UNIVERSITY OF CALIFORNIA, SAN DIEGO

The Spatial Organization of Economic Activity in Cities

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Economics

by

Wei You

Committee in charge:

Professor Gordon Hanson, Chair
Professor Roger Gordon
Professor Ruixue Jia
Professor David Lagakos
Professor Victor Shih

2017

The dissertation of Wei You is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____

____

Chair

University of California, San Diego

2017

DEDICATION

To my family and friends.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I acknowledge Professor Gordon Hanson for his support and guidance as the chair of my committee. I appreciate Professor Roger Gordon, whose efforts have significantly improved the first chapter of my dissertation. I thank Walker Hanlon, Leah Boustan, Allison Shertzer, and Carlos Eduardo Hernandez for offering hints about where to find historical data. I also thank Professor Ruixue Jia, David Lagakos, Marc Muendler, Paul Niehaus, Tommaso Porzio, Noam Yuchtman, Gregory Clark, Martha Olney, Dan Bogart, David McKenzie, for offering many helpful suggestions. Special thanks to my wife, Jing Shi, who supports me in all aspects of my life.

Chapter 1, in full, is currently being prepared for submission for publication of the material. The dissertation author, Wei You, was the sole author of this paper.

Chapter 2, in full, is currently being prepared for submission for publication of the material. The dissertation author, Wei You, was the sole author of this paper.

Chapter 3 is coauthored with Ran Goldblatt, Gordon Hanson, and Amit Khandelwal, and in full, is a reprint of the material as it appears in: "Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine", *Remote Sensing*, 2016, 8(8), 634. I thank my coauthors for the permission to use this chapter in my dissertation. The dissertation author, Wei You, was the primary author of this paper.

VITA

2009	B. A. in Economics and Mathematics, Renmin University of China
2011	M. A. in Economics, Peking University
2017	Ph. D. in Economics, University of California, San Diego

PUBLICATIONS

Wei You (joint with Ran Goldblatt, Gordon Hanson, and Amit Khandelwal), “Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine”, *Remote Sensing*, 2016, 8(8), 634.

ABSTRACT OF THE DISSERTATION

The Spatial Organization of Economic Activity in Cities

by

Wei You

Doctor of Philosophy in Economics

University of California, San Diego, 2017

Professor Gordon Hanson, Chair

The three chapters of this dissertation examine the spatial organization of economic activities on a micro scale in a growing historical city, and on a national scale in contemporary India. The first chapter focuses on firms, investigating the question why small family firms were so prevalent in historical cities. I test whether this phenomenon is caused by a lack of technological capability to move goods and people. I exploit the natural experiment that Boston quickly electrified its previous horse-drawn streetcar system between 1889 and 1896. Analyzing new data transcribed from Boston business records from 1885 to 1905, I find that rail-connected locations experienced a 5.3-percentage point relative drop in the share of sole proprietorship establishments after the streetcar electrification, indicating that improved market access leads to an increase in average firm size. The second chapter focuses on individuals, investigating the question whether immigrants stay in ethnic enclaves due to a lack of information about outside communities within the same city. Based on individual records linked between city directories and the decennial census in Boston from 1885 to 1900, I track within-city migrations of immigrants in response to the same transport upgrade event in Boston. I find that immigrant enclave residents who worked within 25 meters of the streetcar rails in 1885 were much more likely to move to a less segregated neighborhood in 1900, compared to enclave residents who worked between 25 and 50 meters away from the rails. Evidence suggests interactions in workplace had an impact on the choices of the residential locations fifteen years after. In the third chapter, my coauthors and I perform a large scale classification of satellite imagery into "built-up" or "not built-up" areas to measure the urbanization process in India, using a reliable and comprehensive ground-truth data set we construct and a cloud-based computational platform - Google Earth Engine (GEE). Our methodology yields a classification accuracy rate between 70% and 85%, and can easily be applied to other countries and regions.

Chapter 1

The Economics of Speed: The Electrification of the Streetcar System and the Decline of Mom-and-Pop Stores in Boston, 1885-1905

1.1 Abstract

Small family firms dominated the American economy in the nineteenth century, and still dominate in many developing economies today. A long-conjectured cause of this phenomenon, represented by Chandler (1977), is that a lack of technological capability to move goods and people precludes the emergence of modern firms. This paper provides the first causal evidence in support of this hypothesis, exploiting the natural experiment that Boston quickly electrified its previous horse-drawn streetcar system between 1889 and 1896 while keeping the preexisting transit routes almost unchanged. The inference comes from comparing changes in firm size in rail-connected locations to changes

in neighboring unconnected locations. Analyzing new data transcribed from Boston business records from 1885 to 1905, I find that rail-connected locations experienced a 5.3-percentage point relative drop in the share of sole proprietorship establishments after the streetcar electrification.

1.2 Introduction

Before the 1840s, “mom-and-pop” businesses dominated the American economy. This type of firm was typically owned and managed by an individual or a small number of family members, dealt in a single product line, and served a highly localized market. Chandler (1977) observed that the size and nature of firms in America remained relatively unchanged between 1790 and 1840, despite the substantial growth in population and the total volume of trade during this period. The increase in market size translated into a parallel increase in the number of firms, but not the size of firms. He proposed that the traditional sources of energy - wind and animal power - “simply could not generate a volume of output in production and number of transactions in distribution large enough to require the creation of a large managerial enterprise or to call for the development of new business forms and practices.” The revolution in transportation technology since the mid-nineteenth century, first brought on by railroads, made it possible to move goods at a steady, high volume and at high speeds, which demanded organizational innovations within production units. It is in this period that we saw the rise of modern firms.

Chandler’s hypothesis is relevant to understand the prevalence of micro and small enterprises in today’s developing countries. Using increasingly available cross-country data, researchers have established the stylized fact that the self-employment rate, a measure that captures the prevalence of micro and small enterprises, decreases as income per capita increases (La Porta and Shleifer, 2008, 2014; Gollin 2008; Jensen, 2016). In the bottom income quartile of countries, nearly one half of the labor force is self-employed, while the fraction of the self-employed in the labor force is around 10% in the top income quartile of countries (La Porta and Shleifer, 2008).[1] In terms of firm size, Hsieh and Olken (2014, p.93) and McKenzie (2016, p.2) report that close to 100% of firms have fewer than 10 workers in India, Indonesia, and Nigeria. In contrast, the modal manufacturing firm in the United States has 45 workers (Hsieh and Klenow, 2014). This is a puzzle to standard models of the firm size distribution (Lucas, 1978) unless the distribution of entrepreneurial talent is very limited in developing countries. Chandler pointed out that a potentially important impediment to firm size growth in developing countries is the higher transportation costs that segment markets, which cause firms to primarily serve a highly localized market. Similar explanations are also proposed by Lagakos (2016), Tybout (2000), Hsieh and Klenow (2014), Holmes and Stevens (2014), and Ziv (2016). In this paper, I call these explanations the market segmentation hypothesis.

Empirically testing the market segmentation hypothesis is difficult for both identification and data availability reasons. The causal evidence that a reduction in transportation costs leads to an increase in firm size has been missing in the literature thus far, both in the historical US setting and in today's developing country setting. One needs to find a transportation technology shock that took place in a short period of time and made differential impacts on different geographic areas, and with the intensity of the impacts independent of the possible trends in the outcome of the areas. Ideally, we need within-city cross-neighborhood variation in transport costs, as the majority of firms employs fewer than 10 workers. Only a neighborhood-level market segmentation seems to be consistent with the firm size distribution. Correspondingly, one needs to find a firm-level dataset that contains detailed geographic information, which is not left-tail

[1] This is true even in the non-agricultural sector.

truncated in the firm size distribution. However, such data sets are rare in developing economies (Hsieh and Olken, 2014).

In this paper, I exploit the natural experiment that Boston quickly electrified its previous horse-drawn streetcar system during 1889-1896 in order to identify the causal effects of an upgrade of transportation infrastructure on the presence of sole proprietorships. This upgrade increased the speed of the best means of intra-city transportation from 4-5 mph to 8-10 mph, tripled transportation capacity, and enabled services to be provided at lower fares (Warner, 1962).[2] More importantly, the majority of the electric streetcar routes were upgraded from previous, long-existing horsecar routes, which avoided non-randomness in the placement of the new transit lines. While the electrification of the streetcar system in Boston was a positive shock to market access citywide, it also improved market access more strongly near the streetcar rails. According to the market segmentation hypothesis, firms that can reach more distant markets should be larger in size. Consequently, we should be able to observe a relative decline in the share of small firms near the streetcar rails compared to declines in off-rail areas.

While the electrification of streetcar systems took place in virtually every major US city in the late-nineteenth and early-twentieth centuries, a number of features make Boston particularly well suited for my study. First, Boston followed a distinct direction when modernizing its commuting system. Unlike other major US cities that adopted a mixture of cable car systems and horsecar systems in the late-nineteenth century, Boston went directly to a more advanced, entirely electric streetcar system from a completely horse-drawn streetcar system. The pace of this electrification was very quick, beginning in 1889 and ending in 1896, making Boston the first major city in the US that adopted a citywide electric streetcar system. Second, Boston has rich historical data sources,

[2] In 1900, the five-cent fare was almost universal, and no additional charge was needed for transfer rides. In contrast, during the 1870s and 1880s, two full eight-cent fares were typical if riders took two cars run by different companies.

including property tax ledgers, which contain assessed property values for every building unit in Boston in my study period. Third, there are two Boston peninsulas - Charlestown and East Boston - that were similar in size and proximity to the city center (defined as City Hall) but that differed sharply in their connectivity to the city center in the study period. This contrast is helpful for testing the role of market access in determining firm size.

I assemble a novel data set to conduct the empirical analysis. I digitize a data set of the universe of the businesses from the top 25 $retail/wholesale$ services/products (accounting for 20% of all businesses) in the *Boston Directories*-a source that resembles today's business yellow pages-for each five-year period between 1885 and 1905. [3] The original function of these directories was to provide information about every citizen and business in the city.[4] The main pieces of information I obtain are the firms' addresses and names, from which I distinguish three types of firms: sole proprietorships, partnerships, and companies/corporations. I collect supplemental data from the R. G. Dun & Co's credit rating records in 1885 and 1899, which shows that the sole proprietorships were, in terms of estimated net worth,[5] the smallest type. The median net worth of the sole proprietorships was only one-tenth of the median net worth of the second smallest firm type. I then correspond the sole proprietorships to the "mom-and-pop stores," and use a sole proprietorship dummy as the main outcome of interest in the empirical analysis. In order to analyze the spatial patterns of firms and individuals in my study context, I also georeference 1,660 plot-level historical city maps to identify the geographic coordinates

[3] I focus on analyzing the responses of the firms in the retail/wholesale sector, because the transactions in this sector mostly involved the movement of people, to which the upgrade of the commuter rails was highly relevant. The retail/wholesale sector is also interesting in itself because it features a particularly high self-employment rate. In 1910 - the first year in which the census asks subjects their occupation types - the self-employment rate in the retail/wholesale sector was 0.32, while it was 0.17 in the aggregate non-agricultural sector.

[4] In the 1789 Boston Directory - the first issue, the cover shows that it contained "a list of the merchants, mechanics, traders, and others, of the town of Boston; in order to enable strangers to find the residence of any person."

[5] called "pecuniary strength" in the credit rating books.

of all the addresses on the 1885-1905 *Boston Directories*. These results allow me calculate the distance between each firm and the nearest streetcar route to measure the intensity of shocks associated with the streetcar electrification.

I use constant geographic areas - plots - as the units of the regression analysis. The outcome variable is defined as the share of establishments that were sole proprietorships in each plot. To identify the causal effect of an upgrade of transportation infrastructure on the presence of sole proprietorships, I apply a difference-in-differences strategy, comparing changes in outcomes in the plots directly connected to the streetcar rails (treatment) to changes in neighboring unconnected plots (control), controlling for 200 m × 200 m block by year fixed effects. The identification assumption is that by removing time and geographical variations at the block level, both the rail-connected plots and the unconnected plots would have changed similarly in the absence of the streetcar electrification.

My baseline results show a striking treatment effect of the electrification of the streetcar system. Before the upgrade of the streetcar rails in 1885, the rail-connected plots had, on average, a 4.7-percentage point smaller share of sole proprietorships than the neighboring unconnected plots. After the electrification in 1905, this gap widened to 10-percentage points. The effect is statistically significant, and is robust to different controls, distance thresholds for defining treated locations, and block sizes. The magnitude of the treatment effect is striking considering that in my regression sample, the rail-connected plots are, on average, within 25 m of the rails while the neighboring plots are, on average, between 25 and 100 m away from the rails.

I then examine the mechanisms of the treatment effect. Among three potential advantages of proximity to streetcar rails - access to inputs or supplies, access to workers, and access to consumers - I show evidence that access to consumers is the most plausible advantage. Thus, a critical condition to generate my effect sizes is that consumers must be highly sensitive to commuting costs, such that being located a short distance away from the rails makes a significant difference in market access for firms. I provide three pieces of evidence suggesting that this could be the case in my study context. First, the spatial distribution of employment gravitated highly toward streetcar rails: over half of the businesses were located on rail-connected streets. The density of employment declined sharply, with movement just one block away from the rail-connected streets. Second, the treatment effect is much larger among - and is in fact, mostly driven by - food grocery products, which feature higher commuting costs per dollar. Third, the street blocks contained highly diverse businesses. A typical 200 m × 200 m block covered about 30% of the top 16 most-common products, and this coverage rate was higher for food-related products. The data implies that a single small block can provide most products necessary for daily life.

Finally, I consider the alternative mechanisms that could also have explained the treatment effect. While I cannot rule out all of these alternative mechanisms, I show that none of the alternative mechanisms alone is able to account for all of the patterns observed in the data, leaving the mechanism in my model as the most plausible one.

This paper contributes to three bodies of literature. First, it adds to a greater understanding of the prevalence of micro and small enterprises in the process of economic development. Besides the market segmentation hypothesis, existing explanations to this phenomenon include more limited entrepreneurial talent or managerial capital (Lucas, 1978; Bloom et al, 2013), higher regulatory and institutional barriers (Lewis, 1954; Harris and Todaro, 1970; De Soto, 1989; Rauch, 1991; Levy, 2008), and more severe contracting problems for hiring outside managers (Akcigit, Alp, and Peters, 2016) in developing countries than in the developed. By looking at a case when there was a discrete change in transport costs and presumably not in the other factors, this paper identifies the important role played by transportation technology in determining firm size. The dense streetcar network and the high decay rate of treatment effect as one moves away from streetcar rails signify a very high degree of market segmentation in the context of historical Boston. This is consist with the fact that the vast majority of firms employed fewer than 10 workers, and lends strong support to the market segmentation hypothesis.

Second, this paper is related to the literature on market integration and economic growth. Studies have exploited large shocks to transport costs, typically in the form of large-scale investments in inter-city transport infrastructure, and found that the market integration process is associated with changes in relative demand for skilled workers across cities (Michaels, 2008), reduced regional price dispersion and improved welfare (Donaldson, 2012), increases in disparities in GDP growth between peripheral and metropolitan regions (Faber, 2014), and increases in land values in areas with improved market access (Donaldson and Hornbeck, 2016). In this paper, I show that market integration process can also directly affect the organizational form of the basic economic units - firms. This finding implies that a more comprehensive evaluation of the benefits of market integration needs to take account of increases in the productivity of firms.

Finally, this paper is related to research on the effects of urban rail infrastructure on the internal structure of cities. Studies in this literature have primarily focused on property values around transit stations (Bowes and Ihlanfeldt, 2001; Cervero and Duncan, 2002; Gibbons and Machin, 2005; Ryan, 2005; Billings, 2011; Hewitt and Hewitt, 2012; Ko and Cao, 2013). A sharp contrast between this paper and the existing literature is that I use highly detailed micro-geographic data to show how businesses and residents reacted to a quick upgrade of their urban rail infrastructure, and my context is a historical city. As it pertains to methodology, this paper contributes a new identification strategy that exploits the upgrade of existing transit routes to address the endogeneity associated with the non-randomness in new route placement. This strategy is distinct from three currently popular identification strategies in the literature, as reviewed in Redding and Turner (2014).

The remainder of this chapter is organized as follows. Section 1.3 provides the historical background on the electrification of the streetcar system in Boston. Section 1.4 presents a model that determines firm sizes in a non-tradable service sector in a city. Section 1.5 describes the data. I provide summary statistics in Section 1.6. Section 1.7 introduces my empirical strategy and presents the benchmark regression results. Section 1.8 examines the mechanism underlying the treatment effect. In Section 1.9, I consider alternative mechanisms that might explain the treatment effect. Lastly, in Section 1.10, I conclude.

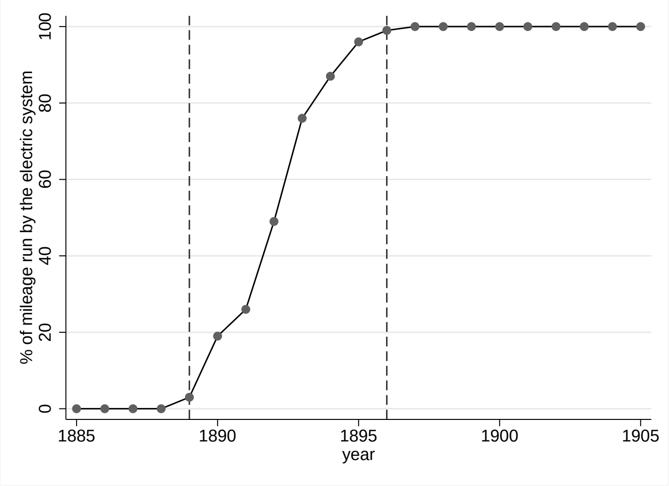
1.3 Historical Background

Up until the 1880s, most cities in the world relied on horsecar for intra-city transportation. There are four most commonly cited disadvantages for this mode of transit. First, horsecars were extremely slow. Even if rails were laid on streets to eliminate a great deal of friction, horsecars ran only at a speed of 4 to 5 *mph*, equivalent to the speed of a brisk walk. Second, the animals had to be fed and cared for, which involved considerable expense. Third, horsecars were very unreliable under bad weather conditions. Finally, disposing of the huge quantity of excrements that the horses deposited on the city's streets was a huge problem in sanitation. Some historians attribute a rise in the incidence of tuberculosis in nineteenth century American cities to residents breathing in the dried air-borne germs from animal excrements.

