and wage distributions in an open economy. *American Economic Review*, 106(3):625–663.

Crafts, N. (2004). Steam as a general purpose technology: a growth accounting perspective. *The Economic Journal*, 114(495):338–351.

Crafts, N. F. (1977). Industrial revolution in england and france: Some thoughts on the question, 'why was england first?'. *The Economic History Review*, 30(3):429–441.

Cronon, W. (2009). *Nature's metropolis: Chicago and the Great West*. WW Norton & Company.

Damron, W. (2023). Gains from factory electrification: Evidence from north carolina, 1905-1926. *Working Paper*.

Dano, K. (2023). Transition probabilities and moment restrictions in dynamic fixed effects logit models. *Working Paper*.

David, P. A. (1969). *A Contribution to the Theory of Diffusion*. Research Center in Economic Growth Stanford University.

David, P. A. (1985). Clio and the economics of qwerty. *American Economic Review*, 75(2):332–337.

David, P. A. (1990). The dynamo and the computer: an historical perspective on the modern productivity paradox. *American Economic Review*, 80(2):355–361.

David, P. A. and Bunn, J. A. (1988). The economics of gateway technologies and network evolution: Lessons from electricity supply history. *Information economics and policy*, 3(2):165–202.

Davis, D. R. and Dingel, J. I. (2019). A spatial knowledge economy. *American Economic Review*, 109(1):153–170.

Davis, D. R., Dingel, J. I., Monras, J., and Morales, E. (2019). How segregated is urban consumption? *Journal of Political Economy*, 127(4):1684–1738.

Dedrick, B. (1931). Plan and power of a flour mill of many years ago. *The Millers Review and Feed Forum*, XCV(2).

Del Tufo, M. V. (1947). *Malaya Comprising the Federation of Malaya and the Colony of Singapore: A Report on the 1947 Census of Population*. Crown Agents for the Colonies.

Dell, M., Lane, N., and Querubin, P. (2018). The historical state, local collective action, and economic development in vietnam. *Econometrica*, 86(6):2083–2121.

Dell, M. and Olken, B. A. (2020). The development effects of the extractive colonial economy: The dutch cultivation system in java. *The Review of Economic Studies*, 87(1):164–203.

Delle Donne, C. (1973). *Federal census schedules, 1850-1880: Primary sources for historical research.* National Archives and Records Service.

Denison, E. F. (1964). The unimportance of the embodied question. *American Economic Review*, 54(2):90–94.

Department of Statistics Malaysia (1980). Report of the household expenditure survey: Peninsular malaysia (1980), sabah & sarawak (1982). Technical report.

Devine, W. D. (1983). From shafts to wires: Historical perspective on electrification. *The Journal of Economic History*, 43(2):347–372.

dhu Renick, R. (1965). The emergency regulations of malaya causes and effect. *Journal of Southeast Asian History*, 6(2):1–39.

Diamond, R. (2016). The determinants and welfare implications of us workers' diverging location choices by skill: 1980–2000. *American Economic Review*, 106(3):479–524.

Dippel, C. (2014). Forced coexistence and economic development: evidence from native american reservations. *Econometrica*, 82(6):2131–2165.

Dixit, A. K. and Pindyck, R. S. (1994). *Investment under uncertainty.* Princeton university press.

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

Dondlinger, P. T. (1919). *The Book of Wheat: An Economic History and Practical Manual of the Wheat Industry.* Orange Judd Company.

Duflo, E. and Pande, R. (2007). Dams. *The Quarterly Journal of Economics*, 122(2):601–646.

Duranton, G. and Puga, D. (2004). Micro-foundations of urban agglomeration economies. In *Handbook of regional and urban economics*, volume 4, pages 2063–2117. Elsevier.

Eberle, U. J., Henderson, J. V., Rohner, D., and Schmidheiny, K. (2020). Ethnolinguistic diversity and urban agglomeration. *Proceedings of the National Academy of Sciences*, 117(28):16250–16257.

