Spring 2022

Essays in the Industrial Organization of Regulatory Policy

Jonathan Thomas Hawkins-Pierot
Yale University Graduate School of Arts and Sciences, jonathan.hawkinspierot@gmail.com

Abstract

Essays in the Industrial Organization of Regulatory Policy

Jonathan T. Hawkins-Pierot
2022

Governments use a wide variety of policy instruments to achieve their goals, including price signals, constraints on firms' behavior, and direct action. The consequences of such policies depend on how they interact with the underlying economic system. Measuring the impact of an intervention can be difficult, especially when the policy was not designed or implemented with evaluation in mind. Moreover, it is often *ex ante* prediction of a policy's costs and benefits, rather than *ex post* evaluation, which is relevant for decision making. This dissertation examines issues related to both evaluating and predicting the effects of regulations in two important contexts: financial inclusion and industrial carbon emissions.

Mandates requiring banks to open a minimum share of their new branches in unbanked villages have been a pillar of the Indian government's rural financial inclusion strategy for decades. By explicitly linking branch licenses in banked municipalities to rural branch expansion, these mandates increase the costs of entry in banked markets and may reduce access there. In the first two chapters of this dissertation, I study the impact of a 25% unbanked share mandate implemented in July 2011 on the size, geographic distribution, and profitability of the national branch network.

In the first chapter, I describe the context of the reform and use novel, comprehensive records of branch licenses, to document the scope of the post-reform rural branch expansion. Over 11,000 unbanked villages, home to more than 40 million people, were entered in the five years post-reform. These villages are substantially smaller, poorer, and more remote than those entered prior to the reform. In the second chapter, I use an economic model of branch entry to estimate banks' profits, compute their regulatory compliance costs, and simulate equilibrium entry and profits under counterfactual policies. Compared to a free-entry counterfactual, the mandate reduces total profits from new branches by about 26% and shifts entry from banked to unbanked markets roughly one-for-one, with disproportionate losses in smaller banked markets. These costs increase rapidly in the mandatory unbanked share. Allowing banks to comply by trading permits in a competitive market modestly increases profits but does not result in net new entry.

In the final chapter, co-authored with Katherine Wagner, we study the implications of low energy prices today for industrial energy efficiency and climate policy in the future. If adjustment costs mediate manufacturing plants' responses to increases in energy prices, incumbents may be limited in their ability to re-optimize energy-inefficient production technologies chosen based on past market incentives. Using U.S. Census data and quasi-experimental variation in state energy prices, we first show that the initial electricity prices that manufacturing plants pay in their first year of operations are important determinants of long-run energy intensity. Plants that open when the prices of electricity and fossil fuel inputs into electricity are low consume more energy throughout their lifetime, regardless of current electricity prices. We then measure the relative contributions of initial productivity and capital adjustment frictions to creating this "technology lock-in" by estimating a model of plant input choices. We find that lock-in can be largely explained by persistent differences in the relative productivity of energy inputs chosen at entry. We discuss how these long-run effects of low entry-year energy prices increase the emissions costs of delayed action on carbon policy.

Cost-benefit analysis of existing and proposed regulations is central to the policymaking process. This dissertation aims to provide useful insights on how recent advances in industrial organization can inform these analyses.

Essays in the Industrial Organization of Regulatory Policy

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Jonathan T. Hawkins-Pierot

Dissertation Director: Nicholas Ryan

May 2022

Contents

List of Figures

List of Tables

Acknowledgements

It was a great privilege to work with the members of my dissertation committee, Nick Ryan, Steve Berry, and Rohini Pande. Their insight and encouragement kept me on track, even as the world was shutting down. I am deeply grateful for all they have taught me.

The project which makes up the first two chapters of this dissertation would not have been possible without Rakesh Mohan, whose deep knowledge of the Indian financial system and the Reserve Bank of India was an invaluable resource at all stages of the project. I also appreciate the insights of Michi Igami, Mark Rosenzweig, and Katja Seim, whose valuable comments have shaped the project at key junctures.

I am indebted to the Yale Economic Growth Center and the SYLFF Association for their financial support. Thanks to their support, I had the good fortune to work with Alex Wilkinson, who provided excellent research assistance tracking down bank branches across India. I am also grateful to Andrew Tang and Benji Reade Malagueno for their excellent research assistance. I am grateful to Sneharthi Gayen and Dan Keniston, who provided critical data. At the U.S. Census Bureau, I appreciate Stephanie Bailey's guidance and most of all her good humor while resolving my problems with the ambush alarm.

The research community at Yale, and the Economics Department in particular, have been indispensable sources of fellowship, inspiration, and commiseration. First among equals is my friend and co-author Katherine Wagner, who has been there for all the ups and downs of the dissertation, providing insightful comments, steadfast support, and limitless quantities of Jarlsberg cheese. I am also grateful to Zach Barnett-Howell, Paula Calvo, Oscar Soler Sanchez, Suk Joon Son, Allen Vong, Stephanie Weber, and Lucas Zavala for helpful discussions and support, as well as participants in many breakfasts, lunches, and seminars over the years.

I am thankful for all the people outside of Yale who made it possible for me to reach this point. I never would have made it this far without Charlene Brown and the Wells Scholars community, Gerhard Glomm, who first inspired me to study economics, or all the friends I've made along the way. I am endlessly appreciative of my parents, who instilled a love of learning which has sustained my academic career, and my brother, whose intellectual curiosity never ceases to bring about interdisciplinary insights.

Most of all, I am grateful to Agathe, my partner in all things. Your faith in me, your humor and your grace - despite endless monologues about my research - are priceless gifts.

Chapter 1

Constraining Entry to Improve Access: Unbanked Share Mandates in India

Abstract

Providing widespread access to the formal financial system is a priority for many governments and multilateral organizations, yet nearly one third of adults globally do not have a bank account. Mandates requiring banks to open a minimum share of their new branches in unbanked villages have been a pillar of the Indian government's rural financial inclusion strategy for decades. By increasing the cost of entry in banked municipalities, these mandates may reduce access elsewhere. In this chapter, I study the impact of a 25% unbanked share mandate implemented in July 2011. Using novel, comprehensive records of branch licenses, I show that the mandate binds and dramatically increases entry into unbanked villages. In the five years post-reform, banks entered roughly 10,000 additional unbanked villages relative to pre-reform trends. These newly banked villages are home to over 40 million people, and are substantially smaller and poorer than those entered prior to the reform. However, 90% of these villages were within 7km of a market with a preexisting branch, which may attenuate the increases in access. Finally, time-series evidence of reductions in banked entry is mixed, which motivates the model-based analysis of the second chapter.

1.1 Introduction

Governments often rely on private firms to provide access to essential services such as electricity, telecommunications, housing, healthcare, and financial services. However, private firms may not be able to profitably serve all households. This is often the case for rural areas, where low population density makes it difficult to cover the fixed costs of developing the necessary infrastructure. One common solution is to require firms to serve these populations in return for privileged access to more profitable markets. These mandates can be a powerful tool for promoting equity and economic development. However, by forcing firms to cross-subsidize unprofitable markets, mandates may harm households in profitable ones. They may also have important impacts on competition between firms, especially when there is substantial heterogeneity in the costs of complying with the mandate. This opacity can make it hard to assess who bears the costs of a mandate's benefits.

Providing widespread access to the formal financial system is a priority for many governments and multilateral organizations, yet nearly one third of adults globally do not have a bank account (Demirgüç-Kunt et al., 2020). In 2011, over 600 million Indians lived in municipalities without formal bank branches. Brick-and-mortar branches are critical access points to the financial system, particularly for the rural poor, who are less able to travel and less likely to have access to or trust in mobile banking technology (Dupas et al., 2016). Policymakers in high-income countries such as the United States have also raised concerns about the availability of bank branches, particularly in low-income and minority neighborhoods (Morgan et al., 2016).

Since the 1960s, the Indian government has promoted rural branch expansion by requiring banks to open a minimum share of their new branches in unbanked rural villages. That is, in order to obtain licenses for new branches in profitable banked municipalities, banks must open a corresponding number of branches in rural villages without preexisting bank branches.

In this chapter, I document how a 2011 mandate requiring commercial banks in India to open 25% of their branches in unbanked villages changed the size and geographic distribution of the national bank branch network. I use the complete municipality-level bank branch network in each year from 2006-2019, based on administrative records from the Reserve Bank of India's Master Office File, a database of all bank branch licenses. The public version of the Master Office File contains data for existing branches but not for branches which have closed or been acquired in a merger. Entry dates for 46,426 branches were affected by a merger between 2006 and 2019, rendering the public data unsuitable for calculating the number or parent bank of branches operating in a market in a given year. The RBI provided access to the confidential merger and closure files for 2006-2019, which enabled me to reconstruct the complete branch network over time. I link the branch network to the Socioeconomic High-resolution Rural-Urban Geographic platform, which provides a consistent crosswalk for a variety of municipality-level datasets, including the Population and Economic Censuses (Asher et al., 2019).

I begin by describing the Indian retail banking industry and the context in which the reform was implemented. I document differences between the branch networks and business activities of public and private sector banks. At the time of the reform, private sector banks are remarkably concentrated in large urban and metropolitan markets. This suggests that regulations requiring banks to open a share of their new branches in unbanked rural villages will be relatively more constraining for the private sector banks, and therefore tend to shield incumbent public sector branches from competition.

I then turn my attention to the 2011 reform. I show that the mandate was binding, well-enforced, and led to a large rural branch expansion. After 2011, the rate of entry into unbanked villages jumped from virtually zero to around 1,200 villages per year, and continued to increase in line with accelerating entry into banked markets. In the five years following the reform, commercial banks opened 11,542 branches in previously unbanked markets, home to over 42 million people. A comparison with pre-reform trends suggests that very few of these branches would have opened without the mandate.

The branch-level data provides a unique opportunity to explore the spatial structure of branch networks. Indian banks typically have a limited geographic scope, meaning that they tend to concentrate their operations in a few states or districts. Even national banks, like the State Bank of India, tend to open new branches in close proximity to their existing networks. This may undermine the reach of the post-reform rural branch expansion: 90% of newly banked villages are within 7km of a pre-existing branch. However, due to the scale of the rural branch expansion, about 2 million people saw their distance to the nearest branch reduced by at least 10km.

The sudden and dramatic expansion of the rural branch network was costly. Banks pay a direct cost of opening and operating unprofitable branches in rural villages. The mandate is a binding constraint, so losses in unbanked villages represent a shadow cost on entry into banked markets. This leads to an additional loss of profits in banked markets where branches would not be sufficiently profitable to cover the compliance costs. These indirect costs depend on the marginal compliance cost, relative to profits in banked markets. As banks enter more banked markets, they must enter increasingly unprofitable unbanked villages, driving up compliance costs. If banks rapidly run out of relatively cheap unbanked villages, there can be large reductions in banked entry – and large indirect losses – even when the direct cost of entering unbanked villages is modest.[1]

Unlike the change in unbanked entry following the mandate, the results for entry

[1] This is clearest in the extreme case: a 100% unbanked share would lead to no direct losses at all, as banks would have no incentive to enter unprofitable unbanked villages, and massive indirect losses, since entry in banked markets would be prohibited.

in banked markets are very sensitive to the number of periods used to compute the pre-reform trend. Compared to the four years before the reform, there appears to be a sharp slowdown in banked entry, leading to a cumulative deficit of nearly 14,000 branches by the end of 2016. However, post-reform entry is right on trend if we take 2001 as our baseline year, implying that the effect was minimal. Without a clear control group, is difficult to interpret these trend-breaks causally. It is possible that entry in banked markets would have slowed even without the mandate.

The key question is therefore: how (un)profitable are these new branches? In the remainder of the chapter, I present evidence that branches opened in unbanked villages are substantially less profitable than those opened in banked markets or in unbanked villages prior to the reform. For example, the median village banked after the reform has a 40% higher poverty rate than rural villages first entered in the five years prior. Measures of branch activity also fall following the reform. In 2011, a bank branch in a one-branch rural village typically collects about 100 million rupees in deposits and disburses about 60 million rupees worth of loans. Despite rapid economic growth, credit and deposits per branch in the average one-branch village declines following the reform, and does not fully reach pre-reform levels until after Demonetization five years later. In the second chapter, I use an economic model to recover branch profitability from banks' observed entry decisions. This provides an alternative, highly granular measure of the impact of the mandate.

1.1.1 Related Literature

There is a robust literature on the role of financial development, and bank branching in particular, in economic growth (King and Levine, 1993; Rajan and Zingales, 1998; Burgess and Pande, 2005; Bruhn and Love, 2014; Young, 2019; Nguyen, 2019; Cramer, 2021). Compared to the literature measuring the benefits of bank branches for the communities which have them, there is little evidence on the costs associated with the policies which deliver them. By considering the effect of the mandate on the branch network as a whole, as well as the tradeoffs involved in different versions of the policy, this paper provides a broader basis on which to assess the welfare consequences of unbanked share mandates. Furthermore, much of the existing literature considers the effect of rural branch expansion policies at the state- or district-level, whereas my paper offers a municipality-level measure of exposure to the policy treatment and lays the groundwork for a more granular analysis of the effects of such policies. A few recent papers, notably Garg and Gupta (2020) and Garg and Gupta (2021), have used the public version of the RBI's branch-level Master Office File to conduct municipality-level analysis, but are limited by the missing records of closed and merged branches which I was able to acquire through collaboration the RBI. This is in the same vein as a growing body of work using high-resolution data to study economic development (Asher and Novosad, 2020).

This paper also contributes to the literature on the use of mandates to promote access to essential services. This paper presents an analysis of a distinctive approach to the ubiquitous problem of ensuring widespread access to essential services. Much of this literature focuses on regulated monopolies subject to Universal Service Obligations (USO) (Geddes, 2005; Cremer et al., 2008). Under a USO, a firm is often granted a monopoly in return for providing universal access, with the expectation that monopoly rents in profitable markets will be sufficient to cross-subsidize operations in unprofitable ones. The regulated firm may face unconstrained competitors which can "skim the cream" by serving only the most profitable customers. This is particularly common in settings where public and private sector firms interact, like the Indian banking system, and has been studied in telecommunications (Laffont and Tirole, 1990), healthcare (Barros and Siciliani, 2011), and education (Altonji et al., 2015). If universal access is funded through cross-subsidization, cream skimming may undermine the regulated firm's ability to provide high-quality universal service.

Policymakers must weigh this possibility against the welfare gains from increased competition or innovation in the profitable markets. An alternative approach is to simply subsidize rural access, as for U.S. hospitals (Murphy et al., 2018) and broadband (GAO, 2021). The mandate studied in this paper is similar in spirit to a Universal Service Obligation, in that it forces banks to cross-subsidize branches rural villages with profits from urban markets. However, the obligation is proportional, rather than universal, and binds both public and private sector firms. Because banks incur obligations to serve rural villages in proportion to their entry in banked markets, this creates an incentive to underinvest in banked markets not present for universal service obligations. On the other hand, it avoids both the efficiency losses of a regulated monopolist and the large fiscal outlays necessary to subsidize universal access.

The remainder of the paper is structured as follows. Sections 2 describes the Indian banking sector and the regulatory context. Section 3 presents the data. Section 4 documents the effects of the policy in the reduced form. Section 4 introduces the model of bank branching decisions under free entry and the unbanked share constraint, and Section 5 concludes.

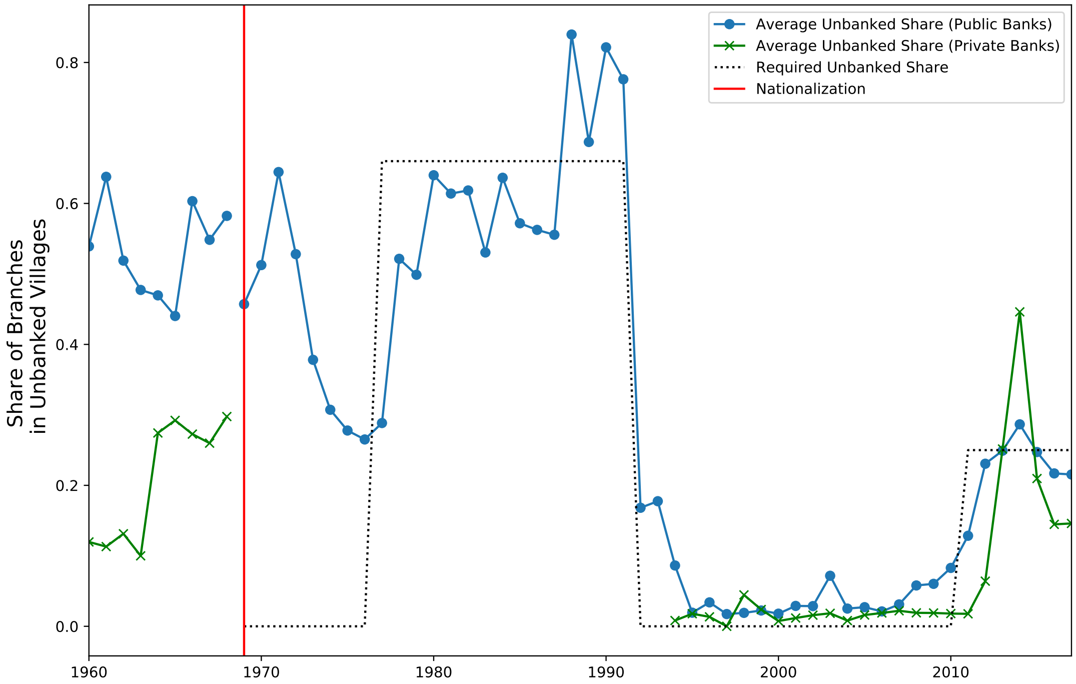
1.2 Institutional Context

I begin with a brief description of the Indian retail banking sector prior to the 2011 reform, discuss the context in which the reform was implemented, and describe the policy in detail.

1.2.1 Bank Branch Authorization Policy

Rural financial inclusion has been a policy priority for the Indian government since before Independence in 1947. In 1969, the 14 largest commercial banks were nationalized with the goal of improving the cost and availability of credit in rural areas. While

Figure 1.1: Unbanked Share Over Time



Notes: This figure reports the average share of new branches opened in unbanked villages in each year, separately for public and private sector banks. The dotted line plots the mandated share of branches in unbanked villages following nationalization. *Source: RBI Master Office File* a few private banks continued to operate, they remained small and highly specialized and did not play a major role in the banking system for the next two decades. At the time of nationalization, rural lending was dominated by informal moneylenders, who were widely regarded as inadequate for financing the Green Revolution at best and usurious at worst (Mohan, 2006). Following nationalization, interest rates were strictly regulated and banks were required to direct a specified proportion of their total lending to priority sectors, such as agriculture. In 1977, the government intensified these efforts by requiring banks to allocate two thirds of their new branches to rural areas. Banks were permitted to open one branch in a metropolitan market and one in any non-metropolitan banked market for every four branches opened in unbanked rural villages (Panagariya, 2006). For the purposes of unbanked share mandates, the RBI defines a market as a Census village or town, except for a few large metropolitan areas, which are disaggregated by the Census but not by the RBI. Throughout the paper, I use "market" and village/town interchangeably.

This 66.7% unbanked share mandate led to a massive expansion into unbanked villages. However, the strict regulations of the "Social Banking Period" undermined the health of the banking system and contributed to an economy-wide liberalization in the early 1990s. From 1991 to 2005, branch licenses were issued based on commercial viability (Narasimham, 1992). There was essentially no entry into unbanked villages during this period, although banks were prohibited from closing rural branches.

Beginning in 2005, there were a series of reforms which streamlined the licensing process and established incentives for banks to open branches in districts with below-average branches per capita. Overall bank branch entry accelerated during this period – which also saw strong economic growth – but the impact on branching in unbanked villages was limited for public sector banks and negligible for their private sector counterparts. This led to series of updates to the rules governing bank branch authorization designed to encourage rural branch expansion, culminating in a restored version of the unbanked share constraint enforced in the 1970s and 80s.

In July 2011, the RBI announced that banks were required to open 25% of their new branches in unbanked rural villages. Banks were allowed to smooth compliance by carrying surpluses and shortfalls in unbanked entry for up to two years. Branches must satisfy basic requirements regarding opening hours and other activities to count. In 2017, the definition of a branch was expanded to count satellite counters, ATMs, and traveling banking correspondents as partial branches. Since data on these activities is not available, I terminate my analysis at the end of 2016. How this increased flexibility impacts the costs and benefits of the mandate is an interesting avenue for future work.

Figure 1.1 plots the evolution of these policies, and the observed share of new entry in unbanked villages, over time. On average, the mandates were binding. More importantly, there was virtually no entry in unbanked villages during the 20 year period where this was not required.

1.2.2 Public and Private Sector Banks

Another key component of the post-1991 liberalization program was that several new private sector banks were allowed to enter. These banks grew rapidly: By 2006, private sector banks represented 12% of total branches. Five years later, the number of private bank branches had doubled and their share increased to nearly 16%. Over the same period, private sector banks' total nominal deposits more than tripled, bringing their market share from 20% to 24%.

This rapid growth, concentrated in the most profitable urban areas, led to concerns about the effect of increased competition on public sector banks' ability to both serve policy goals and remain profitable. (Thorat, 2009). Table 1.1 shows the extent of private sector banks' concentration in urban and metropolitan areas. In 2011, 32% of private sector branches were in metropolitan markets, with population over 1 million, compared to about 20% and 24% for the SBI and Nationalized banks, respectively[2]. On the other hand, SBI and Nationalized banks had 32% and 30% of their branches in rural villages with population under 10,000, whereas less than 10% of private sector banks were in rural villages. The gap is even larger in terms of banks' customer bases. Only 8% and 5% of private sector banks deposit and credit accounts, respectively, come from rural branches. Private sector banks have an astounding 80% of credit accounts in metropolitan markets. For public sector banks, about 27% of deposit accounts and 39% of credit accounts are in rural villages.

Private sector banks' relative lack of experience has significant consequences for the effect of the 2011 mandate on competition. In section 1.4.4, I discuss the role of geography in banks' entry decisions. Pre-existing branches in rural and semi-urban areas provide beachheads from which banks can cheaply expand into nearby unbanked villages. Without these branches, private sector banks face a higher cost of complying with the mandate and suffer relative to their public sector counterparts.

[2] The State Bank of India, the largest public sector bank, is treated separately from Nationalized banks, which were founded as private sector banks and taken over by the government in 1969.

Table 1.1: Banking Activities by Banking Group (March 2011)

	State Bank of India	Nationalized	Private Sector	Total
Total Branches	**18,308**	**43,877**	**11,682**	**73,867**
Rural	5,849	13,113	1,129	20,091
Semi-Urban	5,687	10,681	3,738	20,106
Urban	3,187	9,458	2,994	15,639
Metropolitan	3,585	10,625	3,821	18,031
Deposit Accts ('000s)	**204,814**	**405,095**	**87,823**	**697,732**
Rural	52,475	112,330	7,037	171,812
Semi-Urban	72,912	96,895	18,509	188,316
Urban	44,567	91,877	25,580	162,024
Metropolitan	34,869	103,992	36,697	175,558
Credit Accts ('000s)	**23,216**	**38,652**	**33,300**	**95,168**
Rural	7,623	16,427	1,606	25,656
Semi-Urban	9,863	11,162	3,055	24,080
Urban	4,014	7,020	4,867	15,901
Metropolitan	1,716	4,043	23,772	29,531

Source: RBI Basic Statistical Returns, Vol. 40 (March 2011)

1.3 Data

I use public and private data on bank branching and other activities from the Reserve Bank of India (RBI). For village and town level information, I use the Population and Economic Censuses linked with consistent identifiers by the Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG). I further combine this with geospatial data from the NASA Socioeconomic Data and Applications Center (SEDAC).

Data on the bank branch network comes from the RBI Master Office File (MOF). The MOF is a continuously updated directory of bank branches, based on administrative documents banks must file with the Bank Branch Statistics Division of the RBI when they open a branch or change its status (e.g. closing, merging with another branch.) The MOF covers all bank branches, and includes parent bank, opening date, exit date (if applicable), and street address. The publicly available version of the MOF does not include closed branches, and branches acquired through mergers have opening dates equal to the date of the merger, not the date when the original branch opened.

The RBI provided records containing these missing or incorrect dates which I merged with the public data in order to construct a complete dataset of the bank branch network from 2006-2019.[3] The branch network over time cannot be reliably reconstructed without these additional files. When calculating annual entry at the municipality level, about one-third of the entry dates in the public 2016 MOF are incorrect, often off by decades. Closures are rare, particularly in rural villages where they are essentially prohibited. Out of the 28,169 branches operating in rural villages in 2006, only 106 had closed by 2019. However, 46,426 branches across all population groups were affected by a merger during this period. The median acquired branch had been operating for slightly over 23 years. Using only the public data creates many problems for understanding entry into unbanked rural villages over time. Years with major mergers will spuriously appear to be years with extensive entry in unbanked villages. If the first branch to open in a market was later acquired, it will appear as though that market was unbanked for years. A second bank which entered in the interim may appear to have "banked" the market instead. To my knowledge, this is the first paper to use the complete branch-level MOF.

Town and village-level data are provided by the SHRUG database, which links a variety of Census and remote sensing data across years (Asher et al., 2021). This includes demographics, amenities, and information about local economies, such as the share of workers in agriculture and proxies for wealth. The spatial structure of the branch network is an important determinant of entry costs. It is also important for interpreting the impact of rural branch expansion because it allows us to measure changes in proximity to bank branches outside of newly banked villages. The SHRUG

[3] In fact, the dataset goes back to the late 19th century. I validate the branch-level panel constructed from the MOF to various aggregates published annually by the RBI starting in 2006. Pre-2006 branch data is based on entry dates for branches operating in January 2006. Without merged and closed branches, becomes is less reliable as it goes further back in time.

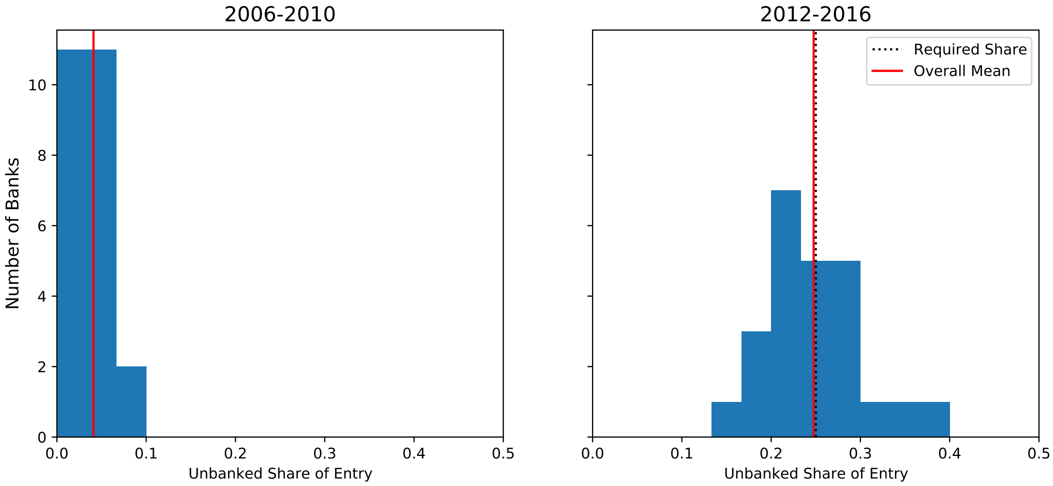
does not include geographic information systems (GIS) data, so I have merged this with the NASA-SEDAC geospatial data based on 2001 Census maps (Meiyappan et al., 2018). For branches which could not be matched with the NASA-SEDAC data, I obtained GPS coordinates from branch street addresses using the Google Maps API.

Credit and deposits data are from the Quarterly Statistics on Deposits and Credit of Scheduled Commercial Banks, Basic Statistical Returns - 7. These are branch-level reports on aggregate deposits and bank credit filed by banks each quarter and verified by the RBI using other data sources. Due to data confidentiality restrictions, only credit and deposit data from villages and towns with three or more branches are publicly available. The RBI provided special tabulations of quarterly district-wise aggregate deposits and credit for markets with one and two branches, respectively. For further information on banks activities, I use bank-level data derived from banks' annual accounts available from the Database on Indian Economy. This includes bank-level deposits and income and expenses used to compute average return on deposits and deposit rate paid.

1.4 2011 Reform and Rural Branch Expansion

In this section, I show that the 2011 mandate requiring banks to open 25% of new branches in unbanked villages was binding and that the rate of entry in unbanked villages increased substantially. I characterize the types of unbanked villages most likely to benefit, and document aggregate changes in the types of market entered. I then discuss the implications for entry into banked markets.

Figure 1.2: Unbanked Shares of Major Banks, Before and After Reform



Notes: This figure presents the histogram of bank-level unbanked shares during the five years before and after the 2011 mandate. The red line indicates the overall mean for the period, and the black dotted line indicates the mandatory share.

1.4.1 An Unbanked Share Constraint

I first show that the 25% constraint is both enforced and binding, in the sense that banks meet but do not choose to exceed their obligations.[4] Figure 1.1 shows that on average both public and private banks increase their unbanked share in compliance with the constraint. This is consistent with banks' taking advantage of permission to smooth compliance over time. In the first few years after the policy was implemented, private sector banks in particular took advantage of the opportunity to front-load their entry into banked markets. Private sector banks are slow to comply, with very small unbanked shares in the first two years, but eventually make up this deficit in 2014. Public sector banks were more active in unbanked rural villages prior to the reform, but still saw a large increase after the reform.

The aggregate changes shown in Figure 1.1 are not driven by a subset of banks. The same change occurs for individual banks. Figure 1.2 presents the histogram of

[4] For example, there was a contemporaneous regulation requiring banks to open at least 50% of their branches in markets with population under 100,000. Both before and after the reform, almost all banks exceeded this level and I therefore treat it as non-binding in the analysis.

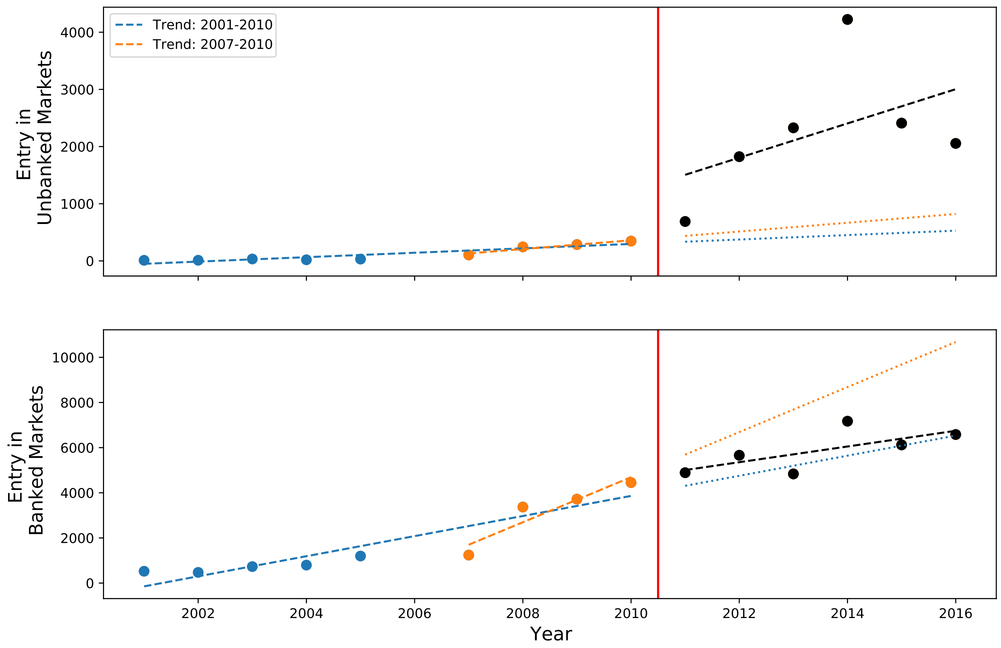
22 major banks' unbanked shares in the five years before and after the reform. The distribution moves from near-zero to being clustered around 25%. On average, the unbanked share during the five years after the reform was just a few tenths of a percent short of the regulatory minimum.[5] The very low share of unbanked entry between 1992 and 2011, the fact that private sector banks' postponed entry into unbanked rural markets until absolutely necessary, and the almost exact average compliance are strong evidence that the 25% unbanked share constraint drove the post-2011 wave of rural branch expansion.

1.4.2 The 2011 Reform Increased Rural Branch Expansion

Banks have two margins through which to increase their unbanked share. They can either increase their entry into unbanked rural villages or decrease entry into banked markets. The efficacy of an unbanked share mandate depends on the extent to which the former dominates the latter. Between 2011 and 2016, the State Bank of India, nationalized banks, and private sector banks opened 11,542 branches in previously unbanked villages. In contrast, 503 unbanked villages were entered in the five years before the reform. In terms of population, that is an increase of 42.5 million and 2.3 million people living in newly banked markets, respectively. The top panel of Figure 1.3 plots annual entry into unbanked villages from 1995 through 2016. There is some additional unbanked entry starting after 2006, which is consistent with the rapid expansion of bank branch networks overall. Although the coefficients are imprecisely measured, the line of best fit indicates that roughly 85% of the new entrants were due to the reform. Figure 1.4 maps these changes for all of India, and shows visually the dramatic increase in entry into unbanked villages across all regions. There is a large literature documenting the importance of distance to nearest branch for financial

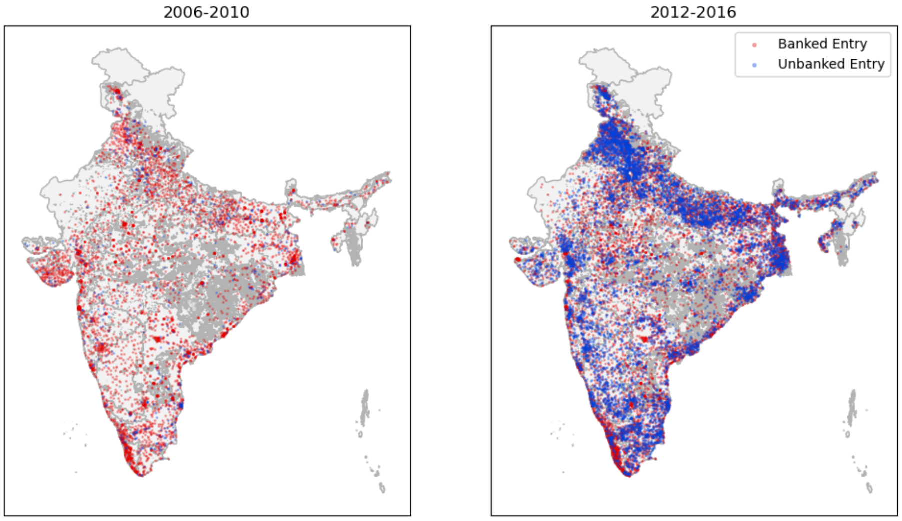
[5] There are several reasons that banks may persistently fall slightly short of 25%, including deferring compliance into the future, regulatory forbearance, and incentives for entering in larger markets of underbanked districts such as the NE Territories.