Because of these disadvantages of horsecars, in the late-nineteenth century, almost every major American city put efforts into modernizing intra-city transit systems. However, Boston was the winner of this race, in the sense that it was the first to build a large-scale city-wide electric streetcar system. There are two main driving factors to Boston's success. The first is the narrow, winding streets of Boston that discouraged the use of cable-cars. In the 1880s, the cable-car system had been set up in several US cities despite the high expense and complex maintenance and operation. However, the difficulty in implementation of cable-cars in Boston invigorated the development of a more efficient system, which culminated in the significant advancement of electric streetcars. The second driving force is the great entrepreneur and president of the West End Street Railway Company, Henry Whitney (Most, 2014). In 1888, after consolidating Boston's horse-drawn street railway companies under one company, Whitney was ready to modernize the horsecar system. His initial plan was also to install a cable system. However, in the same year, Whitney was invited by an engineer, Frank Sprague, to Richmond, Virginia to see a demonstration of an electric street rail. Whitney was very impressed and quickly abandoned the cable car idea. The West End Street Railway Company then pioneered in meeting the engineering challenge to design and construct safe, economically viable, and reliable electric power for Boston's rapid transit. The research and development progressed very rapidly, and attracted the attention of electrical engineers all across the country. As one major electric journal put it at the time:

> *The West End Street Railway company of Boston is making rapid progress in the equipment of its line with the Thomson-Houston system and work this winter. The permanent power plant will be a model of its kind, and when completed the largest and best equipped in the world. $\cdots$ before long the electric car will be a familiar sight in the heart of the city.*

The fast pace of this work can also be seen from the percent of mileage that was run as an electric system, presented in the *Annual Report of the West End Street Railway Company*. Figure 1.1 documents this statistic annually in my study period, from 1885 to 1905. Starting from an entirely horse-drawn system in 1888, the company completely electrified the system over the next eight years. Another indicator for this fast pace is the transportation horse population in Boston, which dropped from 7,684 in 1888 to 487 by 1897, as shown in Figure 1.2.

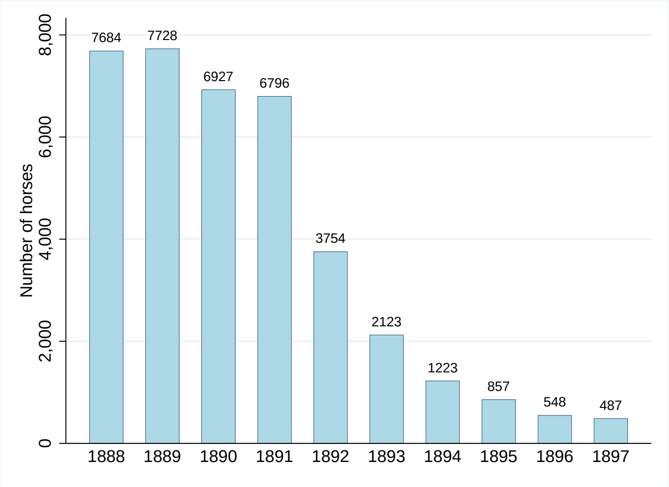


Source: The Annual Reports of the West End Street Railroad Company.

Figure 1.1: Pace of the Electrification

Compared to the horse-drawn system, the new electricity powered streetcar system had a number of advantages: First, as mentioned earlier, electric cars ran much faster - 8 to 10 miles per hour compared to 4 to 5 miles per hour for horse-powered vehicles.[6] Second, the electric system was much more reliable in bad weather. Third, the carrying capacity tripled compared to that of horsecars. Fourth, the city was able to avoid the pollution generated from animals, making the streets much cleaner than before. Fifth, the marginal costs of the services were lower, so that the company was able to offer lower fares to the public while simultaneously expanding services. Compared to the cable car system of other cities, the electric streetcar system was cheaper, more practical, and avoided the imperfections and dangers of cable haulage. By 1905, as a consequence of Boston's success, most cable car systems in other US cities were converted to electric traction or abandoned altogether (Vuchic, 2007).

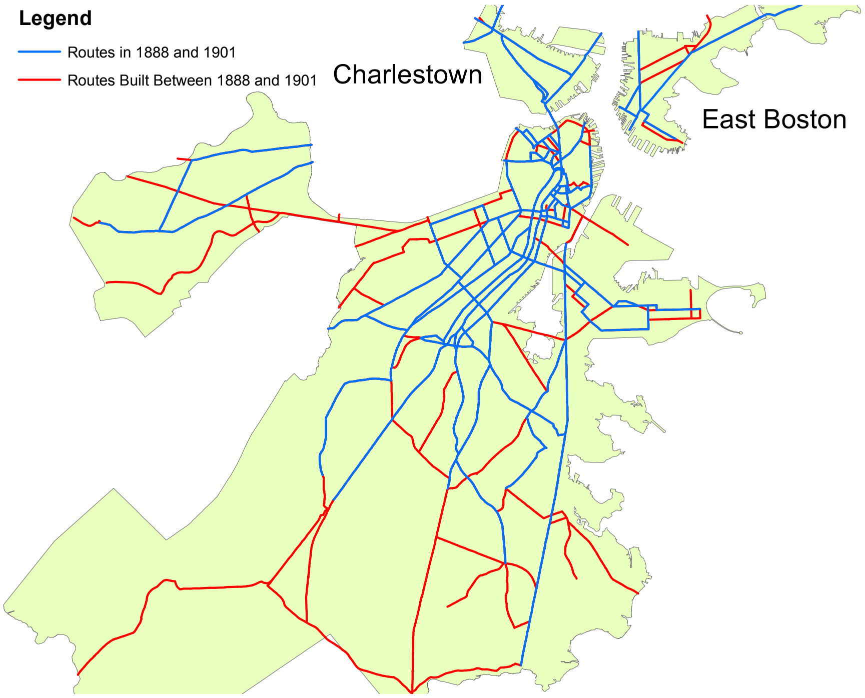
[6] The speed takes into account average traffic conditions.



Source: The Annual Reports of the West End Street Railroad Company.

Figure 1.2: Horse Population in Boston

Figure 1.3 shows the streetcar routes in Boston at two points of time: one in 1888 in blue, which is one year before the electrification, and hence, using an entirely horse powered system; and the other in 1901 in blue and red, which is four years after complete electrification. Not surprisingly, the technology upgrade was associated with the substantial expansions of previous streetcar lines. For identification, I use the preexisting routes before the electrification as the basis for calculating proximity to the rails, as the placement of the new lines may have been non-random. Figure 1.3 demonstrates that the 1888 routes were already extensive and that they covered the core areas of Boston. The new lines put in between 1888 and 1901 were primarily placed in suburban residential areas. I lose only 0.4% of business establishment observations by excluding those near the newly expanded lines .



Source: Digitized Boston city maps.

Figure 1.3: The Streetcar Routes in 1888 and 1901

1.4 A Model of Firm Size in the Urban Non-Tradable Service Sector

I formalize the market access hypothesis in the context of a non-tradable service sector in a city. My model is a modified version of Ziv's (2016) model, and can be thought of as a Melitz (2003) model with spatial interactions between agents in a city. In my model, the services are differentiated, and each firm provides one variety of service. Consumers have a love of variety. Firms and workers/consumers make endogenous location choices across city neighborhoods and have an outside option. Workers are homogeneous. They supply labor in the local neighborhood, so they do not commute between their residences and workplaces. However, as consumers, they can incur commuting costs in order to obtain goods or services provided in nonlocal neighborhoods, where the commuting costs depend on the network of rails and the speed of the streetcars. Firms are heterogeneous in terms of productivity. In choosing across locations, firms exchange market access for local labor costs and land rents.

A key element of the model is the bilateral commuting costs for shopping across locations. I assume that consumers from nonlocal neighborhoods take streetcars to shop. To find a store in an off-rail location, they need to incur additional commuting costs. The event of streetcar electrification is modeled as a reduction in commuting costs along the streetcar rails while keeping all other parameters-including the commuting costs to reach off-rail locations-constant. The new equilibrium features a redistribution of firms and workers as well as heterogeneous changes in firm size across space. The mathematical characterization of this model is found in Appendix 1.

While there is no closed-form solution to equilibrium, this model characterizes the conditions under which we observe a treatment effect, defined as the difference in firm size changes between rail-connected and neighboring unconnected locations. Assume that rail-connected locations and neighboring unconnected locations are sufficiently small, which is consistent with the empirical setting.[7] The model predicts that we will observe a relative increase in firm size in rail-connected locations when the following two conditions are met: first, that there is an increase in the nonlocal market size; and, second, that the commuting costs to reach off-rail locations is substantial. The intuition is that if the upgraded streetcar system brings in more nonlocal consumers, and that it is very costly for nonlocal consumers to reach off-rail locations, then the increase in nonlocal market size will be discounted for the firms in off-rail locations.

The model has additional predictions. In the model, land is a fixed input in production. Production features economies of scale. Improved market access will be capitalized into land prices. Thus, we should observe a relative increase in land prices

[7] On average, each "location" is half of a 200m×200m block. The entire Boston area in any year during the study period consists of 1,540 such locations.

in rail connected locations. Moreover, more productive firms can better take advantage of market access. They will be able to outbid less productive firms in land rents in rail-connected locations, and thus, spatial sorting will occur. This sorting mechanism magnifies the treatment effect.

Section 1.6 and 1.7 provide evidence for the basic prediction of the model. Section 1.8 examines the mechanism in the model more thoroughly.

1.5 Data Construction

In this section, I provide details on the data used, describe its sources, and assess its validity for my study purpose.

1.5.1 Streetcar Routes

I obtained digital city maps of Boston in 1888 and 1901 from the online David Rumsey Historical Map Collection, which contained streetcar routes and legible street names. I then georeferenced the two maps such that the points of each of the two city maps were geographically aligned with a common 1930 street centerline shapefile, which I retrieved from the Historical Urban Ecological data set, created by the Center for Population Economics. By overlaying the street centerline shapefile with the georeferenced city maps, I extracted the portions of the streets that coincided with the streetcar routes on the city maps and digitized them into new shapefiles. The routes of the two years are shown in Figure 1.3. I then used the streetcar route shapefiles to calculate the distance between each business establishment and the nearest streetcar line.

1.5.2 The Boston Directory

I digitized the primary data source for the firms from the *Boston Directories* published by the Sampson, Murdock, & Company, and printed annually. Each of these volumes consists of two main sections. The first section lists the names of the inhabitants and firms, their occupations/products, and the places of the business and dwelling houses. Generally, the inhabitants in the directories were in the labor force. For firms, the names of partners are typically listed. The second main section of each volume is the business directory. It uses only the firms from the first section, categorizing them according to product/occupation (e.g., lawyers, grocers, bakers, etc.), and then providing street addresses for each. A small portion of the firms (4%) have multiple addresses/establishments. In my empirical analysis, I treated establishments as the basic units of the analysis, thus I use the term "establishment" henceforth. I obtained scanned images of the full directories for the years 1885, 1890, 1895, 1900, and 1905 from the genealogy Web site Ancestry.com. I digitized a 1% random sample of the individuals from the first main section,[8] and all of the establishments for the 25 most frequent retail/wholesale products from the business directory section. These 25 products are listed in Table 1.11 in Appendix 2.

1.5.3 Historical Credit Ratings of Businesses

The third data source I used was the *Mercantile Agency Reference Book*, published by the R.G. Dun & Co.[9] The need for credit ratings stems from the first half of the nineteenth century, when commission merchants based in large urban cities were increasingly providing goods and supplies to rural merchants, jobbers, and general stores,

[8] I first randomly selected a 5% sample of the scanned images, digitized the full text in these images, and then randomly selected 20% entries (individuals or firms) from the digitized text.

[9] The title of this book has changed from time to time. After 1925, the title read *R.G. Dun & Co. Reference Book*, but in 1960, it changed to *Reference Book of Dun & Bradstreet*, and finally, in 1991, to *The Dun & Bradstreet Reference Book of American Business*.

but were unable to discriminate their credit-worthiness. Credit rating agencies established a network of local correspondents, who gathered business information on merchants and jobbers in their areas and reported it to the rating agency's headquarters.[10] The agency then sold this credit information to subscribers for a fee. R.G. Dun & Co was one of the most successful credit rating agencies of the era, and it merged with the company J.M. Bradstreet in 1933 to form the Dun & Bradstreet Corporation.

R.G. Dun & Co's reference books cover a wide range of businesses in the United States and Canada, containing their names, main product lines, pecuniary strengths (i.e., estimated net worth, grouped into 17 size categories), and credit ratings (8 classes). [11] These books were published bimonthly, and most of the issues are found in the Library of Congress, with the exceptions of those published between 1889 and 1898. I digitized the Boston sections of these books for September 1885 and July 1899.

1.5.4 Boston Property Tax Ledgers

Historically, the City of Boston sent tax assessors to each building to collect information for annual real estate and personal property taxes. The critical information for the current study is the assessed value of the building, the assessed value of the land, plot size, and street name and number for each building unit. [12] The genealogy Web site *FamilySearch.org* contains scanned images of the handwritten ledgers from 1822 to 1918, and these are publicly available. I digitized a 10% random sample of the building units' data for 1885 and 1898. I chose 1885 as the initial year to match the first year of the business data. I chose 1898 as the final year because it is immediately after the completion the electrification of the streetcar system while still before the announcement

[10] Initially, there were no direct employees of these firms, but instead, the firms often used lawyers or postmasters who lived in the particular area. Later, the system relied on paid reporters, who worked exclusively for a particular agency.

[11] For a more detailed discussion of this data source, see Sarada and Ziebarth (2015).

[12] A more detailed description about this data source can be found in Hornbeck and Keniston (2016).

of subsequent major transportation infrastructure projects, including the construction of subways and elevated railways, which could have affected the real estate prices.

1.5.5 Comprehensiveness of the Business Directory and Credit Rating Data

I performed an assessment of the comprehensiveness of the business directory and the credit rating data by matching the two data sources in two directions. First, I randomly selected an 8% sample of firms from the credit rating reference books in 1885 and 1899, totaling 1,935 firms. I then manually matched these firms to those in the first main section of the *Boston Directories* in the corresponding years by both their names and products (occupations). I was able to match 1,736 of the total 1,935 firms, yielding a matching rate of 89.7%. Next, I randomly selected 826 firms from the business section of the *Boston Directories* in 1885 and 1899, and then matched them to the credit rating reference books, also by their names and products (occupations). 287 of these 826 firms could be matched, yielding a matching rate of 34.8%. [13] The much higher matching rate in the first direction suggests that the *Boston Directories* contain a more comprehensive list of firms, while the credit rating books probably selected businesses that catered to the needs of their subscribers. For this reason, as well as for the fact that the credit rating reference books are missing for the years 1889-1898, I used the *Boston Directories* as the main data source and drew on the credit rating records as a supplemental data source.

[13] The matching rates in both directions are positively correlated with firm size. In the first direction, 100% of the firms with the highest net worth class in the credit rating reference books were matched to the *Boston Directories*, and this matching rate fell down to 76% for the lowest net worth class. In the second direction, 27.7% of the sole proprietorships in the *Boston Directories* were matched to the credit rating reference books, while this matching rate was 46.6% for the other legal forms.

1.5.6 Measurement of Firm Size

From the names of the establishments in the *Boston Directories*, we can distinguish three legal forms of establishments: (1) sole proprietorships, identified as those listings showing names of individuals rather than business names; (2) partnerships, defined as the names in the format of *A* & *B* (e.g., *Whitcher* & *Emery*), *A Bros* (e.g., *Abbott Bros*), or *A* & *Sons* (e.g., *Reynolds S. H.* & *Sons*); and (3) companies (corporations), identified as those with the word "Company (Corporation)" in their names (e.g., *Gilchrist Co*). I used a sole proprietorship dummy to proxy for establishment size, and corresponded these establishments to "mom-and-pop" stores.

To verify that the legal forms are informative about establishment size and productivity, I used two sources of data to compare establishment size and productivity via the different legal forms. In the first verification, I used the first direction-matched data (from the credit rating to the *Boston Directories*) mentioned above, to document the estimated net worth for each legal form of establishment. Using the rating key (shown in the Appendix Figure 1.20), I converted the letter ratings into numeric values. [14] The mean, 25th-, 50th-, and 75th percentiles of the estimated net worth are shown in Table 1.1. We can see that there was a very sharp contrast in the estimated net worth between the sole proprietorships and the other two legal forms: the median net worth of the sole proprietorships was only one tenth of the median net worth of the second smallest legal form, partnerships, but there was no significant difference between the companies (corporations) and the partnerships. The net worth of a median sole proprietorship was $1,500$$, or 5.5 times of the gross domestic product (GDP) per capita in 1900.

In the second verification, I used the *1954 Census of Retail Trade* national summary statistics to compare establishment employment size and productivity (sales per

[14] I assigned the mean of the value range for each letter rating of pecuniary strength. For example, the letter *K* stands for estimated pecuniary strength of 1000$ – 2000$. I assigned 1500$ to every *K* rating.

Table 1.1: Estimated Net Worth by Type

Type	mean	p25	p50	p75
Companies/Corporations	82,401	7,000	27,000	100,000
Partnerships	78,031	4,000	15,000	60,000
Sole Proprietorships	11,600	300	1,500	7,000

Notes: The estimated net worth refers to the pecuniary strength in the credit rating reference books. The value is measured by current-price USD. The statistics are calculated by pooling 1885 and 1899 data.

worker) by legal form. 1954 is the earliest year for which we have data on both measures. As shown in Table 1.2, sole proprietorships were the least productive and employed the fewest workers compared to the other two legal forms. Typically, a sole proprietorship retail establishment in 1954 had only one active proprietor and employed less than two paid workers. By contrast, a typical company/corporation was 28% more productive and had six times as many workers (active proprietors plus paid employees). These sharp contrasts lend creditability for treating sole proprietorships as a qualitatively different business form.