Eckert, F. and Peters, M. (2023). Spatial structural change. *Working Paper.*

Ellis, F. and Hungerford, A. N. (1886). *History of that Part of the Susquehanna and Juniata Valleys: Embraced in the Counties of Mifflin, Juniata, Perry, Union and Snyder, in the Commonwealth of Pennsylvania*, volume 1. Everts, Peck & Richards.

Emery, C. E. (1883). The cost of steam power. *Transactions of the American Society of Civil Engineers*, 12(1):425–431.

Evans, O. (1795). *The Young Mill-Wright and Miller's Guide*. Printed For, and Sold By the Author.

Fajgelbaum, P. D. and Gaubert, C. (2020). Optimal spatial policies, geography, and sorting. *The Quarterly Journal of Economics*, 135(2):959–1036.

Farrell, J. and Klemperer, P. (2007). Coordination and lock-in: Competition with switching costs and network effects. *Handbook of industrial organization*, 3:1967–2072.

Feigenbaum, J. (2016). Automated census record linking: A machine learning approach. *Working Paper*.

Feigenbaum, J. and Gross, D. (2023). Organizational and economic obstacles to automation: A cautionary tale from at&t in the twentieth century. *Working Paper*.

Feigenbaum, J., Lee, J., and Mezzanotti, F. (2022). Capital destruction and economic growth: The effects of sherman's march, 1850–1920. *American Economic Journal: Applied Economics*, 14(4):301–342.

Felbermayr, G., Impullitti, G., and Prat, J. (2018). Firm dynamics and residual inequality in open economies. *Journal of the European Economic Association*, 16(5):1476–1539.

Fell, H. (1960). Population census of the federation of malaya. Technical report.

Fellegi, I. P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical association*, 71(353):17–35.

Fenichel, A. H. (1966). *Growth and Diffusion of Power in Manufacturing, 1838–1919*, pages 443–478. NBER.

Fenneman, N. M. (1946). Physical divisions of the united states. Technical report, US Geological Survey.

Fernihough, A. and O'Rourke, K. H. (2021). Coal and the european industrial revolution. *The Economic Journal*, 131(635):1135–1149.

Ferrie, J. P. (1996). A new sample of males linked from the public use microdata sample of the 1850 us federal census of population to the 1860 us federal census manuscript schedules. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29(4):141–156.

Fisher, R. (1845). Advantages of manufactories in the city of new york. *Fisher's National Magazine and Industrial Record*, 1(4):358–362.

Floud, R. and McCloskey, D. N. (1981). *The economic history of Britain since 1700*, volume 1. Cambridge University Press.

Flour and Feed (1945). *Devoted to the Interests of the Flour and Feed Trade*, volume 22. Milwaukee, Wisconsin.

Fogel, R. W. (1964). *Railroads and American Economic Growth: Essays in Econometric History*. Johns Hopkins University Press.

Foster, L., Grim, C., and Haltiwanger, J. (2016). Reallocation in the great recession: cleansing or not? *Journal of Labor Economics*, 34(S1):S293–S331.

Franck, R. and Galor, O. (2021). Flowers of evil? industrialization and long run development. *Journal of Monetary Economics*, 117:108–128.

Franck, R. and Galor, O. (2022). Technology-skill complementarity in early phases of industrialisation. *The Economic Journal*, 132(642):618–643.

Frankel, M. (1955). Obsolescence and technological change in a maturing economy. *American Economic Review*, 45(3):296–319.

Gerschenkron, A. (1962). *Economic Backwardness in Historical Perspective (Belknap Press S)*. Belknap Press: An Imprint of Harvard University Press.

Gershman, B., Ashraf, Q. H., Cinnirella, F., Galor, O., Hornung, E., et al. (2022). Structural change, elite capitalism, and the emergence of labor emancipation. *Working Paper*.

Gilchrist, S. and Williams, J. C. (2000). Putty-clay and investment: a business cycle analysis. *Journal of Political Economy*, 108(5):928–960.

Ginsburg, N. S., Roberts, C. F., and Comber, L. (1958). Malaya. *(No Title)*.