Figure 1.3: Entry into Banked and Unbanked Markets



Notes: This figure plots the number of branches opened in banked and unbanked markets. The blue and orange lines plot the linear trend from 2001-2010 and 2007-2010, respectively. The black line plots the post-reform linear trend. Coefficients are reported in Table 3. The post-reform increase in unbanked entry is similar on the two time-scales. Post-reform banked entry is either significantly slower or unchanged relative to short- and long-run trends, respectively.

Figure 1.4: Entry into Banked and Unbanked Markets Before and after reform



Notes: This figure maps entry in banked and unbanked markets before and after the reform. Red dots mark banked markets with additional entry during the period. Blue dots market unbanked villages entered for the first time.

inclusion. Over 100 million people saw their distance to the nearest branch reduced by at least 2km, an order of magnitude larger than the population with the same reduction in the five years prior to the reform.

A branch in an unbanked rural village typically collects 60-120 million rupees worth of deposits and disburses 40-90 million rupees of credit. Given that this includes one-branch markets banked prior to the reform, it is an upper bound on the credit and deposits of branches opened in unbanked villages due to the policy. On the other hand, these data are snapshots of branches' accounts at the end of each quarter and may not capture ephemeral deposits, such as government transfers.

It is also worth observing that there is a large spike in average deposits in the fourth quarter of 2016, which coincides with Demonetization, which required people to deposit any high-denomination banknotes in bank accounts. Post-reform entry in unbanked rural villages may have helped mitigate the local economic disruptions which followed. (**?**)

The 2011 reform also led to more entry into smaller, poorer, and more remote communities. Figure 1.5 plots the median population and literacy rate of markets entered over time.There is a general trend towards entering smaller and poorer markets, which is consistent with economic growth. However, there is visibly a sharp acceleration in these trends after the reform. This is the result of two important changes in entry patterns: banks enter more unbanked markets, which are smaller and poorer than the typical banked markets entered, and the unbanked markets entered after the reform are smaller and poorer than those entered beforehand.

The compositional effect of increased entry into unbanked villages is visible in the solid blue line, which plots the median across all markets. Before the reform, the median of all markets entered is very close to the median for banked markets entered because almost entry was in markets with preexisting branches. After the reform, unbanked markets represent a much larger share of total entry which pulls down the median. The second effect - the movement of banks into smaller and poorer unbanked villages - is more difficult to see from the aggregates but equally important. Table 1.2 presents the median characteristics of markets entered in the five years before and after the reform. This shows that the changes shown in Figure 1.5 are replicated across a wide range of characteristics. The median unbanked village entered after 2011 was about 30% smaller than the median of those entered before. It was also about 4p.p. less literate and had 5p.p. more people living in poverty than those entered before. p-values from permutation tests show that these differences are statistically significant. That is, the post-2011 rural branch expansion was deep as well as broad, and reached rural communities that almost surely would not have had formal branch branches without the reform.

1.4.3 Proximity to Existing Branch Network

For branch profitability, geography is as important as market characteristics such as population and wealth. It is more profitable for a bank to operate branches which are close to its existing network because this reduces logistics costs, allows for sharing of managerial overhead, and facilitates the acquisition of private information. Geography may be particularly important for unbanked entry if rural villages are less easily differentiated than urban markets.

Figure 1.7 shows the importance of proximity: Following the reform, 88% of entry by public sector banks in unbanked villages was within 20km of their 2006 branch network. Figure 1.7 also illustrates the differences between public and private sector banks.

This tendency to comply with the mandate by opening new branches in unbanked villages close to existing branches may have limited the reach of the program. Figure 1.8 presents the increase in access by plotting the total population of villages experiencing at least a Xkm decrease in the distance to the nearest branch. Nearly

200 million people live in villages which became somewhat closer to a formal bank branch, but only about 2 million became at least 10km closer. While substantial improvement in financial access for 2 million people is a success for financial inclusion, the accounting for proximity to the existing network suggests that the benefits of the reform may be somewhat more modest than the headline rural branch expansion would suggest.

1.4.4 Entry in Banked Markets

The rural branch expansion documented above was the result of a constraint on banks' branching decisions. The price of this increased financial inclusion was paid through two mechanisms: Lost profits and lost branches in already banked markets. Because bank branches are long-term investments, the profits lost from foregone opportunities in urban markets are unlikely to show up in the accounting profits reported in banks' annual reports. Urban branching is directly constrained, and is the most promising place to look for costs in the data. Furthermore, if redistribution of additional branches from urban markets to rural areas decreases financial inclusion or development in rapidly growing cities, this may be a more important consideration for public policy than reduced profitability of the banking sector.

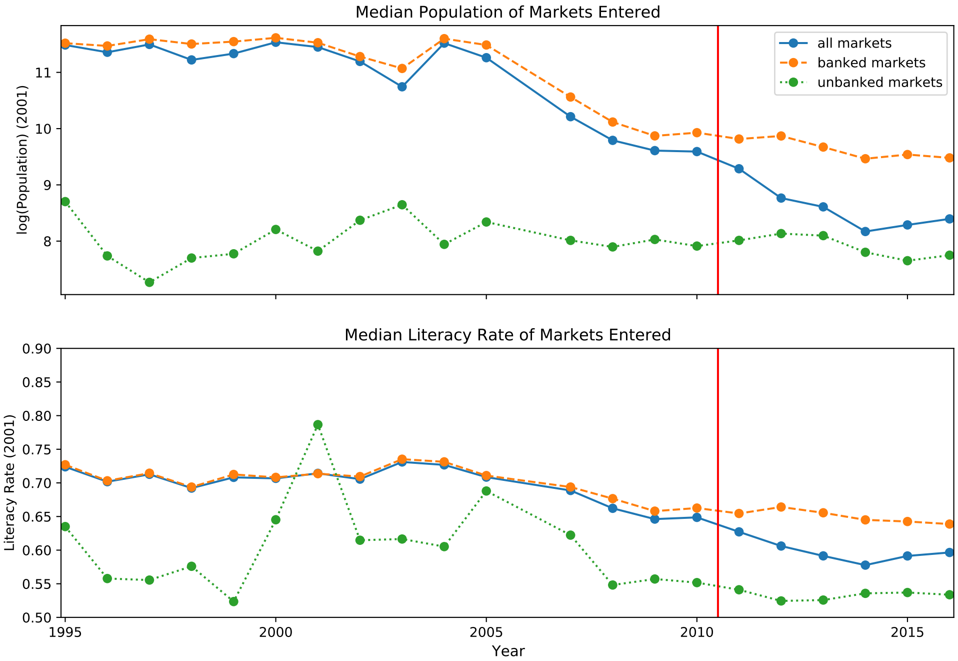
The bottom panel of Figure 1.3 plots the number of branches opened in banked markets in each year, along with fitted trends before and after the reform for two different time horizons. Again, the primary coefficient of interest is $post$, the level change in annual entry following the reform, with the slope, $postXyear$, of secondary interest. Unlike for unbanked markets, there is no clear break in the year of the reform. Depending on the number of years used to compute the pre-reform trend, we might conclude that the reform increased, decreased, or did not change the rate of entry into banked markets. Table 1.3 reports the associated coefficients and standard errors. Given the relatively small number of observations, the coefficients are necessarily imprecisely estimated. A short-run model, extrapolating banked entry from the four years following a streamlining of the licensing process in 2005-6, shows considerable divergence. The gap in the year following the reform is about 700 branches, growing to about 4,000 fewer new branches in banked markets annually by 2016, for a total cumulative deficit of slightly under 14,000. This would imply a high price for the rural branch expansion: Each new branch in an unbanked rural village requires about 1.4 fewer branches in banked markets.

An extrapolation based on a slightly longer time horizon, given in column (2), provides a more sanguine view - growth in annual banked entry continues right on trend. This pattern of uninterrupted urban branch growth is consistent with relatively small losses in unbanked markets leading to a negligible constraint on entry elsewhere. Under this hypothesis, entry into banked markets was largely unaffected by the reform and banks' compliance with the reform came almost entirely through increases in rural branch expansion. This would be a best case scenario, redirecting banks' economic rents in profitable urban markets to achieve a major policy goal without significant losses elsewhere.

All else equal, specification (1) might be preferred. Reforms to the bank branch licensing process in 2005-06 substantially reduced the regulatory burden of opening new branches in banked markets. Subject to the unbanked share mandate, the costs of opening a new branch in a banked market after the reform are therefore most similar to the post-2006 period.

However, the extent to which pre-mandate trends are a good counterfactual for unconstrained post-2011 entry is unclear. It is possible that bank branching would have slowed down even without the mandate, or accelerated even faster. Measures of the relative (un)profitability of branches in unbanked villages vs. banked markets provide an alternative source of information about the shadow cost the constraint imposed on banks and the likely effects on branching decisions.

Figure 1.5: Characteristics of Markets Entered Before and After the 2011 Reform



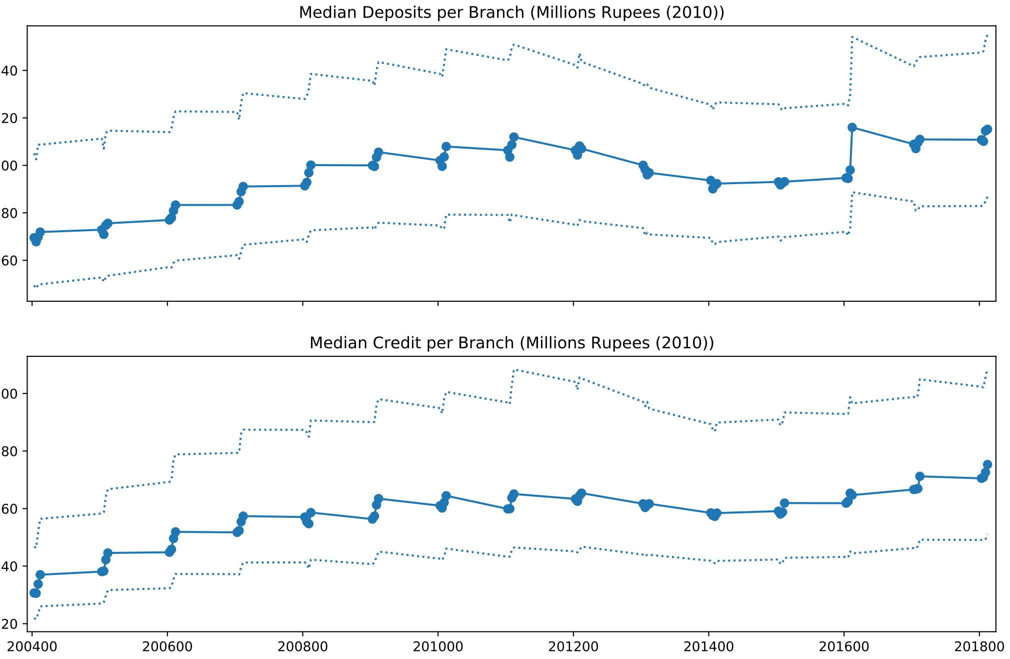
Notes: This figure plots the median population and literacy rate for markets entered before and after the 2011 reform, indicated by the solid red line. Over time, banks move into smaller and less literate banked markets. Post-2011, the median for all markets diverges from the median for banked markets due to increased entry in unbanked villages.

Table 1.2: Summary Statistics

	All Market Median			Unbanked Market Median		
Name	Pre-Reform	Post-Reform	p-value	Pre-Reform	Post-Reform	p-value
Population (2001)	25952.0	5403.0	0.0	3180.0	2713.0	0.0
km to nearest town	6.9	15.6	0.0	15.1	16.4	0.074
km to nearest city	19.1	25.2	0.0	22.6	24.6	0.062
% Literate (2001)	67.5	61.0	0.0	58.8	54.8	0.0
Nightlights (2006)	25.4	14.9	0.0	14.3	10.6	0.0
% Services (2005)	8.6	5.4	0.0	4.5	3.2	0.0
% Ag. Employment (2011)	0.0	10.0	0.0	12.1	23.0	0.0
Consumption (2011)	25.2	21.4	0.0	21.3	19.2	0.0
Poverty Rate (2011)	9.1	14.3	0.0	14.5	20.3	0.0
km to nearest branch	11.9	12.2	0.172	7.9	8.8	0.012
branches in 50km	18.0	15.0	0.0	29.0	21.0	0.0
Total Entry:	6071	22859		537	8815	
Unique Markets:	3204	14613		504	8305	

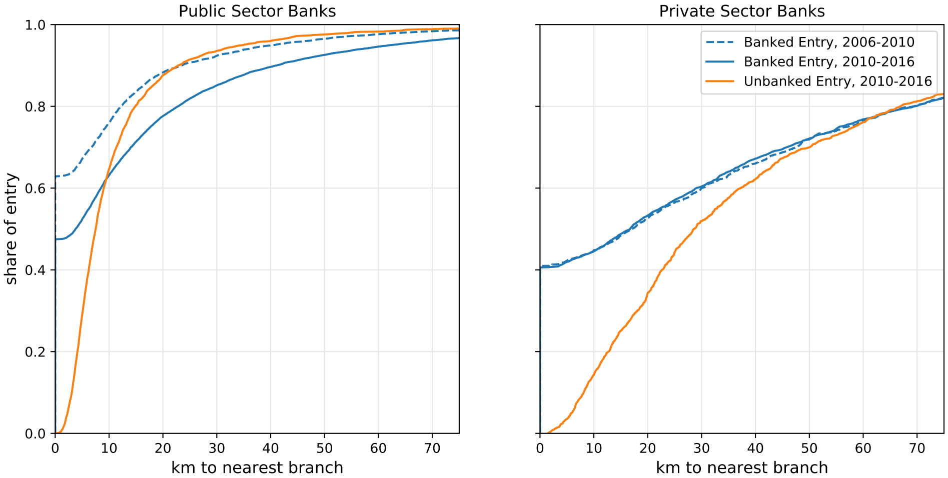
Notes: Columns are median characteristics for markets entered by commercial banks in the five years before and after the reform. A town and a city are defined as Census municipalities with at least 50,000 and 100,000 people, respectively. See appendix for details on variable construction. Note that some villages and towns are missing data and are not included in these statistics. p-values are computed using a permutation test with 10,000 trials.

Figure 1.6: Average Credit and Deposits per Branch in One-Branch Markets



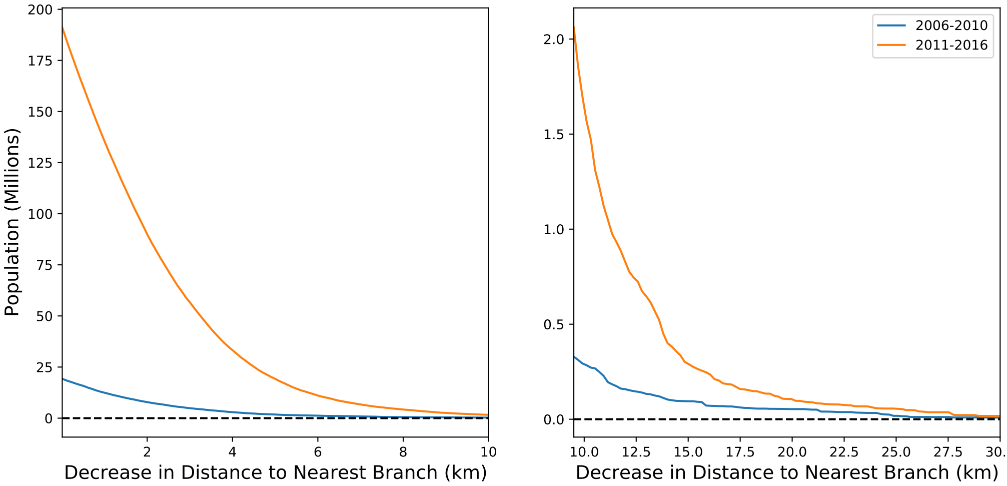
Notes: Average quarterly credit and deposits in markets with a single bank branch are available at the district level. The solid line plots the median across districts and the dotted lines mark the 25th and 75th percentiles. Units are million rupees, deflated to 2010 constant rupees using the World Bank's CPI series.

Figure 1.7: Branches Opened by Proximity to Nearest Branch



Notes: The figure plots the empirical CDF of the distance from new branches to the nearest branch belonging to the same bank, separately for public and private sector banks and for banked and unbanked entry. The curve for pre-reform unbanked entry is not shown, because there is insufficient entry during this time.

Figure 1.8: Financial Access: Proximity to Nearest Branch



Notes: This figure plots the total population of villages which experienced a decrease in the distance to the nearest branch of at least Xkm, in the periods before and after the reform.

Table 1.3: Time Trends in Banked and Unbanked Entry

	Unbanked Entry		Banked Entry	
	(1)	(2)	(1)	(2)
constant	130.2	-51.2	1698.4	-145.1
	(31.0)	(31.7)	(543.9)	(320.3)
year	76.7	38.6	997.9	445.3
	(12.6)	(5.8)	(218.3)	(68.0)
post	1068.8	1170.6	-682.4	700.1
	(644.4)	(583.7)	(552.5)	(586.7)
postXyear	222.9	261.0	-650.2	-97.6
	(243.1)	(219.7)	(232.1)	(98.5)
Obs. - Pred.	9756	10938	-13848	2737
	(4375)	(3756)	(3385)	(3652)
R^2	0.692	0.782	0.892	0.936
Years	10.0	15.0	10.0	15.0

Notes: Specifications (1) and (2) fit linear trends with pre-periods beginning in 2007 and 2001, respectively. 2006 is omitted from the analysis due to a change from the 1991 to 2001 Census location definitions, which creates spurious unbanked entry as locations are split. Change from trend reports the cumulative difference between observed and predicted entry based on pre-reform trends, for the years 2011-2016.

1.5 Conclusion

Mandates which force firms to provide socially beneficial but unprofitable services can be useful tools, particularly when government capacity to directly administer or adequately subsidize those services are limited. Mandates do not require a financing mechanism and can often be built onto existing licensing infrastructure. Unbanked share mandates, like those used in India, also encourage spatial and economic diversification in bank branch networks, which may have important benefits for the allocation of capital and the stability of the banking system.

This paper considers the effects of a 2011 reform requiring banks to open 25% of their new branches in unbanked villages on banks' branching decisions. I find that the mandate was binding and led to entry into roughly 10,000 additional unbanked villages relative to the pre-reform trend. These newly banked villages were smaller and poorer than typical villages entered prior to the reform, which further supports the conclusion that these branches were not profitable and would not have been opened without the mandate.

While rural branch expansion *per se* was a goal of this mandate, the welfare effects of these branches is an important question for future research. A natural question is whether the unbanked villages entered as a result of the mandate are those which would benefit most from access to a formal bank branch. On one hand, banks have an incentive to seek out those villages where demand for financial services is highest. On the other, banks may comply with the mandate by opening branches in the lowest cost locations, regardless of local demand. Quantifying consumer surplus is beyond the scope of this chapter, but there is some evidence that the allocation of branches resulting from the mandate may have more modest welfare benefits than the scale of the rural branch expansion would suggest. Banks tend to comply with the mandate by opening branches close to their existing network, and when other banks' networks are also included most newly banked villages were close to preexisting branches. Relatively few people saw large decreases in their distance to the nearest branch.

The primary costs of the mandate are lost profits and lost consumer surplus from deterred entry in banked markets. The trend-break evidence on the effect on banked entry is ambiguous. Relative to the short-run trend, there is a massive decline in entry post-reform. However, this may just have been a reversion to longer term trends. In the next chapter, I address this question in more detail using an economic model, which also quantifies the effect on profits.

Bibliography

Aguirregabiria, V., Clark, R., and Wang, H. (2016). Diversification of geographic risk in retail bank networks: evidence from bank expansion after the riegle-neal act. *The RAND Journal of Economics*, 47(3):529–572.

Altonji, J. G., Huang, C.-I., and Taber, C. R. (2015). Estimating the cream skimming effect of school choice. *Journal of Political Economy*, 123(2):266–324.

Andrews, D. W., Stock, J. H., et al. (2005). Inference with weak instruments. *Cowles Foundation Discussion Papers*, (1530).

Andrianova, S., Demetriades, P., and Shortland, A. (2012). Government ownership of banks, institutions and economic growth. *Economica*, 79(315):449–469.

Asher, S., Lunt, T., Matsuura, R., and Novosad, P. (2019). The Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG). Working paper.

Asher, S., Lunt, T., Matsuura, R., and Novosad, P. (2021). Development Research at High Geographic Resolution: An Analysis of Night Lights, Firms, and Poverty in India using the SHRUG Open Data Platform. World Bank Economic Review.

Asher, S. and Novosad, P. (2020). Rural Roads and Local Economic Development. *American Economic Review).*

Barros, P. P. and Siciliani, L. (2011). Public and private sector interface. In *Handbook of health economics*, volume 2, pages 927–1001. Elsevier.

Berry, S. and Reiss, P. (2007). Empirical models of entry and market structure. *Handbook of industrial organization*, 3:1845–1886.

Bresnahan, T. F. and Reiss, P. C. (1991). Entry and competition in concentrated markets. *Journal of political economy*, 99(5):977–1009.

Bruhn, M. and Love, I. (2014). The real impact of improved access to finance: Evidence from mexico. *The Journal of Finance*, 69(3):1347–1376.

Burgess, R. and Pande, R. (2005). Do rural banks matter? evidence from the indian social banking experiment. *American Economic Review*, 95(3):780–795.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2019). Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies*, 86(5):1867–1900.

Chernozhukov, V. and Hansen, C. (2008). The reduced form: A simple approach to inference with weak instruments. *Economics Letters*, 100(1):68–71.

Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828.

Cramer, K. F. (2021). Financial development and health. *Available at SSRN 3917526*.

Cremer, H., De Donder, P., Boldron, F., Joram, D., and Roy, B. (2008). Social costs and benefits of the universal service obligation in the postal market. *MA Crew et al., Competition and Regulation in the Postal and Delivery Sector, Cheltenham and Northampton, MA*, pages 23–35.

Demirgüç-Kunt, A., Klapper, L., Singer, D., Ansar, S., and Hess, J. (2020). The global findex database 2017: measuring financial inclusion and opportunities to expand access to and use of financial services. *The World Bank Economic Review*, 34(Supplement_1):S2–S8.

Dupas, P., Green, S., Keats, A., and Robinson, J. (2016). *2. Challenges in Banking the Rural Poor: Evidence from Kenya's Western Province*. University of Chicago Press.

Galichon, A. and Henry, M. (2011). Set identification in models with multiple equilibria. *The Review of Economic Studies*, 78(4):1264–1298.

GAO (2021). Broadband: Fcc is taking steps to accurately map locations that lack access. *Government Accountability Office Publication 21-104447*.

Garg, S. and Gupta, S. (2020). Financial access of unbanked villages in india from 1951 to 2019: A spatial approach. *IEG Working Paper No. 403*.

Garg, S. and Gupta, S. (2021). Financial access and gender gap in entrepreneurship and employment: Evidence from rural india. *Available at SSRN 3923735*.

Geddes, R. R. (2005). Policy watch: Reform of the u.s. postal service. *Journal of Economic Perspectives*, 19(3):217–232.

King, R. G. and Levine, R. (1993). Finance and growth: Schumpeter might be right. *The quarterly journal of economics*, 108(3):717–737.

Kuehn, J. (2018). Spillovers from entry: the impact of bank branch network expansion. *The RAND Journal of Economics*, 49(4):964–994.

La Porta, R., Lopez-de Silanes, F., and Shleifer, A. (2002). Government ownership of banks. *The Journal of Finance*, 57(1):265–301.

Laffont, J.-J. and Tirole, J. (1990). Optimal bypass and cream skimming. *The American Economic Review*, pages 1042–1061.

Meiyappan, P., Roy, P., Soliman, A., Li, T., Mondal, P., Wang, S., and Jain, A. (2018). India village-level geospatial socio-economic data set: 1991, 2001. Technical report, NASA Socioeconomic Data and Applications Center.

Mohan, R. (2006). Economic growth, financial deepening and financial inclusion. *Reserve Bank of India Bulletin*, 1305.

Morgan, D. P., Pinkovskiy, M. L., Yang, B., et al. (2016). Banking deserts, branch closings, and soft information. Technical report, Federal Reserve Bank of New York.

Murphy, K. M., Hughes, L. S., and Conway, P. (2018). A path to sustain rural hospitals. *Jama*, 319(12):1193–1194.

Narasimham, M. (1992). *Narasimham Committee Report on the Financial System, 1991*. Standard Book Company.

Nguyen, H.-L. Q. (2019). Are credit markets still local? evidence from bank branch closings. *American Economic Journal: Applied Economics*, 11(1):1–32.

Panagariya, A. (2006). Bank branch expansion and poverty reduction: A comment.

Rajan, R. and Zingales, L. (1998). Financial dependence and growth. *The American Economic Review*, 88(3):559–586.

Sanches, F., Silva Junior, D., and Srisuma, S. (2018). Banking privatization and market structure in brazil: a dynamic structural analysis. *The RAND Journal of Economics*, 49(4):936–963.

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *The RAND Journal of Economics*, 37(3):619–640.

Thorat, U. (2009). *Report of the High Level Committee to Review Lead Bank Scheme*. Reserve Bank of India.

Young, N. (2019). Banking and growth: Evidence from a regression discontinuity analysis. Working paper.

Appendices

Appendix: Data

Matching

Table A.1 reports the matching rate between Reserve Bank of India data and the SHRUG database. Place names in the RBI dataset have a variety of suffixes (e.g. "(m)" or "ct") appended to the end of village/town names which are not included in the SHRUG names. Excluding these from the place names considerable improves the match rate. The MOF uses 2019 district names. Where possible, these are harmonized with 2011 Census district boundaries.

Branches in the MOF which are not matched by this process are linked to their SHRUG identifiers by matching the GPS coordinates of the listed address on the Google Maps API to the centroid of the nearest Census village or town.

Unmatched branches are dropped from the analysis.

Table A.1: Matching to SHRUG

Item	In RBI Data	Matched	Match Rate:
District-level Deposits (2010)			
Districts	556	524	94.2%
1 Branch Deposits (billion rs.)	3,256	3,134	96.3%
2 Branch Deposits (billion rs.)	1,275	1,236	96.9%
Center-level Deposits (2010)			
Centers	6,462	5,285	81.8%
Deposits (billion rs.)	49,008	47,818	97.6%
Branch Data (2019)			
Branches	187123	185647	99.2%
Centers	48055	45596	94.9%
of which Rural	*39379*	*37537*	*95.3%*
of which Semi-Urban	*7575*	*7334*	*96.8%*
of which Urban	*439*	*435*	*99.1%*
of which Metropolitan	*41*	*41*	*100.0%*

Chapter 2

The Price of Financial Inclusion: Equilibrium Effects of Rural Bank Branching Policy in India

Abstract

In this chapter, I use an economic model of branch entry to estimate banks' profits, compute their regulatory compliance costs, and simulate equilibrium entry and profits under counterfactual policies. Compared to a free-entry counterfactual, the 25% unbanked share mandate reduces total profits from new branches by about 26% and shifts entry from banked to unbanked markets roughly one-for-one, with disproportionate losses in smaller banked markets. Higher mandatory unbanked shares can further increase unbanked entry, with rapidly escalating losses in banked markets. In equilibrium, compliance costs are higher for private vs. public sector banks. This tends to benefit public sector banks by shielding their incumbent branches from competition. Equalizing these costs by allowing banks to trade branch licenses in a competitive market modestly increases profits for private sector banks but does not result in net new entry.

2.1 Introduction

Governments use a wide variety of policy instruments to achieve their goals, including price signals, constraints on firms' behavior, and direct action. The consequences of such policies depend on how they interact with the underlying economic system. Measuring the equilibrium impact of an intervention can be difficult, especially when the policy was not designed or implemented with evaluation in mind. However, it is often ex ante prediction of a policy's costs and benefits, rather than ex post evaluation, which is relevant for decision making. Furthermore, there are often a range of alternative policies under consideration. To select the most effective version of the policy, it is often useful to have a model of how the policy affects the relevant economic agents and, after estimation, permits an analysis comparing simulated outcomes under the various alternatives.

In this chapter, I develop a model of bank branch entry which incorporates the unbanked share mandates described in the previous chapter. I compute the effect of the 2011 reform by comparing the equilibrium outcome under the 25% unbanked share mandate to the outcome with free entry. I then explore the efficacy of the 25% mandate, and the tradeoffs involved, by comparing it to a variety of alternative regulations.

The sudden and dramatic expansion of the rural branch network following the implementation of the 2011 reform was costly. Banks pay a direct cost of opening and operating unprofitable branches in rural villages. The mandate is a binding constraint, so losses in unbanked villages represent a shadow cost on entry into banked markets. This leads to an additional loss of profits in banked markets where branches would not be sufficiently profitable to cover the compliance costs. These indirect costs depend on the marginal compliance cost, relative to profits in banked markets. As banks enter more banked markets, they must enter increasingly unprofitable unbanked villages, driving up compliance costs. If banks rapidly run out of relatively cheap unbanked villages, there can be large reductions in banked entry – and large indirect losses – even when the direct cost of entering unbanked villages is modest.[1]

The key question is therefore: how (un)profitable are all these new branches? I measure expected profits by estimating a model in which observed branch entry is a Nash equilibrium of an entry game. Within the estimated model, I compute (marginal) compliance costs as the expected losses in unbanked villages required to obtain an additional license in a banked market. I use these compliance costs, along with profits in banked markets and the model of bank conduct, to simulate banks' competitive decisions under alternative policy scenarios. I estimate the model on entry data from the five years prior to the reform, when banks could freely enter banked markets. To do so, I use the complete municipality-level bank branch network in each year from 2006-2019, based on administrative records from the Reserve Bank of India's Master Office File, a database of all bank branch licenses. The public version of the Master Office File contains data for existing branches but not for branches which have closed or been acquired in a merger. The RBI provided access to the confidential merger and closure files for 2006-2019, which enabled me to reconstruct the complete branch network over time. I link the branch network to the Socioeconomic High-resolution Rural-Urban Geographic platform, which provides a consistent crosswalk for a variety of municipality-level datasets, including the Population and Economic Censuses (Asher et al., 2019).

Within the model, banks' profits are identified by observed entry probabilities in markets with different characteristics. One challenge arises because entry games often have multiple equilibria. For example, multiple banks may be profitable as monopolists but not as duopolists. In such cases, the model does not deliver a well-defined

[1] This is clearest in the extreme case: a 100% unbanked share would lead to no direct losses at all, as banks would have no incentive to enter unprofitable unbanked villages, and massive indirect losses, since entry in banked markets would be prohibited.

probability distribution over outcomes. Instead, the model bounds the probability of a given outcome, based on the probability that it is either the unique equilibrium or not an equilibrium at all (Ciliberto and Tamer, 2009). This requires using moment inequalities for inference. A second challenge is identifying the effect of competition on profits in the presence of market-level unobservables. If multiple banks enter the same market, it may be because the branches do not strongly compete or because some unobserved factor makes the market particularly profitable. I use variation in banks' historical presence in nearby markets to instrument for their own entry. For reasons related to logistics, management, or information, branches may be more profitable if located close to branches belonging to the same bank, whereas other banks' branches will not obtain these same benefits. This isolates variation in market structure not related to local profit shocks, which identifies the causal relationship between competition and profits.

I first consider the magnitude and distribution of the reduction in entry in banked markets resulting from the mandate requiring banks to open 25% of their new branches in unbanked villages. By comparing equilibrium entry under the baseline mandate to a free entry counterfactual, I find that the mandate reduced entry in banked markets by around 21%, resulting in a roughly one-for-one reallocation from banked to unbanked markets. This is a slightly lower price than is implied by post-reform deviations from pre-reform trends in entry, although the difference is not statistically significant. Smaller markets were more likely to lose branches than larger metropolitan areas: the median banked market predicted to lose a new branch under the baseline policy is 23% smaller than the median banked market entered under free entry. This disproportionate impact may be an important consideration for policymakers seeking to ensure an equitable distribution of branches.

I then consider the mandate's effects on banks. The baseline mandate reduces total bank profits from new branches by about 26% relative to free entry.[2] Only about one-third of the total losses are direct compliance costs, with the rest coming from foregone profits in banked markets. This represents considerable inefficiency: for every dollar of profit diverted to pay for rural branch expansion, banks lose two dollars from deterred entry in banked markets. One reason that the mandate's costs are large is that marginal compliance costs are not equalized across banks. In equilibrium, the typical public sector bank faces substantially lower compliance costs than private sector banks, primarily due to more extensive preexisting rural branch networks. The mandate therefore disadvantages private sector banks, which bear a disproportionate share of the policy-induced reduction in entry and profits. One way to equalize these costs is to allow banks to acquire permits by opening branches in unbanked villages and trade them in a competitive market.