Table 1.2: Productivity and Establishment Size by Legal Form

Legal Form	$\frac{Sales(\$)}{workers}$	$\frac{Workers}{est.}$	$\frac{Employees}{est.}$	$\frac{Proprietors}{est.}$
Companies/Corporations	21.6	16.7	16.7	
Partnerships	17.5	5.8	3.8	2.0
Sole Proprietorships	16.9	2.8	1.8	1.0

Notes: The statistics represent national average. The value is measured by current-price USD.
Data Source: *1954 Census of Retail Trade*.

Finally, sole proprietorship status is not only informative for establishment size and productivity but also interesting in itself. The literature on tax recognizes sole proprietorships as a legal form that is particularly prone to under-report taxable income, and thus, evade taxes.[15] Understanding the causes of the changes in the share of sole

[15] See Slemrod (2007) for a review.

proprietorships is closely linked to understanding the income tax capacity of a state (Jensen, 2016).

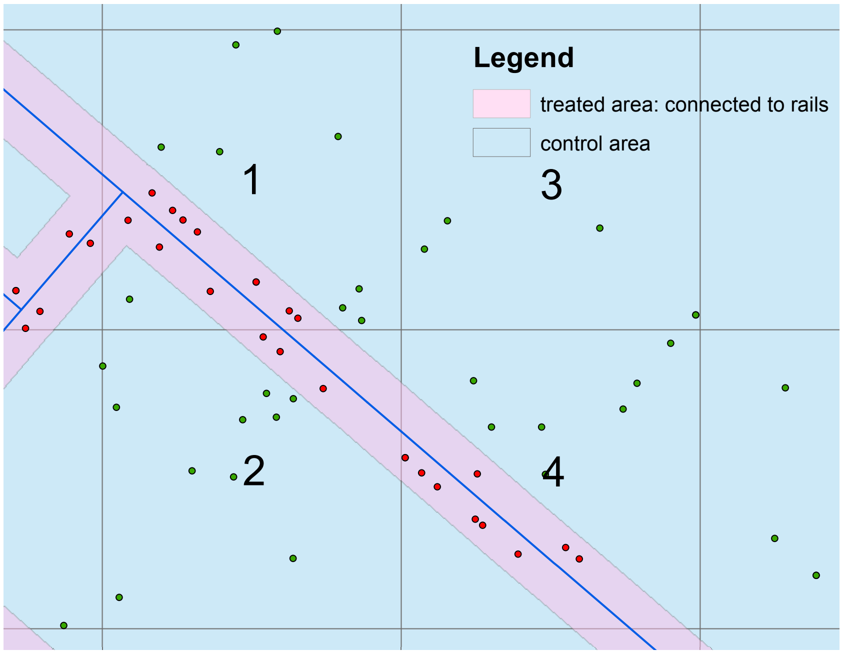
1.5.7 Plot-Level City Maps

The key to combining the digitized streetcar routes data and the establishment-level data is to geocode the addresses of the establishments using contemporaneous city maps. I georeferenced 1,660 plot-level *Sanborn Fire Insurance Maps* of Boston published during the period 1895-1900, which, altogether, covered the entire Boston area. I then manually extracted the street name and number of every building on the maps to a GIS shapefile, generating a point shapefile of 100,743 buildings (Figure 1.22 in Appendix 4 shows a sample map). The geographic coordinates of each building were calculated in ArcGIS and then matched to the addresses in the *Boston Directories* by street name and number. [16] For all of the establishments in this study, 95% of them could be geocoded.

The empirical analysis benefited from geocoding the addresses in the *Boston Directories* in four specific ways. First, I could calculate the distance of each establishment to the nearest streetcar rails. This distance allowed me to define whether an establishment was treated or not. Second, I could create a panel data set of fixed geographic locations over time. These fixed geographic locations served as the units of the regressions. Third, the exact location information allowed me to control for the fixed effects of a larger geographic area and adjust for spatial correlations in the error term. Finally, the geocoded residential and commercial addresses for the 1% random sample of individuals allowed me to recover commuting patterns and the distribution of residents and employment in the study period.

[16] For special addresses, such as “Street A corner Street B,” I manually located them on the georeferenced maps.

In the main regressions, I used fixed geographic locations as the units of the regressions, which I call *plots* hereafter. The outcome variable was the share of establishments (restricted to the 25 wholesale/retail products) that were the sole proprietorships in each plot. The main empirical analysis compared the changes in plots with direct rail connections (treatment) to the changes in neighboring unconnected plots (control).



Notes: The above figure illustrates the definition of the treatment plots and control plots, as well as the blocks. The grids in this figure are $200m \times 200m$, called blocks. If a block is passed through by a streetcar rail, such as blocks 1, 2, and 4, it is then separated into two plots: the rail-connected plot, indicated by the purple areas, and the unconnected plot, indicated by the light blue areas. I define the rail-connected plots as the treatment locations, and the unconnected plots as the control locations. Block 3 is dropped from the regressions.

Figure 1.4: Illustration of Treatment and Control "Plots", and "Blocks"

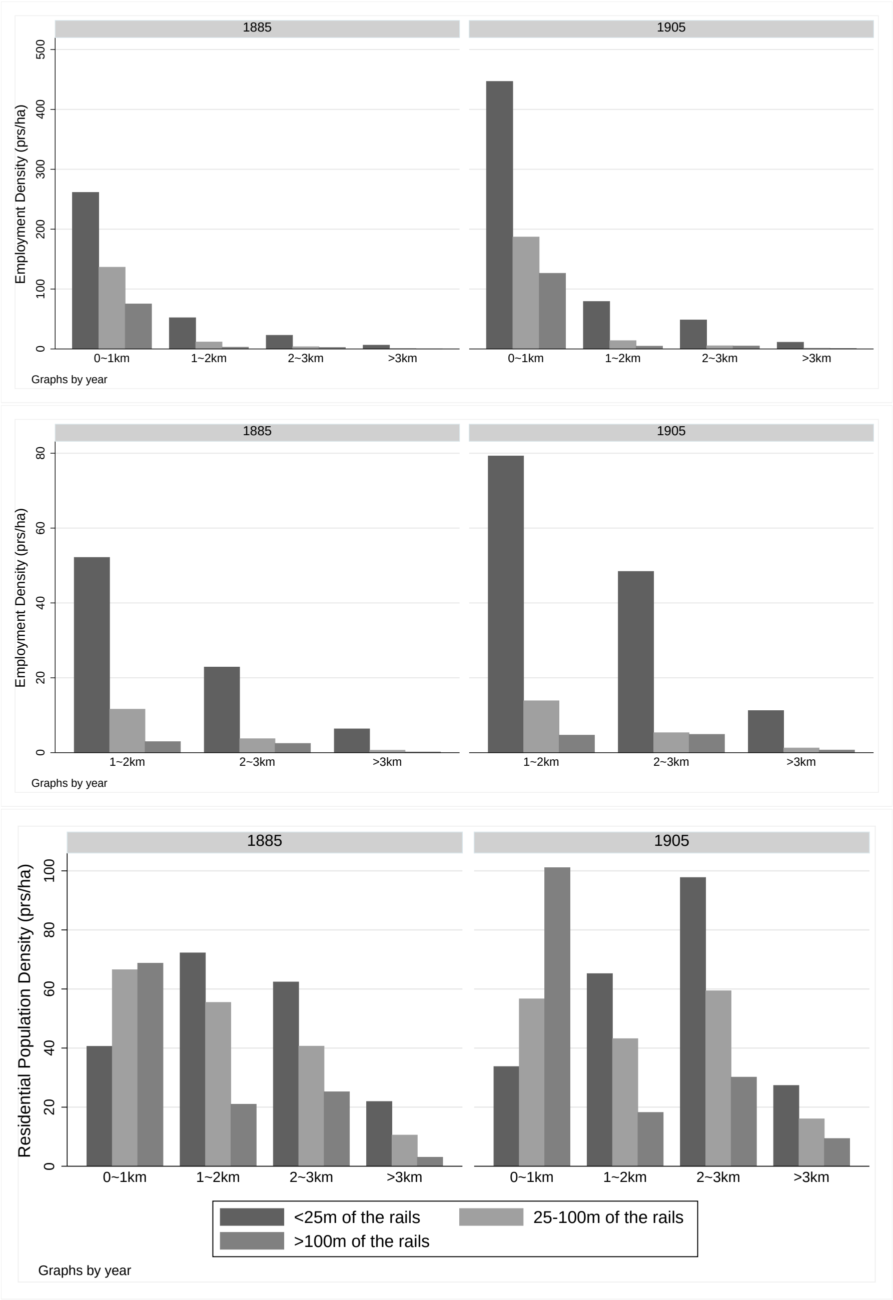
I use Figure 1.4 to illustrate my construction of a *plot*. Here, I first divided the entire Boston area into 770 blocks of size $200m \times 200m$. A small block size increases the number of observations and the significance of the statistical inferences, but the number of establishments in some blocks could fall to zero, which would result in an unbalanced panel data set. I chose 200 m as the block size to balance the trade-off. In robustness checks, I also tried block sizes of 300 m and 400 m. I then dropped the blocks that did not intersect with any portion of the 1888 streetcar rails. For example, I dropped Block 3 in Figure 1.4. The rest of the blocks were then separated by the rails into two plots: the first plot was a bin enclosing all of the establishments on the rail-connected streets, indicated by the purple areas. The second plot was the remainder of the areas within the block, indicated by the light blue areas. I defined the first type of plots as the treatment locations, and the second type of plots as the control locations. I chose the narrowest possible bin to define the treatment locations because 50.8% of the establishments (across all years) were located exactly on the 1888 streetcar routes. Finally, to keep a quasi-balanced panel of observations, I dropped the blocks within which there were no establishments in either the pre-electrification or post-electrification periods. This left me 1,632 observations/plots across all the years for my regressions. On average, each plot contained 19 establishments. The geographic areas in the regressions covered 28,209 establishments out of a total of 41,174 establishments.

1.6 Descriptive Statistics

1.6.1 Distribution of Residential Population and Employment

How did the distribution of the residential population and employment change in the study period? Using the coordinates of both the residential places and the commercial places for the 1% representative sample of the individuals who commuted in the *Boston Directories*,[17] which covered all occupations and industries, I plot the distributions of both residential population and employment by distance to City Hall and their distance to

[17] One half of the individuals in the *Boston Directories* had a residential address but no commercial address, and thus, not a commuter. These people probably either worked from home (e.g. as a grocer) or did not have a fixed workplace (e.g., day laborers, peddlers). The fraction of commuters is stable over the study period.



Notes: The horizontal axis represents the distance from the city center. The bars in different colors indicate different distances from the streetcar rails. The 1885 distances from the rails were calculated using the 1888 rails. The 1905 distances from the rails were calculated using the 1901 rails. The spatial data of population came from the geocoded 1% random sample of the inhabitants in the 1885 and 1905 *Boston Directories*.

Figure 1.5: The Distribution of Population in 1885 and 1905

the streetcar rails in 1885 and 1905, seen in Figure 1.5. We find that the spatial patterns of employment growth and residential population growth are quite different. From the first row of Figure 1.5, the majority of employment growth took place in the city center, most often within 1 km of City Hall. In contrast, we see from the last row of Figure 1.5 that the residential population density increased primarily in the areas 2-3 km and >3 km away from the city center. It is worth mentioning that despite the fact that the absolute increase in residential *density* is not substantial in the area >3 km away from the city center, this area is so vast that the majority of the increase in residential *population* took place there. Table 1.3 reports the evolution of the commuting distances of these individuals; the commuting distances are calculated as the distance between their residential addresses and their commercial addresses. The median commuting distance increased from 2.2 km in 1885 to 4 km in 1905, indicating that after the electrification of the streetcar, the population was more mobile. These facts provide quantitative evidence for Warner's (1962) observation that Boston saw the emergence of "streetcar suburbs" in this period, in the sense that more and more people began to live in the suburbs and commute to their workplace in the Central Business District (CBD, defined as the areas within 1 km of City Hall in Boston).

Table 1.3: The Centiles of The Commuting Distances (km)

year	p25	p50	p75
1885	0.50	2.19	4.74
1890	0.83	2.90	5.27
1895	0.75	3.03	5.83
1900	1.12	3.95	6.44
1905	1.07	3.97	7.09

Notes: Only one half of the people in the Boston Directories were commuters, and this ratio is stable over time. Commuting distance is defined as the distance between the residence and the workplaces of the worker's main occupation. Percentiles were calculated only for the commuters.
Source: The geocoded 1% random sample of the inhabitants in the *Boston Directories* between 1885 and 1905.

The facts documented in this section serve as a verification of the first condition for generating the treatment effect: it shows that population density increased almost everywhere in Boston in my study period, whether measured by employment density or residential density. Moreover, this population was more mobile than it was before. Thus, nonlocal market size probably increased for each location.

1.6.2 Pre-Electrification Plot Characteristics

Table 1.4 documents the summary statistics of the outcome variable between rail-connected plots and unconnected plots prior to the electrification, i.e., in year 1885 and 1890. By construction, the number of plots in these two groups is identical. From columns (1) and (2), the (unweighted) average shares of sole proprietorships are 77.6% and 83.6% in the connected plots and unconnected plots, respectively. From column (3), we can see that the average difference (weighted by the average number of establishments from 1885 to 1905) in the levels of the share of sole proprietorships is 3.9%. The sign of the coefficient suggests that there is a positive relationship between establishment size and rail-connection before the electrification. This fact is consistent with the model, which predicts that in a static equilibrium, the more productive firms sort into locations with better market access, such as transit hubs. Importantly, from column (4), we find that there are no significantly different trends between the two groups of plots prior to the electrification.

1.6.3 Evolution of Outcomes

I illustrate my main findings in Figure 1.6. This figure depicts the time trends in the weighted average of the share of sole proprietorships across rail-connected plots and unconnected plots in dashed lines and solid lines, respectively. The pre-trends had

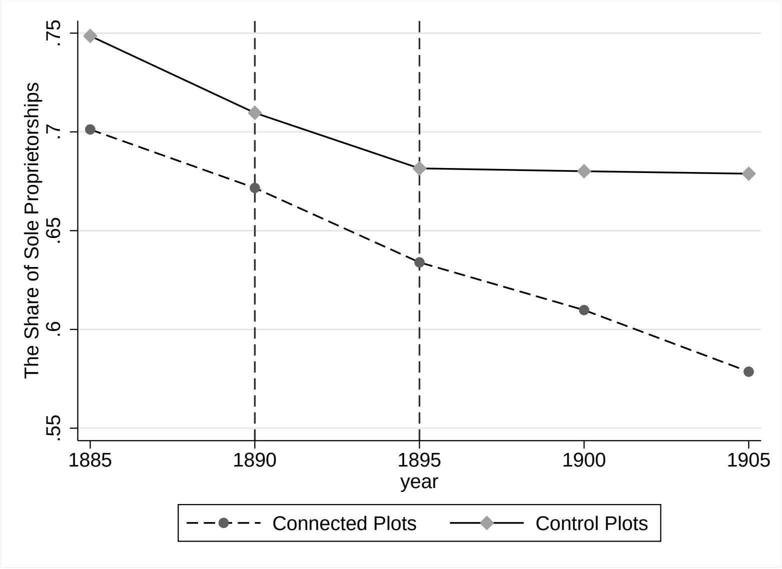
Table 1.4: Plot Characteristics prior to the Electrification

	Unweighted Average		Weighted Difference	Weighted Difference
	Connected Plots	Control Plots	in Levels: (1) - (2)	in Trends 1885-1890
	(1)	(2)	(3)	(4)
Share of Soles	0.776	0.836	-0.039**	0.011
	(0.231)	(0.253)	(0.015)	(0.015)
Number of Plots	370	370	370	332

Notes: Summary statistics of the outcome variable are reported for the treatment and the control plots prior to the electrification. Columns (1) (2) use unweighted average, while columns (3) (4) use average weighted by the average number of establishments from 1885 to 1905 in the plot.

been declining prior to the electrification. By visual inspection, we see that the pre-trends were parallel between the two groups of plots, consistent with the statistical results in column (4) in Table 1.4. Thus, the electrification of the streetcar system can be thought of as an event that disrupted an ongoing process of increasing establishment size.

The model predicts that average establishment size will diverge between rail-connected plots and unconnected plots after the electrification as long as the commuting costs for consumers to reach off-rail locations are positive. The trends in Figure 1.6 confirm this to be the case. Prior to the electrification, the gap in the outcome between the two groups of locations was 4.7-percentage points. At the end of the study period, this gap had widened to 10-percentage points, or a 5.3-percentage point relative drop in the rail-connected plots. Considering that the rail-connected plots were, on average, within 25 m of the rails, while the unconnected plots were, on average, between 25 m and 100 m away from the rails, this magnitude of relative drop implies that the commuting costs for consumers to reach off-rail locations must have been very large. I use regressions in Section 1.7 to confirm that the economically large relative drop in the rail-connected plots is statistically significant, and that it is robust to different econometric specifications and different definitions of the treatment and the control groups. In Section 1.8, I provide evidence suggesting that the consumers could be highly sensitive to commuting costs in my study context.