Giorcelli, M. (2019). The long-term effects of management and technology transfers. *American Economic Review*, 109(1):121–152.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., and Shleifer, A. (1992). Growth in cities. *Journal of political economy*, 100(6):1126–1152.

Gordon, R. B. (1983). Cost and use of water power during industrialization in new england and great britain: a geological interpretation. *Economic History Review*, pages 240–259.

Greenberg, D. (1982). Reassessing the power patterns of the industrial revolution: An anglo-american comparison. *The American Historical Review*, 87(5):1237–1261.

Greenstone, M., Hornbeck, R., and Moretti, E. (2010). Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy*, 118(3):536–598.

Greenwood, J., Seshadri, A., and Yorukoglu, M. (2005). Engines of liberation. *The Review of Economic Studies*, 72(1):109–133.

Greenwood, J. and Yorukoglu, M. (1997). 1974. In *Carnegie-Rochester conference series on public policy*, volume 46, pages 49–95. Elsevier.

Gregg, A. G. (2020). Factory productivity and the concession system of incorporation in late imperial russia, 1894–1908. *American Economic Review*, 110(2):401–427.

Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica*, 25(4):501–522.

Guilfoos, T. (2022). The value of water power during the american industrial revolution. *Working Paper*.

Gyourko, J., Saiz, A., and Summers, A. (2008). A new measure of the local regulatory environment for housing markets: The wharton residential land use regulatory index. *Urban studies*, 45(3):693–729.

Habakkuk, H. (1967). *American and British technology in the nineteenth century*. Cambridge University Press.

Haggard, S. (1990). The politics of growth in the newly industrialized countries.

Haines, M. R. (2010). Historical, demographic, economic, and social data: The united states, 1790–2002. *ICPSR*, 2896.

Hall, B. H. (2004). Innovation and diffusion.

Handel, B. R. (2013). Adverse selection and inertia in health insurance markets: When nudging hurts. *American Economic Review*, 103(7):2643–2682.

Hanlon, W. W. (2022). The rise of the engineer: Inventing the professional inventor during the industrial revolution. Technical report, NBER.

Hansen, G. D. and Prescott, E. C. (2002). Malthus to solow. *American economic review*, 92(4):1205–1217.

Hardesty, R. D. (2019). Wild confusion in every direction.

Heblich, S., Redding, S. J., and Sturm, D. M. (2020). The making of the modern metropolis: evidence from london. *The Quarterly Journal of Economics*, 135(4):2059–2133.

Henderson, R. (1995). Of life cycles real and imaginary: The unexpectedly long old age of optical lithography. *Research Policy*, 24(4):631–643.

Hilhorst, D. and Leeuwen, M. v. (2000). Emergency and development: The case of imidugudu, villagization in rwanda. *Journal of Refugee Studies*, 13(3):264–280.

Hirschman, A. O. (1958). *The Strategy of Economic Development*. Yale University Press.

Holmes, T. J., Levine, D. K., and Schmitz Jr, J. A. (2012). Monopoly and the incentive to innovate when adoption involves switchover disruptions. *American Economic Journal: Microeconomics*, 4(3):1–33.

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica*, 60(5):1127–1150.

Hornbeck, R. (2010). Barbed wire: Property rights and agricultural development. *The Quarterly Journal of Economics*, 125(2):767–810.

Hornbeck, R. and Keniston, D. (2017). Creative destruction: Barriers to urban growth and the great boston fire of 1872. *American Economic Review*, 107(6):1365–1398.

Hornbeck, R. and Rotemberg, M. (2024). Growth off the rails: Aggregate productivity growth in distorted economies. *Journal of Political Economy*, Forthcoming.

Hotz, V. J. and Miller, R. A. (1993). Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, 60(3):497–529.

Howes, A. (2022). Age of invention: How the steam engine was invented.

Hsieh, C.-T., Hurst, E., Jones, C. I., and Klenow, P. J. (2019). The allocation of talent and us economic growth. *Econometrica*, 87(5):1439–1474.