I find that a competitive market paired with the baseline mandate, would reduce banks' losses by about 11%. These benefits go disproportionately to private sector banks. Tradable permits reduce the mandate's cost to banks, but do not lead to net new entry. Instead, the branch network becomes more stratified, with private sector banks increasingly dominating the most profitable urban markets. This may be undesirable from the governments' perspective, especially when not compensated for by increasing branching overall.

Finally, I compare the results of the 25% unbanked share mandate implemented in 2011 to the stricter 67% unbanked share required during the "Social Banking" era (Burgess and Pande, 2005). If implemented in 2006, this would have resulted in a roughly 50% larger rural branch expansion than the baseline mandate, with catastrophic effects on banked entry and profitability. As in the late 1980s, such a strict mandate would likely cripple the Indian banking system. Because higher

[2] This captures only the expected profits earned by new branches. With fewer new competitors, incumbent branches are likely to be more profitable, and therefore the 26% reduction in profits is an upper bound on bank losses due to the mandate.

shares both require more unprofitable entry per new branch and force banks into less profitable villages more quickly, their costs can escalate rapidly. At shares higher than 67%, banked entry decreases so rapidly that both total and unbanked entry decline.

2.1.1 Related Literature

My analysis of the mandate builds on the literature on empirical entry games (Bresnahan and Reiss, 1991; Berry and Reiss, 2007; Ciliberto and Tamer, 2009) These methods have recently been applied in the context of bank branching by Aguirregabiria et al. (2016) and Kuehn (2018), among others. As in these papers, I use a moment inequality framework for estimation, although I implement it by an alternative approach, outlined in Chernozhukov et al. (2019). In order to study the regulations of interest, I augment these models with a constraint on the types of markets firms can enter. I also emphasize differences between public and private sector banks, as in Sanches et al. (2018). There is a long tradition of studying the role of public banks in economic and financial development (La Porta et al., 2002; Andrianova et al., 2012). One potential interpretation of the unbanked share mandate is as a mechanism by which public sector banks may hinder financial development: encouraging governments to enact regulations which favor public banks over their private sector rivals.

The remainder of the paper is structured as follows. Section 2 introduces the model of bank branching decisions under free entry and the unbanked share constraint. Section 3 describes identification and the estimation strategy. Section 4 presents the parameter estimates and interprets them in the context of the model. Section 5 presents results from counterfactual simulations, and section 6 concludes.

2.2 Model

In this section, I develop the structural model of branch entry that I use to estimate the expected profits each bank would earn from entering in each market and to simulate outcomes under counterfactual policy scenarios.

2.2.1 A Model of Bank Branch Networks

Bank Branch Profits

Banks earn profits through intermediation, that is, by collecting deposits as cheaply as possible, pooling those deposits, and investing them in higher return projects or financial assets. To accomplish this objective, banks choose their branch networks to maximize the expected return on deposits collected through their branches, minus the costs of opening and operating these branches. Banks are risk-neutral and consider the potential payoff and cost of a branch in each village or town individually.[3] At the market level, profits for bank j in market m are:

$$\pi_j(\cdot;\theta) = (\bar{r}_j - r_j^{dep})D_j(y_{jm}, y_{-jm}; X_m^d; \theta) - C_j(y_{jm}, X_m^c, W_{jm}) + \epsilon_{jm}(\theta) \tag{2.1}$$

where $D_j(\cdot;\theta)$ is the total deposits collected in market m by bank j, $C(\cdot;\theta)$ is the cost of operating y_{jm} branches. The shock, $\epsilon_{jm}(\theta)$, is an idiosyncratic unobservable profit shock, with a mean-zero distribution parametrized by θ and potentially correlated across banks within a market. Profit earned in a market is therefore the deposits collected, D_j, times the spread between the banks' rate of return on capital, $\bar{r}_j$ and

[3] While banks do extend credit through their branches and this is likely an important benefit to consumers, it a minor source of revenue for most Indian banks based on their annual reports. There is some work, including Aguirregabiria et al. (2016), which considers bank branching decisions as a portfolio choice problem, where risk averse banks seek geographically diverse branch networks to reduce the volatility of their deposit base. In the empirical specification, returns from local lending and other sources of value, including diversification, are captured in the reduced-form profit function.

the interest rate paid to depositors, r_j^{dep}, minus the cost of entry, plus an unobservable profit shock.

Bank j's deposits are a function of j's branches in that market, y_{jm}, and branches operated by competitors, y_{-jm}. The deposit functions are heterogeneous by bank. For example, a new public sector branch may divert more market share from other public sector banks than from private sector banks. Finally, deposits depend on the characteristics of the market itself, X_m, which determine pool of potential deposits or the availability of informal savings instruments.

Costs depend on j's own entry decision but not the choices of other banks. Market-level characteristics, X_m, may affect both deposits and costs. For example, larger towns may have both more potential deposits and higher real estate costs. Bank-market characteristics in W_{jm} shift costs, and at least some must be excluded from demand and other banks' profits in order to identify competitive effects when profits shocks are correlated between banks within a market.

I distinguish between *legacy* branches, y_{jm}^0, which are present at the beginning of the game, and new branches, y_{jm}^1 which are opened by banks during the game. In India, most banks have been operating for decades and have opened branches under multiple regulatory regimes. Modeling entry decisions for the entire historical branch network is outside the scope of the model and not necessary for evaluating banks' reactions to the 2011 mandate.

Finally, most Indian banks have historically operated in particular states or regions, so in the model bank j only considers entering in a subset of markets where it is "active", $\mathcal{A}_j$. This both adds realism and substantially eases the computational burden of enumerating Nash equilibria, which increases exponentially in the number of players per market. To facilitate modeling unbanked share mandates, I partition these active sets into banked markets, $\mathcal{A}_j^B$, and unbanked villages, $\mathcal{A}_j^U$.

Nash Equilibrium Conditions

Banks play static Nash equilibrium strategies which maximize their total profits, subject to a regulatory constraint, given other banks' choices. Their total profits are given by:

$$\Pi(Y_j, Y_{-j}) = \sum_{m=1}^{M} \pi_j(y_{jm}, y_{-jm}; \cdot)$$

The constraint is equivalent to a shadow price on entry in banked and unbanked markets. A Nash equilibrium of the entry game with mandatory unbanked share λ is a set of entry decisions $\{y^1_{jm}\}_{m=1}^{M}$ and marginal compliance costs C_j for each bank j satisfying the following conditions:

1. For all banked markets $m \in \mathcal{A}_j^B$:

$$y^1_{jm} = y^0_{jm} + 1 \iff \pi_j(y^0_{jm} + 1, y_{-jm}; \cdot) - C_j \geq \pi_j(y^0_{jm}, y_{-jm}; \cdot)$$

2. For all unbanked markets $m \in \mathcal{A}_j^U$:

$$y^1_{jm} = 1 \iff \pi_j(1, y_{-jm}; \cdot) + \frac{1-\lambda}{\lambda} C_j \geq \pi_j(0, y_{-jm}; \cdot)$$

3. The marginal compliance cost, C_j, enforces the unbanked share constraint:

$$\sum_{m \in \mathcal{A}_j^B} y^1_{jm} \leq \frac{1-\lambda}{\lambda} \sum_{m \in \mathcal{A}^{\mathcal{U}}} y^1_{jm}$$

4. Bank j maximizes total profits, subject to condition 3.

$$\sum_{m \in \mathcal{A}_j} \pi_j(y^0_{jm} + y^1_{jm}, y_{-jm}; \cdot) \geq \sum_{m \in \mathcal{A}_j} \pi_j(y^0_{jm} + \hat{y}^1_{jm}, y_{-jm}; \cdot)$$

In equilibrium, all branches opened in banked markets must earn enough profit to cover the compliance costs which could be saved by opening one less branch in a banked market, denoted by C_j. A bank will open a branch in an unprofitable unbanked village if and only if the value of the permits gained covers its losses. If a bank enters an unbanked village, they can open $\frac{1-\lambda}{\lambda}$ additional branches in banked markets without violating the constraint. From condition 1, we know that in equilibrium these marginal branches in banked markets would make at most C_j in profit, because j has already entered all markets with profits greater than C_j.

Banked and unbanked entry, on the left- and right-hand sides of condition 3, are weakly decreasing and increasing in C_j, respectively. Holding other banks' entry constant, the difference between the permits used (LHS) and permits obtained (RHS) is therefore monotonic in C_j and there will be a unique minimal C_j^* such that condition 3 holds. Under the additional assumption that branches in banked markets are profitable and branches in unbanked markets are unprofitable on the margin, this minimal C_j^* will maximize profits, satisfying condition 4. In equilibrium, conditions 1-4 must be satisfied for all J banks simultaneously. Because one bank's entry reduces its competitors' profits, the equilibrium branch networks and compliance costs may not be unique.

2.2.2 Identification

The primary object of interest is bank branch networks under counterfactual unbanked share constraints. This requires identifying the (relative) profitability of branches in each market. Following Seim (2006), Ciliberto and Tamer (2009), and others, I work with a reduced-form payoff function instead of the full structural profit function. This reduced-form payoff function captures the relative profitability of markets and the competitive effects of entry on rivals' profits. Collapsing demand and supply factors into a single measure of profitability simplifies identification and is sufficient for an analysis of the effect of bank branching regulation on the supply of branches.

There are two main challenges for identification. First, competitors entry decisions are correlated with the unobserved profit shock, which is itself correlated across banks. Second, entry models frequently have multiple equilibria, which may lead to partial identification.

Correlated Shocks and Endogenous Entry

Markets may have unobserved characteristics which make them more or less profitable for all banks. In the language of the model, this corresponds to correlation in the profit shocks, ϵ_{jm}. This will tend to bias estimates of the negative competitive effect that one branch has on another upwards - in an extreme case, the correlation in banks' entry decisions driven by unobserved profitability may be spuriously interpreted as a positive effect of a branch on a competitor's profits.

The solution lies in bank-by-market variables that are excluded from other banks' profit functions. These operate as instruments for branch entry. The effect of bank A's branch on the profitability of bank B's branches is identified by comparing B's probability of entry in otherwise similar markets where A's branches are exogenously more or less profitable. Formally, we need some cost-shifter, W_{jm}, such that:

$$\mathbb{E}[\epsilon_{km}|W_{jm}] = 0, \ \ \forall k \neq j$$

That is, W_{jm} affects bank k's entry probability only through its effect on bank j's entry.

One potential source of such variation is the legacy network, y^0. It may be cheaper for a bank to open and operate branches near its preexisting branches, for logistical, managerial, or information reasons which do not benefit rival banks. In the empirical specification, I use two measures of the legacy network: proximity, captured by distance between the market and the banks' nearest legacy branch, and density, captured by the number of the banks' branches within 50km of the market. Clearly, past entry is likely correlated with unobservable characteristics. Because these legacy-network based instruments are a key source of variation, it is worth discussing the assumptions necessary for their validity. y^0_{jm} and ϵ_{jm} are likely to be correlated if unobservable profits are persistent over time. Unless only the idiosyncratic component is persistent, y^0_{jm} is presumably also correlated with other banks' shocks, ϵ_{km}. This does not necessarily violate the exclusion restriction, because legacy-network based instruments rely branches in *other* markets. The assumption necessary for these instruments to be valid is that past entry, y^0_{jm} is uncorrelated with current unobservable profits of other banks, in other markets:

$$\mathbb{E}[\epsilon_{km}|y^0_{jn}] = 0, \ \ \forall j \neq k, n \neq m$$

The fact that exclusion has to hold only across banks means that violations of this assumption need to occur through persistent spatial correlation in the common component of the profit shocks. For example, if branches serve sufficiently large areas, unobserved characteristics of market m might directly affect entry in market n, and this correlation could persist over time. As long as this "catchment area effect" attenuates sufficiently quickly with distance, this type of concern can be mitigated by using only legacy branches sufficiently far from a market to construct instruments.

Multiple Equilibria

The second complication for identification is the fact that entry games frequently support multiple equilibria. For example, several banks may be profitable as monopolists, but none as duopolists. In such cases, the model does not generate a well-defined probability distribution over outcomes, which prohibits methods based on a likelihood function or moment equalities. Intuitively, identifying branch profits as a function of observable characteristics depends being able to match the observed probability of entry in different types of markets. When the model does not produce a well-defined probability, it is not obvious that this is possible.

There are a number of ways to resolve this difficulty.[4] Following Ciliberto and Tamer (2009), I exploit the fact that although the model does not provide a unique probability of a given market-level outcome, Y, it does provide *bounds* on that probability. For some realizations of the shocks, Y will be one of several equilibria. For others, it will be the unique equilibrium or not an equilibrium at all. Integrating over the shocks, the probability that Y is observed cannot be more than the probability that it is an equilibrium (i.e. supposing that the unknown selection mechanism always leads to Y when Y is an equilibrium) and cannot be less than the probability that it is the unique equilibrium.

Formally, define $R_1(y;\theta,X)$ and $R_2(y;\theta,X)$ to be the sets of shocks ϵ such that y is the unique Nash equilibrium and one of multiple Nash equilibria, respectively. Then, the model requires that:

$$P_L(y|X;\theta) = \int_{R_1(y;\theta,X)} dF \leq P(y|X) \leq \int_{R_1(y;\theta,X)} dF + \int_{R_2(y;\theta,X)} dF = P_U(y|X;\theta) \tag{2.2}$$

where $P(y|X)$ is the conditional probability of y observed in the data.

Across outcomes y and market characteristics X, these inequalities identify the set of parameter vectors corresponding to economic models consistent with the observed data. The identified set is $\Theta_I = \{\theta | \forall y, X, \text{ inequalities (2) hold at } \theta, \text{ almost surely}\}$

[4] One approach, followed by (Bresnahan and Reiss 1991) and (Berry 1992), is to impose sufficient assumptions on bank profits so that the equilibrium number of entrants is unique, if not their identities. Then identification can proceed based on the likelihood of the number of entrants. Unfortunately, this rules out most kinds of heterogeneity, which could lead to multiple equilibria with, say, either one "large" branch or two "small" branches. Another possibility is to complete the model through assumptions about the equilibrium selection mechanism. This can be achieved through an ad hoc rule (e.g. pick the equilibrium maximizing total profits) or by parametrizing a probability distribution over outcomes as in (Bajari, Hong, Ryan 2010).

In general, multiple parameter vectors θ satisfy (2). That is, Θ_I is not a singleton and the model is not point identified. Variation in the market characteristics shrinks this identified set. Heuristically, the model is point identified if there exists variation in characteristics which can independently drive the entry probabilities of any given bank to 0 or 1. In other words, if the excluded cost instruments are sufficiently strong. (Ciliberto and Tamer, 2009)

2.3 Estimation and Inference

These model-implied upper and lower bounds on the conditional probability of different market structures lead naturally to estimators based on moment *inequalities*, which reduces to a standard GMM estimator in the case where all equilibria are unique. Inference proceeds based on minimizing violations of bounds on the observed probabilities of equilibrium outcomes conditional on observables X. The upper and lower bounds on the probability of each equilibrium provide moment *inequalities* which must be satisfied for the true parameters.

Constructing these moment inequalities requires computing the probability of an outcome implied by the model for a given parameter vector, θ, and the empirical analog which the model should bound. To simulate the upper and lower bounds on equilibrium probabilities for each market, I draw S entry cost shocks. For each shock, I enumerate all possible pure-strategy Nash equilibria. I compute $\hat{P}^U_m(Y)$ and $\hat{P}^L_m(Y)$ as the fraction of simulation draws in which a given equilibrium is an equilibrium and the unique equilibrium, respectively.[5]

[5] It is possible that no pure-strategy Nash equilibrium exists, because entry by a competitor may raise the return to entry for a bank with a pre-existing branch. A bank may optimally defend its market share by opening another branch if and only if another bank enters. If another bank will enter if and only if the incumbent does *not* open a second branch, then there will be a cycle in the best response mapping and a pure strategy equilibrium will not exist. In the rare event that this occurs, I use the uninformative bounds of [0,1] on the entry probabilities for all equilibria. While inference will be correct, this may result in confidence sets that are larger than necessary. Tighter bounds could be obtained by computing the support of the mixed strategies or, for example, iterated

With a finite number of markets, it is necessary to estimate the empirical probability of an outcome based on market characteristics. Following the literature, I use a non-parametric frequency estimator, where the probability of outcome Y_i for a market of type k is the share of markets of type k where outcome Y_i occurs. That is:

$$P_k(Y_i) = \frac{1}{\#B_k} \sum_{m \in B_k} 1\{y_m \in Y_i\} \tag{2.3}$$

where B_k is the set of markets of type k and y_m is the observed equilibrium in market m. Note that some moments are based on sets of equilibria. Galichon and Henry (2011) demonstrate the importance of the probability of sets of equilibrium outcomes, in addition to specific equilibria, for identification. For example, in the case where the number, but not identity, of entrants is unique the probability of the event that any equilibrium with n entrants occurs will contain information not captured in the probabilities of the individual equilibria which constitute the event.[6] This implies that the number of potentially informative moments grows at the rate of 2^{2^J},

The choice of market types, B_k, used to form moments conditional on characteristics, is analogous to the choice of interactions in a standard method of moments estimator. Heuristically, the effect of a characteristic on profits is identified by correctly matching entry probabilities across bins with different levels of that characteristics. In general, there is a tradeoff between the number of bins, which adds inequalities and shrinks the identified set, and the number of observations per bin, which determines the error in the first-stage estimates. In practice, bins need not be mutually exclusive and there may be some gains from including both coarse and fine-grained bins. (Cite Berry Compiani) The need to satisfy conditional moment inequalities further multiplies the number of moments.

elimination of dominated strategies. In practice, non-existence is very rare and the loss of power is minimal.

[6] Apart from identification, aggregated equilibrium outcomes occur more frequently and thus estimated probabilities are more precisely estimated or simulated.

Chernozhukov et al. (2019) (CCK) propose a method for constructing confidence regions that performs well with large numbers of inequalities. Formally, I define N sets of equilibria, Y_i and K subsets of markets. From the data, I compute $P_k(Y_i)$ for each outcome and bin, as defined in equation 5. Given parameters θ, I compute $\hat{P}^L_m(Y_i;\theta)$ and $\hat{P}^U_m(Y_i;\theta)$ by simulation.

For each market m, bin k, and outcome i, I compute:

$$(X^L_{ikm}, X^U_{ikm}) = \begin{cases} \Big(\hat{P}^L(Y_i|m) - P_k(Y_i), P_k(Y_i) - \hat{P}^U(Y_i|m)\Big) & \text{if } m \in B_k \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

Note that $X^L_{ikm} > 0$ if the model-implied lower bound is greater than the observed probability and $X^U_{ikm} > 0$ if the model-implied upper bound is less than the observed probability. I then construct $2 \times N \times K$ moment means and standard deviations, $\hat{\mu}^L_{ik}, \hat{\mu}^U_{ik}$ and $\hat{\sigma}^L_{ik}, \hat{\sigma}^U_{ik}$

$$\hat{\mu}^b_{ik} = \frac{1}{M}\sum_{m=1}^{M} X^b_{ikm} \quad \text{and} \quad \hat{\sigma}^b_{ik} = \sqrt{\frac{1}{M}\sum_{m=1}^{m}(X_{ikm} - \mu_{ik})^2}, \quad b \in \{L, U\}$$

I then construct a test statistic for the null hypothesis that θ is consistent with the observed data based on the maximum of the normalized moment inequality violations:

$$T(\theta;\cdot) = \max\{\frac{\sqrt{M}\hat{\mu}_n}{\hat{\sigma}_n}\} \tag{2.5}$$

If $T(\theta;\cdot) \leq C_\alpha(\theta;\cdot)$, then I do not reject that θ is in the identified set, where $C_\alpha(\theta;\cdot)$ is a critical value with size α. CCK gives an analytic expression for these critical values, based on the properties of self-normalized sums.[7] I construct $1-\alpha\%$ confidence sets

[7] In particular, I use their 2-step Self-Normalized method. They propose more powerful tests based on bootstrap methods, which take advantage of the correlation structure of the moments. These are substantially more computationally intensive, so I use the more conservative analytic bounds instead.

by first minimizing the difference between the test statistic and the critical value, $f(\theta) = T(\theta; \cdot) - C_\alpha(\theta; \cdot)$, and then exhaustively sampling the surrounding region for values of θ for which $f(\theta) \leq 0$. I report maximum and minimum values in each dimension, so reported confidence intervals form a hypercube containing the true confidence set.

The resulting confidence sets are robust in the sense that they do not conflate proximity to the minimum with proximity to the data generating process. The CCK asks, directly, whether we can reject that, for parameter value θ, the model is the true data generating process. It is therefore possible for this method to reject the model altogether. In addition to confidence sets, I also report the $\hat{\theta}$ which minimizes $T(\theta; \cdot) - C_\alpha(\theta; \cdot)$ as the "best estimate," and use this value to compute counterfactuals. I can obtain confidence intervals for counterfactual outcomes by finding the maximum and minimum value on the full confidence set.

2.3.1 Empirical Specification

Branch Payoffs

The empirical specification of the payoff function is:

$$\begin{aligned}\pi_j(y_{jm}, y_{-jm}; Z_{jm}, Z_{-jm}, X^d_m, X^c_m, W_{jm}, \epsilon_{jm}; \theta) = D_j(y_{jm}, y_{-jm}; Z_{jm}, Z_{-jm}, X^d_m; \theta) \\ - C(y_{jm}, X^c_m, W_{jm}) - \epsilon_{jm}\end{aligned}$$

It consists of a nonlinear component, D_j, which is analogous to variable profits from deposits, and a linear component

$$C(y_{jm}, X^c_m, W_{jm}) = y_{jm} * (\beta X^c_m + \gamma W_{jm}) \tag{2.6}$$

analogous to the fixed costs of entry. As discussed above, it is important that some subset of bank-specific characteristics, W_{jm}, are exogenous and excluded from other firms' profits. Fixed costs can also depend on common market-characteristics.

I assume that the idiosyncratic profit shock has a common component $\epsilon_{jm} = \tilde{\epsilon}_{jm} + \bar{\epsilon}_m$, where the purely idiosyncratic component, $\tilde{\epsilon} \sim N(0,1)$ has its variance normalized to 1 in order to fix the scale of the parameters and the common shock $\bar{\epsilon}_m \sim N(0, \sigma_m)$, where $\sigma_m \in \{\sigma_{rural}, \sigma_{urban}\}$, depending on the population of market m.

Without incorporating additional data and imposing additional assumptions, I cannot identify deposit demand or revenues separately from costs and so D_j should not be given a structural interpretation. Nevertheless, the parametric specification may better approximate the true payoff function if derived from an underlying model of demand, especially when incorporating competitive effects across both large and small markets. The functional form described below is derived from a 3-layer nested logit discrete choice demand model, where substitution between branches of the same bank and between branches of the same banking group are allowed to be flexible.

I use the following specification for D_j:

$$D_j(y_{jm}, y_{-jm}; X_m^d; \theta) = \exp(\alpha X_m^d) \left(\frac{y_j^{\frac{\lambda_0}{\lambda_g}} \exp(\delta_g) \left(\sum_{k \in J_g} y_k^{\frac{\lambda_0}{\lambda_g}} \right)^{\lambda_g - 1}}{1 + \sum_{h \in G} \exp(\delta_h) \left(\sum_{k \in J_h} y_k^{\frac{\lambda_0}{\lambda_h}} \right)^{\lambda_h}} \right) \tag{2.7}$$

The first component, $\exp(\alpha X_m^d)$, is analogous to market size, and the larger expression is analogous to market share. X_m^d is a vector of market-level characteristics which determine the revenue potential of the market. The second component is analogous to market shares. In order to reduce the dimension of the parametrization, all branches belonging to banking group $g \in G = \{\text{SBI, Other Public, Private}\}$ are assumed to have the same branch quality, δ_g. This parameter governs the share of the potential revenue captured by banks in group g in the absence of competition.

When other branches are present, bank market shares also depend on $\lambda_0, \lambda_g \in (0,1)$. As $\lambda_0 \to 0$, additional branches belonging to the same bank do not increase that banks' market share. λ_{public} and $\lambda_{private}$ similarly allow public or private sector banks to compete more fiercely with other banks of the same type. For example, as $\lambda_{private} \to 0$, a new private sector branch will increasingly draw its market share from other private sector banks instead of public sector banks or the outside option.

Active Entrants and the Non-competitive Fringe

Indian banks are typically limited in geographic scope, and I take advantage of this by defining banks' consideration sets. That is, banks are not modeled as being active participants in all markets. This increases the accuracy of the model because geographic scope is an important determinant of entry. It also considerably eases the computational burden, because the complexity of enumerating Nash equilibria increases exponentially in the number of potential entrants.

To define the set of players in each market, I first select the 12 largest banks by pre-existing branch network or observed entry during the 2006-2010 period. This includes the State Bank of India, eight other public sector banks, and three private sector banks. Together, these represent roughly 70% of total entry during the period. The remaining banks' are aggregated into a public and private fringe and their entry is treated as exogenous.

I then define zones of activity as cities with over 500,000 people or districts, excluding these cities. I designate a bank as active in a zone if a large share of the banks' entry is in the zone or if the bank represents a large share of total entry in the zone. Specifically, I assign banks to be active in the minimal set of districts representing at least 80% of their own total 2006 branches and such that 80% of all branches in each district belong to an active bank. If less than three banks are active in a district, I assign the banks with the next-largest market share in the state in which the district is located. The result covers roughly 95% of branches at baseline and 90% of subsequent entry. Banks' entry decisions are treated as exogenous in markets where they are not active.

The median district has 4 active banks. Because the number of inequalities to check increases exponentially in the number of potential entrants, this considerably reduces the computational complexity of enumerating the equilibria. To further reduce the computational burden, I restrict banks' entry decisions to whether to open a single branch or not enter at all. That is, $y^1_{jm} \in \{y^0_{jm}, y^0_{jm} + 1\}$.

Tables 5 and 6, which report the confidence sets, provide the full list of included covariates.

2.3.2 Construction of moments

There are two key steps for selecting the moments used for inference. First, I must choose the sets of equilibria whose probabilities I will bound. Then, I must choose the bins within which to bound them. These bins are analogous to moments based the covariance of a residual and an exogenous variable commonly used in GMM estimators. Because entry outcomes are discrete, "observed" probabilities must also be estimated and are not easily defined at the market level. Instead of matching the covariance between an outcome and population, I match the predicted and observed outcome across markets in different population bins.

With 11 banks, there are up to 2,048 individual equilibrium outcomes which can occur, and $2^{2,048}$ sets of these equilibria. However, not all of these events are equally informative. Intuitively, an event is most informative if its probability is clearly linked to a particular parameter. For example, an increase in the parameter determining the effect of population on profits has an ambiguous effect on the probability that some individual bank j, specifically, enters a market. If the market is more profitable for all banks, j may be less likely to enter because some competitor k is now more likely to enter. A moment based on the probability that a given number of banks enter will not be subject to this ambiguity and may be more informative for that parameter.

With this in mind, I target 4 sets of equilibria: 1) Those in which each bank j enters, regardless of other banks' entry decisions. 2) Those in which at least one public or private bank enters. 3) Those in which multiple public banks, multiple private banks, or both public and private banks enter, and 4) those in which exactly 0,1,2,3,4, or 5+ banks enter. This results in a total of 22 outcome probabilities to bound for each market.

The other component is to define the bins over which these probabilities are bounded. More granular bins tends to increase power, because it makes more use of the data. All else equal, more bins means more inequalities which must be satisfied, and a (weakly) smaller identified set. Requiring the model to fit outcomes well for each percentile of the population distribution exploits more information about the effect of population on outcomes than an objective function which distinguishes only between markets above and below the median.

Of course, all else is not equal. More granular bins are necessarily smaller and therefore less precisely estimated. An objective function which matches outcomes in markets above or below median will be less noisy than one which matches for every percentile. Entry is relatively rare even in large markets, so large bins are necessary to obtain precise estimates of the outcome probabilities. In practice, some characteristics such as population, are highly skewed, which makes bins based on quantiles inefficient.

For each of the 8 market or bank characteristics used in the model, and for all characteristics, demand characteristics, cost-characteristics, and excluded cost shifters as groups, I use the K-means algorithm to partitions data into 4 clusters which minimize the within-cluster variance on the relevant characteristics.[8]

The full set of moments is the upper and lower bounds on 21 sets of equilibrium outcomes across these $12 \times 4 = 48$ bins, plus the set of all markets, for a total of 2,058 moments.

2.4 Results

In this section I discuss the estimated model. The estimated structural parameters underline the importance of geographic scope as a determinant of entry and indicate significant differences between public and private sector banks. I then evaluate the model's fit and present the implied compliance cost curves associated with unbanked share constraints.

2.4.1 Parameter Confidence Sets

As described in the previous section, the CCK method involves direct inference, rather than estimation followed by an inference step. For each parameter value, it produces a test statistic and a critical value associated with a given confidence level, α. If the test statistic is less than the critical value, then we cannot reject that those parameters generate the data. The result is a confidence set of all parameters which cannot be rejected at the specified level.[9]

Table 4 reports the maximum and minimum values for each parameter within the confidence set. The reported intervals take each parameter individually, so it may be possible to reject that some pair of parameters are both at the high end of their reported range, for example. Although the CCK method is not intended to produce

[8] I use K-means clustering rather than quantiles because it produces more informative bins for characteristics, like population, with highly skewed distributions. Furthermore, unlike quantiles, K-means generalizes naturally to forming clusters of bins which are similar across a set of characteristics.

[9] This procedure is similar in spirit to methods from the weak instruments literature, as described in Andrews et al. (2005) and Chernozhukov and Hansen (2008)

Table 2.1: Parameter Estimates

"Revenue" Parameters (β)			"Cost" Parameters (γ)		
Parameter	**Best**	**95% CI**	**Parameter**	**Best**	**95% CI**
intercept	1.12	(0.98, 1.19)	total preexisting	1.49	(1.35, 1.58)
total preexisting	0.27	(0.03, 0.34)	log(population)	0.00	(-0.03, 0.04)
log(population)	0.1	(0.04, 0.17)	dist. to urban	0.1	(0.01, 0.2)
nightlights	0.79	(0.72, 0.86)			
% literate	0.83	(0.74, 0.91)	**nearest branch**	-0.76	(-0.84, -0.67)
% service	0.52	(0.45, 0.59)	**branch density**	0.15	(0.07, 0.2)
δ_{SBI}	0.64	(0.56, 0.71)	C_{SBI}	-2.70	(-2.78, -2.59)
δ_{Public}	-0.16	(-0.23, -0.05)	C_{Public}	-3.86	(-3.92, -3.73)
$\delta_{Private}$	0.71	(0.61, 0.78)	$C_{Private}$	-2.93	(-3.05, -2.88)
λ_{own}	0.93	(0.84, 0.99)	σ_{rural}	0.01	(0.00, 0.13)
λ_{Pub}	0.06	(0.01, 0.22)	σ_{urban}	1.15	(1.09, 1.32)
λ_{Pri}	0.31	(0.24, 0.40).			

Notes: The Best column reports the parameters minimizing the maximum normalized violation of the moment inequalities. The parameters on the left-hand side enter the (non-linear) revenue function D_j. Parameters on the right hand side enter the linear cost function C_j and are in units of standard deviations of the idiosyncratic profit shock. σ_{rural} and σ_{urban} are the standard deviation of the market-level common shock. Variables in bold are bank-specific cost instruments.

Table 2.2: Competitive Effects on Marginal Increase in Market Share From Entry

	SBI	Nationalized Banks	Private Sector Banks
Monopoly			
First Branch	0.671	0.486	0.677
Second Branch	0.126	0.16	0.125
Third Branch	0.055	0.082	0.054
Single Competitor			
SBI	X	0.237	0.408
Nat. Bank	0.512	0.251	0.519
Private Bank	0.397	0.234	0.363
Multiple Competitors			
Two	[0.055, 0.42]	[0.001, 0.161]	[0.102, 0.426]
Three	[0.031, 0.357]	[0.0, 0.122]	[0.038, 0.363]
Four	[0.02, 0.311]	[0.0, 0.099]	[0.018, 0.317]
Five	[0.014, 0.276]	[0.0, 0.083]	[0.01, 0.282]

Notes: This table reports the increase in banks' share of the overall market size from opening one additional branch for a variety of competitive scenarios. For the case of multiple competitors, I report the minimum and maximum increase in market share across the 3^N possible market structures.

a point-estimate in the usual sense, a "best" estimate is a useful focal point for discussing the results. I therefore also report the parameter values, $\hat{\theta}$ which minimize the maximum normalized violation of the moment inequalities. The best estimates are the values which produce simulated moment inequalities most consistent with the data, according to the weighting used by the CCK method. Roughly speaking, $\hat{\theta}$ yields the largest p-value for the null hypothesis that the data was generated by the model with parameters $\hat{\theta}$.

Recall that the model is only identified up to scale, since multiplying all payoffs by a scalar will not affect entry/exit decisions. The units of profit are standard deviations of the bank-specific idiosyncratic shock, and can only be meaningfully interpreted in relative terms. The revenue function, D_j, is non-linear, and these parameters are best interpreted in context. The δ and λ parameters, corresponding to the mean quality and substitution parameters in the nested-logit market share equation, parametrize the competitive effects. Table 2.2 presents the implied market shares for different types of banks under a variety of market structures. One important featured of these estimated competitive effects is the asymmetry between public-sector nationalized banks and private sector banks. The presence of a private sector branch reduces a nationalized bank's "revenue" by more than half, whereas a nationalized bank branch reduces a private sector bank's revenue by only about 20%.