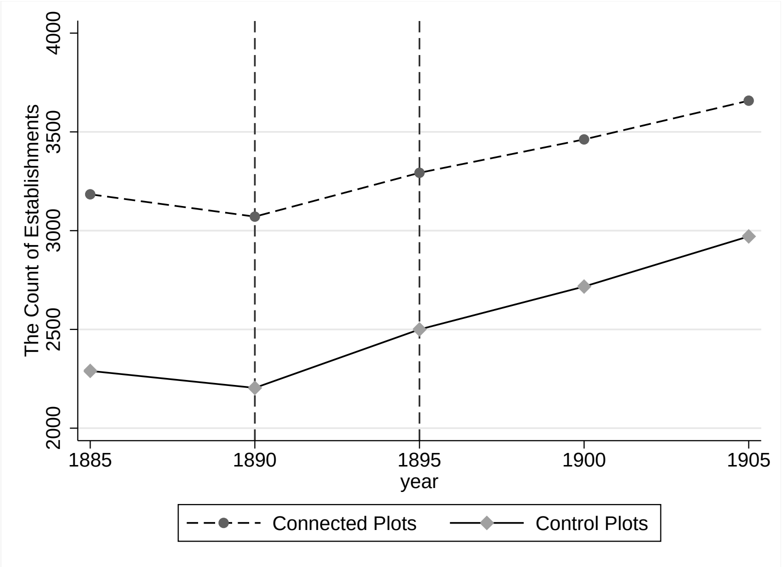


Notes: The time trends in the weighted average of the share of sole proprietorships across the rail-connected plots and the unconnected plots are plotted here in dashed lines and solid lines, respectively. The average for a group of plots is weighted by the average number of establishments from 1885 to 1905 in each plot.

Figure 1.6: Trends in Outcome between the Treatment and the Control Plots

The trajectories of the outcome are different from standard DID results. The trajectories of standard difference-in-difference results would have been that the outcome in the control group fell as fast as before, while the outcome in the treatment group fell at a faster rate. Conversely, here, we observe that the outcome in the control group stopped falling after the treatment, but that the outcome in the treatment group continued to fall as fast as before.

To shed more light on these non-standard trajectories, I document the time trends in the count of all establishments (including sole proprietorships and the other legal forms) in Figure 1.7. In both location groups, we observe that right after the beginning of the streetcar electrification process, the count of business establishments reversed the previous declining trend, and started to increase until the end of the study period. The trends between the two groups were almost parallel.

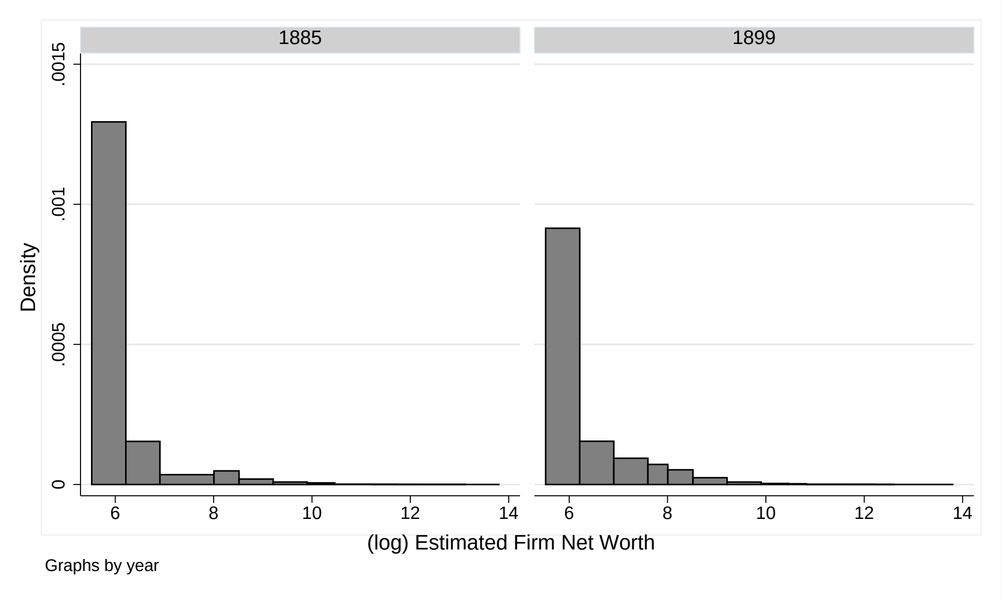


Notes: The time trends in the count of establishments located in the rail-connected and the unconnected areas are plotted here in dashed lines and solid lines, respectively. The count includes both sole proprietorship establishments and the other establishments.

Figure 1.7: Trends in the Count of Establishments Located in the Connected and Unconnected Areas

The facts documented in Figures 1.6 and 1.7 can be understood using my model. Due to the endogenous entry and exit of establishments, the mass of establishments will change in response to transport cost shocks. A reduction in transport costs induced business net entry. The bounce-back in the share of sole proprietorships in the control plots in Figure 1.6 could reflect that these locations became more favorable for the operation of small firms. In Section 1.8, I provide more details on the business dynamics in this period.

Finally, using the random sample of firms (restricted to the 25 retail/wholesale products) from the credit rating reference books in 1885 and 1899 (without matching them to the *Boston Directories*), I plot the firm size distributions in these two years in Figure 1.8. The assumption on the data is that the sample selection rules used by R.G. Dun & Co were similar in these two years, so that the firm size distributions are comparable across these two years. Under this assumption, we can see that during this



Notes: The estimated firm net worth refers to the pecuniary strength in the credit rating reference books. The firms are NOT matched to the *Boston Directories*. I restricted to the 25 retail/wholesale products in order to be consistent with the rest of my empirical analysis. The sample sizes are 442 firms and 363 firms in 1885 and 1899, respectively.

Figure 1.8: Firm Size Distributions Before and After the Electrification

period, there was a substantial shift in the firm size distribution to the right. The left tail became much thinner in 1899 than in 1885. This contrast corroborates the overall declining trend in the share of sole proprietorships in Figure 1.6.

1.7 Empirical Methodology and Main Results

1.7.1 Econometric Specifications

To estimate the causal effects of the upgrade of the transportation infrastructure on the share of sole proprietorships, I first estimate the following econometric specification

$$\text{Sole}_{ijt} = \beta_0 \text{Post}_t + \beta_1 T_i + \beta_2 \text{Post}_t \times T_i + \beta_3 t + \text{Controls}_j + \varepsilon_{ijt} \qquad (1.1)$$

where i denotes the plots, j denotes the blocks, t denotes the years, and $Sole$ is the share of establishments that were sole proprietorships. T_i is the treatment dummy: $T_i = 1$ indicates that i is a connected plot, and $T_i = 0$ indicates that i is a neighboring unconnected plot. $Post_t = 1$ indicates a post-electrification period, i.e. 1895, 1900, and 1905, and $Post_t = 0$ otherwise. t is a linear time trend. $Controls_j$ include the distance of block j to City Hall, or time-invariant block fixed effects, which capture block characteristics that have a constant impact on establishment size. β_1 measures the average difference in outcome between the connected plots and unconnected plots before the treatment. β_2 is the coefficient of interest, which measures how much more the outcome changed between the connected plots and unconnected plots after the treatment.

Alternatively, I estimate a more restrictive specification

$$\text{Sole}_{ijt} = \beta_1 T_i + \beta_2 \text{Post}_t \times T_i + \gamma_j \times \theta_t + \varepsilon_{ijt} \tag{1.2}$$

Compared to equation (1.1), this specification controls block-by-year fixed effects, which capture differential time trends in each block. Since each block contains exactly one pair of plots, the identification of β_2 comes only from the differential time trends between the plots within the same block.

The identification assumption of equation (1.2) is that, the pair of plots within each block would have undergone similar time trends in the absence of the electrification of the streetcar system. In the most restrictive specification, I also control for pre-electrification plot industrial composition, which may be associated with plot-level differential changes after the electrification

$$\text{Sole}_{ijt} = \beta_1 T_i + \beta_2 \text{Post}_t \times T_i + \gamma_j \times \theta_t + \beta_3 Food_{ij,t0} + \beta_4 Clo_{ij,t0} + \varepsilon_{ijt} \tag{1.3}$$

Here $Food_{ij,t0}$ is plot i's share of food-related establishments in 1885, and $Clo_{ij,t0}$ is plot i's share of clothing-related establishments in 1885. Therefore, the coefficient of interest β_2 in equation (1.3) reflects the difference in changes for connected plots and unconnected plots within the same block and with similar industrial composition.

In all specifications, the standard errors are clustered by block to adjust for serial correlation and within-block spatial correlation. The regressions weight each block by the average number of establishments from 1885 to 1905.

1.7.2 Benchmark Regressions

Table 1.5 reports estimated impacts on the plot-level share of sole proprietorships in the rail-connected areas relative to the plots in the unconnected areas. Columns (1) and (2) report estimates from the initial specification (equation (1.1)), using every five years between 1885 and 1905 and only 1885 and 1905, respectively. In column (1), the coefficient before *Treatment* is -0.039, suggesting that the average share of sole proprietorships in 1885 and 1890 is 3.9% lower in the rail-connected plots than in the unconnected plots. The key coefficient of interest, $Treatment * Post1895$, is -0.032, which is statistically significant. The magnitude implies that the average share of sole proprietorships across 1895, 1900, and 1905 is $3.2\% + 3.9\%$ lower in the rail-connected plots than in the unconnected plots. In other words, there is a break in the time trend of the outcome difference between the treatment and the control in 1895, with a magnitude of 3.2%. The coefficient before *Trend* is -0.02, suggesting that the share of sole proprietorships declined by 2% every 5 years, which occurred both in the connected and unconnected plots. The close to zero and insignificant coefficient before *Post*1895 indicates that there is no break in this overall time trend. The coefficient before *Distance to CBD* is 0.097, meaning that in an average year, the share of sole proprietorships increased by 9.7% as we move 1 km further away from the city center.

Column (2) uses the same specification as column (1), but uses only year 1885 and 1905 observations. The estimated coefficients now capture cumulative changes over 20 years. The coefficient before $Treatment$ shows that the gap in outcome between treatment and control is 4.7% in 1885. This number corresponds to the height gap between the solid and dashed lines in 1885 in Figure 1.6. The coefficient before $Post1895$ reflects that over the 20 years, there was a 7.7% overall drop in outcome in both the treatment and the control, which corresponds to the decline between 1885 and 1905 along the solid lines in Figure 1.6. The coefficient before $Treatment * Post1895$ is -0.053, which shows that the gap in the share of sole proprietorships between the treatment and the control widened to $4.7\% + 5.3\%$ in 1905, which corresponds to the height gap between the solid and dashed lines in 1905 in Figure 1.6.

From columns (3)-(5), I use increasingly restrictive specifications, and use only year 1885 and 1905 observations to capture the cumulative effects over an extended time period. In column (3), I control for the block fixed effects to capture block characteristics that have a constant impact on the outcome. Column (4) controls for the block by year fixed effects to capture differential time trends in each block, which corresponds to equation (1.2). Column (5) controls for the block by year fixed effects as well as the pre-electrification, plot-level industrial composition to take into account the possibility that there could be plot-level differential time trends associated with the initial industrial composition, which corresponds to equation (1.3). The key coefficients are statistically significant at least at a 10% level across all these specifications. The coefficient before $Treatment * Post1895$ is constant at -0.053 across columns (2) to (5), implying that the relative drop in outcome in the rail-connected locations is robust to various controls.

Table 1.5: Benchmark: All the 25 Retail/Wholesale Products

Dep. Var.: Sole	(1)	(2)	(3)	(4)	(5)
Treatment	-0.039**	-0.047***	-0.047**	-0.047**	-0.040*
	(0.016)	(0.016)	(0.018)	(0.023)	(0.024)
Post1895	-0.003	-0.077***	-0.077***		
	(0.016)	(0.017)	(0.020)		
Treatment*Post1895	-0.032**	-0.053***	-0.053**	-0.053*	-0.053*
	(0.015)	(0.020)	(0.023)	(0.028)	(0.028)
Distance To CBD	0.097***	0.093***			
	(0.014)	(0.014)			
Trend	-0.020***				
	(0.005)				
200m-Block FE			YES		
200m-Block*Year FE				YES	YES
Init. Industrial Comp.					YES
Year	1885-1905	1885,1905	1885,1905	1885,1905	1885,1905
Observations	1,632	680	680	680	680
R-squared	0.221	0.240	0.754	0.857	0.859

Notes: For all specifications, the outcome variable is the share of sole proprietorship establishments of the plot. Every plot is weighted by its average number of establishments across 5 years (every 5 year between 1885 and 1905). Standard errors clustered by block are reported in parentheses: *** indicates statistical significance at the 1% level, ** at the 5% level and * at the 10% level.

Table 1.6: Regressions by Different Block Size and Treatment Definitions

Dep Var: shr of sole	(1)	(2)	(3)	(4)	(5)	(6)
Block Size	200m	200m	300m	300m	400m	400m
Treatment Threshold	25m	50m	25m	50m	25m	50m
Treatment	-0.047**	-0.025	-0.029	-0.003	-0.009	0.016
	(0.018)	(0.027)	(0.023)	(0.027)	(0.045)	(0.043)
Post1895	-0.077***	-0.076***	-0.071***	-0.068**	-0.069***	-0.063***
	(0.020)	(0.027)	(0.017)	(0.026)	(0.016)	(0.020)
Treatment*Post1895	-0.053**	-0.040	-0.055**	-0.048	-0.069*	-0.050**
	(0.023)	(0.028)	(0.022)	(0.029)	(0.037)	(0.022)
200m-Block FE	YES	YES	YES	YES	YES	YES
Observations	680	572	496	452	396	380
R-squared	0.754	0.759	0.809	0.803	0.817	0.803

Notes: For all specifications, the regression sample includes only the years 1885 and 1905. Every plot is weighted by its average number of establishments across 5 years. Standard errors clustered by block are reported in parentheses: *** indicates statistical significance at the 1% level, ** at the 5% level and * at the 10% level.

1.7.3 Robustness to Block Size and Treatment Threshold

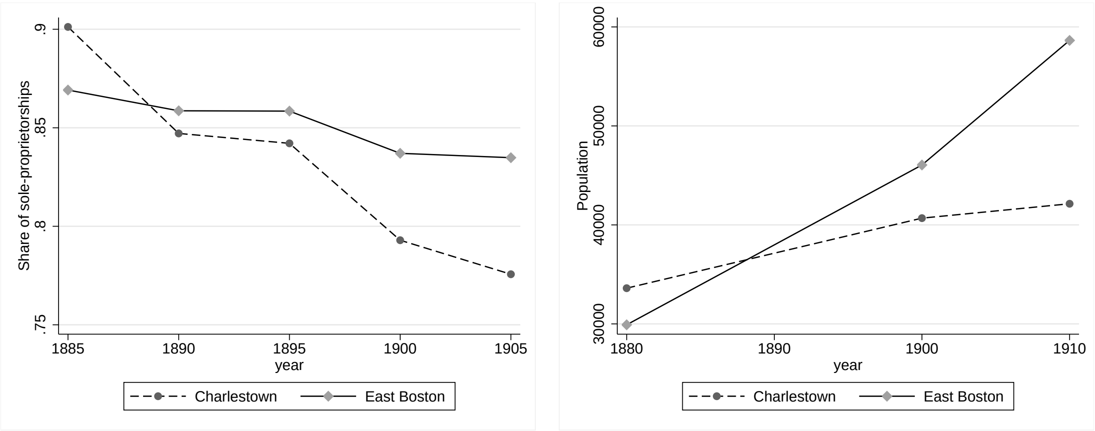
Table 1.6 examines the robustness of the benchmark results to different block sizes and distance thresholds to define the treatment group. For conciseness, only the estimation results corresponding to column (3) in Table 1.5 are reported. Column (1) in Table 1.6 is taken exactly from column (3) in Table 1.5 for the ease of comparison.

From columns (2)-(6), I vary the size of blocks between 200 m, 300 m, and 400 m, and vary the distance threshold between 25 m and 50 m. I find that the key coefficients are consistently negative. The magnitude of the coefficients before the interaction term increases with the block size, and it is slightly sensitive to the treatment threshold. Comparing this coefficient between (1) and (2), (3) and (4), and (5) and (6), respectively, I discover that adopting a wider "bin"-within 50 m of the rails versus a direct connection-attenuates the treatment effect. There are two possible causes for this finding. First, by demarcating a smaller area for the control plots, the number of establishments in some control plots falls to zero, so their corresponding blocks are dropped from the regression sample. This introduces a sample selection bias. Second, 25 m is the threshold for immediate proximity to the streetcar rails, while 50 m is further away and an arbitrary threshold. There could be a discontinuity in consumers' costs to access stores between locations with direct and indirect rail connections, but no such discontinuity at the random distance threshold. In Section 1.8, I discuss evidence that is consistent with the second interpretation.

Taken together, Tables 1.5 and 1.6 suggest that the estimated relative decrease in the share of sole proprietorships in the rail-connected plots is a robust result.

1.7.4 Heterogeneity by Geography

In this section, I examine the heterogeneity of the treatment effect by geography. The specific geography of Boston provides a useful case study to identify the importance of market access in affecting the outcome variable. In Boston, there are only two peninsulas, Charlestown and East Boston. (Their precise locations are found in Figure 1.3.) These two peninsulas were similar in population size, geographic area, and distance to the city center. However, since the late-eighteenth century, Charlestown was connected to central Boston by bridges, [18] and the streetcar electrification included the portion of the streetcar rails on the bridges. However, East Boston was not connected to central Boston by any walkable roads until the opening of the East Boston (streetcar) Tunnel in 1904. Thus, the streetcar electrification in the early 1890s shortened the distance between Charlestown and central Boston, while East Boston remained largely isolated from the city center.



Time trends in firm size proxy.
Data source: The *Boston Directories*.

Time trends in population.
Data source: full-count census data, IPUMS.

Figure 1.9: Comparison of Charlestown and East Boston

Figure 1.9 shows the overall trends in the share of sole proprietorships and population in these two neighborhoods. The left subfigure documents that the share of sole

[18] The first bridge in this area was the old Charles River Bridge, chartered in 1785 and opened on June 17, 1786

proprietorships-a proxy for average establishment size-grew much more in Charlestown than it did in East Boston. Between 1885 and 1905, this share declined by 12.5-percentage points in Charlestown, while there was only a 3.5-percentage point drop in East Boston. From the right subfigure, we can see that this occurred at the same time that there was a relative population growth in East Boston.