Hubbard, W. (1861). *Maine Reports: Cases in Law and Equity Determined by the Supreme Judicial Court of Maine*, volume XLVI. Hallowell: Masters, Smith & Co.

Huesler, J. and Strobl, E. (2023). The creative-destructive force of hurricanes: Evidence from technological adoption in colonial jamaican sugar estates? *Working Paper*.

Humlum, A. (2022). Robot adoption and labor market dynamics. *Working Paper*.

Humphrey, J. W. (1971). *Population resettlement in Malaya*. Northwestern University.

Hunter, L. C. (1979). *A History of Industrial Power in the United States, 1780-1930. Volume 1: Waterpower in the Century of the Steam Engine*. Published for the Eleutherian Mills-Hagley Foundation by the University Press of Virginia.

Hunter, L. C. (1985). *A History of Industrial Power in the United States, 1780-1930. Volume 2: Steam Power*, volume 2. published for the Hagley Museum and Library by the University Press of Virginia.

Huntington, H., Samaniego dela Parra, B., and Shenoy, A. (2023). Measuring and estimating retail productivity. *Working Paper*.

International Publishing Co (1887). *Quarter-century's Progress of New Jersey's Leading Manufacturing Centres: Dover*. New York.

Jackson, D. C. (1975). Historic american engineering record: Luckenbach flour mill. *National Park Service*.

Jaremski, M. (2014). National banking's role in us industrialization, 1850–1900. *The Journal of Economic History*, 74(1):109–140.

Jones, J. E. (1914). *A History of Columbia County, Wisconsin: A Narrative Account of Its Historical Progress, Its People, and Its Principal Interests*, volume 2. Lewis Publishing Company.

Jovanovic, B. and Lach, S. (1989). Entry, exit, and diffusion with learning by doing. *American Economic Review*, 79(4):690–699.

Jovanovic, B. and MacDonald, G. M. (1994). Competitive diffusion. *Journal of Political Economy*, 102(1):24–52.

Jovanovic, B. and Nyarko, Y. (1996). Learning by doing and the choice of technology. *Econometrica*, 64(6):1299–1310.

Jovanovic, B. and Rousseau, P. L. (2005). General purpose technologies. In *Handbook of economic growth*, volume 1, pages 1181–1224. Elsevier.

Jovanovic, B. and Yatsenko, Y. (2012). Investment in vintage capital. *Journal of Economic Theory*, 147(2):551–569.

Juhász, R., Lane, N. J., and Rodrik, D. (2023a). The new economics of industrial policy. *Working Paper*.

Juhász, R., Squicciarini, M. P., and Voigtländer, N. (2023b). Technology adoption and productivity growth: Evidence from industrialization in france. *Working Paper*.

Keane, M. P., Todd, P. E., and Wolpin, K. I. (2011). The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications. In *Handbook of Labor Economics*, volume 4, pages 331–461. Elsevier.

Kennan, J. and Walker, J. R. (2011). The effect of expected income on individual migration decisions. *Econometrica*, 79(1):211–251.

Kermani, A. and Ma, Y. (2023). Asset specificity of nonfinancial firms. *The Quarterly Journal of Economics*, 138(1):205–264.

Klemperer, P. (1995). Competition when consumers have switching costs: An overview with applications to industrial organization, macroeconomics, and international trade. *The Review of Economic Studies*, 62(4):515–539.

Kline, P. and Moretti, E. (2014). Local economic development, agglomeration economies, and the big push: 100 years of evidence from the tennessee valley authority. *The Quarterly journal of economics*, 129(1):275–331.

Kuhlmann, C. B. (1929). *The development of the flour-milling industry in the United States: With special reference to the industry in Minneapolis*. Houghton Mifflin.

Kuznets, S. (1967). *Modern Economic Growth: Rate, Structure, and Spread, (Study in Comparative Economics)*. Yale University Press.

Lagakos, D. and Waugh, M. E. (2013). Selection, agriculture, and cross-country productivity differences. *American Economic Review*, 103(2):948–980.