The parameters which enter the linear "cost" function, C_j, are more easily interpretable. For example, the coefficient on the log km to nearest bank branch implies that doubling this distance increases costs by roughly three-quarters of a standard deviation of profits. The cost coefficients imply that the State Bank of India has the lowest fixed costs of entry, and is therefore more likely to enter, all else equal. The coefficients on the market-level common shocks, σ_{rural} and σ_{urban}, imply that almost all unobserved variation in unbanked villages is idiosyncratic, whereas slightly more than half is due to market-level unobservables in urban markets.

Table 2.3: Model Fit: Number of Entrants per Market and Total Entry by Bank Group

	Lower Bound	Observed	Upper Bound
Number of Entrants			
No Entry	94110.6	93963.0	94111.3
1 Entrant	1804.4	1976.0	1805.1
2 Entrants	417.1	391.0	417.8
3 Entrants	214.1	194.0	214.9
4+ Entrants	137.8	148.0	138.6
Entry by Banking Group			
State Bank of India	2097.1	2025	2098.9
Other Public Banks	2436.8	2437	2450.4
Private Sector Banks	1469.5	1622	1476.9

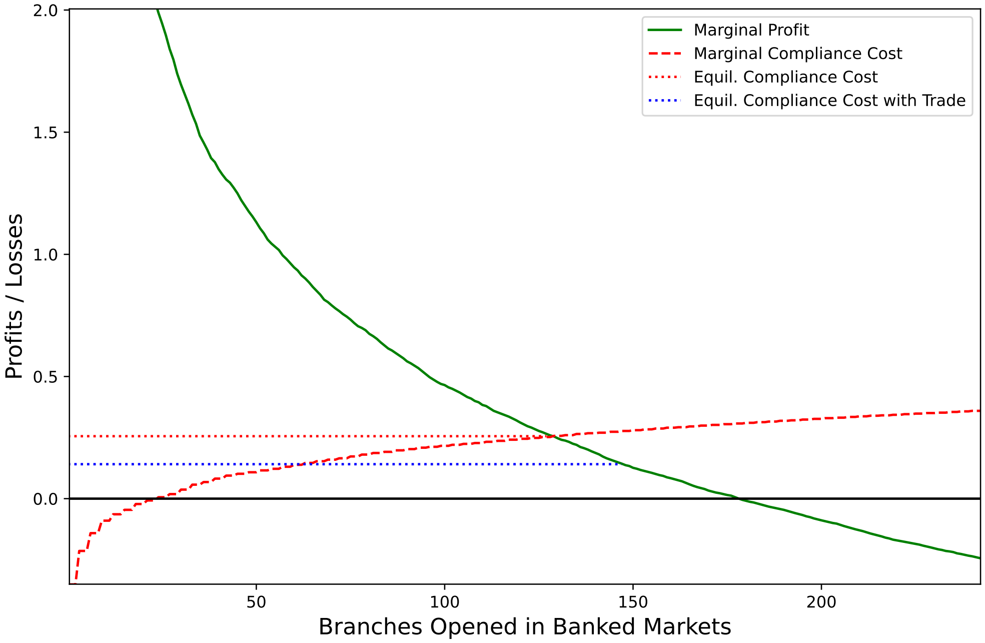
2.4.2 Model Fit

Due to the computational complexity of enumerating Nash equilibrium, it is necessary to have a parsimonious model of branch payoffs. A natural question is how well such a model can replicate the observed patterns of the data. Table 5 compares the observed data to the bounds on predicted entry decisions at the best estimate, which minimizes violations of the moment inequalities.

Fit is, overall, quite good. The model underpredicts entry for private sector banks, which implies that estimated profits are lower for private sector banks than they were in the data. In order to mitigate the role of model-misspecification in the counterfactuals, I use the posterior distribution of profit shocks, conditional on observed entry, in the counterfactual exercises. This procedure is described in detail in Section 7.

Note that the lower bounds is very close to the upper bound. Because these bounds are aggregated up from the market-level, this indicates that, despite their theoretical ubiquity, multiple equilibria are uncommon at the estimated parameters. The fact that almost all equilibria are unique is very useful for the counterfactuals, as ignoring multiplicity greatly simplifies computing the counterfactual equilibria.

Figure 2.1: Estimated Monopoly Profits and Compliance Cost Curves for a Private Sector Bank



Notes: The green solid line plots ICICI Bank's monopoly profits in banked markets, sorted from most to least profitable. The red dashed line plots the compliance cost of opening the N^{th} branch in a banked market. In a simplified model without competitive effects, the unconstrained equilibrium entry occurs where the profits curve crosses the x-axis and constrained equilibrium entry occurs where the profit curve intersects the compliance cost curve. The equilibrium compliance cost is indicated by the red dotted line. The blue dotted line indicates the equilibrium compliance cost in when banks are allowed to trade permits. Note that ICICI Bank both increases entry into banked markets and decreases entry into unbanked markets.

2.4.3 Estimated Compliance Cost Curves

The success of any branch expansion program depends on the unprofitability of entry into the target markets. For unbanked share constraints, such as those repeatedly implemented in India, success depends on the magnitude of these losses *relative* to the potential profits from branches in banked markets.

I use the estimated profit function compute bank-specific compliance cost curves. For each banked market they enter, banks must pay the cost of opening a branch in the least-unprofitable unbanked market still available, amortized over the permitted number of branches in banked markets. For example, under a 25% unbanked share constraint, a bank must pay 1/3 the cost of entry in the most profitable unbanked market for their first 3 branches in banked markets, 1/3 the cost of entry in the second-most profitable unbanked markets for the next 3, and so on. Legally, these compliance costs are discontinuous - the whole price must be paid for the first branch in a banked market, regardless of whether the bank opens a second or third. However, given that banks are allowed to smooth compliance across time, this is a close approximation.[10]

Figure 2.1 presents the estimated compliance cost curves for ICICI Bank, the largest private sector bank. Banks monopoly profits are plotted as solid lines. Note that there are some profitable unbanked markets, so it is possible in theory that the constraint is not binding. On the other hand, even at low levels, an unbanked share constraint may substantially reduce entry if the bank has many banked market which are only marginally profitable.

In a simplified model without any strategic interactions, the expected number of markets entered would correspond to the intersection between the solid monopoly profit curve and the dashed compliance cost curve for each bank. The dotted line marks the equilibrium marginal compliance cost. Bank profits would be the area between the profit curve and the compliance cost curve, to the left of their intersection. Although competitive effects are an important part of the equilibrium effects of the policy, Figure 2.1 provides some insight about how the reform affects different types of banks differently.

On average, private sector banks like ICICI Bank have a steep gradient of profits in both banked markets and unbanked markets. As a result, ICICI Bank is more constrained than average, with an estimated equilibrium compliance cost about 3x as high as that of the State Bank of India, for example. The blue dotted line indicates

[10] That is, the exact condition is that the sum of the profits in the last three banked markets entered must exceed the compliance costs. It is analytically simpler to use an amortized compliance cost, and the two ways of calculating the compliance costs can diverge by at most two branches per bank.

the estimated price for permits in banked markets in a competitive equilibrium. As shown by the intersection of this line with ICICI Bank's profit and compliance cost curves, access to cheaper permits will lead ICICI Bank to both open more branches in banked markets and fewer branches in unbanked villages. Banks which are net suppliers of permits will both open more branches in unbanked villages and fewer branches in banked markets. This makes the net effect on total entry ambiguous, even before strategic interactions between banks are taken into account. In the next section, I explore these effects within the full model.

2.5 Counterfactuals

In this section, I describe how I use the estimated model to simulate counterfactual entry decisions. I first quantify the effects of the reform by comparing simulations of free entry and the baseline policy. I then

2.5.1 Computing Counterfactuals

I report counterfactual simulations for the best estimate of the model parameters. Confidence intervals for counterfactual outcomes can be obtained by computing the maximum and minimum values over the entire confidence set.[11] For counterfactuals without an unbanked-share constraint, entry in each market is considered independently. I simulate entry patterns by enumerating the Nash equilibrium in each market. As in the previous section, I report the upper and lower bounds on selected equilibrium outcomes, averaged over simulation draws.

For counterfactuals with an unbanked-share constraint, I exploit the fact that for the profit-maximizing branch network, all branches opened in a banked market must be profitable enough to cover the compliance costs of opening a marginal branch.

[11] Computing the equilibrium branch networks under the branching constraint is computationally intensive, so I do not report confidence intervals for all counterfactual outcomes.

Given equilibrium compliance costs, I compute the equilibrium branch network as above. However, equilibrium compliance costs are endogenous. They are a function on the total number of branches a bank opens in banked markets, which depends on other banks' entry decisions, where all entry decisions depend on the equilibrium compliance costs. The equilibrium compliance costs are thus a fixed point.

I compute them by iterating computing the equilibrium number of branches opened in banked markets determines the marginal compliance cost, which can then be used to recompute the equilibrium entry decisions until equilibrium entry is consistent with the marginal compliance costs.[12]

In order to fully exploit the available information, I draw profit shocks from the posterior distribution of shocks, conditional on observed entry. To the extent that the estimated model does not perfectly fit entry patterns, comparisons between observed entry decisions and counterfactual simulations will be biased. Conditioning on observed entry patterns mitigates this bias and improves the precision of counterfactuals in general. Concretely, instead of drawing S sets of profit shocks from the (estimated) unconditional distribution, I sample from the conditional distribution of shocks consistent with the observed equilibrium by rejection sampling. I report simulation outcomes averaged across these shocks.

2.5.2 The Effect of the 2011 Reform on Entry

I obtain estimates of the effect of the reform on bank branch networks by comparing the simulated constrained equilibrium with a free-entry counterfactual. This produces a comparison of upper and lower bounds on the probability of each possible equilibrium for every village and town in India.

I first focus on the primary question: How many branches in unbanked markets

[12] The potential for multiple equilibria complicates the question of how to compute the compliance costs. I compute compliance costs given each banks' minimum or maximum entry across equilibria and report the average. In practice multiple equilibria are sufficiently rare that this has a small impact on equilibrium compliance costs.

were gained due to the policy, how many branches in banked markets were lost, and what was the overall effect on bank profits? Table 2.4 presents comparisons between the baseline 25% unbanked share constraint and free entry. The constraint increases unbanked entry by 608%, decreases banked entry by 21% and decreases profits from new branches by 26%.

Note that profits reported here are the total payoffs from entry, including losses in unbanked markets. It is an upper bound on the total losses, because the denominator does not include profits from pre-existing branches or other revenue streams. On the other hand, decreases in total profits in banked markets are a *lower* bound on the deadweight losses imposed by the policy, as they count foregone producer surplus but not lost consumer surplus.

These headline increases are not evenly distributed between types of banks. The reform reduces private sector entry by 29%, about twice as much in percentage terms as for the State Bank of India. The relative effect on profits is similar. In both cases, other public sector banks fall in between these two extremes. The bottom panel in Figure 2.2 decomposes total profits into profits in banked markets and losses in unbanked villages. The direct costs of opening and operating branches in unbanked villages make up about 30% of the reduction in profits under the baseline policy, with the rest coming from foregone opportunities in banked markets.

The reform's costs and benefits were also not equally distributed among markets. Table 2.5 reports the median characteristics of unbanked markets which became banked as a result of the reform ("Winners") compared all unbanked markets in the sample. Unsurprisingly, the median winner was almost 50% larger than the median unbanked village. The median winner was also substantially closer to preexisting branches, had more branches within 50km, and was closer to the nearest town or city.

The opposite was true for banked markets which lost branches. Unlike unbanked markets, it is not a priori obvious whether big cities are more likely to lose branches than small towns. On one hand, metropolitan areas are likely more profitable overall than smaller or lower-income markets. On the other, there is more competition in metropolitan areas and if many banks are entering there is in some since more opportunity to lose branches. The second two columns of Table 6 compare median characteristics in all markets entered under free-entry to those entered under free entry but not the baseline policy.

2.5.3 Alternative Unbanked Share Constraints

During the Social Banking Period of the 1970s and 80s, the Reserve Bank of India required banks to open 4 branches in unbanked villages for every 2 in banked markets, or a 67% unbanked share constraint, as shown in Figure 1.1.

Compliance costs increase rapidly as the required share increases. There is a direct effect - banks have to open more unprofitable branches for each profitable one they open - and an indirect effect as banks are pushed down the gradient into more unprofitable unbanked villages more quickly. If compliance costs increase rapidly enough, banked entry decline fast enough to reduce unbanked entry as well, despite the increased required share. By definition, if banks were required to open 100% of their branches in unbanked markets, only outright profitable unbanked markets would be entered. There would be no additional rural branch expansion compared to free entry, but the costs to household in banked markets - and the banking system itself - would be very high.

This is illustrated visually in the top panel of Figure 2.2, which plots expected banked and unbanked entry as the required unbanked share increases. For required shares up to about 33%, declines in banked entry are modest and gains in rural branching are roughly equal. Total entry remains flat, but new branches shift from banked to unbanked markets. Beyond this point, unbanked entry continues to rise but the total branch network begins to shrink. The price of an additional branch in an unbanked village rises above one branch in a banked market.

Rural branch expansion is maximized at 66.7% unbanked share. Beyond this point, higher required shares do not offset the reductions in banked entry and rural branch expansion decreases. This results in about 50% more rural branch expansion as the 25% unbanked share, at the price of a roughly 85% reduction in banked entry and a nearly 90% reduction in profits.

2.5.4 Efficiency of Unbanked Share Constraints

By effectively paying for rural branch expansion through reduced entry in banked markets, unbanked share mandates create deadweight loss. For the baseline policy, the direct cost of operating unprofitable branches in unbanked villages is about half the profits lost from reduced entry into banked markets.[13] That is, only about one out of every three rupees of profit lost by banks goes towards providing access.

If it were possible to collect the total profit losses under the baseline policy through a lump-sum tax - preserving banked entry - and use this revenue to provide the minimal subsidy to induce entry in as many unbanked villages as possible, about twice as many unbanked villages could be banked. An analogous scheme could achieve the same scale of rural branch expansion as the policy at nearly a quarter the cost.

Such as scheme is not feasible because the minimal subsidy to induce entry is known only to the banks' themselves. However, one notable feature of unbanked share constraints as opposed to ideal subsidy schemes is that the constraint is not equally binding for all banks. The estimated equilibrium marginal compliance cost for ICICI Bank, the largest private sector bank, is about three times larger than for the State Bank of India. These differences in equilibrium marginal compliance costs implies that the same level of rural branch expansion could be achieved with lower overall losses if banks with low marginal compliance costs entered more unbanked

[13] This is *total* profits, and therefore is net of the positive effect of reduced entry on other entrants' profits.

villages and those with higher costs entered fewer.

Unbanked share constraints may therefore be an inefficient tool for promoting financial inclusion. A feasible way of addressing this limitation would be allowing banks to buy and sell permits for branches in banked markets. Banks with low compliance costs relative to their opportunities in banked markets could profit by selling permits to banks with high compliance costs. While this would not eliminate deadweight losses entirely, relaxing the branching constraints for some banks might reduce deadweight losses and even result in more rural branch expansion by increasing entry into banked markets.

I implement a policy counterfactual in which banks trade permits in a competitive market. There is a single price for a permit to enter a banked market. Banks obtain permits by entering the appropriate number of unbanked villages when their losses in these markets are less than the permit price. Banks demand permits for branches in all banked markets where profits exceed the cost of the permit. At the market clearing price, net suppliers of permits - notably the State Bank of India - earn a producer surplus on the difference between the cost of obtaining the permits at the revenue from selling them. Net buyers earn a producer surplus on the difference between the cost of permits and the profits from branches in banked markets.

Table 2.4 reports the equilibrium outcome for the scenario at the baseline 25% unbanked share mandate. Bank profits increase by about 4% relative to the baseline policy, reducing overall losses by about 11%. Remarkably, there is a slight decrease in overall entry. SBI finds it more profitable to sell its permits than to use them itself, and opens about 11% fewer branches in banked markets as a result. Private sector banks, which were most constrained under the baseline policy, are the primary beneficiaries - their entry rises by about 15%. As shown in Table 2.2, private sector banks have larger competitive effects on other branches than the State Bank of India. As private sector branch networks expand, they deter entry by other banks. On average, slightly more than one other branch does not enter for each additional branch opened by private banks.

Furthermore, one of the original intentions of the 2011 policy was to reduce the geographic diversity of bank branch networks. Tradable permits works against this policy goal, as it incentivizes SBI to leave profitable banked markets and largely removes the need for private sector banks to serve the rural poor. One positive effect of the increased role of the State Bank of India in rural branching is that, under the tradable permit scheme the unbanked markets which are entered tend to be substantially smaller, poorer, and more remote, as reported in Table 2.6. For stricter unbanked share mandates, the efficiency gains can be larger. As shown in Figure 2.2, at the unbanked-entry maximizing share of 66.7%, allowing banks to trade permits can nearly double profits and increases unbanked entry by about 8%. However, at the baseline policy, allowing banks to trade entry permits does not achieve further gains in rural branch expansion and works against other policy goals, the efficiency gains may not justify such a reform.

As a final note, it may be unrealistic to assume that such a market would be competitive if established. The State Bank of India is essentially the sole net supplier of permits, and would likely assert some market power. Furthermore, it would require somewhat myopic management to sell permits to other banks which may use them to open competing branches. In addition to market power, SBI would have an incentive to restrict sales of permits in order to benefit from reduced competition in banked markets. How such a market would function, as the effect it would have on welfare, are interesting questions outside the scope of this paper.

Table 2.4: Entry and Profits under Counterfactual Policies

	25% Mandate	Free Entry	25% Mandate with Trade
Banked Entry	**2243**	**2830**	**2203**
State Bank of India	1246	1462	1106
Nationalized Banks	486	652	508
Private Sector Banks	511	717	589
Unbanked Entry	**743**	**105**	**730**
State Bank of India	415	71	588
Nationalized Banks	159	22	102
Private Sector Banks	169	11	40
Relative Profits	**100.0**	**135.9**	**103.8**
State Bank of India	44.8	55.2	42.2
Nationalized Banks	27.7	37.4	29.1
Private Sector Banks	27.4	43.3	32.5

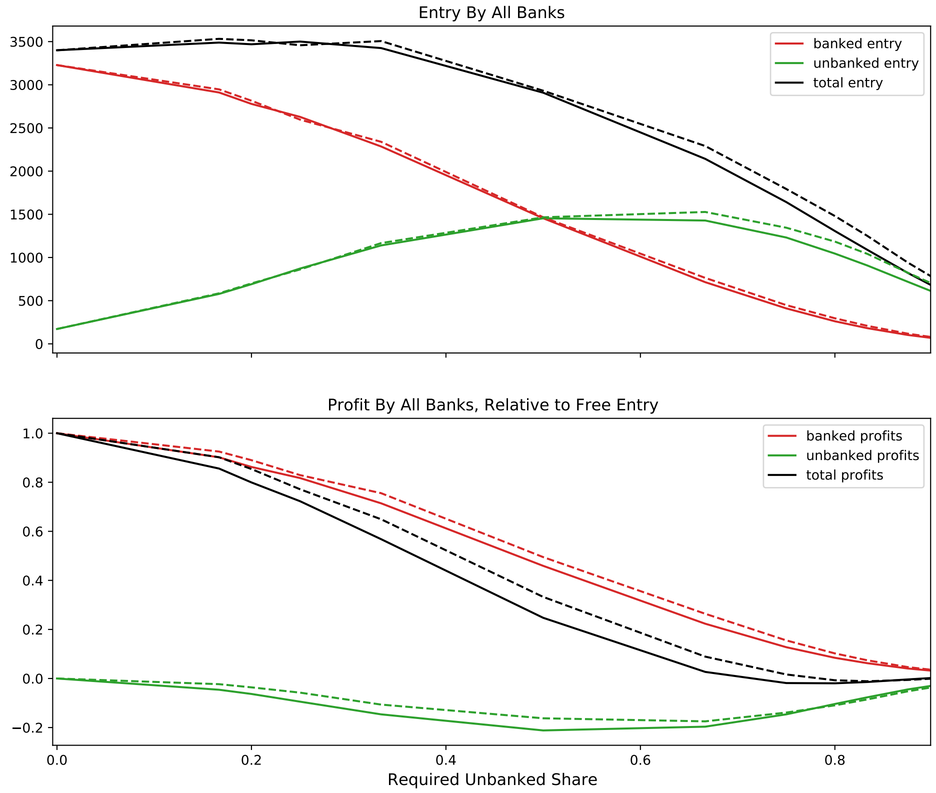
Notes: This table reports simulated banked and unbanked entry for each major banking group. Profits are expressed as percentages of total profits under the baseline 25% unbanked share mandates. Totals reported here are for the smaller sample of markets used in estimation and may not match summary statistics reported elsewhere.

Table 2.5: Median Characteristics of Winners and Losers From 25% Unbanked Share

Name	**Unbanked**	**Winners**	**Entered**	**Losers**
Branches (2005)	0.0	0.0	2.0	2.0
Population (2001)	2886.0	4256.0	13696.0	10537.0
km to nearest town	20.5	13.7	16.2	18.5
% Literate (2001)	47.4	59.2	64.7	63.5
Nightlights (2006)	7.7	13.5	17.0	13.1
% Services (2005)	2.2	3.5	8.1	7.4
km to nearest branch	33.3	22.1	27.8	26.4
branches in 50km	13.9	16.1	13.6	14.1
Unique Markets:	77581	664	2860	283

Notes: This shows the characteristics of unbanked villages which received additional branches, i.e. “winners,” and and banked markets which lost entry, i.e. “losers,” compared to all unbanked markets and markets with entry under free entry, respectively. Some banked markets may lose multiple branches, so the number of unique markets in the “Losers” column will be smaller than the total reduction in banked entry, reported in Table 2.4.

Figure 2.2: Entry and Profits by Unbanked Share, with and without Tradable Permits



Notes: This plots expected entry and expected profits, summed across all banks for required unbanked shares between 0 and 90%. The solid lines correspond to the policy as implemented, and the dashed lines plot outcomes under tradable permits. A market for permits always increases profits but only substantially increases branching for very strict mandates.

Table 2.6: Characteristics of Markets Entered Under Counterfactual Policies

	All Market Median			Unbanked Market Median		
Name	Baseline	Free Entry	Tradable	Baseline	Free Entry	Tradable
Population (2001)	10928.0	15949.0	9858.0	4293.0	4587.0	3099.0
km to nearest town	14.3	15.1	15.9	14.2	16.8	19.6
km to nearest city	25.1	26.6	26.7	23.5	24.8	27.9
% Literate (2001)	64.4	65.6	61.6	58.7	56.3	47.3
Nightlights (2006)	18.1	18.8	15.2	13.3	12.9	7.9
% Services (2005)	7.0	8.3	6.3	3.6	4.6	2.4
km to nearest branch	27.1	28.4	29.9	22.9	25.3	33.6
branches in 50km	13.8	12.9	13.1	15.8	14.2	13.8
Unique Markets:	2546	2159	2505	768	102	693

Notes: Columns are median characteristics for markets entered by the 12 major banks whose entry is explicitly modeled under three scenarios. Market characteristics are not weighted by number of branches opened. In the baseline, banks are required to open 25% of their new branches in unbanked villages. In the Tradable counterfactual, banks are allowed to exchange permits in order to meet the 25% minimum. Under Free Entry, banks have no obligation to enter unbanked villages.

2.6 Conclusion

Mandates which force firms to provide socially beneficial but unprofitable services can be useful tools, particularly when government capacity to directly administer or adequately subsidize those services are limited. However, the costs of such policies can be opaque. To clarify the equilibrium consequences of the 25% unbanked share mandate implemented in India, I develop an economic model of bank branch entry and estimate it using novel municipality-level data on bank branch networks. I find that the policy transfers new branches from banked to unbanked markets at a rate of about one for one, with smaller banked markets bearing the brunt of the losses. The Reserve Bank of India has identified the need to address the extent of financial exclusion in urban areas. (Thorat, 2009) This poses a question for policymakers: Does a branch in an unbanked rural village generate more or less social value than a second branch in a small town, or the first neighborhood branch in a sprawling city?

Even if this tradeoff is acceptable, policymakers should consider alternatives which are less costly to banks or people living in banked markets. Lost entry in banked markets also represents the majority of banks' lost profits. If those losses could be collected as a lump-sum tax and used to provide a minimal subsidy for unbanked entry, the government could achieve almost twice as much rural branch expansion. Clearly such a policy, which relies heavily on banks' private information about profitability, is not feasible in practice. The competitive market for permits does not, in itself, solve this problem, because on net entry into banked markets remains constrained. However, the equilibrium price of permits represents the per-branch subsidy that would achieve the same extent of unbanked entry. This presents an alternative option for policymakers: Are the deadweight losses associated with taxation greater or less than those arising from the lost branches?

This paper considers the effects of the 2011 reform on the availability of bank branches across India and on bank profitability. It characterizes how unbanked share mandates and related policies can trade off rural branch expansion and further financial development in urban markets. This is necessary but not sufficient to establish whether the benefits of the reform outweighed its costs. The analysis presented here is therefore strongly complementary to other work on the role of bank branches in a variety of economic outcomes, which can inform judgements about how tradeoffs in the number and location of commercial bank branches affect welfare.

I can obtain a rough estimate of the effect of the 2011 reform on economic outcomes by combining my results with existing estimates of the effect of bank branches. For example, Young (2019) finds that one additional (private-sector) branch increases district GDP by about 0.33%. Because these branches were overwhelmingly in banked markets, this represents the impact of bank branches in marginally profitable banked markets and is a useful benchmark for the value of branches lost due to the 2011 mandate. My estimates imply the policy reduced entry into banked markets by about 12 branches per district, which would correspond to about a 4% reduction in GDP - or about $100 billion in 2016. By comparison, total operating profits of commercial banks in India totaled roughly \$37 billion.

This paper also provides a novel source of variation in the supply of bank branches, with the potential to improve our understanding of the impact of bank branches on local economic activity. Municipality-level measures of the change in the probability of entry resulting from the 2011 reform, which by construction depend only on observable characteristics, may form the basis for future work estimating the local economics impacts of bank branches in both rural and urban areas.

Bibliography

Aguirregabiria, V., Clark, R., and Wang, H. (2016). Diversification of geographic risk in retail bank networks: evidence from bank expansion after the riegle-neal act. *The RAND Journal of Economics*, 47(3):529–572.

Andrews, D. W., Stock, J. H., et al. (2005). Inference with weak instruments. *Cowles Foundation Discussion Papers*, (1530).

Andrianova, S., Demetriades, P., and Shortland, A. (2012). Government ownership of banks, institutions and economic growth. *Economica*, 79(315):449–469.

Asher, S., Lunt, T., Matsuura, R., and Novosad, P. (2019). The Socioeconomic High-resolution Rural-Urban Geographic Dataset on India (SHRUG). Working paper.

Berry, S. and Reiss, P. (2007). Empirical models of entry and market structure. *Handbook of industrial organization*, 3:1845–1886.

Bresnahan, T. F. and Reiss, P. C. (1991). Entry and competition in concentrated markets. *Journal of political economy*, 99(5):977–1009.

Burgess, R. and Pande, R. (2005). Do rural banks matter? evidence from the indian social banking experiment. *American Economic Review*, 95(3):780–795.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2019). Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies*, 86(5):1867–1900.

Chernozhukov, V. and Hansen, C. (2008). The reduced form: A simple approach to inference with weak instruments. *Economics Letters*, 100(1):68–71.

Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828.

Galichon, A. and Henry, M. (2011). Set identification in models with multiple equilibria. *The Review of Economic Studies*, 78(4):1264–1298.

Kuehn, J. (2018). Spillovers from entry: the impact of bank branch network expansion. *The RAND Journal of Economics*, 49(4):964–994.

La Porta, R., Lopez-de Silanes, F., and Shleifer, A. (2002). Government ownership of banks. *The Journal of Finance*, 57(1):265–301.

Sanches, F., Silva Junior, D., and Srisuma, S. (2018). Banking privatization and market structure in brazil: a dynamic structural analysis. *The RAND Journal of Economics*, 49(4):936–963.

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *The RAND Journal of Economics*, 37(3):619–640.

Thorat, U. (2009). *Report of the High Level Committee to Review Lead Bank Scheme*. Reserve Bank of India.

Young, N. (2019). Banking and growth: Evidence from a regression discontinuity analysis. Working paper.

Appendices

Appendix: Geographic Scope of Banks

Indian banks are typically limited in geographic scope. Two of the largest public sector banks, Punjab National Bank and Canara Bank, are concentrated in Northern and Southern India, respectively. The share of a bank's network in a district in 2005 is a strong predictor of subsequent entry, as shown in Fig. 7, with a correlation coefficient of 0.74. Incorporating this limited geographic scope into the model can aid in both identification and computation. Excluding some markets from banks' consideration sets functions as an (infinitely) powerful cost-shifter. This also allows me to limit the number of potential entrants in each market, and thus the number of possible equilibria. Since the complexity of enumerating Nash equilibria increases exponentially in the number of potential entrants, this is allows me to expand the number of banks overall and substantially improve the realism of the model.

I define the set of active banks at the market level for cities with 2011 Census population over 500,000. For smaller markets, it are defined at the district level. In what follows, the decision rule for defining active banks is the same for both metro areas and smaller markets within a district.

I first restrict attention to 12 "major" banks, which are either in the top quintile by branches at the end of FY05-06 or by entry between FY05-06 and the end of FY10-11.[14] Together, these banks represent 69.4% of the 2005 branch network and 73.6% of entry during the 5 years prior to the reform (2006-2010)

I then assign banks to be active in minimal set of districts representing 75% of their pre-existing branches by computing the cutoff for each bank such that districts with more branches that that cutoff represent at least 75% of the banks' total branches and designating the bank as active in those districts. I use the analogous procedure so

[14] These banks are: ['ALLAHABAD BANK', 'UNION BANK OF INDIA', 'CENTRAL BANK OF INDIA', 'BANK OF INDIA', 'STATE BANK OF INDIA', 'PUNJAB NATIONAL BANK', 'BANK OF BARODA', 'CANARA BANK', 'SYNDICATE BANK', 'HDFC BANK LTD.', 'ICICI BANK LIMITED', 'AXIS BANK LIMITED']

that in each district, banks representing at least 75% of the district's 2005 branches are active.

The active bank assignments from these two procedures frequently overlap, but are complementary. The first assignment ensures that small banks are treated as active in districts where they are concentrated, even if they are a not a large share of total branches there. The second assignment ensures that there are a sufficient number of active banks in small districts which are not "important" for any given bank.

These two steps assign active banks to districts that cover 90.2% of pre-existing branches and 74.1% of '06-'10 entry by the 12 major banks. The mean and median active banks per district or metro area are 3.9 and 3, respectively. However, there are 100 districts with only 1 active bank. Given that computational complexity grows exponentially in the number of active banks, it is relatively easy to accomodate additional active banks in markets with few active banks. Furthermore, private sector banks were growing rapidly during this period.

Therefore, I include a final step: For any district with 3 or fewer active banks according to the two above rules, I designate the private sector bank with the largest presence in that state as active.

The resulting distribution of active banks per district/metro area is shown in Figure 8. Table X shows the number of districts in which each bank is active. Active banks cover 90.4% of preexisting branches and 76.1% of subsequent entry.

To summarize the procedure:

1. Select the 12 major banks, in the top quintile either by pre-existing branches or by entry.

2. Give each bank the districts (or metro areas) representing 75% of its preexisting branches

3. Give each district the banks representing 75% of its preexisting branches

4. For districts with ≤ 3 active banks according to 1) and 2), designate the largest private sector bank in that state as active

Appendix: Model Solution / Simulations

Enumerating equilibria

The algorithm I use for computing all Nash equilibria of the entry game is as follows. For each market:

1. Compute the J-vector of bank profits, $\pi(Y_i, X, \theta)$, for each of the 2^J possible market structures, denoted Y_i.

2. For each i, check the Nash Equilibrium conditions: $\pi_i(y^i_j, y^i_{-j}) \geq \pi_i(1 - y^i_j, y^i_{-j})$ for each bank j

3. Return a 2^J vector E, where $E_i = 1$ if i is an equilibrium and 0 otherwise.

Appendix: Identification

In this section I discuss identification of the counterfactual outcomes in the context of a 2-player entry game.

Suppose that branch profits take the form:

$$\pi_i(y_i, y_{-i}; X, W, Z, \epsilon) = y_i \left[f(X) + y_{-i} g(X, Z) + W_i + \lambda_i + \epsilon_i \right] \tag{8}$$

where $y_i \in \{0, 1\}$ and X, Z, and W are vectors of characteristics, f and g are unknown functions, λ_i is a bank-specific fixed effect, and ϵ_i is an unobserved mean-zero profit shock, potentially correlated across players.

The counterfactuals of interest involve translations of branch profits, equivalent to manipulating λ_i. For example, the conditions for the constrained equilibrium shift the entry threshold from 0 to the equilibrium compliance cost.

The object we are trying to identify is bounds on the probability of any given equilibrium market outcome Y, conditional on observables X, Z, and W, as a function of the λ_i's.

Proposition: If $\epsilon \perp\!\!\!\perp W$, and W has large support, then the bounds $P_l(Y|X,Z,W;\lambda_i)$ and $P_l(Y|X,Z,W;\lambda_i)$ are identified from the joint distribution $f(Y,X,Z,W)$.

Proof:

Setting the coefficient on the cost shifter W_i to one is a scale normalization and therefore without loss.

First, consider the observed probability that neither bank enters the market. We have:

$$\begin{aligned} P(Y=(0,0)|X,Z,W) =& P(f(X)+\lambda_i+\epsilon_i \leq -W_i, \forall i|X,Z,W) \\ =& P(f(X)+\lambda_i+\epsilon_i \leq -W_i, \forall i|X,Z) \end{aligned}$$

where the second equality follows from the exogeneity of W.