Table 1.7: Regressions by Geography

	(1)	(2)	(3)
Geographic Coverage	Charlestown	East Boston	Central Boston
Treatment	-0.025	-0.075	-0.048**
	(0.059)	(0.078)	(0.019)
Post1895	-0.032	-0.138	-0.079***
	(0.048)	(0.084)	(0.021)
Treatment*Post1895	-0.165**	0.080	-0.049**
	(0.063)	(0.088)	(0.024)
200m-Block FE	YES	YES	YES
Observations	76	60	556
R-squared	0.472	0.239	0.765

Notes: For all specifications, the regression sample includes only the years 1885 and 1905. Every plot is weighted by its average number of establishments across five years. Standard errors clustered by block are reported in parentheses: *** indicates statistical significance at the 1% level, ** at the 5% level and * at the 10% level.

Next, I rerun the benchmark regression in column (3) in Table 1.5, dividing the sample into three areas: Charlestown, East Boston, and central Boston. I show the results for each area in columns (1)-(3) in Table 1.7, respectively. Again, the regression results reveal a stark contrast between Charlestown and East Boston. While the coefficient before $Treatment * Post1895$ is estimated to be -0.165 for Charlestown, it is 0.08-a positive number and imprecisely estimated-for East Boston. Quite plausibly, the bridges enabled Charlestown businesses to reach consumers from central Boston, and rail connections were particularly important to reach such a consumer base. For businesses located in East Boston, the streetcar rails could not reach nonlocal markets, and thus, they were less important there.

1.8 Main Mechanism

In this section, I examine the mechanism behind the treatment effect. I proceed in three steps. In Section 1.8.1, I show that in my study period the consumers could be highly sensitive to commuting costs, and therefore, being located a short distance away from the rails could make a large difference in market access for firms. In Section 1.8.2, I test the first implication of the model, namely, that market access will be capitalized into land prices. In Section 1.8.3, I show that spatial sorting between firm productivity and market access occurred in the data.

1.8.1 Evidence on the Magnitude of Commuting Costs

The estimated large treatment effect between immediately connected and neighboring unconnected locations hinges critically on consumers being highly sensitive to commuting costs, especially to the costs of going to off-rail locations. I provide three pieces of evidence.

The Spatial Distribution of Employment

First, I illustrate employment distribution by distance to the streetcar rails and by distance to the city center in the first two rows of Figure 1.5. In the first row of subfigures, I show the results for all distances to the city center. Since the employment density within 1 km of the city center is so high that the results for the rest of distances are not clearly visible, I zoom in on the results for distances further than 1 km away from the city center in the second row of subfigures. Here, we can see that employment was highly concentrated around the rails, and that the concentration of employment around the rails became stronger after the electrification of the streetcar. In particular, in the areas further than 1 km away from the city center, there was a big jump in the employment density at a distance of 25 m (immediate proximity) from the rails. These patterns suggest that it was hard to operate a business in the off-rail areas in the suburbs in my study period.

Heterogeneity of Treatment Effects by Product

To shed more light on whether the treatment effect is driven by consumers' sensitivity to commuting costs, I exploit variations across products. For products featuring a higher ratio of the cost of commuting to the cost of goods, we expect that the treatment effect will be stronger among these products.

I impute the cost of commuting to the cost of goods ratio for each product using the 1996 Consumer Expenditure Survey[19]. I match the 25 products in my Boston data to the corresponding products in the 1996 Consumer Expenditure Survey. The details of this match are provided in Appendix 3. The data allow me to calculate four statistics that are relevant to the cost of commuting to the cost of goods ratio for each product: the costs per item, the number of items purchased per week, the number of trips consumers took each week to purchase any item of this product, and the number of trips the consumers made for purchasing every 100$ of this product. I take the last variable as the most relevant measure of the cost of commuting to the cost of goods ratio, and thus order the products by this measure in Table 1.8, from highest to lowest.

I find that food-related products feature a low value per item, a high purchase frequency, and more trips made by consumers for every 100$ purchase. These patterns suggest that the cost of commuting to the cost of goods ratio is higher for food for consumers today. In my study period, there was no domestic refrigerator, and food stores were more specialized than they are today. Thus, the purchase frequency for food at that time could have been even higher.

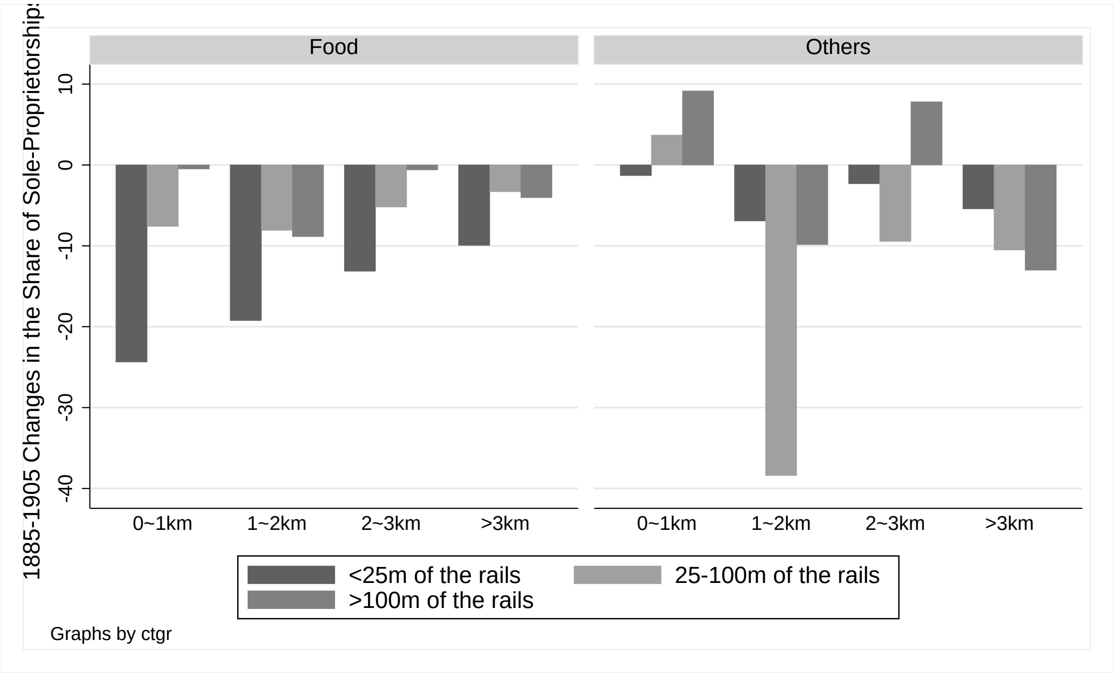
[19] 1996 was the first year for which this survey is publicly available.

Table 1.8: Consumption Behavior Statistics

Pshr	costs($)/item	items/week	trips/week	trips/100$
Confectioners^F	2.30	1.02	0.68	28.74
Bakers^F	1.99	2.87	1.28	22.45
Fruits^F	1.72	4.76	1.29	15.70
Fish^F	5.03	0.36	0.28	15.38
Cigars & Tabaccos	6.46	0.61	0.54	13.78
Produce^F	2.25	5.32	1.56	13.06
Books & Publishers	6.79	0.58	0.46	11.72
Liquors^F	5.85	0.97	0.48	8.44
Restaurants^F[20]	4.76	6.09	2.44	8.43
Provisions^F	3.06	8.84	1.77	6.56
Hats, Caps, & Furs	14.45	0.06	0.05	5.83
Milliners	17.52	0.08	0.07	4.81
Apothecaries & Drugs	15.64	0.63	0.43	4.30
Dry Goods	18.57	0.12	0.09	3.82
Hardwares	19.52	0.18	0.11	3.20
Boots & Shoes	34.52	0.18	0.14	2.29
Clothing	21.28	0.98	0.44	2.10
Jewelry & Watches	48.66	0.07	0.06	1.62
Leather	70.99	0.04	0.03	1.20
Furnitures	191.40	0.04	0.03	0.38
Piano	>93.43	0.00	0.00	1.07
Grocers^F	N/A	N/A	N/A	N/A
Tailors	N/A	N/A	N/A	N/A
Men's Furnishings	N/A	N/A	N/A	N/A

Source: 1996 Consumer Expenditure Survey from the Bureau of Labor Statistics.

Guided by Table 1.8, I present the treatment effect by the food-related products and other products. To best visualize the results, I plot the average changes in the share of sole proprietorships by product and geography in Figure 1.10. One can find the corresponding regression results of the treatment effect in Appendix 2.



Notes: The horizontal axis represents the distance from the city center. The bars of different colors indicate different distances from the streetcar rails. The distances were calculated using the 1888 rails. The food-related products did not include restaurants because consumption is primarily done on-site compared to the other food-related products. I examined the results by including restaurants in food and found similar patterns.

Figure 1.10: 1885-1905 Changes in the Share of Sole Proprietorships by Product and Geography

The left subfigure of Figure 1.10 suggests that the significant treatment effect found in the baseline regressions in Table 1.5 is largely driven by the food-related businesses. The decline in the outcome variable is sensitive to the distance to the streetcar rails: the decline is sharpest within 25 m of the rails, and then becomes much smaller between 25 and 100 m of the rails. We see that the distinction between the areas 25-100 m and further than 100 m away from the rails is not substantial. In contrast, we find that the patterns for the other products are mixed.

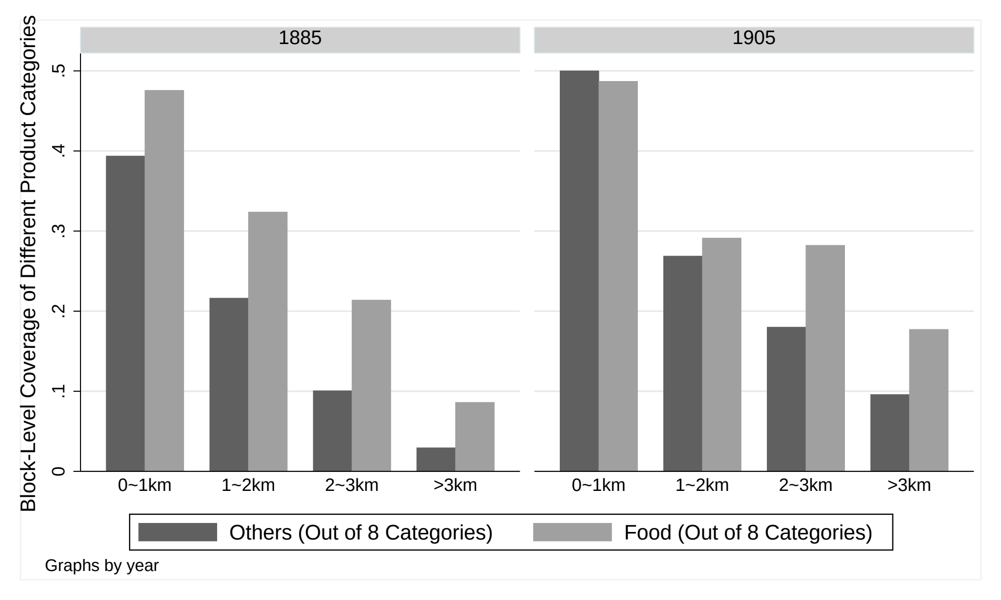
The stark contrast in the treatment effect between the food products and the other products is consistent with consumers being more sensitive to commuting costs to shop for food. However, there could have been other factors that drove this difference. While I cannot rule out all of the other possible factors, in Section 1.9.2, I show that this difference was unlikely to have been caused by the difference in the land-use intensity in production technology by industry.

Product Diversity at Block Level

The last piece of evidence that suggests high commuting costs for shopping is the spatial distribution of businesses. I divide the Boston area into $200m \times 200m$ blocks. I then calculate the coverage of the top eight products in each broad product category (i.e., food products and the other products) in each block in 1885 and 1905. This index reflects the accessibility to a wide range of products within a small neighborhood. I calculate the average of this index across blocks by distance to City Hall, as well as by year, and plot the results in Figure 1.11. We can see that a typical $200m \times 200m$ block covered approximately 30% of the top eight products in either year. Such a high coverage of products is an indicator that consumers needed to save time when shopping in that period. In both years, and at almost all distances from City Hall, there is a higher coverage of the food-related products than the other products in a typical block. This is consistent with consumers being more sensitive to commuting costs for food shopping.

1.8.2 Capitalization of Market Access into Land Prices

In Section 1.7.4, I showed that the streetcar rails mattered for businesses to access nonlocal consumers. In Section 1.8.1, I discussed the fact that the consumers could be



Notes: The horizontal axis represents the distance from the city center. I calculated the coverage of the top eight products in each broad product category (food and the others) in each $200m \times 200m$ block in 1885 and 1905. I then calculated the average of this index across blocks by distance to the city center and by year-product. The food-related products did not include restaurants.

Figure 1.11: 200m-Block-Level Product Diversity by Sector

highly sensitive to commuting costs, especially to the cost of going to off-rail locations. Taking these two facts together, we can infer that market access was much more strongly improved in rail-connected locations relative to neighboring unconnected locations. Both this and the following one test the implications of the model, most specifically: (1) that improved market access will be capitalized into land prices; (2) there is sorting between firm productivity and market access.

In the upper subfigure of Figure 1.12, I document the real estate prices (both land and buildings) in Boston by geography in the years 1885 and 1898, which I calculated using the digitized Boston property tax ledgers. However, any correspondence between market access improvement and real estate appreciation needs to be interpreted with caution. Conditional on the fixed land supply, real estate appreciation might be driven by an increase in demand for either commercial use or residential use or both. In my model, market access refers to firms' access to consumers, which corresponds to the demand for commercial use. From Figure 1.5, we can see that in the areas within 25 m of the streetcar rails, the increase in population density between 1885 and 1905 was primarily driven by the increase in employment density. The residential density in this area stayed almost the same over this same period. Therefore, I interpret the real estate appreciation in this area as driven mostly by the demand for commercial use. Consistent with the model, we observe in Figure 1.12 that real estate prices in this area increased during this period, suggesting that market access was capitalized into real estate value.

The bottom subfigure of Figure 1.12 plots the time trends of real estate values between 1885 and 1898 for both Charlestown and East Boston. Starting from similar levels in 1885, by 1898, the real estate value almost doubled in Charlestown, whereas it declined by 43% in East Boston. Since East Boston was a more isolated neighborhood in this period, thus having poorer market access, this fact provides another piece of evidence that market access was capitalized into land prices.

1.8.3 Firm Sorting

To identify the existence of sorting, I exploit the feature of the *Boston Directories* that allow us to track individuals and firms over time. Since we have information on firm owners, the firms can be tracked over time, even if they changed their names. I do over-time match in two directions. In the first direction, I select a random sample from the years 1890, 1895, 1900, and 1905, totaling 764 firms, and tracked them *backward* to five years before. In the second direction, I select a random sample of 728 firms in 1888, and track them $forward$ to 1899. I provide the details of this match in Appendix 3.

In the first direction tracking (backward tracking), I distinguish two groups of incumbent firms: those who survived from past five years, and those who entered in the past five years. For the former group, the survivors, they can be categorized into "moved

Table 1.9: Locational Choices of Incumbent and Entrant Firms

	Incumbents in 1890 and 1895		Incumbents in 1900 and 1905	
	Co./Partnerships	Sole-Prop.	Co./Partnerships	Sole-Prop.
	(1)	(2)	(3)	(4)
Survived from the Past 5 Yrs				
Connected to Connected	30.4%	28.1%	30.3%	28.7%
Connected to Unconnected	4.3%	3.5%	2.8%	2.5%
Unconnected to Connected	5.2%	4.8%	8.5%	4.3%
Unconnected to Unconnected	28.7%	17.1%	31.7%	17.2%
Entered in the Past 5 Yrs				
Connected	17.4%	24.6%	21.8%	25.8%
Unconnected	13.9%	21.9%	4.9%	21.5%
Observations	115	228	142	279

Notes: The data came from a random sample of backward linked establishments in the *Boston Directories*. For each incumbent establishment in the sample, I distinguished it between a survivor from the past five years, and an entrant in the past five years. For survivors, I tracked their relocation patterns (four categories, which are shown in the above table). For entrants, I distinguished them between those who located in connected areas and those who in unconnected areas.

from connected to connected locations," "moved from connected to unconnected locations," "moved from unconnected to connected locations," or "moved from unconnected to unconnected locations." For the latter group, the entrants, they can be categorized into "entered into connected locations" or "entered into unconnected locations." I show the share of each type of incumbent by five-year cohort as well as by legal form in Table 1.9. The main information to be drawn from this table is that we find evidence of spatial sorting. Moreover, we find that the sorting was not driven mainly by the relocation of the survivors; instead, it was driven primarily by the new company/partnership entrants, who suddenly exhibited a stronger preference for locating near rail-connected locations after the electrification. This fact is shown in the last two rows of columns (1) and (3) in Table 1.9. On the other hand, sole proprietorship entrants did not exhibit a change in preference over locations before and after the electrification, which is shown in the last two rows of columns (2) and (4) in Table 1.9.

Table 1.10: The Dynamics of the Establishments Between 1888 and 1899

	Unconnected to Rails in 1888		Connected to Rails in 1888	
	Sole-Prop.	Co./Partnerships	Sole-Prop.	Co./Partnerships
	(1)	(2)	(3)	(4)
Survived Between 1888 and 1899				
Connected to Connected			36.5%	41.2%
Connected to Unconnected			6.4%	6.7%
Unconnected to Connected	9.6%	16.1%		
Unconnected to Unconnected	23.9%	48.3%		
Exited Between 1888 and 1899				
Exited	58.9%	29.7%	50.0%	47.9%
Occupation Changed	7.7%	5.9%	7.1%	4.2%
Observations	209	118	282	119

Notes: The data came from a random sample of forward linked establishments in the *Boston Directories*. I distinguished the incumbent establishments in 1888 between those who survived over 1888-1899, and those who exited during this period. For survivors, I tracked their relocation patterns (four categories, which are shown in the above table). For those who exited, I distinguished them between disappearing from the directories and changing their occupation.