Lancaster, T. and Nickell, S. (1980). The analysis of re-employment probabilities for the unemployed. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 143(2):141–152.

Landes, D. S. (1969). *The unbound Prometheus: technological change and industrial development in Western Europe from 1750 to the present*. Cambridge University Press.

Lane, N. (2022). Manufacturing revolutions: Industrial policy and industrialization in south korea. *Available at SSRN 3890311*.

Lanteri, A. (2018). The market for used capital: Endogenous irreversibility and reallocation over the business cycle. *American Economic Review*, 108(9):2383–2419.

Le Bris, D., Goetzmann, W. N., and Pouget, S. (2019). The present value relation over six centuries: The case of the bazacle company. *Journal of Financial Economics*, 132(1):248–265.

Lee, K. H. and Tan, C.-B. (2000). *The Chinese in Malaysia*. Oxford University Press.

Lim, C.-Y. (1964). *Economic development of modern Malaya*. PhD thesis, University of Oxford.

Lipscomb, M., Mobarak, A. M., and Barham, T. (2013). Development effects of electrification: Evidence from the topographic placement of hydropower plants in brazil. *American Economic Journal: Applied Economics*, 5(2):200–231.

Loh, F. K.-W. (1988). Beyond the tin mines: coolies, squatters and new villagers in the kinta valley, malaysia, c. 1880-1980.

Lu, Y., Phillips, G. M., and Yang, J. (2023). The impact of cloud computing and ai on industry dynamics and competition. *Working Paper*.

Manuelli, R. E. and Seshadri, A. (2014). Frictionless technology diffusion: The case of tractors. *American Economic Review*, 104(4):1368–1391.

Margo, R. A. (2002). The north-south wage gap, before and after the civil war.

Marshall, A. (1890). *Principles of economics*. London: Macmillan.

Mayr, O. (1975). Yankee practice and engineering theory: Charles t. porter and the dynamics of the high-speed steam engine. *Technology and culture*, 16(4):570–602.

McElheran, K., Li, J. F., Brynjolfsson, E., Krof, Z., Dinlersoz, E., Foster, L., and Zolas, N. (2023). Ai adoption in america: Who, what, and where. *Working Paper*.

McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., and Rea, A. (2012). User's guide for the national hydrography dataset plus (nhdplus) high resolution. *US Environmental Protection Agency.*

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6):1695–1725.

Melo, P. C., Graham, D. J., and Noland, R. B. (2009). A meta-analysis of estimates of urban agglomeration economies. *Regional science and urban Economics*, 39(3):332–342.

Mian, A. and Sufi, A. (2014). What explains the 2007–2009 drop in employment? *Econometrica*, 82(6):2197–2223.

Miller, N., Osborne, M., Sheu, G., and Sileo, G. (2022). Technology and market power: The united states cement industry, 1974-2019. *Working Paper.*

Mokyr, J. (2016). *A culture of growth: The origins of the modern economy.* Princeton University Press.

Monroe, J. (1825). *Message from the President of the United States, transmitting a report of the commissioners appointed under the act of the third of March, 1823, to establish a national armory on the western waters.* Washington, D.C.

Montalvo, J. G. and Reynal-Querol, M. (2005). Ethnic diversity and economic development. *Journal of Development economics*, 76(2):293–323.

Monte, F., Redding, S. J., and Rossi-Hansberg, E. (2018). Commuting, migration, and local employment elasticities. *American Economic Review*, 108(12):3855–3890.

Montgomery, J. (1840). *A Practical Detail of the Cotton Manufacture of the United States of America: And the State of the Cotton Manufacture of that Country Contrasted and Compared with that of Great Britain; with Comparative Estimates of the Cost of Manufacturing in Both Countries.* J. Niven.

Moore, R. B., McKay, L. D., Rea, A. H., Bondelid, T. R., Price, C. V., Dewald, T. G., Johnston, C. M., et al. (2019). User's guide for the national hydrography dataset plus (nhdplus) high resolution. *Open-File Report-US Geological Survey, 2019-1096.*

Moretti, E. (2004a). Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *Journal of econometrics*, 121(1-2):175–212.