Note that this is simply the conditional joint CDF of $f(X)+\lambda_i+\epsilon_i$. Large support for the W_i therefore allows us to trace out the entire (joint) distribution of $f(X)+\lambda_i+\epsilon_i$, conditional on X and Z. Exogeneity of X is not required, because market characteristics are held fixed in the counterfactuals. We can easily calculate $P(Y=(0,0)|X,Z,W;\hat{\lambda})$ by translating the original distribution by the new bank-specific means.

Next, we consider identifying each banks' profits when one of their competitors is active. The threat here is selection: a competitor's entry decision, y_{-i}, is potentially correlated with a banks' own profit shock, ϵ_i, because the profit shocks are correlated -

perhaps due to a common market-level component. In particular, $\mathbb{E}[\epsilon_i|y_{-i}, X, Z, W] \neq 0$.

The key observation is that as $W_{-i} \to \infty$, the probability that $y_{-i} = 1 \to 1$. Large support on the exogenous cost shifters provides observations where the probability of entry is arbitrarily large, breaking the correlation between y_{-i} and ϵ_{-i}, and therefore between y_{-i} and ϵ_i. This is a consequence of Bayes' Rule:

$$P(\epsilon_i|y_{-i} = 1) = \frac{P(y_{-i} = 1|\epsilon_i)P(\epsilon_i)}{P(y_{-i} = 1} \to \frac{1 * P(\epsilon_i)}{1} = P(\epsilon_i) \text{ as } W_{-i} \to \infty$$

Now, we have that:

$$P(y_i = 1, y_{-i} = 1|X, Z, W_i, W_{-i} \to \infty) = P(f(X)+g(X, Z)+\lambda_i+\epsilon_i \leq -W_i|X, Z, W_i, W_{-i} \to 1)$$

Given that we already know f, λ_i, and the distribution of ϵ, we can recover λ_i from the observed probability that $P(y_i = 1, y_{-i} = 1$. Note that multiple equilibria are not an issue, because the extreme values of W_{-i} ensures that $y_{-i} = 1$, so bank i effectively faces a single agent problem.

Counterfactuals

In the actual counterfactuals, the new bank-specific fixed costs λ_i are endogenous objects, which depend on the profitability of banked markets and the unprofitability of unbanked markets. The prior section establishes that we can identify the bounds on the entry probabilities, conditional on observed characteristics.

If the probabilities were known exactly (as opposed to bounded), solving for the equilibrium compliance costs, λ_i, would be a straightforward fixed point problem. The multiplicity of equilibrium implies that there may be no well-defined λ_i, and instead I need to calculate bounds on the equilibrium compliance costs.

The naive upper and lower bounds, obtained by taking the upper and lower bound for each bank i in each market, will not be tight, because it is not feasible for every bank to be at its upper bound simultaneously. I deal with this by [computing the bounds for a variety of equilibrium selection rules] and then reporting maximum and minimum counterfactual outcomes across these potential λ_i's.

Chapter 3

Technology Lock-In and Optimal Carbon Pricing

Abstract

This paper studies the implications of low energy prices today for energy efficiency and climate policy in the future. If adjustment costs mediate manufacturing plants' responses to increases in energy prices, incumbents may be limited in their ability to re-optimize energy-inefficient production technologies chosen based on past market incentives. Using U.S. Census data and quasi-experimental variation in state energy prices, we first show that the initial electricity prices that manufacturing plants pay in their first year of operations are important determinants of long-run energy intensity. Plants that open when the prices of electricity and fossil fuel inputs into electricity are low consume more energy throughout their lifetime, regardless of current electricity prices. We then measure the relative contributions of initial productivity and capital adjustment frictions to creating this "technology lock-in" by estimating a model of plant input choices. We cannot reject that aggregate lock-in is entirely explained by persistent differences in the relative productivity of energy inputs chosen at entry. This leaves limited scope, on average, for policies which seek to reduce lock-in by reducing capital adjustment frictions. We discuss how these long-run effects of low entry-year energy prices increase the emissions costs of delayed action on carbon policy.

3.1 Introduction

Does the lack of carbon pricing today mediate the effectiveness of carbon pricing in the future? Abundant fossil fuel resources priced below their social cost have set industrial economies on a path of energy-inefficient development and rising anthropogenic carbon emissions. Current global energy infrastructure comprises tens of trillions of dollars of assets and reflects two centuries of technological innovation—and approximately 80% of energy produced comes from burning fossil fuels that contribute to climate change (Seto et al., 2016). Climate change impacts such as extreme temperatures, hurricanes, and wildfires are now causing billions of dollars of economic damage annually, but carbon pricing policies intended to curb greenhouse gas emissions continue to face global opposition. In the United States, over fifty carbon pricing bills have been introduced by Congress in the last three decades; none has passed. Jurisdictions that have successfully implemented carbon pricing schemes, such as the European Union and Canada, struggle to set prices that fully internalize the social costs of energy consumption. Some policymakers have despaired at the political feasibility of such reforms, instead proposing alternative policies such as clean energy subsidies and technology standards (Shearer and Nace, 2010).

In the absence of such policies, global energy usage is projected to increase by more than 50% by mid-century. The largest consumer of this energy is the industrial sector, and the durable nature of capital means that many energy-inefficient manufacturing plants that open when energy is dirty and cheap will contribute to global emissions for many years (EIA, 2019). The increasing trend in energy usage is even steeper in developing countries such as India and China, which are opening the equivalent of one new coal power plant every week (Myllyvirta and Shearer, 2021). Carbon emissions from existing coal power plants are already 150% higher than permissible in optimistic climate scenarios that limit global temperature increases to 2 degrees

Celsius above pre-industrial levels—even before accounting for planned construction (Shearer and Nace, 2010).

This paper quantifies the extent to which the energy prices that manufacturing plants pay in their first year of operations determine their future energy usage and the outcomes of subsequent climate policy. When a plant enters the market, it chooses a combination of factor inputs to use in production based on entry-year prices and beliefs about future prices. We explore the extent to which these initial prices have persistent effects on long-run energy usage, which we refer to as "technology lock-in", and mechanisms for these effects. If adjustment costs mediate responsiveness to changes in input prices, incumbent plants may be limited in their ability to reoptimize their energy usage when energy prices increase. Such a constrained response could cause low energy prices today to undermine the effectiveness of future carbon pricing policies, and also increase the importance of technology subsidies to encourage turnover of energy-inefficient capital.

The first part of this paper provides empirical evidence of technology lock-in. We assess how both initial and contemporaneous electricity prices affect manufacturing plants' energy intensity, defined as energy use per dollar of revenue. We measure plants' energy intensities and input prices using restricted-access microdata from the U.S. Census of Manufactures (CMF) and the Annual Survey of Manufacturing (ASM) for the years 1976 to 2011. Since electricity prices may be correlated with other shocks to manufacturing plants' input demands, we use shift-share instruments to isolate plausibly exogenous price variation (Bartik, 1991; Goldsmith-Pinkham et al., 2020). The instruments exploit national changes in coal, natural gas, and petroleum prices, weighted by each state's use of these fuels to generate electricity in a base year (Ganapati et al., 2020). As an alternative measure of lock-in, we also directly examine how the prevailing prices of these fuels in plants' entry year affect subsequent energy intensity. We show that the results are robust to estimation using alternative energy intensity definitions, different data subsamples, and different energy data sources.

Motivated by this empirical evidence, the second part of the paper explores the extent to which this lock-in arises due to differences in production technologies chosen at entry. To do so, we estimate the parameters of plants' production technologies and the relative productivities of different manufacturing inputs at entry and in subsequent years. The model allows us to quantify the efficiency of plants' energy inputs relative to labor in each year of operations. Using these estimates, we assess whether entry-year energy prices lead to different initial production choices and whether any differences persist over time.

These analyses yield two primary results. First, technology lock-in is important in manufacturing production. We show that plants' entry-year electricity prices are significant determinants of current energy intensity, even conditional on current prices. While energy intensity declines when contemporaneous electricity prices increase, we estimate an initial electricity price elasticity of approximately -0.20—25% of the elasticity with respect to current electricity prices. In addition, we show that the entry-year prices of fuel inputs into electricity generation themselves have persistent effects on manufacturing energy intensity. Separate analysis of the contributions of the prices of different raw fuels reveals that entry-year coal and petroleum prices continue to be important determinants of energy use. Specifically, manufacturing plants established when coal and petroleum were cheap are still consistently more energy-intensive. The persistent effect of these fuel prices on manufacturing energy intensity is surprising because electricity generation in the U.S. is much less reliant on these fuels today. These findings underscore the long-run effects of development based on cheap fossil fuel energy and the emissions implications of expansion of fossil fuel power plants: dirty capital investments undertaken in response to current cheap coal prices around the world seem likely to lock in higher emissions levels in the future.

This lock-in has the potential to increase emissions if carbon pricing is delayed:

entrants who choose production technologies based on current prices choose dirtier technologies than they otherwise would, and cannot subsequently fully adjust them. We find limited heterogeneous effects of initial electricity prices by plant age, which suggests that these entry-year prices remain important throughout a plant's lifetime. Of course, energy-inefficient plants may close if prices increase substantially, which motivates using the model of plants' input choices to directly assess the extent to which plants can adjust their production processes over time.

The second main results show that persistent differences in the relative productivity of energy inputs appear to explain much of the effect of initial electricity prices on subsequent manufacturing energy use. We estimate that a 10% increase in entry-year electricity prices increases relative energy productivity by approximately 3% in subsequent years. Conversely, we find no evidence that entry-year electricity prices have long-term effects on total factor productivity. These results suggest that when electricity prices are low, new manufacturing plants chose production technologies that use energy inputs relatively less efficiently compared with their labor inputs, and these productivity differences persist even if electricity prices change in the future.

This paper seeks to make three primary contributions to existing literature. First, we believe that this paper provides the first estimate of the importance of entry-year energy prices for industrial energy intensity in subsequent years. In addition to identifying this technology lock-in, we explain how it arises. Previous "efforts to characterize the types and causes of carbon lock-in, or to quantitatively assess and evaluate its policy implications, have been limited and scattered across a number of different disciplines" (Seto et al. (2016), p. 425).[1] Our findings contribute to a growing literature on how different initial conditions mediate transitions from dirty to clean energy (i.e., path dependence). Several papers in this literature use macroeconomic

[1] In the climate context, the literature refers to technology lock-in as "the inertia of carbon emissions ... associated with the technologies and infrastructure that indirectly or directly emit CO2", which is distinct from carbon lock-in arising from behavioral or institutional constraints more commonly studied by sociologists (Seto et al. (2016), p.427).

dynamic models and more aggregate data to study incentives to develop clean energy technologies, typically simulating how changes in energy prices affect carbon emissions through innovation (Acemoglu et al., 2012; Acemoglu et al., 2019; Atkeson and Kehoe, 1999; Fried, 2018; Hassler et al., 2012). Other work uses microdata, particularly from the electricity sector, to show that initial regulatory structure and fuel mix choices (e.g., coal versus natural gas) are important determinants of subsequent fuel use (Cullen and Mansur, 2017; Knittel et al., 2015; Meng, 2021). One paper shows that entrant and incumbent manufacturing plants respond differently to current energy prices (Linn, 2008). We depart from these studies by quantifying the extent to which initial energy prices matter after a plant's entry year and by analyzing the contribution of initial energy efficiency and technology choices to creating this lock-in.

These dynamics are relevant for policy. Current U.S. government proposals earmark $400 billion for industrial energy efficiency improvements (DNC, 2021). Understanding whether lock-in exists and how it arises is necessary to predict the outcomes of this suite of policies and to efficiently design them. Ignoring the dynamic effect of current energy prices on energy use tomorrow underestimates the benefits of pricing carbon today.

Second, this paper contributes to literature that models the responses of industrial energy use and productivity to environmental regulation. Research using microdata to study energy implications of environmental policy typically analyze the dynamics of one industry (e.g., cement or electricity) over the long-run (Fowlie et al., 2016; Meng, 2021; Ryan, 2012; Clay et al., 2021) or use static models to study important contemporaneous effects across many industries (Ganapati et al., 2020; Greenstone et al., 2012; Shapiro and Walker, 2018). Our contribution is to bring these two literatures together to provide a new generalizable explanation for why some of the dynamic responses arise. Classic "putty-clay" models of capital investment emphasize that capital adjustment frictions may constrain changes in input mix, but we identify that productivity differences appear to be at least as important as this more common explanation for lock-in (Atkeson and Kehoe, 1999). Showing that entry-year electricity prices have persistent effects on the efficiency of manufacturing inputs requires estimating the relative productivity of these inputs over time. Since commonly used models of manufacturing production functions, such as Cobb-Douglas, assume away the possibility of complementarity between inputs that creates lock-in, studying the causes of persistent effects of entry-year energy prices requires extending more general models of production to include energy (Ackerberg et al., 2015; Demirer, 2020; Doraszelski and Jaumandreu, 2018; Olley and Pakes, 1996).

Finally, we provide a new microfoundation for the rate of decarbonization frequently used in standard models of climate-economy interactions. These Integrated Assessment Models (IAMs) are the basis for calculating the full social costs of carbon emissions and for evaluating national and international climate policy recommendations. Despite their widespread use in regulatory analyses, economists have criticized these models for allowing "a great deal of freedom in choosing functional forms, parameter values, and other inputs" and for "lacking transparency in key underlying assumptions, such as energy resource costs, constraints on technology take-up, and demand responses to carbon pricing" (Pindyck, 2020, p.863; Gambhir et al., 2019, p.5). Standard models extrapolate future rates of decarbonization based on past decarbonization trends, which may overestimate attainable emissions reductions if lock-in is important. We provide a novel estimate of the response of industrial carbon emissions to emissions constraints assumed in climate-economy model, which is "the most important calibration for policy purposes" (Nordhaus and Boyer, 2000, p.44).

The rest of this paper proceeds as follows. Section 3.2 provides background on energy use in U.S. manufacturing to contextualize the analysis. Section 3.3 presents a conceptual framework outlining how technology lock-in might arise. Section 3.4 describes the data and section 3.5 presents descriptive statistics and trends in energy use. Section 3.6 discusses our econometric model for identifying technology lock-in and Section 3.7 presents our empirical evidence of it. Section 3.8 discusses the implications for climate change policy. Section 3.9 concludes.

3.2 Institutional Background

Manufacturing accounts for about one-quarter of total U.S. energy consumption and one-quarter of total U.S. greenhouse gas emissions. Energy consumption in the industrial sector, which comprises manufacturing, mining, construction, and agriculture, is increasing both in absolute terms and as a share of total consumption, and this sector accounts for almost all of the predicted increase in U.S. energy use in the next decade (EIA, 2015; EPA, 2021). Most manufacturing energy is consumed as electricity; a subset of manufacturing plants use raw fuels, such as coal, natural gas, and petroleum, as direct inputs. On average during this study's time period, electricity expenditures account for approximately 75% of total energy expenditures and 95% of thermal energy consumed (measured in British thermal units, or BTUs). Only 0.1% of this electricity is produced on-site. By contrast, manufacturing plants in developing countries such as India are typically more reliant on raw fuel inputs and on-site generation of electricity (Allcott et al., 2016).

Although total U.S. manufacturing energy consumption has increased, energy intensity of production has declined during the past thirty years. The adoption of more energy efficient technology by new manufacturing plants explains some of this decline, while energy prices and energy efficiency regulation are weakly correlated with energy efficiency improvements in aggregate (Levinson, 2021; Linn, 2008). Despite recent entrants' higher energy efficiency, manufacturing energy policy typically does not differentially regulate plants depending on their entry date.[2] Manufacturing en-

[2] Vintage-differentiated energy efficiency regulations are more common in other sectors, such transportation and construction (Jacobsen and Kotchen, 2013; Levinson, 2021; Stavins, 2006; West

ergy efficiency is administered by a mix of federal, state, and local governments that usually target specific industries or technologies. Oregon, for example, offers subsidies for the installation of energy efficient manufacturing capital.[3] Landmark federal industrial environmental regulations, such as the Clean Air Act, more commonly target pollution that is the by-product of energy use rather than targeting energy efficiency directly (NREL, 2009).

The amount and type of energy used depend on plants' production processes. Primary uses include powering production machinery and fueling boilers, while secondary uses include heating and cooling, lighting, on-site transportation, and direct inputs into the finished product (Ganapati et al., 2020). Improving the energy of efficiency of many of these processes requires replacing equipment or machinery. For example, upgrading an energy inefficient turbine involves pausing or re-arranging operations to install an expensive replacement, and such capital adjustment costs create the possibility for technology lock-in. If energy inefficient machinery is installed when energy prices are low, incurring these adjustment costs to replace it may only be worthwhile if energy prices increase substantially.

Though raw fuels account for a small portion of direct energy inputs, the production of the electricity consumed by manufacturing involves important indirect use of raw fuels. At the start of our sample in 1976, electric utilities in the U.S. generate electricity using coal (40%), natural gas (12%), petroleum oil (21%), hydro (15%), and other renewable sources. Natural gas and renewables (e.g., solar) have become more important in the last two decades, with a reduction in the use of petroleum and, to a lesser extent, coal. Appendix Figure C.1 shows that the contribution of these different fuel sources to electricity generation varies widely across the U.S. Electric utilities have distinct regional markets that typically comprise a few states, and in 2011 industrial users paid between 0.04 and 0.28 dollars per kWh for electricity on av-

et al., 2017).

[3] See NREL (2009) for a detailed review of federal, state, and local energy efficiency policies.

erage (EIA, 2020). Local electricity rates depend on the national prices for prevailing fuel inputs and distances to procurement sources. In what follows, we exploit variation in the national prices of these raw fuels to construct instruments for electricity prices.

3.3 Conceptual Framework

In this section, we show how technology lock-in operates in a stylized "putty-clay" model of the manufacturing sector (Atkeson and Kehoe, 1999), to which we introduce the importance of entry-year input productivities for energy use. We define the *energy intensity* of a plant as the ratio of energy inputs to output, $\frac{E}{Y}$. Technology lock-in is the elasticity of current energy intensity with respect to entry-year energy prices, conditional on current energy prices. For new entrants, current energy intensity depends on current prices only.

In our model, only new entrants can flexibly choose all inputs without adjustment costs. Incumbent plants can adjust their energy intensity in response to price changes through three margins. First, they can change their static inputs, such as energy and labor, which are chosen in each period. Unlike energy and labor, plants' capital inputs are subject to adjustment costs. The adjustment of potentially sticky capital stocks is the second margin through which plants adjust their energy intensity. Finally, plants enter or exit on the basis of differences in production technology, including the relative productivity of energy inputs. Changes in energy prices will change the composition of plant productivity within entry cohorts, and therefore average energy intensity.[4]

The first margin of adjustment—static reoptimization—operates in the short-run

[4] The energy productivity shocks in our model play a similar role to vintage capital effects in the classic putty-clay model of Atkeson and Kehoe (1999). Our model weakens the assumption of perfect complementarity between capital and energy. More importantly, it emphasizes plant-level cross-sectional differences in productivity which arise through entry and exit decisions, as opposed to differences in the energy efficiency of capital within an individual plants' capital stock.

even for small, temporary fluctuations in relative prices. The other two, capital adjustment and energy productivity, prevent full reoptimization and are sources of technology lock-in. If capital and energy are complementary in the plant's production function, incomplete adjustment of capital stocks attenuates the response of energy intensity to changes in energy prices, leading to lock-in. Empirically, capital investment is characterized by infrequent spikes interspersed with periods of no or minimal investment. Investment is also slow to respond to large changes in economic fundamentals. These stylized facts suggest that both fixed and convex capital adjustment costs are important (Cooper and Haltiwanger, 2006; Khan and Thomas, 2008).

Some plants may also use production technologies which are more energy-efficient than others, leading to lock-in based on productivity differences. Given their expectations of future energy prices, plants choose to enter at a given level of energy productivity. This induces selection. In periods where energy is cheap, it will be profitable for plants with low energy productivity to enter. Since entry costs are sunk, these energy-inefficient plants may continue to operate even when energy prices rise. This generates lock-in by vintage: if energy prices increase, the *average* entrant will have a lower energy intensity than the average incumbent not only because they can flexibly choose their level of capital, but because they have higher energy productivity.

To illustrate how such lock-in might arise, we characterize these three margins of adjustment in a simple two-period model of myopic manufacturing plants. We then describe how these intuitions carry over into the richer model that we estimate.

3.3.1 A Model of Lock-In

To fix ideas, suppose the plant has a constant elasticity of substitution (CES) production technology:

$$Y(K, L, E; \beta) = \alpha \left(K^\rho + L^\rho + (\beta E)^\rho \right)^{\frac{\nu}{\rho}} \tag{3.1}$$

where K, L, and E are the quantities of capital, labor and energy inputs, respectively, $\sigma = \frac{1}{1-\rho}$ is the elasticity of substitution between energy and capital, $\nu \in (0,1]$ is a returns to scale parameter, α is total factor productivity, and β captures the productivity of energy relative to labor.[5] Labor and energy are assumed to be fully flexible static inputs, chosen optimally in each production period. As detailed below, an incumbent plant's level of capital is only partially flexible due to non-linear adjustment costs. The relative productivity of energy, β, is fixed at entry and is fully locked in.[6]

A potential entrant i draws productivity levels α_i and β_i and solves the static optimization problem:

$$\max_{K,L,E} \pi(K,L,E;\alpha_i,\beta_i) = pY(K,L,E;\alpha_i,\beta_i) - rK - wL - p_e E$$

The potential entrant chooses to enter if profits exceed the fixed costs of entry, that is if:

$$\pi^*(p,w,r,p_E;\alpha_i,\beta_i) = pY(K^*,L^*,E^*) - wL - rK - p_E E \geq FC$$

In this equation, L^* and E^* solve the static profit maximization problem and FC is the fixed entry cost. We assume that the capital stock is fully flexible on entry, subject to a linear cost of capital investment. To clarify the role of the relative price of energy p_e in this example, we set the price of output p, the wage w, and the rental rate of capital r to be equal to one.

The first channel through which lock-in arises is the selection effect on the productivity of plants that choose to enter in the first period. Because profits are monotonically increasing in energy productivity, the potential entrant's problem yields a cutoff rule, where all else equal plant i enters if its energy productivity is sufficiently

[5] We estimate a CES production technology rather than the more common Cobb-Douglas function to allow for factor-specific productivities and complementarity between inputs. The Cobb-Douglas specification is equivalent CES where $\rho = 0$. In this case, input expenditure shares are fixed and capital stocks do not affect the optimal energy input.

[6] In the empirical specification, we allow total factor and energy-specific productivities, as well as prices, to evolve over time following AR(1) processes. We discuss estimation details in Section 3.6.2.

high. That is, if:

$$\beta_i \geq \beta^{entry}(p_E)$$

The cutoff, $\beta^{entry}(p_E)$, determines the distribution of energy productivity β_i for plants that entered in a period with energy price p_E. Because the cutoff is increasing in p_E, the cohort-average energy productivity will also be increasing in the initial energy price.[7]

Capital adjustment costs provide the second channel through which lock-in arises. Plants that enter earn their period one profits and continue to the second stage with their current capital stock, K. At the end of period one, plants observe prices in the next period and choose their capital in the next period, K', to solve

$$\max_{K'} \begin{cases} \pi^*(K'; p'_E, \beta_i) - \gamma_0 - r(K' - K) - \gamma_1(K' - K)^2 & \text{if } K' \neq K \\ \pi^*(K; p'_E, \beta_i) & \text{otherwise} \end{cases}$$

Here, $\pi^*(K; p_E, \beta)$ is the maximum profit holding capital fixed at K given prices and productivity. In addition to the cost of capital r, γ_0 and γ_1 are fixed and convex adjustment costs, respectively. The fixed cost to capital adjustment, γ_0, implies that plants will not reoptimize capital at all for marginal changes in the energy price. The convex adjustment cost, γ_1, implies that while plants may invest in response to larger prices changes, they will only partially close the gap relative to frictionless entrants, because large capital investments are increasingly more costly than small ones. One implication of this is that, without policies such as technology subsidies, plants with both fixed and convex adjustment costs may never reach the optimal level of energy intensity.

[7] We can see that $\beta^{entry}(p_E)$ is increasing in p_E because, algebraically, we can write Y in terms of "effective energy", $\hat{E} = \beta E$, and the price of an effective unit of energy will be $\frac{p_e}{\beta}$. This implies that if $p'_E > p_E$, then the distribution of β_i conditional on entry at price p'_E first-order stochastic dominates the distribution of β_i conditional on entry at price p_E. This, in turn, implies that $\mathbb{E}[\beta_i|$ entered at $p'_E] > \mathbb{E}[\beta_i|$ entered at $p_E]$

Incumbent plants will shut down if their scrap value exceeds their profit: $\pi^* < S$. As with entry, there is a cutoff value $\beta^{exit}(p_E)$ such that for a given energy price p_E plants with energy productivity below $\beta^{exit}(p_E)$ will exit.[8] As the least efficient plants exit, cohort-average energy productivity will rise as p_E increases. However, if scrap values are lower than entry costs, $\beta^{exit}(p_E) < \beta^{entry}(p_E)$, and incumbents will, on average, have higher energy intensity than new entrants.

Figure 3.1 plots simulated current energy intensity as a function of energy prices at entry, relative to the energy intensity of a fully flexible entrant. The blue, long-dashed line plots the magnitude of lock-in for plants which cannot adjust their capital stock. The only margins of adjustments are exiting or changing energy and labor inputs. This represents an extreme case of lock-in. The orange, short-dashed line plots lock-in for plants which can partially adjust capital, subject to both fixed and convex capital adjustment costs. The green, solid line isolates the energy productivity effect by setting $\gamma_0 = \gamma_1 = 0$, shutting down capital adjustment frictions. Even without fixed or convex capital adjustment costs, the average plant from a low energy price vintage will be more energy intensive than the average new entrant. The gap is due to the difference between the entry cost and the scrap value. At higher energy prices, it is no longer profitable to open a new energy-intensive plant, but existing plants may continue to operate and pollute.[9]

We conclude this section with a brief discussion of how insights from this highly stylized example carry over into more general models. For energy productivity to play a role, we require selection at entry and persistence over time. Partial irreversibility of entry costs or capital investments is one natural way to generate more intense

[8] In the dynamic model used in the empirical application, we need only substitute the present discounted expected value of future profits, V, for profits, π^*. Since V inherits the same qualitative properties of π^*, all of these results will go through.

[9] Figure 3.1 illustrates the lock-in that arises when plants are myopic. The other extreme, where plants have perfect foresight regarding future prices, looks qualitatively similar but with a smaller difference between the energy intensity of incumbents who entered at high and low energy price (i.e., a less steep slope of energy intensity). Discounting future energy price changes creates lock-in even in the presence of perfect foresight.

selection for entrants than incumbents. Large exit subsidies or buyouts for low-energy productivity incumbents might result in a higher productivity threshold for exit than for entry, which would result in the opposite sign for our estimated entry-year energy price elasticities.

For capital adjustment frictions to generate lock-in, it is sufficient that capital and energy are complements in production. The intuition is that an increase in energy prices lowers the marginal product per dollar of energy and causes the optimal energy input to decrease. If capital and energy are complements, this will decrease the marginal product per dollar of capital and, by extension, the optimal capital stock. If capital can optimally adjust, this drives further decreases in energy inputs. Incomplete capital adjustment will attenuate this change and result in higher energy intensity than for a fully flexible plant. If capital and energy are substitutes in production instead, this logic would be reversed and capital adjustment costs would increase plants' sensitivity to current price changes. Empirically, capital and energy are typically estimated to be complements (Hassler et al., 2012; Ryan, 2018).

3.3.2 From Theory to Data

In the remainder of the paper, we exploit exogenous variation in current and initial electricity prices to measure the persistent effect of electricity prices at entry. Technology lock-in is important if plants facing the same current electricity prices have systematically higher energy intensity if they entered in years when electricity was less expensive. This overall estimate of lock-in is analogous to the orange, short-dashed line in Figure 3.1, which captures lock-in due to both capital adjustment frictions and persistent energy-specific productivity.

The regressions of energy intensity on electricity prices cannot, by themselves, distinguish between these two sources of lock-in. To do so, we estimate a structural production function for each industry and recover the energy-specific productivity shocks for each plant. This allows us to measure the contribution of energy-specific productivity differences to lock-in, which corresponds to green, solid line in Figure 3.1. By comparing the total effect with the productivity estimates, we quantify the relative importance of these two mechanisms. The contribution of capital adjustment costs is then analogous to the residual distance between the green, solid line and the orange, short-dashed line.

3.4 Data

We draw on restricted microdata from the U.S. Census Bureau on manufacturing inputs and outputs and on energy data from publicly available government sources. Additional data details are in Appendix 3.10.

3.4.1 Manufacturing Inputs and Outputs

Our primary sources of data are administrative records on annual plant-level inputs and outputs from the Annual Survey of Manufacturing (ASM) and the Census of Manufactures (CMF) from the U.S. Census Bureau for the years 1976 to 2011. The CMF is conducted in years ending with 2 or 7 and surveys all manufacturing plants in the United States. The ASM annually surveys plants in the years between censuses and comprises a nationally representative sample of approximately 50,000 establishments per year. These surveys report quantity of electricity purchased and expenditures on electricity and raw fuels (e.g., coal, natural gas, petroleum oil) separately. We calculate each plant's annual average electricity price as reported total electricity expenditure divided by electricity purchased.[10] We measure plants' annual capital investment using total capital outlays, materials, electricity, and raw fuels

[10] We verify the reliability of our calculated average electricity prices by comparing against utilities' posted industrial rate schedules, available from the OpenEI rate database, and against state-level electricity prices reported by the Energy Information Administration.

inputs using reported expenditures, and labor inputs using worker hours, available in both the ASM and CMF.[11] The CMF also contains information on plants' capital stocks, measured as reported book values of equipment and machinery.[12]

We supplement these data with the Manufacturing Energy Consumption Survey (MECS) and the ASM Fuel Trailers. Together with the ASM and CMF, these surveys allow us to calculate three measures of energy intensity of production: electricity consumed per dollar of revenue, carbon dioxide (CO2) emissions produced per dollar of revenue, and British thermal units (BTU) of energy consumed per dollar of revenue. The MECS and ASM Fuel Trailers include a probabilistic sample of about 15,000 manufacturing plants, for the years 1976-1981 for the ASM Fuel Trailers and for every three years between 1985 and 1994 and every four years thereafter for the MECS. These more detailed energy surveys provide breakdowns of expenditure on and quantity consumed of raw fuels that are not available from the ASM and CMF, which report detailed quantity and expenditure information on electricity but not other energy sources. We calculate plant-level CO2 emissions and BTU consumption from electricity directly from the ASM and CMF using conversion factors from the U.S. Energy Information Administration (EIA) and from eGRID, which incorporates the carbon intensity of each state's electricity grid. To obtain total CO2 emissions and BTU consumption including raw fuels, we use plant-level annual raw fuel expenditures times industry-average estimates of energy consumption per dollar of raw fuels expenditures from the MECS and the ASM Fuel Trailers (Lyubich et al., 2018). We calculate CO2 emissions and BTU consumption per dollar of raw fuels expenditure by converting quantities of raw fuels into common units using fuel-specific conversion factors from the EIA and the Environmental Protection Agency (EPA). In years in which neither the ASM Fuel Trailer or the MECS surveys are conducted, we linearly

[11] We calculate worker hours as plants' reported production-worker hours times the ratio of total payroll to payroll for production workers (Ganapati et al., 2020; Baily et al., 1992)

[12] Appendix 3.10 describes how we calculate annual capital stocks implied by ASM investment and depreciation.

interpolate these coefficients by six-digit North American Industry Classification System (NAICS) industry. Estimating total BTU consumption and CO2 emissions allows us to measure energy intensity using BTU per dollar of revenue and CO2 per dollar of revenue—measures which incorporate the use of raw fuels in a way that electricity intensity does not.[13]

Our final source of manufacturing data is the Longitudinal Business Database (LBD). This census provides information on all plants' entry year, which we link to the other data sets using unique plant identifiers. We match plants to their own initial electricity prices using these plant identifiers if the plant was surveyed in its initial year of operations. If a plant is not observed in its entry year, we impute its initial electricity price using the average of other contemporaneous entrants in its state and industry where possible, or simply the same year and state if there are no other contemporaneous entrants in its industry. A short-coming of the LBD is that any plant that began operations before the start of the survey (i.e., 1975) is recorded as opening in 1975; we therefore restrict the sample to plants that enter after 1975, for which we observe their entry year. Appendix 3.10 describes additional restrictions imposed during the cleaning of the data, such as excluding observations with missing or negative input values. The primary analysis sample includes approximately 1,294,000 plant-year observations. Throughout, we deflate all monetary values to 2011 dollars using the input- and industry-specific price indices available from the National Bureau of Economic Research-Census of Economic Studies (NBER-CES) Productivity Database.

[13] While measuring energy intensity using CO2 and BTU per dollar of revenue has the benefit of incorporating use of raw fuels, the more intermittent measurement of raw fuels use means that the time series of these energy intensity measures discussed in Section 3.5 are noisier.

3.4.2 State Energy Use and Fuel Prices

The data on state energy input prices and fuel shares in the electricity sector are from the EIA State Energy Data System (SEDS) (EIA, 2020). We use these data to calculate average national prices for coal, natural gas, and petroleum as well as the share of each of these fuels used to generate electricity in each state. We also draw on state-level measures of electricity prices to assess the validity of our calculated, plant-level ones. We deflate prices using the average of the energy deflators from the NBER-CES Database.

3.5 Trends in Energy Use and Prices

This section reports descriptive statistics and discusses trends in energy intensity and the productivity of energy relative to labor. We highlight trends in energy intensity using microdata for a longer time period than previous studies (Linn, 2008; Levinson, 2021; Huntington, 2010; Metcalf, 2008) and estimates of the trend in relative energy productivities based on less aggregated data (Hassler et al., 2012).