In the second direction tracking (forward tracking), I also distinguish two groups of incumbent firms in 1888: those who survived between 1888 and 1899, and those who exited during this period. I show the dynamics of these firms in Table 1.10. The key information of this table is in the third row of columns (1) and (2): among the sole proprietorships who were located in unconnected locations in 1888, 9.6% of them moved to connected locations in 1899. In contrast, among the companies/partnerships who were located in unconnected locations in 1888, 16.1% of them moved to connected locations in 1899. This is consistent with sorting between productivity and market access.

1.9 Alternative Mechanisms

In this section, I consider alternative mechanisms that might also explain the treatment effect.

1.9.1 Other Infrastructures

A natural concern is whether the electrification of the streetcar system was associated with any other improvements along the rail-connected streets, in addition to the improved market access. The first possible improvement is better availability of electricity along the rail-connected streets, which might have benefited the operation of large businesses disproportionately. The extension of electric services in Boston did indeed begin in the 1890s, however, electric utilities were available to users throughout the city at uniform prices (Warner, 1962). Therefore, it is unlikely that the availability of electric utilities drove the differential time trends in the outcomes between the rail-connected and unconnected locations.

Another improvement along the rail-connected streets was better sanitation caused by the elimination of horse pollution. In fact, horse pollution was once an enormous public health and sanitation problem for almost every American city (Morris, 2007). A byproduct of the electrification of the streetcar system was that the streets along the previous horse-drawn streetcar routes became much cleaner than before. Therefore, the heterogeneity of the treatment effect by food and non-food products documented in Figure 1.10 could be partly driven by the fact that the food-related businesses benefited more from the cleaner streets. However, if the sanitation improvement was the only story, one could hardly explain the sorting patterns that were documented in Section 1.8.3; in other words, it is not clear why improved sanitation conditions would have been more relevant to larger firms than to smaller ones.

Finally, beginning in 1897, this period also saw the openings of the Boston subway system.[21] Because subways and electric streetcars were similar in nature-both were commuter rails, and the routes of the subways overlapped with parts of the streetcar routes-I interpret the effects of the subway system as a robustness check of the main

[21] Boston was the first city that opened a subway system in North America.

mechanism. Because subways represent a more advanced type of intra-city commuting practice, [22] we would expect to find a large treatment effect for the subways.

Indeed, I find a significant treatment effect of the opening of the subway. In particular, the subway lines are highly responsible for the anomalies observed in the right subfigure of Figure 1.10, which shows that there was a particularly sharp drop in the share of sole proprietorships in the area between 1 and 2 km away from City Hall. In this area, two subway stations opened in 1898, the Boylston Station and the Pleasant Station, which were located along the electric streetcar routes and were 1 km and 1.5 km away from City Hall, respectively (visualized in Figure 1.23).

To test whether the results in the area between 1 and 2 km away from City Hall are driven by the construction of the subways, I restrict the sample to three regions: the areas within 300 m of the Boylston Station and the Pleasant Station, respectively, and then the area within 300 m of the midpoint of the two stations. I then divide each area into three distance bands from the centroid. I plot the evolution of the share of sole proprietorships in each distance band in each area. From the top two subfigures in Figure 1.13, we find that there was an especially steep, post-1895 drop in the share of sole proprietorships in the areas within 100 m of these two subway stations (plotted in solid lines), relative to the drops in the areas further than 100 m away from the two stations. As a placebo test, the bottom subfigure plots the trends at different distance bands, with the centroid at the midpoint of the two subway stations. I do not find that proximity to the midpoint of subway stations had any significant impact on the outcome trends. These tests suggest that the subway construction was an important cause of the extreme outcomes observed in the areas between 1 and 2 km away from City Hall. Given the similar nature of the subways and the electrical railways, these results lend additional support to the hypothesis that market access was a significant barrier to firm size.

[22] From a contemporaneous engineering journal, subway trains traveled at speeds of approximately 17 mph, which was double the speeds of the electric streetcars.

1.9.2 Explanations for the Heterogeneous Treatment Effects by Product

What explains the treatment effect being mainly driven by food? In Section 1.8.1, I argued that this due to the higher ratio of the cost of commuting to the cost of goods for food. In this section, I consider a second possible mechanism-the differences in production technologies by sector. The food industry might use land more intensively in production, so, in this scenario, the land rents would account for a larger share of the total costs for food-related businesses. An upgrade in transportation infrastructure would improve market access nearer to the infrastructure, and would thus raise the land rents there relative to more distant areas. Because food-related businesses are more sensitive to rents, relatively more small food-related businesses would be driven out of the locations near the infrastructure than small other businesses would be.

I test whether the pre-assumption for the land use intensity mechanism holds, i.e., that food-related businesses use land as inputs more intensively. To my knowledge, there is no data on production technology for the firms in my study sample. To shed some light on this parameter, I calculate a measure of land-use intensity using the *Enterprise Surveys* from the World Bank. I restrict my investigation to country-years that were at a comparable level of GDP per capita as the United Stated was between 1885 and 1905. I provide descriptions of this data set and details of the sample selection criteria in Appendix 3. The land-use intensity is measured by

$$\frac{\text{Cost for Establishment to Repurchase All of Its Land and Buildings}}{\text{Cost for Establishment to Repurchase All of Its Machinery, Land and Buildings}} \tag{1.4}$$

I plot this measure across the ten countries in my sample in Figure 1.14. Relative to the other sectors, the land-use intensity in the food sector is higher in six countries, but lower in the remaining four. Therefore, there is no consistent pattern across countries.

I also run a regression of land-use intensity on the food sector dummy and the country fixed effects, using all the firm observations. I find an insignificant and close to zero coefficient before the food sector dummy. Hence, the evidence from the firms in today's developing countries does not support the pre-assumption that the food sector uses land more intensively.

Even if the share of land rents in total costs in the food industry is higher in my study context, this explanation might not be the whole story. As documented in Figure 1.11, there is a much greater product diversity at the block level for food. This can be explained by consumers at that time being more sensitive to food shopping commuting costs, but it can hardly be explained by the share of land rents in the total costs.

1.10 Conclusion

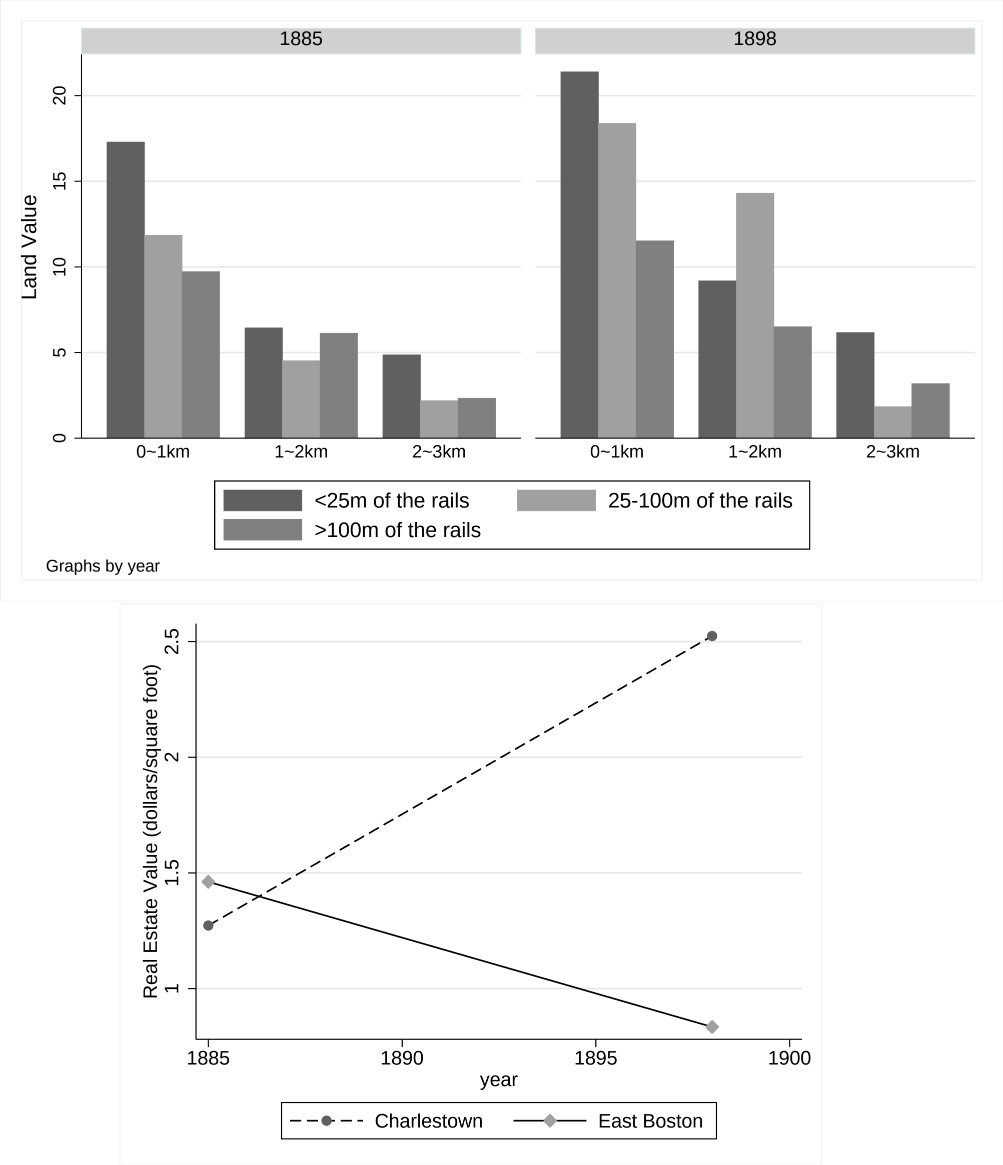
In this paper, I use a natural experiment-the electrification of the streetcar system in Boston between 1889 and 1896-to provide the first causal evidence that an upgrade of transport infrastructure leads to a decrease in the share of sole proprietorships. To do so, I digitized and geocoded business data for the universe of the top 25 retail/wholesale products, as well as the city transit network and land value data in Boston between 1885 and 1905. The identification strategy exploits the fact that the new electric system was quickly upgraded, while keeping the preexisting horse-drawn streetcar routes almost unchanged.

Using a difference-in-differences estimator, I find that rail-connected locations experienced a 5.3-percentage point relative drop in the share of sole proprietorship establishments after the electrification compared to neighboring unconnected locations. The treatment effect is robust to alternative divisions of locations, thresholds for defining treated location, and different controls. The magnitude of the treatment effect is striking considering that the rail-connected plots are, on average, within 25 m of the rails, while the neighboring plots are, on average, between 25 and 100 m away from the rails. Further analysis reveals that consumers could be highly sensitive to commuting costs in my study context, so being only a short distance away from the rails made a significant difference in market access for firms. To a large extent, the treatment effect reflects spatial sorting between firm productivity and market access, which is consistent with more productive entrepreneurs being able to take better advantage of market access.

The results of this paper have implications for the theories of the firm size distribution. Existing explanations for the prevalence of micro and small enterprises in the process of economic development emphasize regulatory and institutional barriers, which distort the firm size distributions by disfavoring either small or large firms. In my study context-historical Boston between 1885 and 1905-institutions had been quite stable, but we still observe a quick shift in the firm size distribution. The evidence in this paper points to the important role played by transport infrastructure improvements, which lends support to Chandler (1977), Tybout (2000), Hsieh and Klenow (2014), and Lagakos (2016).

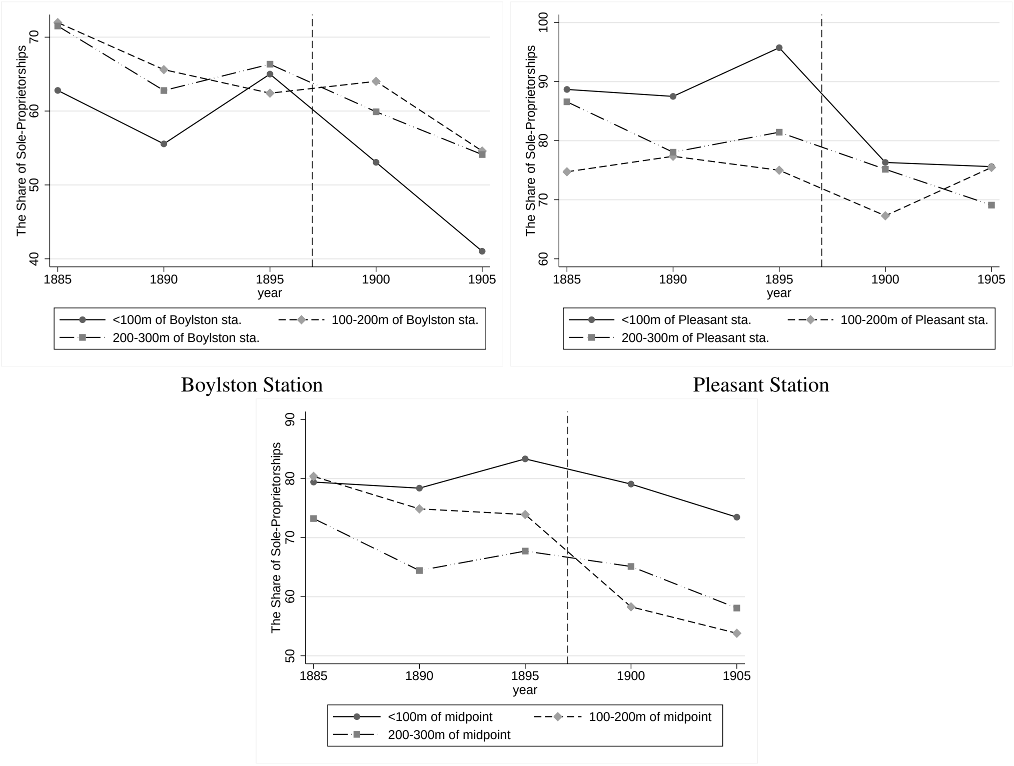
The results on spatial sorting between firm productivity and transit access after the upgrade of a transit system have implications on resource (mis)allocation across firms. Hsieh and Klenow (2009) document that there exists a higher degree of resource misallocation across firms in China and India than in the US. This paper suggests a potential source of misallocation at a low level of economic development: the geographic segregation of markets, which enables a large number of low-productivity entrepreneurs to stay away from competition with more productive, larger firms. An improvement in transport infrastructure could improve the resource allocation across firms by inducing more productive firms to move into more advantageous locations, which enlarges their market shares. The trade-off could be an increase in inequality, as suggested by the higher exit rates among the small firms following the transportation shock in my study event.

Chapter 1, in full, is currently being prepared for submission for publication of the material. The dissertation author, Wei You, was the sole author of this paper.



Notes: In the upper subfigure, the horizontal axis represents the distance from the city center. The bars in different colors indicate different distances from the streetcar rails. The distances from the rails were calculated using the 1888 rails. For both figures, the data came from the 10% random sample of the building units from the 1885 and 1898 Boston property tax ledgers. The real estate values were calculated as the summation of building value and land value, divided by the square feet of the land. The data for the area further than 3 km away from the city center was not digitized because a large fraction of the addresses in this area could not be geocoded.

Figure 1.12: Real Estate Value in 1885 and 1898



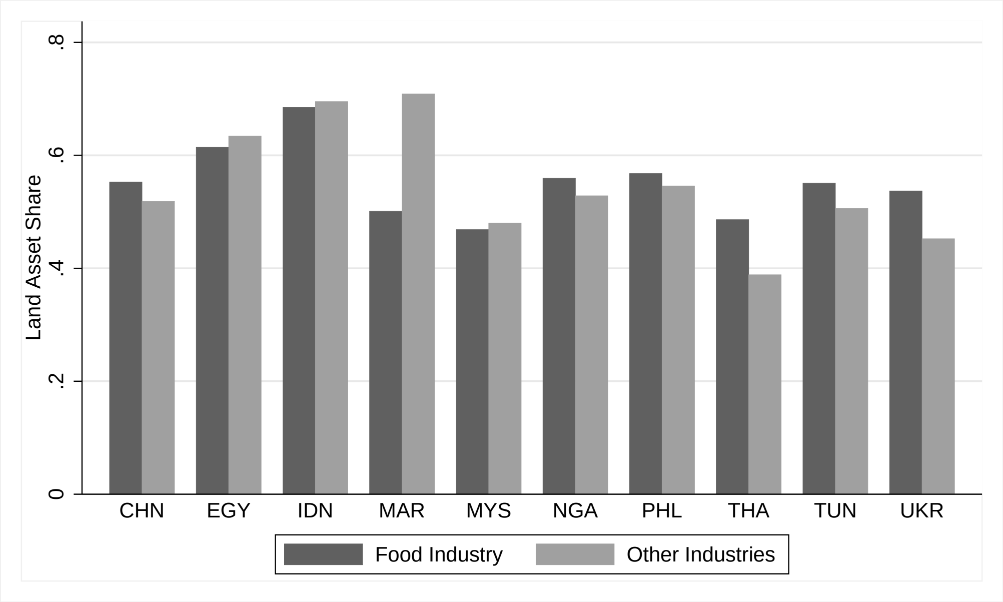
Boylston Station

Pleasant Station

The Midpoint of the Two Stations

Notes: The locations of the Boylston and Pleasant subway stations in 1898 are depicted in Figure 1.23.

Figure 1.13: Time Trends in the Share of Sole Proprietorships near Subway Stations



Notes: The data came from the World Bank Enterprise Surveys for the corresponding countries between 2012 and 2015. The details of the this data set and the sample selection rules are provided in the Appendix 3. The land asset share was calculated using the ratio of the hypothetical repurchase value for land and buildings to the hypothetical repurchase value for land and builings plus captial.

Figure 1.14: The Land Asset Shares by Industry and by Country

1.11 Appendix 1: Details of the Model

This appendix section provides details of the model described in Section 1.4.

1.11.1 Setup

Geography

Consider a city embedded within a wider economy. The city consists of a discrete set of locations or blocks, which are indexed by $s = 1, \ldots, S$. S is large so that the impact of each location on the entire city is ignorable (each location corresponds to a "plot" in the empirical analysis). All economic activities-production and consumption-take place at $s \in S$. The space is further defined by bilateral transport costs, τ_{sd}, for consumers located in s to receive services provided in location d. I interpret source s as the consumers' location and destination d as the firms' location, because the sector of interest is the retail sector. I then impose symmetry and the triangle inequality on τ_{sd}.