Moretti, E. (2004b). Workers' education, spillovers, and productivity: evidence from plant-level production functions. *American Economic Review*, 94(3):656–690.

Morten, M. and Oliveira, J. (2018). The effects of roads on trade and migration: Evidence from a planned capital city. *NBER Working Paper*, 22158:1–64.

Mullin, J. R. and Kotval, Z. (2021). Manufacturing in puritan rural towns in new england 1630–60:'a miller never goes to heaven'. *Rural History*, 32(2):187–196.

Mundial, B. (1993). *The East Asian miracle: economic growth and public policy*. World Bank.

National Industrial Conference Board (1925). Trade associations: Their economic significance and legal status. National industrial conference board, Incorporated.

NOAA (2016). National water model: Improving noaa's water prediction services. Technical report.

North, D. (1958). Ocean freight rates and economic development 1730-1913. *The Journal of Economic History*, 18(4):537–555.

Nunn, N. and Puga, D. (2012). Ruggedness: The blessing of bad geography in africa. *Review of Economics and Statistics*, 94(1):20–36.

Nyce, R. (1973). *Chinese new villages in Malaya: a community study*. Malaysian Sociological Research Institute.

Olmstead, A. L. and Rhode, P. W. (2001). Reshaping the landscape: the impact and diffusion of the tractor in american agriculture, 1910–1960. *The Journal of Economic History*, 61(3):663–698.

Ottonello, P. and Winberry, T. (2020). Financial heterogeneity and the investment channel of monetary policy. *Econometrica*, 88(6):2473–2502.

Parente, S. L. (1994). Technology adoption, learning-by-doing, and economic growth. *Journal of Economic Theory*, 63(2):346–369.

Parente, S. L. and Prescott, E. C. (1994). Barriers to technology adoption and development. *Journal of Political Economy*, 102(2):298–321.

Perren, R. (1990). Structural change and market growth in the food industry: flour milling in britain, europe, and america, 1850-1914. *Economic History Review*, 43(3):420–437.

Peters, M. (2022). Market size and spatial growth-evidence from germany's post-war population expulsions. *Econometrica*, 90(5):2357–2396.

Porter, C. T. (1868). On the allen engine and governor. *Proceedings of the Institution of Mechanical Engineers*, 19(1):50–80.

Price, J., Buckles, K., Van Leeuwen, J., and Riley, I. (2021). Combining family history and machine learning to link historical records: The census tree data set. *Explorations in Economic History*, 80:101391.

Purcell, V. (1947). Chinese settlement in malacca. *Journal of the Malayan Branch of the Royal Asiatic Society*, 20(1 (141):115–125.

Ragnar, N. (1953). *Problems of Capital Formation in Underdeveloped Countries*. Oxford University Press.

Ramey, V. A. and Shapiro, M. D. (2001). Displaced capital: A study of aerospace plant closings. *Journal of Political Economy*, 109(5):958–992.

Reichardt, H. (2023). Scale-biased technical change and inequality. *Working Paper*.

Ridolfi, L., Salvo, C., and Weisdorf, J. (2023). The effect of mechanisation on labour: Evidence from the diffusion of steam. *CAGE Working Paper 689*.

Robinson, J. B. P. (1956). Transformation in malaya. *(No Title)*.

Rohner, D. and Thoenig, M. (2021). The elusive peace dividend of development policy: From war traps to macro complementarities. *Annual Review of Economics*, 13:111–131.

Rosenberg, N. and Trajtenberg, M. (2004). A general-purpose technology at work: The corliss steam engine in the late-nineteenth-century united states. *The Journal of Economic History*, 64(1):61–99.

Rosenthal, S. S. and Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. In *Handbook of regional and urban economics*, volume 4, pages 2119–2171. Elsevier.

Rossi-Hansberg, E., Sarte, P.-D., and Schwartzman, F. (2023). Cognitive hubs and spatial redistribution. Technical report, National Bureau of Economic Research.