Appendix Table C.1 presents summary averages on manufacturing inputs and outputs separately for all industries and excluding industries which use energy sources other than electricity in important ways (i.e., including only industries for which electricity accounts for at least 70% of total energy expenditures). Overall, plants consume approximately 0.2 kWh of electricity, 0.1 kg of CO2, and 0.001 million BTU per dollar of revenue, with about 10% higher energy usage in the electricity-intensive subsample. On aggregate, current electricity prices are slightly lower than prices paid in plants' entry year.

These summary averages mask important heterogeneity in energy prices over the 1976-2011 time period. Figure 3.2 shows that electricity prices paid by the industrial sector vary widely, generally trending downward until the late 1990s before increasing back to their 1976 level. These changes in electricity prices track the trends in the prices of raw fuels used to generate electricity, shown in Appendix Figure C.2. Since 1976, petroleum prices have tripled, while coal and natural gas prices have risen less steeply over this same time period. These fuel price increases appear to have contributed to important changes in the mix of fossil fuel mix used to generate electricity. Appendix Figure C.3 shows that the contributions of coal, natural gas, and petroleum to generating electricity vary substantially across the U.S. at the start of our sample in 1976, while Appendix Figure C.1 shows that this distribution has changed over the past four decades.[14] As a plausible consequence of the rising price of oil, the use of petroleum in electricity generation has declined almost everywhere and is barely used at all today. Coal use has also declined, though less steeply than oil, while natural gas generation has increased substantially after the fracking boom in the 2000s.[15]

Manufacturing energy intensity has also changed in the last four decades. Figure 3.3.a shows that aggregate electricity intensity has declined by approximately 30% since 1976, with comparable changes in CO2 and BTU intensities.[16] Some of this reduction is attributable to energy efficiency improvements, while manufacturing has also shifted toward producing less energy-intensive products locally and more energy-intensive goods abroad (NAM, 2014). Entrants have also adopted more energy efficient technologies over time (Linn, 2008). Of course, if lock-in is important, then we expect that some of this decline could also be driven by the exit of more

[14] Appendix Figure C.4 summarizes the fuel mix changes at the state-level.

[15] We focus on electricity prices as opposed to composite indices of electricity and any raw fuels used for two reasons. First, plant-level prices of these inputs are available only approximately every four years, for a small subset of our full sample. As a result, we almost never observe entry-year raw fuel prices, which require that a plant is surveyed in the ASM Fuel Trailer or MECS in its entry year. Second, electricity accounts for over 95% of BTUs of energy consumed on average and therefore captures most energy used.

[16] Consistent with out results, Linn (2008) and Levinson (2021) document declining energy intensity of manufacturing production over approximately half of our time period. Huntington (2010) and Metcalf (2008) additionally analyze sector- and state-level data, respectively. The CO2 and BTU intensity measures in Figure 3.3.a are more highly variable since these are surveyed less frequently and on fewer plants than the electricity measures, and also reflect the changing composition of inputs into electricity.

energy-intensive plants that entered the market at low energy prices.

Conversely, the relative productivity of energy inputs shows no significant trend over most of this time period. The time series of estimated energy productivity relative to labor productivity, shown in Figure 3.3.b, is relatively constant, with a decline beginning in the mid-2000s. This trend implies similar growth in the productivities of labor and energy inputs over much of this time period.[17] Meanwhile, the total factor productivity trend in this figure shows that the productivity of all inputs has more than doubled over this time period. This result is consistent with prior work using conducted over shorter time periods using similar data (Greenstone et al., 2012).

Overall, this discussion highlights that manufacturing plants beginning production in different years face very different initial electricity prices. We now test whether these price differences have led to persistent differences in energy intensity and productivity.

3.6 Econometric Model

3.6.1 Instrumental Variables Analysis

In this section, we discuss how we assess whether the electricity prices that manufacturing plants pay in their entry year are important determinants of subsequent energy usage and relative energy productivity. We estimate the following equation:

$$y_{it} = \beta_0 p_{it_0} + \beta_1 p_{it} + \alpha_{js} + \tau_{jtt_0} + \epsilon_{it} \tag{3.2}$$

In this equation, y_{it} is an energy outcome for plant i in year t (i.e., the log of energy use per dollar of revenue $\frac{E}{R_{it}}$ or the log of relative energy productivity ω_{it}^{E}), p_{it_0} is the

[17] Doraszelski and Jaumandreu (2018) find that labor productivity increased by roughly 40% relative to materials inputs in aggregate using data from Spanish manufacturing plants from 1991 to 2006; though this other study is conducted in a different context, an implication is that energy productivity may have grown more than other materials inputs.

log of the average price of electricity in the year t_0 that plant i enters the market, and p_{it} is the log of the average price of electricity paid by plant i in year t. Industry × state fixed effects α_{js} control for time-invariant characteristics common to industry j in a given state s, such as geography, industry × year × entry year fixed effects τ_{jtt_0} control for time-variant changes that affect all plants in a given industry that entered the market in the same year, such as new regulation, and ϵ_{it} is the error term. We cluster standard errors at the state-level throughout and we weight regressions using the Census sampling weights.

The main parameter of interest in equation (3.2) is β_0, which measures the effect of initial electricity prices on current energy intensity or current (relative) energy productivity. The second parameter of interest, β_1, measures the effect of contemporaneous electricity prices on these outcomes. If technology lock-in is important, we expect initial electricity prices to affect current energy usage $\frac{E}{R_{it}}$ even conditional on current electricity prices (i.e., $\beta_0 < 0$ in models where energy intensity is the outcome variable). In addition, if lock-in arises through persistent differences in the relative productivity of inputs, then we also expect higher initial electricity prices to lead to higher productivity of current energy inputs relative to labor inputs ω_{it}^E (i.e., $\beta_0 > 0$ in models where relative energy productivity is the outcome variable).

Even conditional on the fixed effects, it is possible that omitted variables or measurement error could introduce bias into the OLS estimation of the price elasticities β_0 and β_1. For example, classic reverse causality would arise if unobserved shocks to plants' aggregate energy demand (e.g., new demand for certain products) also affect electricity prices, leading to estimates of the price elasticities that are biased upward (i.e., less negative). In addition, plants' entry-year electricity prices are, in some cases, measured with error: if a plant is not surveyed in its entry year, we approximate its initial electricity price using the average of other entrants in the same state, industry, and year. As a result, the effect of entry-year prices may be biased toward zero.

To address these concerns, we construct instrumental variables Z_{st} to isolate changes in plants' electricity prices that are uncorrelated with other shocks to energy intensity. These Bartik-style shift-share instruments isolate exogenous variation in electricity prices using the interaction of historical state electricity generation shares and current national fuel prices (Ganapati et al., 2020). Specifically, the instruments are:

$$Z_{st} = [\rho_{-s,f,t} \times \sigma_{s,f,1976}] \tag{3.3}$$

where $\sigma_{s,f,1976}$ is the share of total fuel expenditure of each fuel in electricity generation in state s in 1976, for each fuel $f \in \{\text{coal, natural gas, petroleum oil}\}$, and $\rho_{-s,t,f}$ is the mean of all other states' log fuel price in year t. The intuition is that a plant's electricity price will be more strongly affected by changes in national fuel prices if the electricity sector in its state is more dependent on this fuel source. Appendix Figure C.2 shows that there is significant variation in the prices of these fuels between 1976 and 2011. We find that these instruments are strong predictors of electricity prices (Table 3.1).[18]

The identifying assumption is that plants' differential exposure to changes in national fuel prices are uncorrelated with other production shocks, conditional on the variables in the model:

$$\mathbb{E}[Z_{st} \times \epsilon_{it} | \alpha_{js}, \tau_{jtt_0}] = 0 \tag{3.4}$$

For example, the inclusion of industry × year × entry year fixed effects controls for annual macroeconomic conditions that could affect both plant's production choices and national fuel prices.[19] The identifying assumption would be violated if states' fuel

[18] We focus on electricity generation shares from fossil fuels that are traded in commodity markets, as opposed to fuels without clearly defined market prices (e.g., hydro and nuclear generation).

[19] Specifically, industry × year × entry year fixed effects control for e.g., annual shocks that are common to all cement plants that opened in 1990. The geographic clustering of entrants in specific industries reduces concern about exposure to state × year variation since our instrument precludes the inclusion of state × year fixed effects. Appendix Tables C.7 and C.8 shows that the results are robust to the inclusion of state × year trends.

generation shares in 1976, which determine exposure to national fuel prices changes, are correlated with other factors that affects plants' production decisions. The availability of skilled labor, for instance, is one such factor that could be correlated with shocks to plants' labor demand. We assess the validity of the identifying assumption by examining whether state fuel electricity generation shares are correlated with state characteristics that could suggest other channels through which the instruments could affect the outcomes of interest (Goldsmith-Pinkham et al., 2020).[20] Reassuringly, Appendix Table C.3 shows no evidence of significant systematic relationships between state fuel shares and these characteristics, which supports the identifying assumption.[21] Appendix 3.10 discusses this test of instrumental exogeneity in greater detail.

In equation (3.2), both current electricity prices p_{it} and initial electricity prices p_{it_0} are potentially endogenous. We therefore include instruments Z_{st} based on the contemporaneous fuel prices measured at t as instruments for log current prices p_{it} and Z_{st_0} based on the fuel prices in the year t_0 when the plant opened as instruments for log initial prices p_{it_0}. Specifically, the first stage regression equation for current prices p_{it} is:

$$p_{it} = \gamma_1 Z^{coal}_{-s,t} + \gamma_2 Z^{gas}_{-s,t} + \gamma_3 Z^{oil}_{-s,t} + \gamma_4 Z^{coal}_{-s,t_0} + \gamma_5 Z^{gas}_{-s,t_0} + \gamma_6 Z^{oil}_{-s,t_0} + \alpha_{js} + \tau_{jtt_0} + \psi_{it} \quad (3.5)$$

and the first stage regression equation for initial prices replaces p_{it_0} as the outcome variable.

In some specifications, we also examine whether the importance of initial prices

[20] Jaeger et al. (2019) highlight the importance of controlling for dynamic adjustments to past shocks when using Bartik-style instruments for causal inference. Our inclusion of both initial and current electricity prices in the regression equation (3.2) addresses this issue.

[21] Data on state characteristics are from the Federal Reserve Bank of St Louis database (FRED) and the 5 percent sample of the Integrated Public Use Microdata Series (IPUMS) of US Census Data. We examine the correlation of fuel shares with state characteristics in 1980, rather than in 1976 when our Bartik weights are measured, because 1980 is the closest year for which American Community Survey data from IPUMS are available.

depends on the plant's age. To do so, we extend equation (3.2) by interacting the log of initial electricity prices with the age of the plant in years:

$$y_{it} = \beta_0 p_{it_0} + \beta_1 p_{it} + \beta_3 p_{it_0} \times age_{it} + \alpha_{js} + \tau_{jtt_0} + \epsilon_{it} \tag{3.6}$$

In these heterogeneous effects models, we also include the interaction of the shift-share instruments Z_{st_0} with the variable age in the first stage.

3.6.2 Production Function Estimation

In this section, we estimate a model of plants' production decisions to separately recover plants' total factor and energy-augmenting productivity shocks. We apply approaches measuring relative labor productivity to energy (Demirer, 2020; Doraszelski and Jaumandreu, 2018).

As discussed in Section 3.3, we use a constant elasticity of substitution (CES) production function, which is sufficiently rich to allow for complementarity between inputs and factor-specific productivity while remaining empirically tractable. That is, a plant's output is:

$$Y_{jt} = \exp(\omega^H_{jt}) \left(\beta_K K_{jt}^{\frac{\sigma-1}{\sigma}} + L_{jt}^{\frac{\sigma-1}{\sigma}} + (\exp(\omega^E_{jt}) E_{jt})^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\nu\sigma}{\sigma-1}} \times \exp(\epsilon_{jt})$$

where σ and ν are respectively the elasticity of substitution and returns to scale, $\exp(\omega^H_{jt})$ is the Hicks-neutral total factor productivity, and β_K and $\exp(\omega^E_{jt})$ are the productivity of capital and energy inputs relative to labor inputs, respectively.[22] The two productivity shocks, ω^H_{jt} and ω^L_{jt}, are known by the plant when it choses inputs, whereas ϵ_{jt} represents unanticipated randomness in the output of the production

[22] Note that the level of factor-specific productivities are not separately identifiable from total factor productivity, so without loss of generality we normalize labor productivity to one and express factor productivity relative to labor productivity.

process.[23]

In each period, plants choose their static inputs, labor L_{jt} and energy E_{jt}, given their capital stock, K_{jt}, productivity draws, and prices to maximize their profits:

$$\max_{L,E} p_Y Y(L, E; K_{jt}, \omega_{jt}^H, \omega_{jt}^E) - w_{jt}L - p_{jt}^E E$$

where w_{jt} and p_{jt}^E are the prices of labor and electricity, respectively. By taking the log of the ratio of the first-order conditions for profit maximization we obtain the expression:

$$l_{jt} - e_{jt} = -\sigma(w_{jt} - p_{jt}^E) + (1-\sigma)\omega_{jt}^E \tag{3.7}$$

Given the elasticity of substitution, σ, equation (3.7) allows us to obtain the energy-augmenting productivity shocks ω_{jt}^E from the (log) ratios of static inputs and their prices. Intuitively, if labor and energy are complementary inputs (i.e. $\sigma < 1$), then conditional on prices a higher ratio of labor to energy inputs implies a higher relative productivity of energy.

Conditional on knowing ω_{jt}^E, we can then recover the total factor productivity, ω_{jt}^H, from the first-order condition for energy, given the values for the rest of the production functions' parameters. As in Ackerberg et al. (2015), estimation proceeds based on moment conditions formed by the evolution of these two productivity shocks. We assume that both productivity shocks follow AR(1) processes:

$$\omega_{jt}^H = \alpha_H + \beta_H \omega_{jt-1}^H + \xi_{jt}^H$$

$$\omega_{jt}^E = \alpha_E + \beta_E \omega_{jt-1}^E + \xi_{jt}^E$$

where ξ_{jt}^H and ξ_{jt}^E are unknown by plants at time $t-1$, and therefore uncorrelated with lagged inputs.

[23] For example, unscheduled maintenance or deviations from anticipated product defect rates could introduce unanticipated production fluctuations.

We estimate the model separately for each industry as follows. First, we take candidate parameters of the production technology, $\tilde{\theta} = (\sigma, \nu, \beta_K)$, and use these to recover the productivities $\omega^H_{jt}, \omega^E_{jt}$ from each plants' input choices in each year. Second, we estimate the parameters of the AR(1) processes by ordinary least-squares to obtain the productivity innovations ξ^H_{jt} and ξ^E_{jt}. We form moments based on these innovations:

$$\mathbb{E}[\xi^H_{jt} Z_{jt}] = 0$$

$$\mathbb{E}[\xi^L_{jt} Z_{jt}] = 0$$

where Z_{jt} are a set of instruments. The timing of decisions and the Markov structure for productivity shocks implies that all past input choices are uncorrelated with the productivity innovations, ξ^H_{jt} and ξ^L_{jt}. We use lagged (log) inputs, l_{jt-1}, e_{jt-1}, and k_{jt-1}, and lagged wage and energy prices, w_{jt-1} and p^E_{jt-1}. This forms a total of 10 moments, collected in the vector $g(X_i, \theta)$, where i flattens the time and plant indices. These lagged-input instruments, and the identifying assumption that past inputs and prices are determined before and do not affect the unanticipated innovations to productivity, are standard in the production function literature (Ackerberg et al., 2015; Doraszelski and Jaumandreu, 2018; Olley and Pakes, 1996).

For each industry, we obtain estimates $\hat{\theta}$ and standard errors using the two-step generalized method of moments (GMM) estimator (Hansen, 1982). We minimize the objective function:

$$C(\theta, \cdot) = \left(\frac{1}{N} \sum_{i=1}^{N} g(X_i, \theta) \right)' \hat{W} \left(\frac{1}{N} \sum_{i=1}^{N} g(X_i, \theta) \right)$$

where $g(X_i, \theta)$ are 10x1 vectors defined above and the weight matrix $\hat{W}$ is the inverse covariance matrix obtained using the initial parameters $\hat{\theta}_0$ from the minimization of the objective function using the identity matrix as the weight matrix.

Overall, we find that our estimates of the production function parameters have reasonable signs and magnitudes. Appendix Table C.4 shows that we find that capital, labor, and energy are strongly complementary; our average estimate of σ is around 0.25.[24] Our estimate of the returns to scale parameter ν, which is around 0.65, is also consistent with estimates from the literature.[25] In what follows, we focus on the relative energy productivity and total factor productivity estimates shown in Figure 3.3, and their relationship with initial and current energy prices.

3.7 Results

3.7.1 Energy Intensity

First, Table 3.1 shows that weighted national fuel prices are strongly predictive of both entry-year and current electricity prices, respectively in Columns 1 and 2. These results form a strong first stage for the instrumental variables analyses. Coal prices are the largest determinant of entry-year electricity prices, and are approximately four to five times as important as natural gas and petroleum prices (Column 1). If a state generated its electricity entirely from coal in 1976, then a 10% increase in coal prices in a plant's entry year would increase its entry-year electricity price by approximately 2.2%. In practice, the state average 1976 coal share is approximately 0.40, and so a 10% increase in coal prices increases electricity prices by 0.9%.[26] Reassuringly, fuel prices in the future are not predictive of entry-year electricity prices in the past. Initial fuel prices have small effects on current electricity prices, pos-

[24] There are relatively few estimates of CES production function parameters involving energy inputs. Our results are comparable to Hassler et al. (2012) and Ryan (2012), who also estimate a strongly complementary relationship between energy and other inputs.

[25] Our ν estimates are smaller than Doraszelski and Jaumandreu (2018)'s estimates of around 0.9, which can be explained by the fact that our returns to scale parameter combines the effects of returns to scale and downward-sloping demand that are separately estimated in this other paper.

[26] Appendix Table C.3 shows the 1976 state fuel generation shares that can be used to adjust the parameters in Table 3.1 to interpret them as elasticities.

sibly reflecting some stickiness in electricity prices paid by plants, but current fuel prices become significantly more important (Column 2). Contemporaneous natural gas prices have the largest effect on current electricity prices, reflecting the shift toward natural gas electricity generation in recent years shown in Appendix Figure C.4. Ganapati et al. (2020), who examine the effects of contemporaneous fuel prices on manufacturing marginal costs, similarly highlight the importance of natural gas as a recent determinant of manufacturing costs.

Table 3.2 presents our first evidence of technology lock-in. This table shows that both initial and current fuel prices have significant effects on current energy intensity. We consider effects on four different measures of energy intensity. Column 1 measures energy intensity using electricity consumed (in kWh) per dollar of revenue, which accounts for most energy use by manufacturing plants. Column 2 focuses on this same measure of energy intensity in "electricity-intensive industries", excluding industries that spend more than 30% of total energy expenditures on raw fuels. Columns 3 and 4 use kg CO2 and million BTU per dollar revenue as measures of energy intensity, respectively. These last two measures include energy from raw fuels and therefore account for changes in energy intensity due to any substitution between fuel sources.

We find consistent results across each of these measures of energy intensity. In all models in Table 3.2, the current natural gas price has a larger impact on current energy intensity than contemporaneous coal or petroleum prices, which is indicative of the recent shift toward natural gas electricity generation after the fracking boom in the 2000s. The precisely estimated zero effect of current petroleum oil prices is consistent with the limited use of petroleum in generating electricity today, shown in Appendix Figure C.1.[27] By contrast, despite the declining roles of petroleum and coal in electricity generation, entry-year coal and petroleum prices have persistent effects on energy intensity. This is lock-in: the prices of these fossil fuels continue to

[27] Comparing Appendix Figures C.1 and C.4 shows that most states generating electricity using petroleum oil in 1976 substantially reduce their use of it by 2011.

affect plants' energy intensity even after the economy has transitioned to other fuel sources. The elasticity of energy intensity with respect to entry-year coal prices is more than twice the current natural gas price elasticity even before accounting for the higher 1976 coal generation share. These results underscore that the continued expansion of coal power capacity, particularly in developing countries, could lead to higher manufacturing energy intensity even if these economies eventually transition to cleaner fuel sources.

This evidence of lock-in is also apparent in both the OLS and instrumental variables analyses of the effects of initial and current electricity prices on energy intensity (Table 3.3, Panels A and B respectively). In both analyses and across our four energy intensity measures, entry-year electricity prices have significant effects on energy intensity in subsequent years. In our preferred IV specifications, the initial price elasticity is between -0.14 and -0.35, which is approximately 25% of the elasticity with respect to current electricity prices. As a result, failing to price carbon in plants' entry year leaves an average of 25% of the energy-reduction benefits on the table. Though not statistically different from the elasticity of electricity intensity with respect to initial electricity prices, the slightly larger elasticity of CO2 intensity suggests that plants may slightly increase their use of more CO2-intensive fuels in response to increases in electricity prices. Overall, it is unsurprising that the price effects on energy intensity measures that include and exclude raw fuels are similar because electricity comprises well over 70% of energy expenditures on average; manufacturing plants therefore have more limited ability to substitute toward other raw fuel types than in other sectors, such as electricity generation (Meng, 2021). We highlight that the inclusion of industry × year × entry year fixed effects controls for plant vintage within each industry, so that the estimates comprise the effect of changes in the price of electricity for plants with the same technologies available to them.

The initial electricity price elasticity is larger in the IV models than in the OLS models, consistent with measurement error in the entry-year electricity prices that are annual averages across entrants in each industry and state if a plant is not surveyed in its entry year. Such measurement error biases the estimates against finding evidence that initial electricity prices are persistent. The elasticity of energy intensity with respect to current electricity prices is about -0.80 (i.e., relatively elastic) and is similar in sign, magnitude, and precision in both the OLS and IV models. Current electricity prices are always measured at the plant level and are therefore less likely to be subject to measurement error in the OLS estimates.[28] Appendix Table C.7 shows that both our initial and current price elasticities are robust to estimation using different covariates, data subsamples, and electricity price data sources. We discuss these additional estimates in Appendix 3.10.

We find limited evidence that the importance of entry-year electricity prices declines as plants age, suggesting that lock-in is persistent (Table 3.6, Column 1-3). Each additional year of operations reduces the entry-year price elasticity by 4%, though for most energy intensity measures this small effect of age is not statistically distinguishable from zero. At this rate, it would take 25 years for the effect of entry-year prices to fade, which Appendix Table C.2 shows is 10 years longer than the average plant lifetime of 15 years. Any decline in the average importance of entry-year prices could be due to plants' gradual investments in energy efficiency improvements or due to changes in entry and exit; the IV estimates combine both of these effects for surviving plants, providing an upper bound on plants' ability to respond to energy price changes and mitigate lock-in without ceasing operations. We turn now to assess the effects of initial energy prices on the productivity of plants' inputs to understand whether capital adjustment costs can fully explain the persistent differences in energy

[28] Our estimated elasticity of energy intensity with respect to current electricity prices is somewhat larger than estimates in Linn (2008) using different variation based on fixed weight price indices as instruments for energy prices. Appendix Table 3.10 shows that our estimates of the effects of current prices on quantity of electricity purchased, rather than intensity, are within the range of elasticity estimates in the literature for industrial consumers (Blonz, 2021; Paul et al., 2009). We are unaware of any estimates of entry-year price elasticities against which to compare ours.

use or whether plants that enter at different energy prices inherently choose different production technologies.

3.7.2 Productivity

Table 3.4 begins to show that initial energy prices lead to persistent differences in plants' production technologies. Columns 1 and 2 show that both initial and current energy prices have long-run effects on the energy bias of technological change, for all industries and electricity-intensive industries respectively. Plants that enter when petroleum or coal prices are high consistently use their energy inputs more efficiently relative to their labor inputs; a 10% increase in the entry-year price of one of these raw fuels increases energy productivity by 0.7% and 0.1%, respectively.[29] Similarly to our energy intensity results, we find that contemporaneous natural gas prices are important determinants of relative energy efficiency. Conversely, the effects of initial and contemporaneous fuel prices on total factor productivity are an order of magnitude smaller and are generally statistically indistinguishable from zero: higher entry-year raw fuel prices bias technological change toward energy relative to labor, but do not affect total factor productivity in meaningful ways.

Turning to the OLS and instrumental variables estimates of the effects of electricity prices on productivity, we again find evidence of lock-in of plants' productivity bias (Table 3.5). Plants that pay higher electricity prices in their entry year exhibit persistently higher energy productivity relative to labor productivity, both in the OLS estimates (Panel A) and in the instrumental variables estimates (Panel B). We again find instrumental variables estimates of the relative energy productivity effects in that are larger in magnitude than the OLS estimates, consistent with measurement error in initial electricity prices. Focusing specifically on our preferred instrumental variables estimates, we find that a 10% increase in entry-year electricity prices increases

[29] Similarly to Table 3.2, we adjust the energy elasticity estimates in Table 3.4 by the average 1976 fuel generation shares in Appendix Table C.3 to arrive at the average weighted elasticity.

relative energy productivity by 3%, with no effect on total factor productivity. Taken together, this pattern of results indicates that plants that begin operations at higher electricity prices are not only using fewer energy inputs per dollar of revenue, as we showed above; they are also using these inputs more efficiently.[30] The effect of entry-year electricity prices is almost as important as contemporaneous electricity prices: the elasticities are statistically indistinguishable in Table 3.5. The overall effect of entry-year electricity prices on relative energy productivity is more than five times as large as the same increase in coal transport costs on relative coal capital investment (Meng, 2021) and the effects of air pollution regulation on manufacturing total factor productivity (Greenstone et al., 2012). These economically meaningful estimates highlight the important role of higher energy prices and, by extension, carbon pricing policies in directly incentivizing reductions in energy use.

Our results suggest that persistent differences in the relative productivity of energy inputs chosen at entry can fully explain why technology lock-in arises. The magnitudes of the relative energy productivity effects of initial electricity prices are slightly larger and statistically indistinguishable from the effects on energy intensity in Table 3.3. An implication therefore is that the contribution of capital adjustment costs to creating lock-in appears to be comparatively small on average.[31] Relative to the model in Section 3.3, our estimate of the effect of entry-year electricity prices on energy intensity is analogous to slope of the curve showing lock-in for plants facing capital adjustment costs, averaged across plants. Under the assumption that energy productivity enters multiplicatively with energy inputs, for fully flexible plants, the elasticity of energy productivity will be equal to the elasticity of energy intensity,

[30] Recall that the energy productivity estimate gives the relative productivity of energy inputs to labor inputs, and hence alone does not indicate an overall increase in energy productivity.

[31] The non-linearity of capital adjustment frictions implies that there may be heterogeneous effects depending on the size of the price change. The difference between lock-in for a plant facing adjustment costs and for a hypothetical plant with fully flexible capital is non-monotonic in the price change, and largest for plants which are close to the threshold at which paying fixed adjustment costs is optimal. This implies that targeted capital adjustment subsidies are likely to be more effective than ones applied to all firms.

which corresponds to the slope of the orange, short-dashed line for fully flexible incumbents in Figure 3.1.

In Appendix Table C.9, we also show estimates of the effects of entry-year electricity prices on quantities of energy inputs consumed, as opposed to energy intensity; we find that the effects in levels can also be explained by persistent differences in relative productivity, though the level effects are somewhat less precisely estimated than the intensity elasticities. We discuss these estimates in more detail in Appendix 3.10, and Appendix Table C.8 discusses the robustness of the elasticity estimates to the use of different covariates, data subsamples, and data sources. The results using these alternative models are similar in sign, magnitude, and precision to our main estimates.

Similarly to the energy intensity results, we find that the effects of entry-year electricity prices on relative energy productivity persist throughout a plant's lifetime. Table 3.6 shows that there is limited evidence of a decline in the effects of initial electricity prices as plants age; an additional year of operations reduces the effect of entry-year electricity prices on relative energy productivity by 2%. These results indicate significant path dependence in the productivity bias of energy inputs and the importance of correctly aligning plants' incentives when they choose their production technologies.

3.8 Discussion and Implications for Climate Policy

Overall, we find robust evidence of technology lock-in. However, this lock-in isn't complete: on average, plants' energy intensity and energy productivity also respond to changes in contemporaneous electricity prices, though less than one-for-one (Tables 3.3 and 3.5). These average effects combine adjustment through investment and through entry and exit.

There are at least three reasons why these estimates of lock-in may be a lower bound on the effects of entry-year electricity prices on subsequent energy efficiency. That is, the estimates may underestimate the effect of a carbon tax on the ability of plants to adjust to higher energy prices without ceasing operations. First, these effects are measured on surviving plants, and reduced entry or increased exit may be important channels through which plants respond to higher prices. The effects that we estimate combine adjustment through investment and through entry and exit; if the entirety of the improvements in average energy efficiency are due to changes in entry and exit, then this means that the ability of plants to adjust their energy use through investment while operating is more limited than our estimates suggest.[32]

Second, our use of revenue-based total factor productivity measures also understates the effects of energy prices compared with measures based on quantity produced. Revenue-based productivity measures are standard in the literature due to limitations of most plant-level data sets, which typically do not collect detailed output price and quantity data (Allcott et al., 2016; Ganapati et al., 2020; Greenstone et al., 2012). When marginal costs rise as energy becomes more expensive, standard theory predicts that plants with market power will increase prices for their products and reduce quantities supplied. The revenue-based productivity measures will capture any negative effects of increasing energy costs as well as any positive price change, which could cause us to understate the effect of electricity prices on total factor productivity.

Third, we investigate the persistent effects of short-run electricity price variation resulting from year-on-year variation in raw fuels prices. Conversely, a goal of carbon pricing is to implement long-run increases in energy prices through policy. The responses to the short-run price changes that we study are consistent with firms' basing their best guess of energy prices tomorrow on observed energy prices today (i.e., with prices following a random walk). Our lock-in estimates again may understate the en-

[32] Separately analyzing the importance of investment relative to entry and exit is the focus of on-going work.

ergy efficiency effects of sustained commitment to higher energy prices because plants may initially install more energy efficient capital investment with the knowledge that prices will be higher throughout their lifetime.

This discussion highlights the importance of plants' beliefs about future energy prices when they undertake investments in durable capital. Prior research suggests that plants may adopt new technologies in anticipation of environmental regulation in the future (Clay et al., 2021). Commitment to federal carbon pricing in the U.S. could reduce lock-in when prices increase by correctly aligning plants' beliefs about the future path of energy prices, though some lock-in may still arise given that plants may discount higher energy costs in the distant future in favor of lower costs of investment today.

It is worth noting that our lock-in elasticity estimates are agnostic about plants' beliefs about future prices. Our estimates are conditional on whatever firms' actual beliefs are about the evolution of electricity prices. The interpretation of the empirical results does not require us to take a stand on what these beliefs are.

Overall, these results suggest that delayed action on carbon pricing comes at the expense of significant energy efficiency gains. Timely implementation of carbon pricing is one policy that could incentivize early reductions in energy use. However, our results also suggest that there appears to be a role for vintage energy efficiency regulations. Targeting efficiency mandates or technology adoption subsidies to plants that enter during low energy price regimes could help adjust for relative misalignment of incentives when these plants were established, and therefore help address the inefficiencies resulting from failing to incentivize internalization of greenhouse gas externalities initially.

3.9 Conclusion

This paper provides new evidence of technology lock-in in the manufacturing sector and analyzes its causes and consequences. Using 35 years' worth of U.S. Census microdata, we show two main ways in which technology lock-in arises. First, we estimate that the prices of fossil fuel inputs into electricity generation have persistent effects on manufacturing plants' energy usage—even after the use of these fuels has declined. Second, we show that the prevailing electricity price in a plant's entry year affects their energy usage throughout their lifetime: plants that are established when electricity prices are low, below the full social cost of energy consumption, consume more energy in subsequent years. On average, we estimate that at least 25% of the energy reductions benefits from carbon pricing are lost by failing to implement these policies in a plants' entry year.

By estimating plant-level total factor productivity and the relative productivity of energy to labor, we demonstrate that an initial and persistent effect of electricity prices on energy productivity is a key explanation for this lock-in. Plants may choose not to undertake later energy efficiency improvements due to capital adjustment costs, but we provide new evidence that their production functions are also different to start out. Our results indicate that a 10% increase in entry-year electricity prices improves the productivity of energy relative to labor by approximately 3% in subsequent years. Since the analysis focuses on plants that continue to operate and choose to enter at higher electricity prices, these estimates exclude effects on energy-inefficient plants that cease operations in response to higher prices. As a result, our estimates plausibly provide a lower bound on the energy reductions resulting from increasing electricity prices.

The implications of these results for climate policy are consequential. Ignoring lock-in underestimates the benefits of pricing carbon today. In the absence of current commitments to do so, future policy will have to be more stringent to counteract the current path of energy-inefficient manufacturing production: small carbon taxes or clean technology subsidies may be insufficient to incentivize existing plants to reverse sunk and partially irreversible capital investments or otherwise to exit. Meanwhile, continued expansion of cheap fossil fuel power around the world seems likely to entrench energy-inefficient technologies and lock in higher emissions levels for many years. A major push to increase energy efficiency worldwide is a key part of proposals to constrain carbon emissions to "safe" levels, which will require annual improvements exceeding three times the annual rate achieved in the last two decades (IEA, 2021). The global trend in increasingly severe natural disasters suggests that it would be inadvisable to delay further action on climate change policy.

Bibliography

Acemoglu, D., Aghion, P., Barrage, L., and Hemous, D. (2019). Climate change, director innovation, and energy transition: The long-run consequences of the shale gas revolution. Working Paper.

Acemoglu, D., Aghion, P., Bursztyn, L., and Hemous, D. (2012). The environment and directed technical change. *American Economic Review*, 102(1):131–66.

Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451.

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.

Atkeson, A. and Kehoe, P. J. (1999). Models of energy use: Putty-putty versus putty-clay. *American Economic Review*, 89(4):1028–1043.

Baily, M., Hulten, C., and Campbell, D. (1992). The distribution of productivity in manufactruing plants. *Brookings Papers: Macroeconomics*, 187.