Locations differ by their relative proximity to other locations. Locations also differ by productivity φ_s, which is independent of firm characteristics. The differences in the location-specific productivity φ_s can arise from differences in natural advantages and proximity to fixed urban features, such as wharfs, inter-city train stations. The CBD area, which features a relatively high density of employment and firms in equilibrium, emerges because of the higher φ_s in this area.[23]

Preferences

There are three types of agents in the economy: entrepreneurs, landowners, and workers. Entrepreneurs obtain income from profits, land owners obtain income from

[23] In my model, CBD could also emerge in the absence of differences in φ_s in an asymmetric equilibrium, where some ex ante identical locations would attract relatively more firms and employment than others, and where such asymmetric distribution would be stable due to the "home market effects" (Krugman, 1991).

rents, and workers obtain income from wages. All the agent types have preferences over differentiated goods produced at all locations. These varieties are combined through the CES function form:

$$Q = \left[\int_{\omega \in \Omega} q(\omega)^{\frac{\sigma-1}{\sigma}} d\omega \right]^{\frac{\sigma}{\sigma-1}}$$

The corresponding price index is

$$P = \left[\int_{\omega \in \Omega} p(\omega)^{1-\sigma} d\omega \right]^{\frac{1}{1-\sigma}}$$

The optimal consumption and expenditure decisions for individual varieties ω are

$$q(\omega) = Q \left[\frac{p(\omega)}{P} \right]^{-\sigma},$$

$$r(\omega) = R \left[\frac{p(\omega)}{P} \right]^{1-\sigma},$$

where $R = PQ$ denotes aggregate expenditure.

I assume that transport costs enter the price of goods in standard “iceberg” form, so that the actual price that consumers face is the price of goods *multiplied* by a factor $\tau_{sd} > 1$. Via this assumption, each variety is purchased by consumers from all the sources. In this sense, even the smallest store has a citywide customer base. In this model, small stores are localized not in the sense that they only serve local customers, but in the sense that they are located in places that lack access to nonlocal markets, so a larger portion of their customers comes from the local area.

Production

There are two factors of production, labor (l) and floor space (h). The individual variety production function is

$$q(\omega) = q(\varphi) = \varphi l(\varphi) \nVdash (h(\varphi) = 1)$$

Any positive amount of output requires one unit of floor space as input, which is independent of firm size. In addition, every additional unit of output requires a constant marginal labor requirement, $1/\varphi$, where φ is firm-specific productivity. Each variety is provided by a single firm. There are no economies of scope, so there is a one-to-one relationship between firms and varieties. [24] Thus, the indexes ω and φ are inter-changeable, and hence, I will use φ hereafter.

Worker Location Decision

Each homogeneous worker supplies one unit of labor inelastically in her workplace and does not participate in a real estate market. Thus, workers do not commute between workplace and residence. Instead, workers commute between their workplace and stores for shopping. Workers are free to choose their workplace, trading off local wage rate against consumption amenity, summarized by the price index. In equilibrium, all workers receive identical reservation utility, $\bar{u}$, in the wider economy. Under the specified utility, the worker's free-mobility condition is given by

$$\frac{w_d}{P_d} = \bar{u},\ d \in S \tag{1.5}$$

[24] This assumption is roughly consistent with the data. In the *Boston Directories*, there are over one thousand product categories. If a firm produced multiple products, it appeared under each product category. Based on a random sample of 2000 firms, I verified that only 8% of them produced multiple products. Among the multi-product firms, most of the products a firm produces were similar in nature.

Firm Location and Pricing Decision

There is an endogenous set of entrepreneurs, with a mass of M. Each entrepreneur draws a level of productivity from some distribution $G(\varphi)$, $\varphi \in (\underline{\varphi}, \bar{\varphi})$, which is independent of sector. In equilibrium, the least productive firm in the lowest-profit sector always earns zero profit, so there is no exit after entrepreneurs draw a level of productivity.

Each entrepreneur runs one firm, operates in a single location, [25] and employs labor at their locations, paying location-specific wages. In equilibrium, all firms will choose to service all locations. Access to consumers, local production costs, local productivity, and rents drive the profits of the firms. For entrepreneurs, firm profits normalized by the local price index govern their location decisions.

The timing of the entrepreneur's/firm's decisions are: (1) whether to enter, (2) where to locate, and (3) how to price their product. A firm or entrepreneur's optimal choice can be found by solving the three decisions in reverse: first, by finding the optimal price of the good at each potential location; second, by finding the optimal location, given the pricing rule at each location; and finally, deciding whether to enter, given the profits in the optimal location.

By the demand and production functions in this model, firms' optimal pricing strategy is a constant markup over marginal costs:

$$p_d(\varphi) = \frac{\sigma}{\sigma - 1} \frac{w}{\varphi \varphi_d}, \tag{1.6}$$

where φ is firm-specific productivity and φ_d is location-specific productivity.

[25] In my data, only 4% of the firms operate in multiple locations.

The operational profits (without paying for the fixed land rent costs) from serving all the consumers in location s for a firm of productivity φ located in d is

$$\pi_{sd}(\varphi) = \kappa \left(\frac{\varphi \varphi_d}{w_d} \right)^{\sigma-1} R_s P_s^{\sigma-1} \tau_{sd}^{-\sigma}, \tag{1.7}$$

where $\kappa = \frac{1}{\sigma} \left(\frac{\sigma-1}{\sigma} \right)^{\sigma-1}$, φ_d is the location-specific productivity while φ is the firm-specific productivity. Summing over operational profits made from all the sources, I obtain an expression for the size of firm φ located in d in terms of operational profits:

$$\pi_d(\varphi) = \kappa \left(\frac{\varphi \varphi_d}{w_d} \right)^{\sigma-1} \int_{s \in S} R_s P_s^{\sigma-1} \tau_{sd}^{-\sigma} ds \tag{1.8}$$

Because entrepreneurs are also consumers at the same time, they care about real profits (adjusted for price-index) $\tilde{\pi}_d(\varphi) \equiv \frac{\pi_d(\varphi)}{P_d}$. Defining location-specific advantage, we have

$$\eta_d \equiv \frac{\kappa}{P_d} \left(\frac{\varphi_d}{w_d} \right)^{\sigma-1} \int_{s \in S} R_s P_s^{\sigma-1} \tau_{sd}^{-\sigma} ds$$

A nice feature of equation (1.8) is that the firm-specific productivity φ and the location-specific advantage η_d enter into the profits *multiplicatively*. It can easily be shown that the more productive firms can take better advantage of the location-specific advantage, in the sense that

Lemma 1.

$$\forall \eta_1, \eta_2 \in (\underline{\eta}, \bar{\eta}), \; \textit{if}\, \eta_1 > \eta_2 \textit{ and } \varphi_1 > \varphi_2,$$

$$\textit{then } \tilde{\pi}_1(\varphi_1, \eta_1) - \tilde{\pi}_1(\varphi_2, \eta_1) > \tilde{\pi}_1(\varphi_1, \eta_2) - \tilde{\pi}_2(\varphi_2, \eta_2)$$

Spatial sorting will take place, such that more productive firms occupy more advantageous locations, and this relationship is strictly monotonic.

Landowner Decision

Atomistic land owners decide the density of structures H_d, taking land rent r_d as given. The rent maximization problem of a representative land owner in d can be formalized as

$$\max_{H_d} \pi_l = r_d H_d - c(H_d), \tag{1.9}$$

where $c(H)$ is the construction cost function, which is increasing, twice-differentiable, and convex:

$$c'(H) > 0, c''(H) > 0.$$

The convexity of the construction cost function precludes the possibility of an equilibrium with an infinite density of structures/firms.

The first-order condition of the above problem yields an increasing relationship between rent and structure density:

$$r_d = c'(H_d) \tag{1.10}$$

Land owners collect total land rents in the form of the consumption aggregate, and they consume the profits part. The remainder of the consumption aggregate is used for construction. For any source location s, the expenditure on the aggregate consumption bundle equals the summation of the wages of local workers, the profits of local entrepreneurs, and the rents of local land-owners.

1.11.2 Equilibrium

Sorting

Because strict sorting takes place in equilibrium by Lemma 1, we have simple incentive compatibility conditions that guarantee that the match between location advantage and firm productivity is stable. Order the index $d \in S$ from the least advantageous to most advantageous such that $\eta_{d+1} > \eta_d$. Let $\{\varphi_d : d \in S\}$ be the set of cutoff productivities such that an entrepreneur with productivity φ_d is indifferent between locations d and $d+1$:

$$\tilde{\pi}_d(\varphi_d) - \tilde{r}_d = \tilde{\pi}_{d+1}(\varphi_d) - \tilde{r}_{d+1} \tag{1.11}$$

where $\tilde{r}_d \equiv \frac{r_d}{P_d}$ denotes real rents. Then all the entrepreneurs with productivity $\varphi \in \Phi_d \equiv [\varphi_{d-1}, \varphi_d)$ will choose location d. The zero-profit-cutoff condition is

$$\pi_0(\underline{\varphi}) - r_0 = 0 \tag{1.12}$$

Balanced Trade

In equilibrium, the total revenues of all the firms in each location (which accrue to local workers, entrepreneurs, and land owners), must equal the expenditures of the consumers from all the sources:

$$R_d \equiv \int_{\varphi \in \Phi_d} r_d(\varphi) d\varphi = \left(\frac{\sigma - 1}{\sigma}\right)^{\sigma-1} \int_{\varphi_{d-1}}^{\varphi_d} g(\varphi) M \varphi^{\sigma-1} \left(\frac{\varphi_d}{w_d}\right)^{\sigma-1} \int_{s \in S} R_s P_s^{\sigma-1} \tau_{sd}^{-\sigma} ds d\varphi \tag{1.13}$$

Price Index

The price index in each location and each sector is given by

$$P_s = \left[\int_{d \in S} \int_{\varphi_{d-1}}^{\varphi_d} \left(\frac{\sigma}{\sigma - 1} \frac{w_d \tau_{sd}}{\varphi_d \varphi}\right)^{1-\sigma} g(\varphi) M d\varphi dd\right]^{\frac{1}{1-\sigma}} \tag{1.14}$$

Housing Market Clearing

Finally, the housing markets must clear

$$H_d = c'^{-1}(r_d) = M\left[G(\varphi_d) - G(\varphi_{d-1})\right] \tag{1.15}$$

The equations (1.5), (1.8), (1.11), (1.12), (1.13), (1.14), and (1.15) characterize the equilibrium conditions.

1.11.3 Treatment Effect

There is no analytical solution to this model. However, the model characterizes the conditions under which we can observe the treatment effect.

Consider three locations: 0, 1, and 2. Location 2 is connected to rails, 1 is neighboring 2 but off-rails, and 0 is the rest of all the locations. Locations 2 and 1 correspond to a treatment plot and a control plot in the empirical analysis, respectively. Because locations 1 and 2 are small relative to the entire city, I assume that only the market in location 0 matters for firms. I also assume that the consumers from location 0 need to incur extra commuting costs to reach 1 - the off-rail location, relative to reach 2 - the rail-connected location:

$$\tau_{01} = \lambda\tau_{02}, \lambda > 1, \tag{1.16}$$

where λ represents the extra commuting cost factor. Assume that location advantages φ_d's are equalized and set to 1.

Using equation (1.8), we have an expression for firm size in both locations 1 and 2:

$$\pi_1(\varphi_1) = \kappa\left(\frac{\varphi_1}{w_1}\right)^{\sigma-1} R_0 P_0^{\sigma-1}\tau_{01}^{-\sigma} \tag{1.17}$$

$$\pi_2(\varphi_2) = \kappa\left(\frac{\varphi_2}{w_2}\right)^{\sigma-1} R_0 P_0^{\sigma-1}\tau_{02}^{-\sigma} \tag{1.18}$$

The above two equations have an intuitive interpretation: the operational profits of the firms in either location are positively related to firm raw productivity and nonlocal market size, and negatively related to local wage costs. The difference in firm size between locations 2 (connected) and 1 (unconnected) can be expressed as

$$\pi_2(\varphi_2) - \pi_1(\varphi_1) = \kappa R_0 P_0^{\sigma-1}\left[\left(\frac{\varphi_2}{w_2}\right)^{\sigma-1}\tau_{02}^{-\sigma} - \left(\frac{\varphi_1}{w_1}\right)^{\sigma-1}\tau_{01}^{-\sigma}\right] \tag{1.19}$$

Using the free labor mobility condition - equation (1.5), and the price index in equilibrium - (1.14), we can express the wages in locations 1 and 2 as

$$w_1 = \bar{u}P_1 = \bar{u}\left[\int_0^{\varphi_0}\left(\frac{\sigma}{\sigma-1}\frac{w_0\tau_{01}}{\varphi}\right)^{1-\sigma} g(\varphi)Md\varphi\right]^{\frac{1}{1-\sigma}} \tag{1.20}$$

$$w_2 = \bar{u}P_2 = \bar{u}\left[\int_0^{\varphi_0}\left(\frac{\sigma}{\sigma-1}\frac{w_0\tau_{02}}{\varphi}\right)^{1-\sigma} g(\varphi)Md\varphi\right]^{\frac{1}{1-\sigma}} \tag{1.21}$$

Combining equations (1.16), (1.20), and (1.21), there is

$$w_1 = \lambda w_2 \tag{1.22}$$

The above equation holds because location 2 has better access to nonlocal markets, so workers are willing to accept a lower nominal wage for the higher consumption amenity there.

Equation (1.14) also implies that

$$P_0 = \left[\int_0^{\varphi_0}\left(\frac{\sigma}{\sigma-1}\frac{w_0}{\varphi}\right)^{1-\sigma} g(\varphi)Md\varphi\right]^{\frac{1}{1-\sigma}} \tag{1.23}$$

where I assume that $\tau_{00} = 1$, i.e. there is no within-location commuting costs. Equations (1.20) and (1.23) together imply

$$w_1 = \bar{u}\tau_{01}P_0 \tag{1.24}$$

Combining equations (1.16), (1.19), (1.22), and (1.24), the difference in firm size between locations 2 and 1 can be further simplified to

$$\pi_2(\varphi_2) - \pi_1(\varphi_1) = \frac{\kappa_2 R_0}{\tau_{02}^{2\sigma-1}} \left(\varphi_2^{\sigma-1} - \frac{\varphi_1^{\sigma-1}}{\lambda^{2\sigma-1}} \right), \tag{1.25}$$

where $\kappa_2 = \frac{\kappa}{\bar{u}^{\sigma-1}}$.

From equation (1.25), we see that in a static setting, a necessary condition for observing a positive treatment effect ($\pi_2(\varphi_2) - \pi_1(\varphi_1) > 0$) is $\lambda > 1$, i.e., the extra-commuting costs to reach off-rail locations is positive. Otherwise if $\lambda = 1$, then the two locations are identical, $\varphi_1 = \varphi_2$, the left hand side of equation (1.25) is then zero. Corresponding to the empirical analysis, if $\lambda > 1$, we should be able to observe that the time-invariant treatment dummy predicts a larger firm size. The relatively greater firm size in location 2 than in location 1 comes from three sources: the relatively higher firm productivity, the relatively lower wages, and the better market access.

To be aligned with the difference-in-differences estimator, we are interested in the difference in firm size between two equilibria. The new equilibrium features a lower τ_{02}, and thus a lower τ_{01} through equation (1.16), while keeping all the other parameters, including the extra commuting costs λ, constant. Denote x' as the variables in the new equilibrium, and $\hat{x}$ as the ratio of a variable in the new and old equilibria, $\hat{x} \equiv \frac{x'}{x}$. The equation corresponding to the difference-in-differences estimator in the empirical analysis is then

$$\pi_2(\widehat{\varphi_2) - \pi_1}(\varphi_1) = \frac{\hat{R}_0}{\hat{\tau}_{02}^{2\sigma-1}} \left(\frac{{\varphi'_2}^{\sigma-1} - \frac{{\varphi'_1}^{\sigma-1}}{\lambda^{2\sigma-1}}}{\varphi_2^{\sigma-1} - \frac{\varphi_1^{\sigma-1}}{\lambda^{2\sigma-1}}} \right) \tag{1.26}$$

Equation (1.26) characterizes three possible sources for us to observe a positive treatment effect on firm size: an increase in nonlocal market size, $\hat{R}_0 > 1$, a decrease in transport costs, $\hat{\tau}_{02} < 1$, and a wider gap in firm productivity between the two locations, as summarized in the term in the parenthesis.

1.12 Appendix 2: Supplementary Tables and Figures

Table 1.11: The Top 25 Retail/Wholesale Products

Product	Percent	Product	Percent
Food related			
Grocers	16.51	Fruits	3.00
Liquors & Wines	9.41	Produce	2.63
Provisions	7.74	Fish	2.56
Restaurants	6.46	Confectioners	1.44
Bakers	4.17		
Clothing related			
Tailors	8.80	Clothing	2.78
Boots & Shoes	8.63	Men's Furnishings	1.31
Dry Goods	3.47	Hats, Caps, & Furs	0.72
Milliners	2.95		
Others			
Apothecaries	3.34	Jewelry, Watches, & c	1.76
Cigars & Tobaccos	3.07	Hardwares	1.30
Books'ers & Publ'ers	2.69	Pianos	0.44
Leather	2.41	Drugs & Medicines	0.42
Furnitures	2.00		

Source: The *Boston Directories* in 1885, 1890, 1895, 1900, and 1905.

1.12.1 Regression Results of Heterogeneous Treatment Effects by Product and Geography

This section confirms the results documented in Figure 1.10 using regressions.

To run product-category-specific regressions, I follow the same procedure to construct the units of the observations and the treatment and the control groups described in Section 1.5, using establishments only from that product category. Because business establishments in each product category are more sparse in space than the aggregate, I adopt a wider block size-300 m-and divide the Boston area into only two parts: the area within 1 km of City Hall, and the area further than 1 km away from City Hall. Table 1.12 shows the estimation results from the baseline regressions by product category and geography. The upper panel reports the results for the food-related products, and the lower panel reports the results for the other products. The two left columns define the treatment groups based on direct connections to the streetcar rails (on average 25 m from the rails), while the two right columns adopt a distance threshold of 100 m.