Rostow, W. W. (1975). *How it All Began: Origins of the Modern Economy*. University of Texas Press.

Rotemberg, J. J. (1982). Sticky prices in the united states. *Journal of Political Economy*, 90(6):1187–1211.

Rotemberg, M. and White, T. K. (2021). Plant-to-table (s and figures): Processed manufacturing data and measured misallocation. *Working Paper*.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*.

Rothschild, E. W. (1976). *A History of Cheltenham Township*. Cheltenham Township Historical Commission.

Ruggles, S., Fitch, C. A., and Roberts, E. (2018). Historical census record linkage. *Annual Review of Sociology*, 44:19–37.

Rusanov, V. (2021). Internal migration and the diffusion of schooling in the us.

Russo, A. and Aspelund, K. (2024). Additionality and asymmetric information in environmental markets: Evidence from conservation auctions. *Working Paper*.

Rust, J. et al. (1987). Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica*, 55(5):999–1033.

Saiz, A. (2010). The geographic determinants of housing supply. *The Quarterly Journal of Economics*, 125(3):1253–1296.

Salter, W. E. G. (1960). *Productivity and technical change*. Cambridge University Press.

Sandberg, L. G. (1969). American rings and english mules: the role of economic rationality. *The Quarterly Journal of Economics*, 83(1):25–43.

Sandhu, K. S. (1964). Emergency resettlement in malaya. *The Journal of Tropical Geography*, 18:180.

Sarvimäki, M., Uusitalo, R., and Jäntti, M. (2022). Habit formation and the misallocation of labor: evidence from forced migrations. *Journal of the European Economic Association*, 20(6):2497–2539.

Saxonhouse, G. and Wright, G. (1987). Stubborn mules and vertical integration: the disappearing constraint? *Economic History Review*, 40(1):87–94.

Schaal, E. (2017). Uncertainty and unemployment. *Econometrica*, 85(6):1675–1721.

Scharf, J. T. (1874). *The chronicles of Baltimore : being a complete history of "Baltimore town" and Baltimore city from the earliest period to the present time*. Turnbull Bros, Baltimore.

Schenck, J. and Rann, W. S. (1887). *History of Warren County, Pennsylvania: With Illustrations and Biographical Sketches of Some of Its Prominent Men and Pioneers*. D. Mason & Company.

Schumpeter, J. A. (1942). *Capitalism, Socialism and Democracy*. Harper & Brothers.

Scientific American (1870). The allen engine. 23(24):374.

Scientific American (1871). Statistics of steam boiler explosions for the year 1870. 24(10):153.

Scientific American (1881). Whose boilers explode. 44(12):176.

Sedláček, P. and Sterk, V. (2017). The growth potential of startups over the business cycle. *American Economic Review*, 107(10):3182–3210.

Semple, E. C. (1903). *American history and its geographic conditions*. Boston; New York: Houghton, Mifflin.

Severnini, E. (2023). The power of hydroelectric dams: Historical evidence from the united states over the twentieth century. *The Economic Journal*, 133(649):420–459.

Sharrer, G. T. (1982). The merchant-millers: Baltimore's flour milling industry, 1783-1860. *Agricultural History*, 56(1):138–150.

Short, A. (1975). The communist insurrection in malaya, 1948-1960. *(No Title)*.

Silva, J. S. and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, 88(4):641–658.

Smil, V. (2014). The long slow rise of solar and wind. *Scientific American*, 310(1):52–57.

Smith, C. and Kulka, A. (2023). Agglomeration over the long run: Evidence from county seat wars.

Snow, D. (2004). Extraordinary efficiency growth in response to new technology entries: The carburetor's "last gasp"'. *Academy of Management Proceedings*, 2004(1):K1–K6.

Sokoloff, K. L. (1984). Was the transition from the artisanal shop to the nonmechanized factory associated with gains in efficiency?: Evidence from the us manufacturing censuses of 1820 and 1850. *Explorations in Economic History*, 21(4):351–382.