Bartik, T. J. (1991). Who benefits from state and local economic development policies? *W.E. Upjohn Institute.*

Blonz, J. A. (2021). Making the best of the second-best: Welfare consequences of time-varying electricity prices. Energy Institute at Haas Working Paper 275.

Clay, K., Jha, A., Lewis, J. A., and Severnini, E. R. (2021). Impacts of the clean air act on the power sector from 1938-1994: Anticipation and adaptation. NBER Working Paper 28962.

Cooper, R. W. and Haltiwanger, J. C. (2006). On the nature of capital adjustment costs. *Review of Economic Studies*, 73:611–633.

Cullen, J. A. and Mansur, E. T. (2017). Inferring carbon abatement costs in electricity markets: A revealed preference approach using the shale revolution. *American Economic Journal: Economic Policy*, 9(3):106–133.

Demirer, M. (2020). Production function estimation with factor-augmenting technology: An application to markups. Working Paper.

DNC (2021). The biden plan for a clean energy revolution and environmental justice. Technical report, Demogratic National Committee.

Doraszelski, U. and Jaumandreu, J. (2018). Measuring the bias of technological change. *Journal of Political Economy*, 126(3):1027–1084.

EIA (2015). Barriers to industrial energy efficiency. Technical report, U.S. Energy Information Administration.

EIA (2019). International energy outlook 2019. Technical report, U.S. Energy Information Administration.

EIA (2020). State energy data system.

EPA (2021). Sources of greenhouse gas emissions. Technical report, U.S. Environmental Protection Agency.

Fowlie, M., Reguant, M., and Ryan, S. P. (2016). Market-based emissions regulation and industry dynamics. *Journal of Political Economy*, 124(1):249–302.

Fried, S. (2018). Climate policy and innovation: A quantitative macroeconomic analysis. *American Economic Journal: Macroeconomics*, 10(1):90–118.

Gambhir, A., Butnar, I., Li, P.-H., Smith, P., and Strachan, N. (2019). A review of criticisms of integrated assessment models and proposed approahces to address these, through the lens of beccs. *Energies*, 12(1747):1–21.

Ganapati, S., Shapiro, J., and Walker, R. (2020). Energy cost pass-through in u.s. manufacturing: Estimates and implications for carbon taxes. *American Economic Journal: Applied Economics*, 12(2):303–342.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020). Bartik instruments: What, when, how, and why. *American Economic Review*, 110(8):1586–2624.

Greenstone, M., List, J., and Syverson, C. (2012). The effects of environmental regulation on the competitiveness of u.s. manufacturing. NBER Working Paper 18392.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054.

Hassler, J., Krusell, P., and Olovsson, C. (2012). Energy-saving technical change. NBER Working Paper no. 18456.

Huntington, H. G. (2010). Structural change and us energy use: Recent patterns. *Energy Journal*, 31(3):25–39.

IEA (2021). Net zero by 2050: A roadmap for the global energy sector. Technical report, International Energy Agency.

Jacobsen, G. D. and Kotchen, M. J. (2013). Are building codes effective at saving energy? evidence from residential billing data in florida. *The Review of Economics and Statistics*, 95(1):34–49.

Jaeger, D. A., Ruist, J., and Stuhler, J. (2019). Shift-sare instruments and dynamic adjustments: The case of immigration. Working Paper.

Khan, A. and Thomas, J. K. (2008). Idiosyncratic shocks and the role of nonconvexities in plant and aggregate investment dynamics. *Econometrica*, 76(2):395–436.

Knittel, C. R., Metaxoglou, K., and Trindade, A. (2015). Natural gas prices and coal displacement: Evidence from electricity markets. NBER Working Paper 21627.

Levinson, A. (2021). Energy intensity: Deindustrialization, composition, prices, and policies in u.s. states. *Resource and Energy Economics*, 65(101243):1–17.

Linn, J. (2008). Energy prices and the adoption of energy-saving technology. *Economic Journal*, 118:1986–2011.

Lyubich, E., Shapiro, J. S., and Walker, R. (2018). Regulating mismeasured pollution: Implications of firm heterogeneity for environmental policy. *AEA Papers and Proceedings*, 108:136–142.

Meng, K. (2021). Estimating path dependence in energy transitions. NBER Working Paper 22536.

Metcalf, G. E. (2008). An empirical analysis of energy intensity and its determinants at the state level. *Energy Journal*, 29(3):1–26.

Myllyvirta, L. and Shearer, C. (2021). China dominates 2020 coal plant development. Technical report, Global Energy Monitor.

NAM (2014). Efficiency and innovation in u.s. manufacturing energy use. Technical report, National Association of Manufacturers.

Nordhaus, W. D. and Boyer, J. (2000). *Warming the World: Economic Models of Global Warming*. MIT Press.

NREL (2009). Energy efficiency policy in the united states: Overview of trends at different levels of government. Technical report, National Renewable Energy Laboratory.

Olley, G. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297.

Paul, A., Myers, E., and Palmer, K. (2009). A partial model of u.s. electricity demand by region, season, and sector. Technical report, Resources for the Future.

Pindyck, R. S. (2020). What we know and don't know about climate change, and implications for policy. NBER Working Paper 27304.

Ryan, N. (2018). Energy productivity and energy demand: Experimental evidence from indian manufacturing plants. NBER Working Paper no. 24619.

Ryan, S. P. (2012). The costs of environmental regulation in a concentrated industry. *Econometrica*, 80(3):1019–61.

Seto, K. C., Davis, S. J., Mitchell, R. B., Stokes, E. C., Unruh, G., and Urge-Vorsatz, D. (2016). Carbon lock-in: Types, causes, and policy implications. *Annual Review of Environment and Resources*, 41:425–52.

Shapiro, J. S. and Walker, R. (2018). Why is pollution from u.s. manufacturing declining? the roles of environmental regulation, productivity, and trade. *American Economic Review*, 108(2):3814–54.

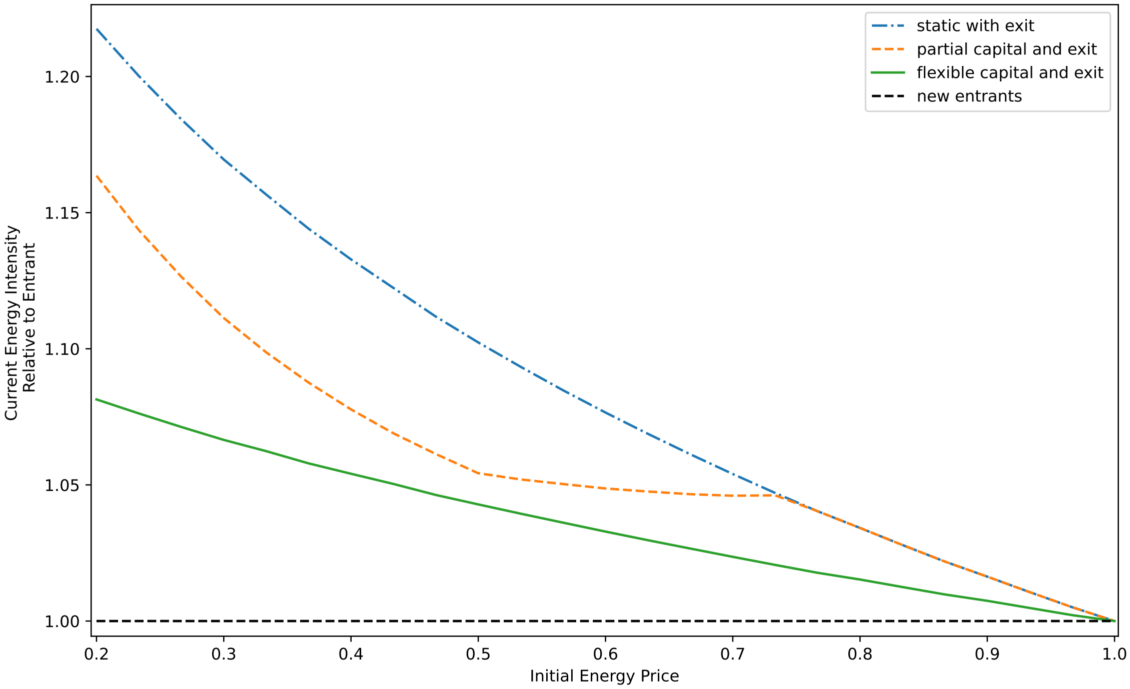
Shearer, C., N. G. L. M. A. Y. and Nace, T. (2010). Boom and bust 2016: Tracking the global coal plant pipeline. Technical report, CoalSwarm, Greenpeace, and Sierra Club.

Stavins, R. N. (2006). Vintage-differentiated environmental regulation. *Stanford Environmental Law Journal*, 25:29–63.

West, J., Hoekstra, M., Meer, J., and Puller, S. L. (2017). Vehicle miles (not) traveled: Fuel economy requirements, vehicle characteristics, and household driving. *The Journal of Public Economics*, 145:65–81.

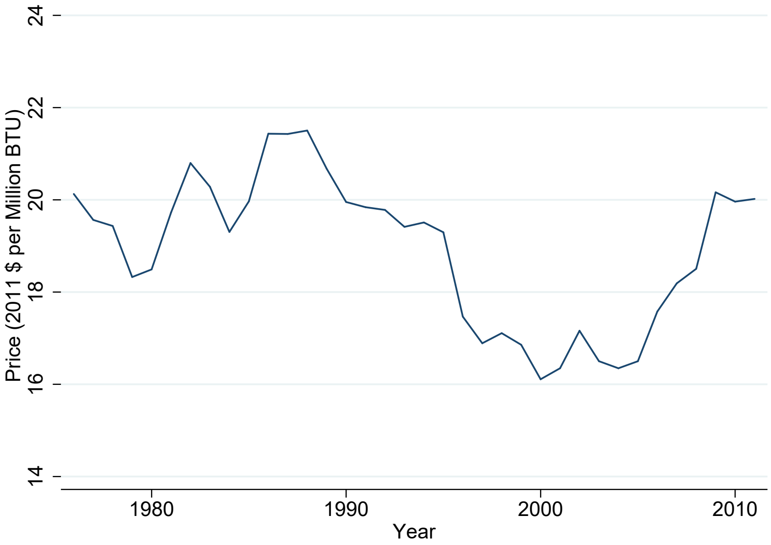
3.10 Figures and Tables

Figure 3.1: Simulated Lock-in



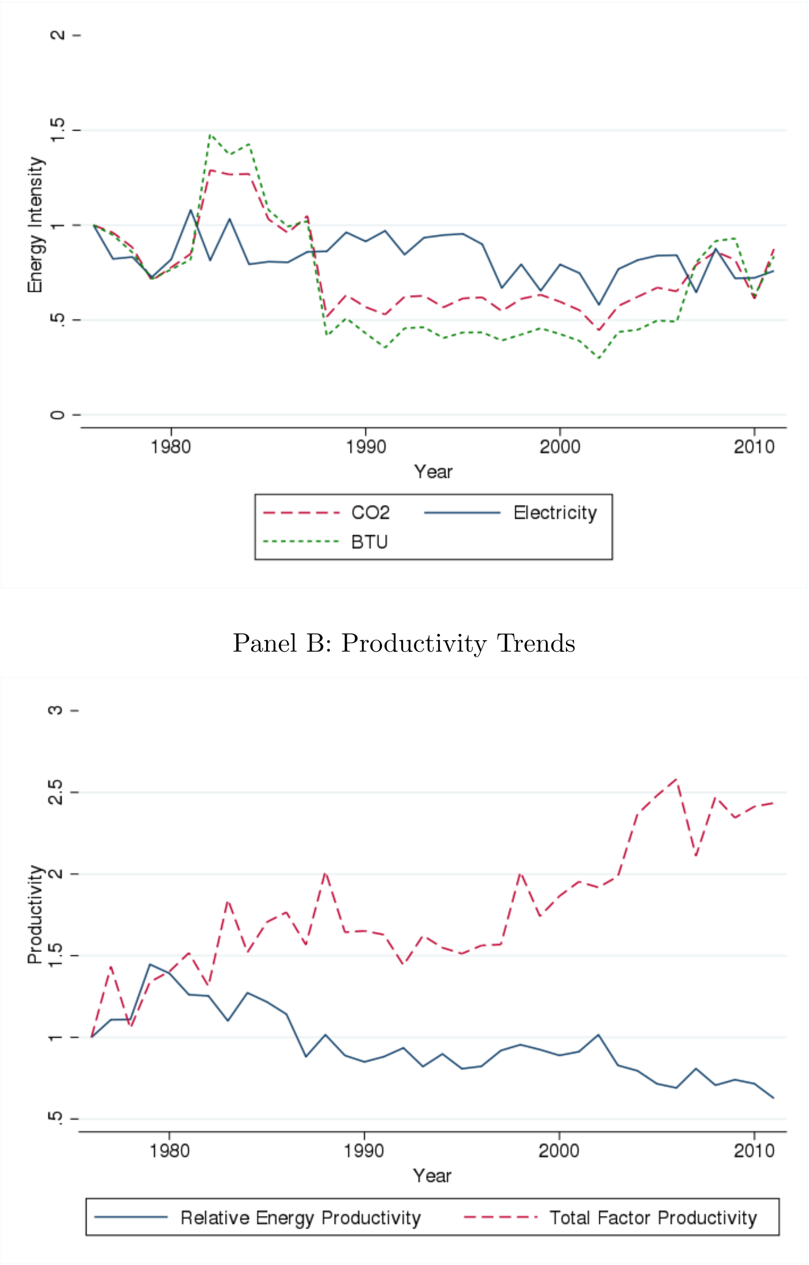
Notes: This figure shows simulation results for the energy intensity of incumbent manufacturing plants relative to entrants as a function of last year's energy price. The x-axis shows initial energy price as a fraction of the current price. "Static with exit" shows relative energy intensity in scenarios where plants cannot adjust their capital stocks after they enter. "Partial capital and exit" shows relative energy intensity in scenarios where incumbents can reoptimize their capital stock subject to fixed and convex adjustment costs. "Flexible capital and exit" shows relative energy intensity in scenarios where all inputs can be reoptimized without adjustment costs. Energy intensity of entrants is normalized to 1.

Figure 3.2: Time Series of Electricity Prices



Notes: This figure shows the time series of average electricity prices paid by the industrial sector in the United States. Prices are in 2011 dollars per million British thermal units (BTU).

Figure 3.3: Time Series of Energy Intensity and Productivity



Panel A: Energy Intensity Trends

Panel B: Productivity Trends

Notes: This figure shows the time series of average energy intensity (Panel A) and relative energy productivity and total factor productivity (Panel B) of the manufacturing sector in the United States. Energy intensity is calculated as electricity consumption (kWh) per dollar of revenue, kg CO2 produced per dollar revenue, and million BTU per dollar revenue. The productivity of energy inputs is measured relative to labor.

Table 3.1: First Stage Effects of Weighted Fuel Prices on Electricity Prices

	$\log(Initial\_Electricity\_Price_{i,t_0})$	$\log(Current\_Electricity\_Price_{i,t})$
	(1)	(2)
$Coal\_Share_{s,1976} \times Current\_Coal\_Price_{-s,t}$	0.013	0.065*
	(0.009)	(0.035)
$Natural\_Gas\_Share_{s,1976} \times Current\_Natural\_Gas\_Price_{-s,t}$	-0.006*	0.058***
	(0.003)	(0.012)
$Petroleum\_Share_{s,1976} \times Current\_Petroleum\_Price_{-s,t}$	0.003	0.012
	(0.003)	(0.010)
$Coal\_Share_{s,1976} \times Initial\_Coal\_Price_{-s,t_0}$	0.220***	0.055**
	(0.049)	(0.023)
$Natural\_Gas\_Share_{s,1976} \times Initial\_Natural\_Gas\_Price_{-s,t_0}$	0.056***	0.011*
	(0.012)	(0.006)
$Petroleum\_Share_{s,1976} \times Initial\_Petroleum\_Price_{-s,t_0}$	0.036***	0.019***
	(0.012)	(0.005)
N	1294000	1294000
Industry × Year × Entry Year Fixed Effects	Yes	Yes
Industry × State Fixed Effects	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the effects of initial and contemporaneous coal, natural gas, and petroleum prices on the log of initial and contemporaneous electricity prices. Fuel prices are calculated as the leave-out-mean log price across states and weighted by the share of each fuel in electricity generation in each state. Electricity prices are measured in USD per kWh (2011). Regressions are weighted using Census sampling weights. Standard errors clustered by state are in parentheses.

Table 3.2: Reduced Form Effects of Weighted Fuel Prices on Energy Intensity

	log($Electricity\_Intensity_{i,t}$)	log($Electricity\_Intensity_{i,t}$)	log($CO_2\_Intensity_{i,t}$)	log($BTU\_Intensity_{i,t}$)
		Electricity-Intensive Industries		
	(1)	(2)	(3)	(4)
$Coal\_Share_{s,1976} \times Current\_Coal\_Price_{-s,t}$	0.043	0.055	-0.019	0.073*
	(0.031)	(0.036)	(0.041)	(0.039)
$Natural\_Gas\_Share_{s,1976} \times Current\_Natural\_Gas\_Price_{-s,t}$	-0.053***	-0.052***	-0.053***	-0.056***
	(0.011)	(0.012)	(0.011)	(0.012)
$Petroleum\_Share_{s,1976} \times Current\_Petroleum\_Price_{-s,t}$	-0.000	-0.004	0.001	0.001
	(0.009)	(0.009)	(0.010)	(0.010)
$Coal\_Share_{s,1976} \times Initial\_Coal\_Price_{-s,t_0}$	-0.121***	-0.123***	-0.163***	-0.129***
	(0.028)	(0.031)	(0.029)	(0.029)
$Natural\_Gas\_Share_{s,1976} \times Initial\_Natural\_Gas\_Price_{-s,t_0}$	-0.003	-0.013	0.001	0.002
	(0.009)	(0.012)	(0.008)	(0.008)
$Petroleum\_Share_{s,1976} \times Initial\_Petroleum\_Price_{-s,t_0}$	-0.028***	-0.031***	-0.028***	-0.024***
	(0.005)	(0.006)	(0.005)	(0.005)
N	1294000	955000	1294000	1294000
Industry × Year × Entry Year Fixed Effects	Yes	Yes	Yes	Yes
Industry × State Fixed Effects	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the effects of initial and contemporaneous coal, natural gas, and petroleum prices on the log of plants' energy intensity. Electricity intensity is measured in kWh per dollar of revenue, CO2 intensity is kg CO2 per dollar of revenue, and BTU intensity is BTU per dollar of revenue. Electricity-intensive industries are industries for which electricity accounts for at least 70% of total energy expenditures. Fuel prices are calculated as the leave-out-mean log price across states and weighted by the share of each fuel in electricity generation in each state. Regressions are weighted using Census sampling weights. All dollar values are in 2011 USD. Standard errors clustered by state are in parentheses.

Table 3.3: Effects of Initial and Current Electricity Prices on Energy Intensity

	$\log(Electricity\_Intensity_{i,t})$	$\log(Electricity\_Intensity_{i,t})$	$\log(CO_2\_Intensity_{i,t})$	$\log(BTU\_Intensity_{i,t})$
		Electricity-Intensive Industries		
	(1)	(2)	(3)	(4)
Panel A: OLS				
$\log(Current\_Electricity\_Price_{i,t})$	-0.851***	-0.831***	-0.824***	-0.807***
	(0.012)	(0.011)	(0.010)	(0.009)
$\log(Initial\_Electricity\_Price_{i,t_0})$	-0.040***	-0.037***	-0.028***	-0.026**
	(0.009)	(0.010)	(0.010)	(0.010)
Panel B: IV				
$\log(Current\_Electricity\_Price_{i,t})$	-0.764***	-0.734***	-0.829***	-0.761***
	(0.090)	(0.104)	(0.072)	(0.087)
$\log(Initial\_Electricity\_Price_{i,t_0})$	-0.165***	-0.232***	-0.289***	-0.144**
	(0.051)	(0.059)	(0.079)	(0.059)
K-P F stat	12.1	11.9	12.1	12.1
N	1294000	955000	1294000	1294000
Industry × Year × Entry Year FE	Yes	Yes	Yes	Yes
Industry × State FE	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the effects of initial and contemporaneous log electricity prices on the log of plants' energy intensity. Models in Panel A are estimated using OLS and models in Panel B are estimated using IV. In IV models, electricity prices are instrumented using initial and contemporaneous prices for coal, natural gas, and petroleum, weighted by the share of each fuel in state electricity generation. Electricity prices are measured in dollars per kWh, electricity intensity is kWh per dollar of revenue, CO2 intensity is kg CO2 per dollar of revenue, and BTU intensity is BTU per dollar of revenue. Electricity-intensive industries are industries for which electricity accounts for at least 70% of total energy expenditures. Regressions are weighted using Census sampling weights. All dollar values are in 2011 USD. Standard errors clustered by state are in parentheses.

Table 3.4: Reduced Form Effects of Weighted Fuel Prices on Productivity

	$\log(Energy\_Productivity_{i,t})$	$\log(Energy\_Productivity_{i,t})$	$\log(TFP_{i,t})$	$\log(TFP_{i,t})$
		Electricity-Intensive Industries		Electricity-Intensive Industries
	(1)	(2)	(3)	(4)
$Coal\_Share_{s,1976} \times Current\_Coal\_Price_{-s,t}$	-0.088*	-0.082	-0.035**	-0.034*
	(0.048)	(0.051)	(0.017)	(0.018)
$Natural\_Gas\_Share_{s,1976} \times Current\_Natural\_Gas\_Price_{-s,t}$	0.037***	0.046***	0.012***	0.005
	(0.013)	(0.012)	(0.004)	(0.004)
$Petroleum\_Share_{s,1976} \times Current\_Petroleum\_Price_{-s,t}$	0.008	0.013	-0.010***	-0.009***
	(0.013)	(0.012)	(0.002)	(0.003)
$Coal\_Share_{s,1976} \times Initial\_Coal\_Price_{-s,t_0}$	0.178***	0.171***	0.037*	0.041*
	(0.041)	(0.044)	(0.022)	(0.025)
$Natural\_Gas\_Share_{s,1976} \times Initial\_Natural\_Gas\_Price_{-s,t_0}$	-0.018	-0.011	0.006*	0.011***
	(0.017)	(0.017)	(0.004)	(0.004)
$Petroleum\_Share_{s,1976} \times Initial\_Petroleum\_Price_{-s,t_0}$	0.035***	0.035***	-0.005	-0.002
	(0.008)	(0.011)	(0.005)	(0.006)
N	1294000	955000	1294000	955000
Industry × Year × Entry Year Fixed Effects	Yes	Yes	Yes	Yes
Industry × State Fixed Effects	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the effects of initial and contemporaneous coal, natural gas, and petroleum prices on the log of plants' productivities. Electricity prices are measured in dollars per kWh, energy productivity is the productivity of electricity relative to labor, and total factor productivity is the productivity common to all manufacturing inputs. Electricity-intensive industries are industries for which electricity accounts for at least 70% of total energy expenditures. Fuel prices are calculated as the leave-out-mean log price across states and weighted by the share of each fuel in electricity generation in each state. Regressions are weighted using Census sampling weights. All dollar values are in 2011 USD. Standard errors clustered by state are in parentheses.

Table 3.5: Effects of Initial and Current Electricity Prices on Productivity

	log($Energy\_Productivity_{i,t}$)	log($Energy\_Productivity_{i,t}$)	log($TFP_{i,t}$)	log($TFP_{i,t}$)
		Electricity-Intensive Industries		Electricity-Intensive Industries
	(1)	(2)	(3)	(4)
Panel A: OLS				
log($Current\_Electricity\_Price_{i,t}$)	0.881***	0.860***	0.060***	0.060***
	(0.023)	(0.020)	(0.007)	(0.009)
log($Initial\_Electricity\_Price_{i,t_0}$)	0.080***	0.093***	-0.037***	-0.047***
	(0.019)	(0.023)	(0.008)	(0.012)
Panel B: IV				
log($Current\_Electricity\_Price_{i,t}$)	0.525***	0.673***	0.088	-0.017
	(0.138)	(0.139)	(0.127)	(0.119)
log($Initial\_Electricity\_Price_{i,t_0}$)	0.349***	0.319***	0.049	0.124
	(0.122)	(0.126)	(0.077)	(0.083)
K-P F stat	12.1	11.9	12.1	11.9
N	1294000	955000	1294000	955000
Industry × Year × Entry Year FE	Yes	Yes	Yes	Yes
Industry × State FE	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the effects of initial and contemporaneous log electricity prices on the log of plants' energy productivity relative to labor productivity and on the log of plants' total factor productivity. Models in Panel A are estimated using OLS and models in Panel B are estimated using IV. In IV models, electricity prices are instrumented using initial and contemporaneous prices for coal, natural gas, and petroleum, weighted by the share of each fuel in state electricity generation. Electricity prices are measured in dollars per kWh, energy productivity is the productivity of electricity relative to labor, and total factor productivity is the productivity common to all manufacturing inputs. Electricity-intensive industries are industries for which electricity accounts for at least 70% of total energy expenditures. Regressions are weighted using Census sampling weights. All dollar values are in 2011 USD. Standard errors clustered by state are in parentheses.

Table 3.6: Heterogeneous Effects of Initial Electricity Prices on Energy Intensity and Productivity, by Plant Age

	$\log(Electricity\_Intensity_{i,t})$	$\log(CO_2\_Intensity_{i,t})$	$\log(BTU\_Intensity_{i,t})$	$\log(Energy\_Productivity_{i,t})$	$\log(TFP_{i,t})$
	(1)	(2)	(3)	(4)	(5)
$\log(Current\_Electricity\_Price_{i,t})$	-0.871***	-1.136***	-0.903***	0.739***	0.109
	(0.061)	(0.097)	(0.069)	(0.131)	(0.096)
$\log(Initial\_Electricity\_Price_{i,t_0})$	-0.134**	-0.140**	-0.076	0.250*	0.097
	(0.052)	(0.061)	(0.054)	(0.130)	(0.095)
$\log(Initial\_Electricity\_Price_{i,t_0}) \times Age_{i,t}$	0.006**	0.006	0.004	-0.006	-0.010***
	(0.003)	(0.004)	(0.004)	(0.007)	(0.003)
K-P F stat	13.3	13.3	13.3	13.3	13.3
N	1294000	1294000	1294000	1294000	1294000
Industry × Year × Entry Year FE	Yes	Yes	Yes	Yes	Yes
Industry × State FE	Yes	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: All models are estimated using IV. Electricity prices are instrumented using initial and contemporaneous prices for coal, natural gas, and petroleum, weighted by the share of each fuel in state electricity generation, and initial electricity prices × plant age is instrumented using the interaction of the initial weighted fuel prices times age. Electricity prices are measured in dollars per kWh, electricity intensity is kWh per dollar of revenue, CO2 intensity is kg CO2 per dollar of revenue, BTU intensity is BTU per dollar of revenue, energy productivity is the productivity of electricity relative to labor, total factor productivity is the productivity common to all manufacturing inputs, and plant age is measured in years since entry. Regressions are weighted using Census sampling weights. All dollar values are in 2011 USD. Standard errors clustered by state are in parentheses.

Appendices

Appendix: Technology Lock-in and Optimal Carbon Pricing

Data

This section provides details on data sources and construction of the primary variables.

We impose several sample restrictions on the measures of firms' inputs and outputs in the ASM and CMF to reduce measurement error. These restrictions closely follow those imposed in other papers using the ASM and CMF (e.g., Ganapati et al., 2020). First, we drop observations for which electricity prices, electricity intensity, capital investment, revenue, labor costs, materials costs, electricity expenditures, or raw fuels expenditures are missing or negative. Second, we exclude observations for which electricity prices, revenue, labor costs, materials costs, or electricity expenditures are equal to 0. Third, we exclude imputed administrative records. Fourth, since some observations still appear to be errors, we drop outliers that have capital stocks, revenue, labor costs, materials costs, electricity expenditures, or raw fuels expenditures that exceed 100 times the 99th percentile of the distribution of these variables. Finally, we exclude observations with electricity prices that are more than ten times or less than one-tenth of the annual median price.

We calculate annual plant-level electricity prices using plants' reported electricity expenditures and purchased quantities from the ASM and CMF, but we do not always observe plants' initial energy prices since only a subset of plants are surveyed in their year of entry. We match plants to their own initial electricity prices using unique plant identifiers where possible. If a plant is not observed in its first year of operations, then we impute its initial electricity price using the average of entrants in the same year, state, and six-digit NAICS industry. For a small number of plants, we use the average of entrants in the same year and state since there are no other entrants in the same industry in the plants' state and year. We also use these year × state × industry (or alternatively year × state) electricity prices to define the electricity prices in the year before a plant enters because we don't have plant-level electricity prices available before a plant opens. We consider that a plant opens during a period of increasing electricity prices if the state average electricity price in its entry year is greater than in the year prior to its entry.

We use the MECS and ASM Fuels Trailers to calculate measures of energy intensity of production that include raw fuels inputs (e.g., coal, natural gas, oil) in addition to electricity. The ASM and CMF include information on total expenditures on raw fuels, but don't include information on how these costs are split between fuels or what quantities are consumed of each. This breakdown is available in the MECS every three years 1985-1994 and every four years 1994-2014 and in the ASM Fuels Trailers for the years 1976-1981. In these surveys, we convert quantities of raw fuels consumed to British thermal units (BTU) using conversion factors from the EIA and to CO2 using data from the EIA where possible and from the EPA for crude oil, biomass, blast furnace gas, coke oven gas, waste gas, and acetylene. We calculate the industry × year BTU and CO2 consumed per dollar of raw fuels expenditure, weighting by the survey weights provided. Expenditures on raw fuels are deflated to 2011 dollars using the industry's annual average energy deflator from the NBER-CES Productivity Database. We exclude fuels used as feedstocks and process emissions in these calculations (Lyubich et al., 2018).

We use these industry average measures of energy consumed per dollar of raw fuels to calculate the total BTU and CO2 implied by each plant's raw fuels expenditures in the ASM and the CMF. To do so, we merge the raw fuels coefficients with the ASM and CMF, and linearly interpolate the coefficients in the missing years separately for each industry. We replace resulting negative coefficients by 0 for 1% of observations; in these cases, all energy consumed comes from electricity. We then calculate the BTU and CO2 embodied in raw fuels as the annual industry average energy coefficient times expenditure on raw fuels in the ASM and CMF, and we calculate the BTU and CO2 embodied in electricity consumption at the plant level using quantities reported in the ASM and MECS. The conversion factors for mWh of electricity to BTU comes from the EIA and the conversion factors for mWh to kg of CO2 come from the EPA's eGRID, which includes separate emissions factors by state that consider the energy mix of each state's electricity grid.

These estimates of BTU and CO2 embodied in energy inputs allow us to calculate measures of BTU and CO2 per dollar of revenue; these alternative measures of energy intensity complement our use of electricity consumed per dollar of revenue in the regression analysis. Since some observations are obvious outliers, we trim the BTU and CO2 intensities that exceed the 99th percentile of the distribution of values. Our energy intensities are comparable to estimates in the literature. For example, the average CO2 intensity of manufacturing that we calculate is within 15% of estimates from Lyubich et al. (2018) using the same MECS year.

A final note about this imputation process is that the ASM Fuels Trailers include substantially less detail than the MECS. Raw fuels are presented at much higher levels of aggregation (e.g., aggregate coal consumption, rather than consumption of different types of coal) and several fuels are grouped into an "others" category, which we exclude. We therefore present results using only the MECS to impute energy consumption from raw fuels and results using both the MECS and ASM Fuels Trailers. We find very similar results using both approaches.

Imputation of Missing Capital Stocks

Capital stocks are a necessary input into the production function estimation, but unlike other inputs are not measured every year. Capital stocks are measured in the CMF in years ending in 2 or 7, and capital investment is measured in both the CMF and, in the intervening years, in the ASM. To obtain estimates of capital stocks in all years, we first calibrate the depreciation rate δ using plants which we observe every year between Censuses. Approximately 12,000 plants are surveyed in the ASM every year between the two most recent Censuses in our sample period (i.e., 2002 and 2007). We iteratively apply the law of motion of capital $K_{i,t+1} = (1-\delta)K_{i,t} + I_{i,t}$ to back out the depreciation rate implied by beginning capital stock $K_{i,t}$ and ending capital stocks $K_{i,t+1}$ and intervening path of investment $I_{i,t}$. Specifically, for each plant i we solve:

$$K_{2007} = (1-\delta)^5 K_{2002} + (1-\delta)^4 I_{2002} + (1-\delta)^3 I_{2003} + (1-\delta)^2 I_{2004} + (1-\delta) I_{2005} + I_{2006} \quad (8)$$

We calculate the average depreciation rate over all plants.

We then use this depreciation rate to recursively calculate capital stocks in the years between Censuses using the law of motion of capital combined with observed investment in the ASM. Specifically, for plants surveyed in the years before and after a Census, we obtain their capital stock using investment from the ASM in those years combined with depreciation. We recursively apply the same approach to plants observed two years before and after a Census. A small number of capital stocks are still missing after applying this procedure. We predict these values using the interaction of (log) total value of shipments with six digit NAICS industry codes. Our results are robust to excluding these observations as well.

Bartik Instruments and Identification

This section comprises a more detailed discussion of testing the validity of the identifying assumption underlying our Bartik-style shift-share instruments.

As Section 3.6 describes, our instrumental variables analysis uses an exposure design that isolates plausibly exogenous variation in electricity prices using states' differential exposure to national changes in the prices of raw fuels (e.g., coal, natural gas, petroleum), where the weights are the shares of electricity generated using each of these fuels. Goldsmith-Pinkham et al. (2020) show that such a research design requires exogeneity of the shares for the identifying assumption to hold because the Bartik instrument is numerically equivalent to a Generalize Method of Moments (GMM) estimator using shares as instruments. As a consequence, our identifying assumption in equation (3.4) is valid if the state fuel generation shares are uncorrelated with shocks leading to changes in the energy intensity outcomes.