Solow, R. M. (1962). Technical progress, capital formation, and economic growth. *American Economic Review*, 52(2):76–86.

Stokey, N. L., Lucas, R. E., and Prescott, E. C. (1989). *Recursive Methods in Economic Dynamics*. Harvard University Press.

Strassmann, W. P. (1959). Creative destruction and partial obsolescence in american economic development. *The Journal of Economic History*, 19(3):335–349.

Survey of India Offices (P.Z.O.) (1944). Malaysia, malaya, yala, series: Hind 1076, 1944, 1:253 440.

Swain, G. F. (1888). Statistics of water power employed in manufacturing in the united states. *Publications of the American Statistical Association*, 1(1):5–36.

Syverson, C. (2004). Market structure and productivity: A concrete example. *Journal of political Economy*, 112(6):1181–1222.

Temin, P. (1966). Steam and waterpower in the early nineteenth century. *The Journal of Economic History*, 26(2):187–205.

Thompson, K. J. and Sigman, R. S. (1999). Statistical methods for developing ratio edit tolerances for economic data. *Journal of Official Statistics*, 15(4):517.

Tombe, T. and Zhu, X. (2019). Trade, migration, and productivity: A quantitative analysis of china. *American Economic Review*, 109(5):1843–1872.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Treasury Department (1838). *Steam Engines: Letter From the Secretary of the Treasury, Transmitting, in Obedience to a Resolution of the House, of the 29Th of June Last, Information in Relation to Steam Engines, &c. (The 'Woodbury Report')*. 25th Congress.

Tsou, Y.-H. (2007). The change of culture landscape of the chinese new village in malaysia—the study of jinjang new village.

United States Census Bureau (1850a). Manufacturers of the united states in 1850.

United States Census Bureau (1850b). Population of the united states in 1850.

United States Census Bureau (1860a). Instructions to u.s. marshals. instructions to assistants.

United States Census Bureau (1860b). Manufacturers of the united states in 1860.

United States Census Bureau (1860c). Population of the united states in 1860.

United States Census Bureau (1870a). *Ninth Census of the United States, 1870: Statistics of the wealth and industry of the United States*.

United States Census Bureau (1870b). Volume 1. the statistics of population of the united states.

United States Census Bureau (1870c). Volume 3. the statistics of wealth and industry of the united states.

United States Census Bureau (1880a). Report on the manufactures of the united states at the tenth census.

United States Census Bureau (1880b). Statistics of the population of the united states at the tenth census.

United States Census Bureau (1900). Volume vii. manufactures, part 1. united states by industries.

U.S. Office of Strategic Services (1942). *Roads and Railroads of Malaya*. OSS Repro. Section, Washington, provisional edition.

U.S. Office of Strategic Services (1944). *Malaya Industrial Facilities*. OSS Repro. Section, Washington, provisional edition.

Veblen, T. (1915). *Imperial Germany and the Industrial Revolution*. Imperial Germany and the Industrial Revolution. Macmillan.

Verhoogen, E. (2023). Firm-level upgrading in developing countries. *Journal of Economic Literature*, 61(4):1410–64.

Vlieland, C. (1931). *British Malaya, A Report on the 1931 Census*. Crown Agents for the Colonies.

Wade, R. (2004). *Governing the market: Economic theory and the role of government in East Asian industrialization*. Princeton University Press.

War Office (1943). *Malaya: Land Utilization Map*. GSGS, London.

Watertown Historical Society (2022). The yellow mill (same as the blanchard mill).

Webb, W. P. (1931). *The Great Plains*. University of Nebraska Press.

Weeden, W. B. (1890). *Economic and Social History of New England, 1620-1789*, volume 1. Houghton, Mifflin.

Whittaker, H. (2012). Forced villagization during the shifta conflict in kenya, ca. 1963–1968. *The International journal of African historical studies*, 45(3):343–364.

Wright, M. N. and Ziegler, A. (2015). Ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*.

Wrigley, E. A. (2010). *Energy and the English Industrial Revolution*. Cambridge University Press.

York Township Historical Society (2023). The early history of york township.