The potential concern in our setting is centered on whether state fuel generation shares might be correlated to other shocks to plants' input mix that affect energy intensity directly, rather than through electricity prices. For example, if fuel generation shares are correlated with the availability of skilled labor, then we might be concerned that the instruments are correlated with unobserved shocks to labor inputs, and thus that the identifying assumption (3.4) would be violated.

We assess the validity of our research design by analyzing whether state characteristics that could be correlated with other input shocks also predict state fuel generation shares (Goldsmith-Pinkham et al., 2020). Appendix Table C.3 reports the results from regressing the shares of electricity generated using coal, natural gas, and petroleum on state characteristics plausibly related to input availability (e.g., unemployment rate, share college educated) and output demand (e.g., mean household income). Reassuringly, this now standard test yields no systematic correlation between the shares and these characteristics, which supports the validity of the research design.[33]

Robustness of Instrumental Variables Analysis

Energy Intensity

Appendix Table C.7 shows that the sign, magnitude, and precision of the main estimates are robust to the use of different covariates, weightings, and data subsamples. For comparison, Panel A reproduces the main estimates of lock-in based on equation (3.2) and shown in Table 3.3.

Panels B and C show that both initial and current electricity prices elasticities are robust to the use of different covariates. Panel B presents results that are almost identical using higher-level fixed effects, which suggests that variation at the year × first year × industry level does not confound the estimates. Panel C includes state × year time trends, which allow for differential energy efficiency trends by state over time. The results again are comparable, with a small increase in the magnitude of the point estimates.

Panel D presents estimates that are not weighted by the Census sampling weights to create a representative sample. Since the ASM oversamples large plants, these plants are assigned higher weight in these regressions relative to the main estimates. The initial price elasticity falls by approximately one-third, suggesting that lock-in is not just driven by large plants, while the sign and significance are generally unchanged.

Panels E, F, and G show estimates using different subsamples of the data. Panel E linearly interpolates CO2 and BTU values from the MECS only, rather than the

[33] We report what we consider more conservative estimates of the significance of these correlations that do not adjust for the multiple hypotheses that we are testing.

MECS and the ASM Fuels Trailers. The MECS contains approximately five times as many fuel categories as the ASM Fuel Trailers (e.g., detailed coal subtypes v. all coal), but the imputing the ASM years using the MECS barely changes the estimated effect of electricity prices on CO2 and BTU intensities at all (Columns 3 and 4). Panel F excludes years in which the ASM Fuels Trailers and the MECS are not collected. This skews the analysis sample toward the early years of the data since the ASM Fuels Trailers were collected every year between 1976 and 1981, while the MECS is subsequently collected every three or four years. The lock-in estimates are larger as a result: the effects of entry-year electricity prices on electricity, CO2, and BTU intensity all increase by approximately one-third. The change in the magnitude of the parameter estimates is the result of the changing time period, rather than the imputation, since the electricity intensities are never imputed. The parameter estimates in Panel G, which exclude the ASM Fuel Trailer years, are closer to the main estimates, though still oversamples the early years of the data.

Productivity

Appendix Table C.8 shows alternative estimates of the effects of initial and current electricity prices on productivity. We analyze the same models as in Appendix Table C.7 using relative energy productivity and total factor productivity as the outcome variables, and we find again find results that are consistent with the main estimates in Table 3.5. We reproduce these results in Panel A of Appendix Table C.8 for comparison.

Panels B and C use different covariates than the models in the main text. We find estimates that are similar in sign, magnitude, and precision to the main estimates using higher level fixed effects (Panel B) and using state × year time trends (Panel C). The magnitude of contemporaneous electricity prices for relative energy intensity increases slightly with the inclusion of state × year trends, suggesting that there may be some differential trends in energy productivity between states, though the estimates are not statistically different from each other. Meanwhile, the entry-year lock-in estimates of the effects of initial electricity prices on electricity intensity are almost entirely unchanged, as are the effects of both initial and contemporaneous prices on total factor productivity.

The estimates in Panel D are unweighted by Census sampling weights. As discussed above, the ASM oversamples large plants; the regression estimates again are similar or, in the case of relative energy productivity, slightly larger, suggesting again that the lock-in estimates are not driven by large plants, or by reweighting.

Panels E, F, and G show estimates using different subsamples of the data. Panel E uses linearly interpolates CO2 and BTU values from the MECS only, rather than the MECS and the ASM Fuels Trailers. The MECS contains approximately five times as many fuel categories as the ASM Fuel Trailers (e.g., detailed coal subtypes v. all coal), but imputing the ASM years using the MECS barely changes the estimated effect of electricity prices on CO2 and BTU intensities at all (Columns 3 and 4). Since the relative energy intensity estimates don't use imputed CO2 or BTU values, this sample is equivalent to the main sample for these productivity models, though is different in the case of Appendix Table C.7.

Panel F excludes years in which the ASM Fuels Trailers and the MECS are not collected. This skews the analysis sample toward the early years of the data since the ASM Fuels Trailers were collected every year between 1976 and 1981, while the MECS is subsequently collected every three or four years. The lock-in estimates again are similar in sign, magnitude, and precision, while the importance of contemporaneous electricity prices falls slightly, perhaps as a result of fewer intervening years between the measurement of initial and current prices. This same pattern is evident in Panel G, which excludes ASM Fuel Trailer years from the Panel F sample, though overall the results are quantitatively and qualitatively similar in all models.

Electricity Price Effects on Other Manufacturing Outcomes

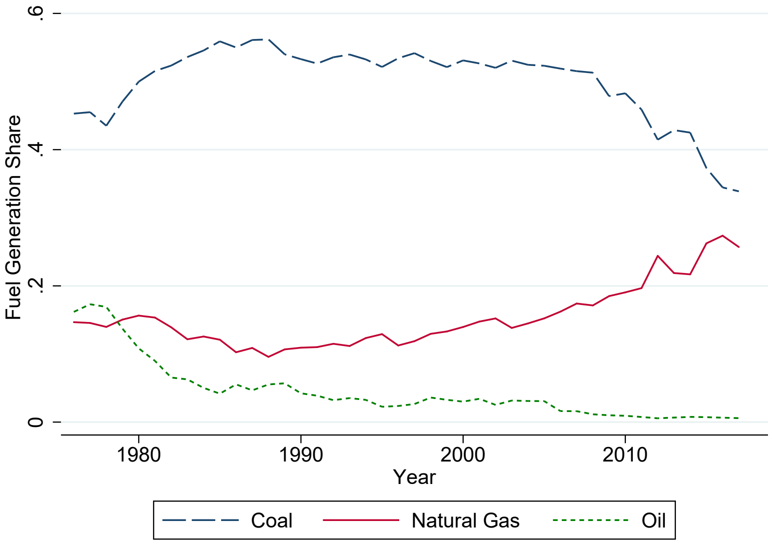
This section discuss the effects of initial electricity prices on manufacturing outcomes other than energy intensity and energy productivity.

Panel A reports the effects of initial electricity prices on the amounts of electricity used, CO2 produced, and BTU consumed, rather than the intensity amounts measured per dollar of revenue. These results are similar in sign and magnitude to the main energy intensity estimates, though in some instances are somewhat less precisely estimated. These estimates are again consistent with lock-in: plants use a greater quantity of all of these energy measures when they begin operations in a low energy price year. The magnitudes of the effects on energy quantities can again be explained by the persistent effect on relative productivity, shown in Table 3.5.

Panel B reports effects on other manufacturing inputs. The effects of initial electricity prices on labor hours, capital outlays, and materials costs (excluding energy) are generally statistically insignificant, with the exception of a weakly positive effect on capital. These findings suggest that while initial electricity prices have important effects on future energy inputs, they have a limited effect on other non-energy inputs. In particular, higher entry-year energy costs appear unlikely to lead to widespread unemployment, as has been raised as a potentially concerning effect of pricing carbon.

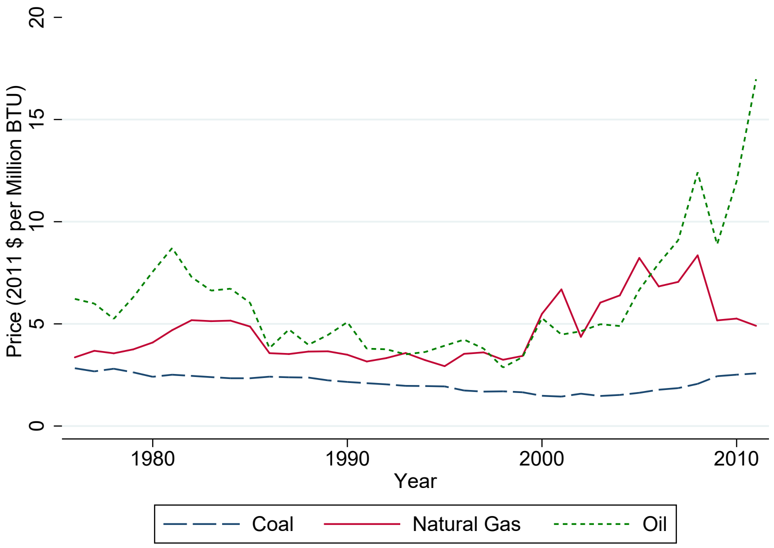
Figures and Tables

Figure C.1: Time Series of Shares of Fossil Fuels used in Electricity Generation



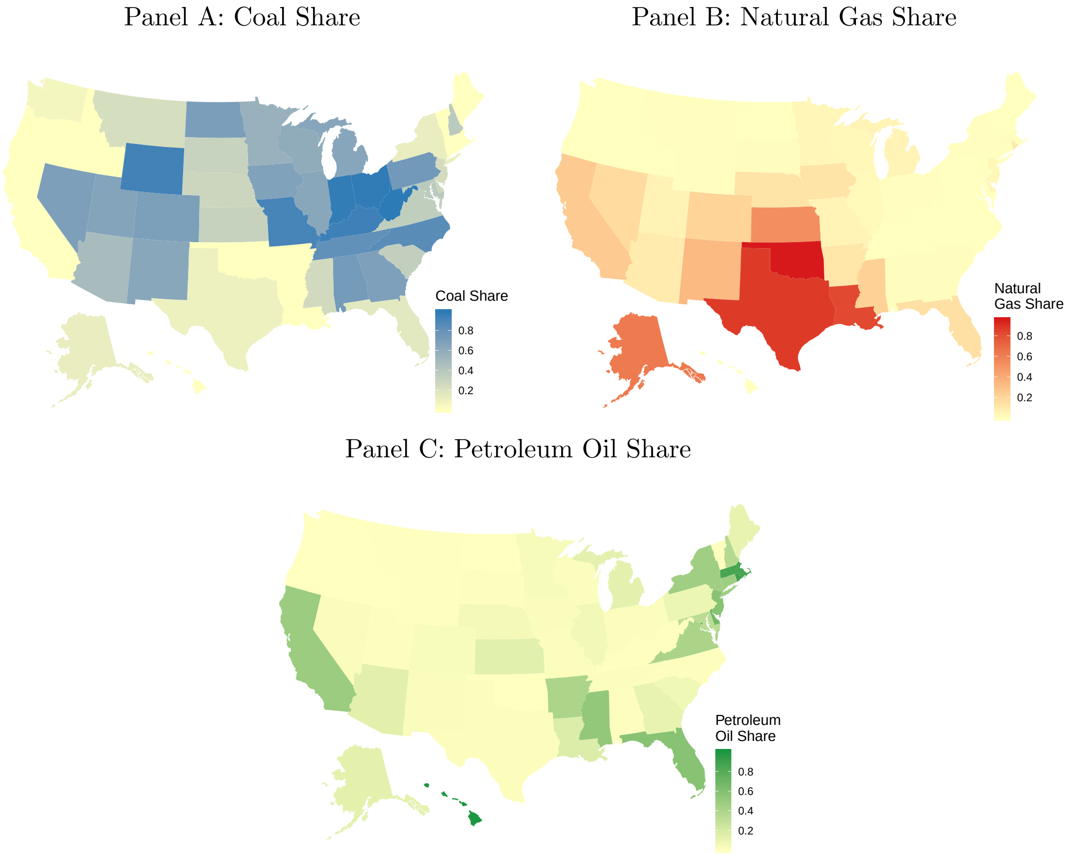
Notes: This figure shows the time series of the fraction of British thermal units (BTU) of electricity generated by coal, natural gas, and petroleum oil in the United States.

Figure C.2: Time Series of Fuel Prices



Notes: This figure shows the time series of average coal, natural gas, and petroleum oil prices paid by electric utilities. Prices are in 2011 dollars per million British thermal units (BTU) of fuel.

Figure C.3: Share of Electricity Generated by Coal, Natural Gas, and Petroleum Oil in 1976



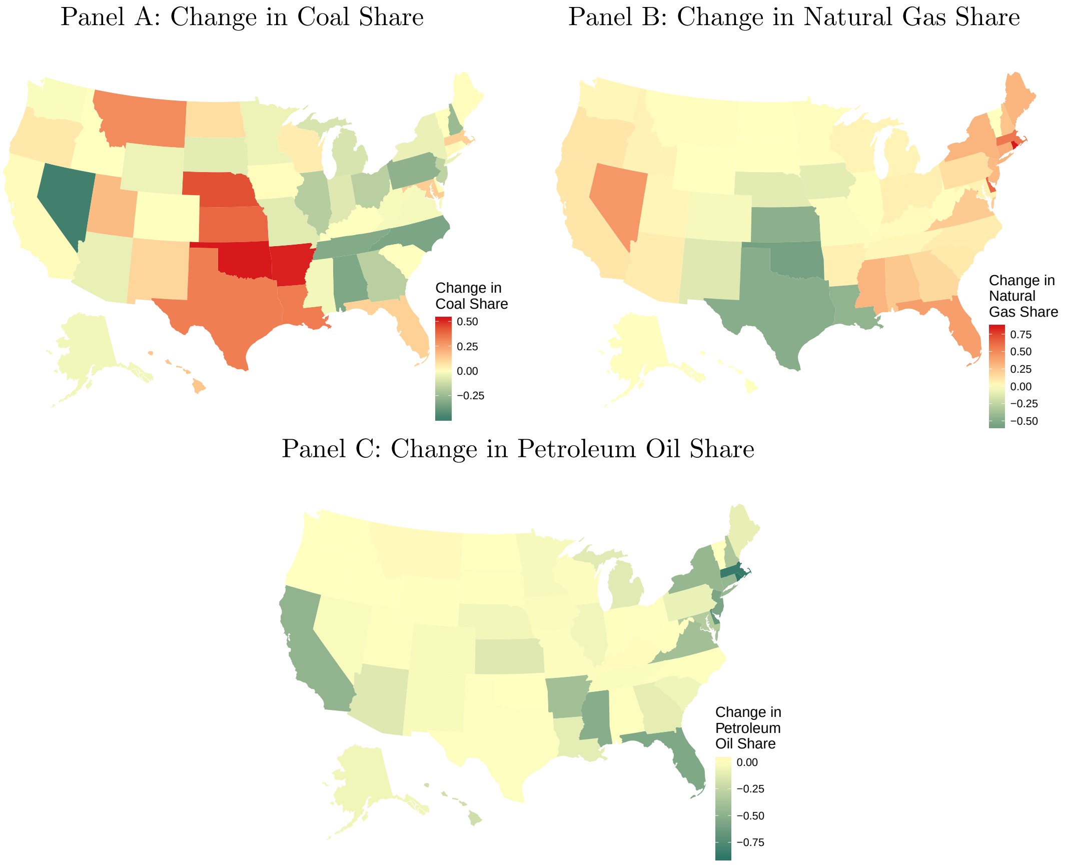
Panel A: Coal Share

Panel B: Natural Gas Share

Panel C: Petroleum Oil Share

Notes: This figure shows the fraction of total British thermal units (BTUs) of electricity generated from coal, natural gas, and petroleum oil in each state in 1976. Shares need not add up to 1 due to the presence of other fuel sources (e.g., nuclear, hydro).

Figure C.4: Change in Share of Electricity Generated by Coal, Natural Gas, and Petroleum Oil, 1976-2011



Panel A: Change in Coal Share

Panel B: Change in Natural Gas Share

Panel C: Change in Petroleum Oil Share

Notes: This figure shows the change in the fraction of total British thermal units (BTUs) of electricity generated from coal, natural gas, and petroleum oil in each state between 1976 and 2011.

Table C.1: Summary Statistics

	All Industries	Electricity-Intensive Industries
	(1)	(2)
Year	1997	1997
	(8.595)	(8.540)
Entry Year	1988	1988
	(8.894)	(8.760)
Plant Age (years)	9.243	9.297
	(8.000)	(7.997)
Current Electricity Price ($ per kWh)	0.087	0.085
	(0.036)	(0.034)
Initial Electricity Price ($ per kWh)	0.088	0.088
	(0.032)	(0.030)
Cost of Purchased Electricity (1000$)	275.4	203.4
	(2065)	(958)
Quantity of Purchased Electricity (1000 kWh)	4411	3175
	(51160)	(22430)
Electricity Intensity (kWh per $ revenue)	0.196	0.208
	(0.500)	(0.500)
CO2 Intensity (kg per $ revenue)	0.122	0.132
	(0.469)	(0.519)
BTU Intensity (million BTU per $ revenue)	0.001	0.002
	(0.006)	(0.007)
N	1294000	955000

Notes: This table shows variable means for U.S. manufacturing plants. Electricity-intensive industries are industries for which electricity accounts for at least 70% of total energy expenditures. All dollar values are in USD (2011). Standard errors are in parentheses.

Table C.2: Entry and Exit Summary Statistics

	All Industries	Elec. Intensive Industries
	(1)	(2)
Entrant Fraction	0.078	0.077
	(0.269)	(0.266)
Exit Fraction	0.004	0.004
	(0.060)	(0.060)
Observations per Plant	4.594	4.466
	(4.765)	(4.466)
Plant Age	9.243	9.297
	(8.00)	(7.997)
Age at Exit	14.960	15.010
	(9.150)	(9.142)

Notes: This table shows summary means for plant entry and exit behavior. Entry and exit fractions are the shares of total plant-year observations in our sample that are entrants or exiters, respectively. Plant age and age at exit are measured in years. Column 1 shows means across all industries and column 2 shows means for electricity-intensive industries. Standard errors are in parentheses.

Table C.3: Relationship between Fuel Generation Shares and State Characteristics

	Coal Share	Natural Gas Share	Petroleum Share
	(1)	(2)	(3)
Unemployment Rate	-0.022	-0.039*	-0.001
	(0.033)	(0.022)	(0.023)
State Per Capita Income (1000s)	0.005	0.048	0.034
	(0.043)	(0.029)	(0.029)
Mean Household Income (1000s)	-0.001	-0.000	0.001
	(0.001)	(0.001)	(0.001)
Share Any College Education	-0.030*	-0.015	0.007
	(0.016)	(0.011)	(0.011)
Share White	0.008	0.001	-0.009*
	(0.007)	(0.005)	(0.005)
Share Black	0.004	0.003	-0.003
	(0.007)	(0.005)	(0.005)
Population (1000s)	-0.004	0.005	0.001
	(0.010)	(0.007)	(0.007)
Household Size	0.506	0.517	-0.215
	(0.673)	(0.454)	(0.458)
Dep. Var. Mean (1980)	0.45	0.12	0.16
Dep. Var. Mean (1976)	0.40	0.12	0.21
R-square	0.266	0.154	0.443
N	51	51	51

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the correlation between state fuel shares in electricity generation and state characteristics in 1980. Standard errors are in parentheses.

Table C.4: Estimated Production Function Parameters

	All Industries	Elec. Intensive Industries
	(1)	(2)
Returns to scale ν	0.620	0.679
	(0.292)	(0.256)
Elasticity of substitution σ	0.260	0.237
	(0.195)	(0.186)
Capital productivity β_K	3.411	3.498
	(2.200)	(2.220)
N	1294000	955000

Notes: This table shows the estimated production function parameters. Column 1 shows mean parameter estimates across all industries and column 2 shows means for electricity-intensive industries (i.e., industries for which electricity accounts for at least 70% of total energy expenditures). Standard errors are in parentheses.

Table C.5: Effects of Current Electricity Prices on Energy Intensity

	$\log(Electricity\_Intensity_{i,t})$	$\log(Electricity\_Intensity_{i,t})$	$\log(CO_2\_Intensity_{i,t})$	$\log(BTU\_Intensity_{i,t})$
		Electricity-Intensive Industries		
	(1)	(2)	(3)	(4)
Panel A: OLS				
$\log(Current\_Electricity\_Price_{i,t})$	-0.855***	-0.835***	-0.828***	-0.810***
	(0.012)	(0.011)	(0.010)	(0.009)
Panel B: IV				
$\log(Current\_Electricity\_Price_{i,t})$	-0.777***	-0.768***	-0.899***	-0.769***
	(0.086)	(0.102)	(0.070)	(0.084)
K-P F stat	11.7	10.9	11.7	11.7
N	1294000	955000	1294000	1294000
Industry × Year × Entry Year FE	Yes	Yes	Yes	Yes
Industry × State FE	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Models in Panel A are estimated using OLS and models in Panel B are estimated using IV. In IV models, electricity prices are instrumented using contemporaneous prices for coal, natural gas, and petroleum, weighted by the share of each fuel in state electricity generation. Electricity prices are measured in dollars per kWh, electricity intensity is kWh per dollar of revenue, CO2 intensity is kg CO2 per dollar of revenue, and BTU intensity is BTU per dollar of revenue. Electricity-intensive industries are industries for which electricity accounts for at least 70% of total energy expenditures. Regressions are weighted using Census sampling weights. All dollar values are in 2011 USD. Standard errors clustered by state are in parentheses.

Table C.6: Effects of Initial Electricity Prices on Energy Intensity

	log($Electricity\_Intensity_{i,t}$)	log($Electricity\_Intensity_{i,t}$)	log($CO_2\_Intensity_{i,t}$)	log($BTU\_Intensity_{i,t}$)
		Electricity-Intensive Industries		
	(1)	(2)	(3)	(4)
Panel A: OLS				
log($Initial\_Electricity\_Price_{i,t_0}$)	-0.194***	-0.183***	-0.178***	-0.172***
	(0.013)	(0.014)	(0.011)	(0.012)
Panel B: IV				
log($Initial\_Electricity\_Price_{i,t_0}$)	-0.554***	-0.589***	-0.707***	-0.538***
	(0.113)	(0.125)	(0.113)	(0.112)
K-P F stat	12.0	12.4	12.0	12.0
N	1294000	955000	1294000	1294000
Industry × Year × Entry Year FE	Yes	Yes	Yes	Yes
Industry × State FE	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Models in Panel A are estimated using OLS and models in Panel B are estimated using IV. In IV models, entry-year electricity prices are instrumented using entry-year prices for coal, natural gas, and petroleum, weighted by the share of each fuel in state electricity generation. Electricity prices are measured in dollars per kWh, electricity intensity is kWh per dollar of revenue, CO2 intensity is kg CO2 per dollar of revenue, and BTU intensity is BTU per dollar of revenue. Electricity-intensive industries are industries for which electricity accounts for at least 70% of total energy expenditures. Regressions are weighted using Census sampling weights. All dollar values are in 2011 USD. Standard errors clustered by state are in parentheses.

Table C.7: Effects of Initial and Current Electricity Prices on Energy Intensity

	log($Electricity\_Intensity_{i,t}$)	log($Electricity\_Intensity_{i,t}$)	log($CO_2\_Intensity_{i,t}$)	log($BTU\_Intensity_{i,t}$)
		Electricity-Intensive Industries		
	(1)	(2)	(3)	(4)
Panel A: Main Results				
log($Current\_Electricity\_Price_{i,t}$)	-0.764***	-0.734***	-0.829***	-0.761***
	(0.090)	(0.104)	(0.072)	(0.087)
log($Initial\_Electricity\_Price_{i,t_0}$)	-0.165***	-0.232***	-0.289***	-0.144**
	(0.051)	(0.059)	(0.079)	(0.059)
K-P F stat	12.1	11.9	12.1	12.1
N	1294000	955000	1294000	1294000
Panel B: Year × Industry, First Year × Industry, State × Industry FE				
log($Current\_Electricity\_Price_{i,t}$)	-0.770***	-0.726***	-0.831***	-0.763***
	(0.097)	(0.106)	(0.080)	(0.095)
log($Initial\_Electricity\_Price_{i,t_0}$)	-0.178***	-0.226***	-0.319***	-0.186***
	(0.051)	(0.070)	(0.074)	(0.057)
K-P F stat	9.7	10.2	9.7	9.7
N	1294000	955000	1294000	1294000
Panel C: State × Year Trends				
log($Current\_Electricity\_Price_{i,t}$)	-0.958***	-1.033***	-1.003***	-1.053***
	(0.127)	(0.144)	(0.081)	(0.129)
log($Initial\_Electricity\_Price_{i,t_0}$)	-0.219***	-0.311***	-0.218***	-0.206***
	(0.062)	(0.075)	(0.062)	(0.059)
K-P F stat	11.9	11.3	11.9	11.9
N	1294000	955000	1294000	1294000
Panel D: Unweighted				
log($Current\_Electricity\_Price_{i,t}$)	-0.821***	-0.809***	-0.883***	-0.827***
	(0.044)	(0.044)	(0.045)	(0.041)
log($Initial\_Electricity\_Price_{i,t_0}$)	-0.109**	-0.153**	-0.172***	-0.074
	(0.052)	(0.067)	(0.059)	(0.045)
K-P F stat	11.6	10.8	11.6	11.6
N	1294000	955000	1294000	1294000
Panel E: Impute CO2, BTU from MECS only				
log($Current\_Electricity\_Price_{i,t}$)	-0.764***	-0.734***	-0.811***	-0.786***
	(0.090)	(0.104)	(0.085)	(0.085)
log($Initial\_Electricity\_Price_{i,t_0}$)	-0.165***	-0.232***	-0.193***	-0.152**
	(0.051)	(0.059)	(0.068)	(0.063)
K-P F stat	12.1	11.9	12.1	12.1
N	1294000	955000	1294000	1294000
Panel F: Exclude Years with Imputed CO2 and BTU Values				
log($Current\_Electricity\_Price_{i,t}$)	-0.625***	-0.585***	-0.800***	-0.564***
	(0.137)	(0.161)	(0.098)	(0.128)
log($Initial\_Electricity\_Price_{i,t_0}$)	-0.368***	-0.451***	-0.715***	-0.256**
	(0.091)	(0.099)	(0.134)	(0.106)
K-P F stat	11.1	11.9	11.1	11.1
N	312000	225000	312000	312000
Panel G: MECS Years Only				
log($Current\_Electricity\_Price_{i,t}$)	-0.685***	-0.696***	-0.722***	-0.725***
	(0.116)	(0.138)	(0.107)	(0.104)
log($Initial\_Electricity\_Price_{i,t_0}$)	-0.343***	-0.424***	-0.370***	-0.367***
	(0.089)	(0.125)	(0.085)	(0.086)
K-P F stat	9.4	10.0	9.4	9.4
N	266000	192000	266000	266000

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: All models are estimated using IV. Initial and contemporaneous electricity prices are instrumented using initial and contemporaneous prices for coal, natural gas, and petroleum, weighted by the share of each fuel in state electricity generation. Electricity prices are measured in dollars per kWh, electricity intensity is kWh per dollar of revenue, CO2 intensity is kg CO2 per dollar of revenue, and BTU intensity is BTU per dollar of revenue. Electricity-intensive industries are industries for which electricity accounts for at least 70% of total energy expenditures. Except for Panel B, all models include industry × year × entry year fixed effects and industry × state fixed effects. Regressions are weighted using Census sampling weights unless otherwise noted. All dollar values are in 2011 USD. Standard errors clustered by state are in parentheses.

Table C.8: Effects of Initial and Current Electricity Prices on Productivity

	$\log(Energy\_Productivity_{i,t})$	$\log(Energy\_Productivity_{i,t})$ Electricity-Intensive Industries	$\log(TFP_{i,t})$	$\log(TFP_{i,t})$ Electricity-Intensive Industries
	(1)	(2)	(3)	(4)
Panel A: Main Results				
$\log(Current\_Electricity\_Price_{i,t})$	0.525***	0.673***	0.088	-0.017
	(0.138)	(0.139)	(0.127)	(0.119)
$\log(Initial\_Electricity\_Price_{i,t_0})$	0.349***	0.319***	0.049	0.124
	(0.122)	(0.126)	(0.077)	(0.083)
K-P F stat	12.1	11.9	12.1	11.9
N	1294000	955000	1294000	955000
Panel B: Year × Industry, First Year × Industry, State × Industry FE				
$\log(Current\_Electricity\_Price_{i,t})$	0.512***	0.671***	0.136	0.047
	(0.169)	(0.150)	(0.118)	(0.113)
$\log(Initial\_Electricity\_Price_{i,t_0})$	0.330***	0.270***	0.021	0.087
	(0.100)	(0.115)	(0.074)	(0.098)
K-P F stat	9.7	10.2	9.7	10.2
N	1294000	955000	1294000	955000
Panel C: State × Year Trends				
$\log(Current\_Electricity\_Price_{i,t})$	0.928***	1.109***	0.040	-0.049
	(0.295)	(0.289)	(0.146)	(0.126)
$\log(Initial\_Electricity\_Price_{i,t_0})$	0.322***	0.352***	0.089	0.141
	(0.140)	(0.156)	(0.082)	(0.093)
K-P F stat	11.9	11.3	11.9	11.3
N	1294000	955000	1294000	955000
Panel D: Unweighted				
$\log(Current\_Electricity\_Price_{i,t})$	0.465***	0.589***	0.174	0.059
	(0.124)	(0.123)	(0.105)	(0.078)
$\log(Initial\_Electricity\_Price_{i,t_0})$	0.425***	0.412***	-0.078	0.017
	(0.087)	(0.092)	(0.049)	(0.055)
K-P F stat	11.6	10.8	11.6	10.8
N	1294000	955000	1294000	955000
Panel E: Impute CO2, BTU from MECS only				
$\log(Current\_Electricity\_Price_{i,t})$	0.525***	0.673***	0.088	-0.017
	(0.138)	(0.139)	(0.127)	(0.119)
$\log(Initial\_Electricity\_Price_{i,t_0})$	0.349***	0.319***	0.049	0.124
	(0.122)	(0.126)	(0.077)	(0.083)
K-P F stat	12.1	11.9	12.1	11.9
N	1294000	955000	1294000	955000
Panel F: Exclude Years with Imputed CO2 and BTU Values				
$\log(Current\_Electricity\_Price_{i,t})$	0.413***	0.421***	0.103	0.049
	(0.150)	(0.188)	(0.099)	(0.095)
$\log(Initial\_Electricity\_Price_{i,t_0})$	0.376***	0.557***	0.091	0.132
	(0.193)	(0.186)	(0.102)	(0.105)
K-P F stat	11.1	11.9	11.1	11.9
N	312000	225000	312000	225000
Panel G: MECS Years Only				
$\log(Current\_Electricity\_Price_{i,t})$	0.411***	0.415***	0.126	0.122
	(0.144)	(0.162)	(0.090)	(0.092)
$\log(Initial\_Electricity\_Price_{i,t_0})$	0.503***	0.573***	0.047	0.097
	(0.177)	(0.193)	(0.101)	(0.111)
K-P F stat	9.4	10.0	9.4	10.0
N	266000	192000	266000	192000

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: All models are estimated using IV. Initial and contemporaneous electricity prices are instrumented using initial and contemporaneous prices for coal, natural gas, and petroleum, weighted by the share of each fuel in state electricity generation. Electricity prices are measured in dollars per kWh, energy productivity is the productivity of electricity relative to labor, and total factor productivity is the productivity common to all manufacturing inputs. Electricity-intensive industries are industries for which electricity accounts for at least 70% of total energy expenditures. Except for Panel B, all models include industry × year × entry year fixed effects and industry × state fixed effects. Regressions are weighted using Census sampling weights unless otherwise noted. All dollar values are in 2011 USD. Standard errors clustered by state are in parentheses.

Table C.9: Effects of Initial and Current Electricity Prices on Manufacturing Outcomes

	(1)	(2)	(3)
Panel A: Energy Inputs (Levels)			
	log($Quantity\_Electricity_{i,t}$)	log($Total\_CO2_{i,t}$)	log($Total\_BTU_{i,t}$)
log($Current\_Electricity\_Price_{i,t}$)	-0.308*	-0.373**	-0.305*
	(0.170)	(0.183)	(0.178)
log($Initial\_Electricity\_Price_{i,t_0}$)	-0.240*	-0.364***	-0.218*
	(0.133)	(0.112)	(0.124)
Panel B: Other Manufacturing Inputs			
	log($Labor\_Hours_{i,t}$)	log($Materials\_Costs_{i,t}$)	log($Capital\_Investment_{i,t}$)
log($Current\_Electricity\_Price_{i,t}$)	0.470***	0.421*	0.391
	(0.153)	(0.239)	(0.287)
log($Initial\_Electricity\_Price_{i,t_0}$)	-0.109	-0.136	0.434*
	(0.124)	(0.146)	(0.247)
K-P F stat	12.1	12.1	12.1
N	1294000	1294000	1294000
Industry × Year × Entry Year FE	Yes	Yes	Yes
Industry × State FE	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: All models are estimated using IV. Initial and contemporaneous electricity prices are instrumented using initial and contemporaneous prices for coal, natural gas, and petroleum, weighted by the share of each fuel in state electricity generation. Electricity prices are measured in dollars per kWh, labor inputs in hours, materials costs and capital investment in 1000s, quantity of electricity purchased in 1000 kWh, quantity CO2 produced in kg, and quantity BTU consumed in million BTU. Regressions are weighted using Census sampling weights. All dollar values are in 2011 USD. Standard errors clustered by state are in parentheses.