Table 1.12: Treatment Effect by Product and Geography and by Distance Threshold

Treatment Threshold	25m		100m	
Distance to City Hall	<1 km	>1 km	<1 km	>1 km
	(1)	(2)	(3)	(4)
Food-Related Products				
Treatment	0.033	-0.039	0.057	-0.012
	(0.041)	(0.025)	(0.057)	(0.026)
Post1895	-0.116***	-0.049*	-0.060	-0.047
	(0.038)	(0.026)	(0.047)	(0.028)
Treatment*Post1895	-0.149**	-0.111***	-0.039	-0.079**
	(0.058)	(0.035)	(0.063)	(0.033)
200m-Block FE	YES	YES	YES	YES
Observations	100	340	80	256
R-squared	0.718	0.478	0.796	0.421
Other Products				
Treatment	0.012	-0.191***	0.111	-0.221***
	(0.047)	(0.032)	(0.120)	(0.065)
Post1895	-0.013	-0.124***	0.014	-0.066
	(0.030)	(0.037)	(0.103)	(0.085)
Treatment*Post1895	-0.054	0.042	-0.092	-0.006
	(0.050)	(0.035)	(0.094)	(0.117)
200m-Block FE	YES	YES	YES	YES
Observations	80	172	56	56
R-squared	0.861	0.591	0.694	0.442

Notes: For all specifications, the regression sample includes only the years 1885 and 1905. Every plot is weighted by its average number of establishments across 5 years (every 5 year between 1885 and 1905). Standard errors clustered by $300m \times 300m$ block are reported in parentheses: *** indicates statistical significance at the 1% level, ** at the 5% level and * at the 10% level.

Comparing the results in columns (1) and (2) in Table 1.12, we can see that the significant and positive treatment effect found in the baseline regressions in Table 1.5

is driven largely by the food-related products. In the area within 1 km of City Hall and the area further than 1 km away from City Hall, the connected plots experienced a 14.9-percentage point and an 11.1-percentage point relative drop in the share of sole proprietorships, respectively. In contrast, there is a much smaller and insignificant treatment effect for the other products. In the area further than 1 km away from City Hall, the sign of the treatment effect is even positive.

Columns (3) and (4) in Table 1.12 show that the treatment effect for different products is sensitive to different treatment thresholds. While there is a huge and significant treatment effect for the food-related products using 25 m as the treatment threshold, the magnitude of the treatment effect decreases sharply when using 100 m as the treatment threshold. On the other hand, for the other products, the treatment effect becomes more negative (expected sign) when using 100 m as the threshold than using 25 m as the threshold.

1.13 Appendix 3: Supplemental Data

1.13.1 Consumer Expenditure Survey Data

The*Consumer Expenditure Survey* data are from the Bureau of Labor Statistics in the U.S. Census Bureau. The*Consumer Expenditure Survey* collects information from the nation's households and families on their buying habits (expenditures), income, and household characteristics. The survey consists of two components: a quarterly Interview Survey and a weekly Diary Survey. In this study, I use the 1996 weekly Diary Survey to get information on the purchase cost, the product code for each item purchased, and the purchase date for each consumer unit. The year 1996 is the first year for which this survey data is available online. There are 551 product codes in the 1996 Survey. I performed a match between the 25 retail/wholesale products in the *Boston Directories* (shown in Table 1.11) and the 551 product codes in the 1996 weekly Diary Survey. There are three products in the *Boston Directories*-grocers, tailors, men's furnishings-where is no reasonable correspondence in the 1996 weekly Diary Survey. I calculate the statistics in Table 1.8 for each of the matched 22 products using the Diary Survey.

1.13.2 Linking Firms Over Time in the Boston Directories

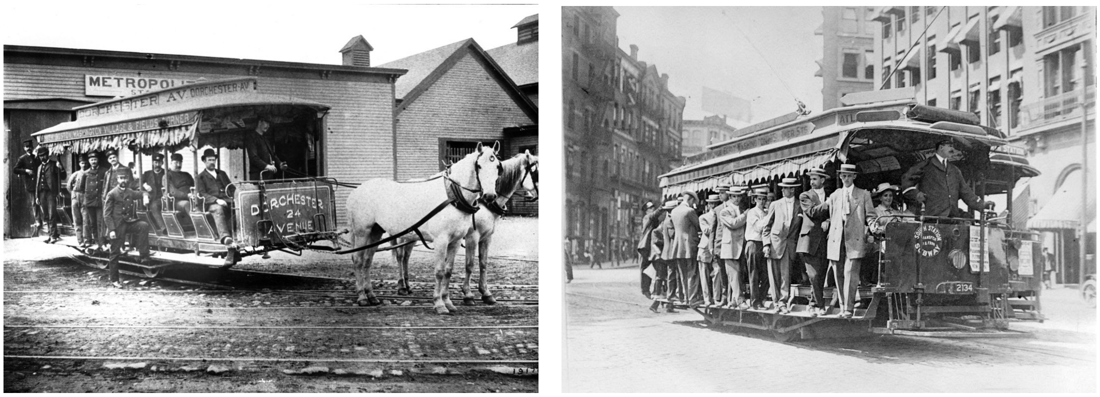
The features of the *Boston Directories* allow me to link firms over time with high accuracy. For a random sample of firms from the business directories, which contain firm names and addresses, I first match them to the main directories, which contain the firms' names, firms' addresses, and the owners' residences. This step provides me with additional information for matching. I then distinguish two types of tracking over time: one for sole proprietorships, i.e., the businesses under individual names; and the other for partnerships or companies, most of which had multiple owners. I match sole proprietorships in two different years by name and by either occupation/product or the owner's residence. Figure 1.16 provides two examples. For partnerships or companies, I make use of the information on the owners. If the business name did not change, then I match two businesses in two years by name and by product. If the business name changed, I then track their owners. If one of the owners remained in business, I consider the original business and the owner as the same entity and a matched case. Figure 1.17 provides an example. In this example, I consider "Billings Bros" in 1890 and "Billings David L. & Co" in 1895 as a matched case.

There are special cases. If there are two entries from two years such that only their names match, but neither their occupations nor their owner's residences match, I treat those two entries as unmatched, although there is a positive probability that the person changed both his/her occupation and residence. Such cases account for 5.8% of all the cases. Another possibility is that there could be multiple entries from two years, which are matched by firm name and product/address. Here, I classify them as unmatched due to the lack of information. Such cases account for 3.3% of all the cases.

1.13.3 World Bank Enterprise Survey

An *Enterprise Survey* collected by the World Bank is a firm-level survey of a representative sample of an economys private sector. The surveys cover a broad range of business environment topics including access to finance, corruption, infrastructure, crime, competition, and performance measures. I use these surveys to calculate a measure of land use intensity in production technology by industry. To be comparable with the level of economic development in Boston between 1885 and 1905, I restrict them to country-years where the GDP per capita (constant 2010 US$) is between $2,000$ and $8,000$. I also restrict them to countries with a population of over 10,000,000 in 2010, which, in the *Enterprise Survey*, typically have a sample of more than 500 firms. Finally, I restrict to the surveys conducted after 2012, which consistently asks a question across countries that are relevant to measure the land-use intensity in production: “Hypothetically, if this establishment were to purchase the assets it uses now, in their current condition, how much would they cost?” One answer calls for information on “Machinery, vehicles, and equipment”, and the other calls for information on “Land and buildings.” I use the formula in Equation 1.4 to measure the land-use intensity in production.

1.14 Appendix 4: Figures for Illustration

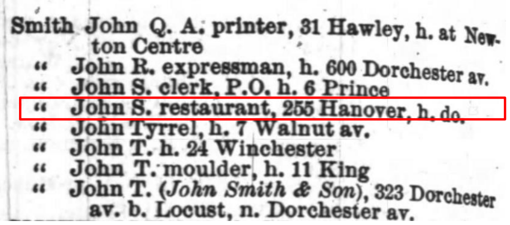


Before: Horse-drawn Streetcars. Speed: 4-5 mph.

After: Electric Streetcars. Speed: 8-10 mph.

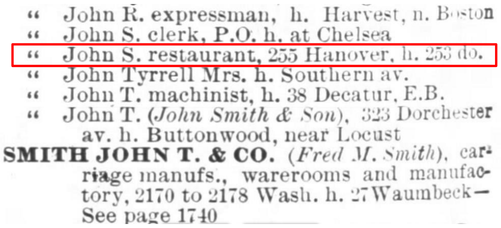
Figure 1.15: The upgrade of streetcars in the 1890's

1885 Boston Directory



Smith John Q. A. printer, 31 Hawley, h. at Newton Centre
" John R. expressman, h. 600 Dorchester av.
" John S. clerk, P.O. h. 6 Prince
" John S. restaurant, 255 Hanover, h. do.
" John Tyrrel, h. 7 Walnut av.
" John T. h. 24 Winchester
" John T. moulder, h. 11 King
" John T. (*John Smith & Son*), 323 Dorchester av. b. Locust, n. Dorchester av.

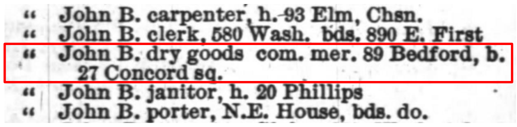
1890 Boston Directory



" John R. expressman, h. Harvest, n. Boston
" John S. clerk, P.O. h. at Chelsea
" John S. restaurant, 255 Hanover, h. 253 do.
" John Tyrrell Mrs. h. Southern av.
" John T. machinist, h. 38 Decatur, E.B.
" John T. (*John Smith & Son*), 323 Dorchester av. h. Buttonwood, near Locust
SMITH JOHN T. & CO. (*Fred M. Smith*), carriage manufs., warerooms and manufactory, 2170 to 2178 Wash. h. 27 Waumbeck—See page 1740

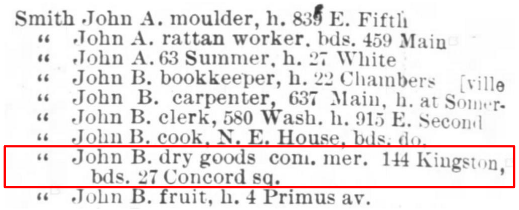
Stayer

1885 Boston Directory



" John B. carpenter, h. 93 Elm, Chsn.
" John B. clerk, 580 Wash. bds. 890 E. First
" John B. dry goods com. mer. 89 Bedford, b. 27 Concord sq.
" John B. janitor, h. 20 Phillips
" John B. porter, N.E. House, bds. do.

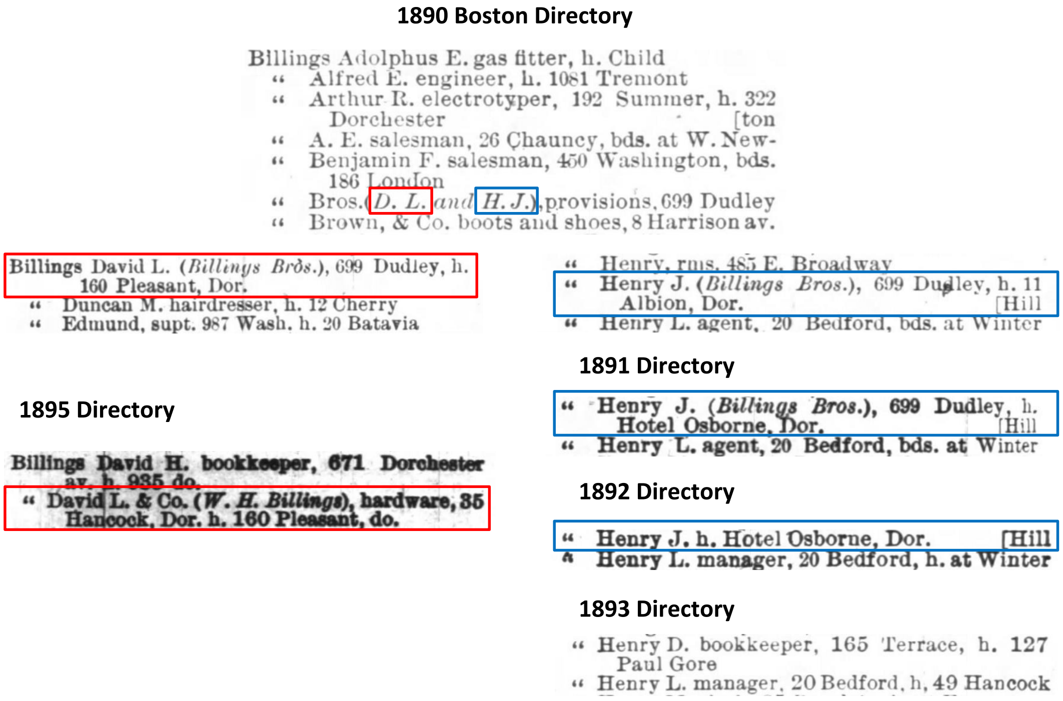
1890 Boston Directory



Smith John A. moulder, h. 839 E. Fifth
" John A. rattan worker, bds. 459 Main
" John A. 63 Summer, h. 27 White
" John B. bookkeeper, h. 22 Chambers [ville
" John B. carpenter, 637 Main, h. at Somer-
" John B. clerk, 580 Wash. h. 915 E. Second
" John B. cook, N. E. House, bds. do.
" John B. dry goods com. mer. 144 Kingston, bds. 27 Concord sq.
" John B. fruit, h. 4 Primus av.

Mover

Figure 1.16: Tracking Sole Proprietorships



1890 Boston Directory

Billings Adolphus E. gas fitter, h. Child
" Alfred E. engineer, h. 1081 Tremont
" Arthur R. electrotyper, 192 Summer, h. 322 Dorchester
" A. E. salesman, 26 Chauncy, bds. at W. Newton
" Benjamin F. salesman, 450 Washington, bds. 186 London
" Bros.(*D. L. and H. J.*), provisions, 699 Dudley
" Brown, & Co. boots and shoes, 8 Harrison av.

Billings David L. (*Billings Bros.*), 699 Dudley, h. 160 Pleasant, Dor.
" Duncan M. hairdresser, h. 12 Cherry
" Edmund, supt. 987 Wash. h. 20 Batavia

" Henry, rms. 485 E. Broadway
" Henry J. (*Billings Bros.*), 699 Dudley, h. 11 Albion, Dor. [Hill
" Henry L. agent, 20 Bedford, bds. at Winter

1891 Directory

" Henry J. (*Billings Bros.*), 699 Dudley, h. Hotel Osborne, Dor. [Hill
" Henry L. agent, 20 Bedford, bds. at Winter

1892 Directory

" Henry J. h. Hotel Osborne, Dor. [Hill
" Henry L. manager, 20 Bedford, h. at Winter

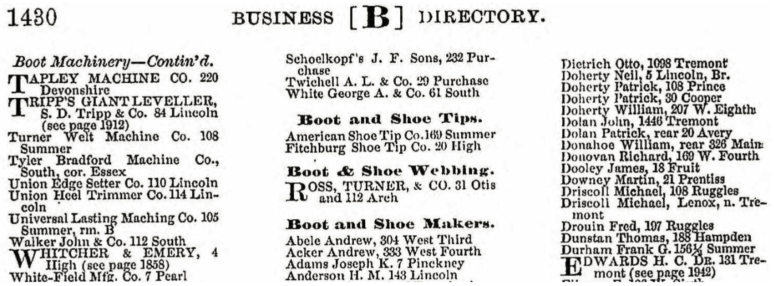
1893 Directory

" Henry D. bookkeeper, 165 Terrace, h. 127 Paul Gore
" Henry L. manager, 20 Bedford, h. 49 Hancock

1895 Directory

Billings David H. bookkeeper, 671 Dorchester av. h. 935 do.
" David L. & Co. (*W. H. Billings*), hardware, 35 Hancock, Dor. h. 160 Pleasant, do.

Figure 1.17: Tracking by the Owners of Firms



1430 BUSINESS [B] DIRECTORY.

Boot Machinery—Contin'd.

TAPLEY MACHINE CO. 220 Devonshire
TRIPP'S GIANT LEVELLER, S. D. Tripp & Co. 84 Lincoln (see page 1912)
Turner Welt Machine Co. 108 Summer
Tyler Bradford Machine Co., South, cor. Essex
Union Edge Setter Co. 110 Lincoln
Union Heel Trimmer Co. 114 Lincoln
Universal Lasting Maching Co. 105 Summer, rm. B
Walker John & Co. 112 South
WHITCHER & EMERY, 4 High (see page 1858)
White-Field Mfg. Co. 7 Pearl

Schoelkopf's J. F. Sons, 232 Purchase
Twichell A. L. & Co. 29 Purchase
White George A. & Co. 61 South

Boot and Shoe Tips.

American Shoe Tip Co. 169 Summer
Fitchburg Shoe Tip Co. 20 High

Boot & Shoe Webbing.

ROSS, TURNER, & CO. 31 Otis and 112 Arch

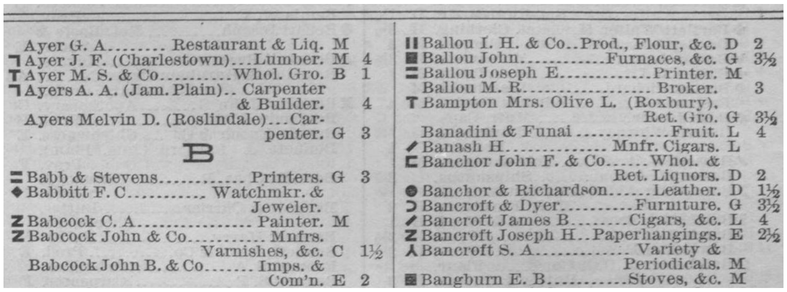
Boot and Shoe Makers.

Abele Andrew, 304 West Third
Acker Andrew, 333 West Fourth
Adams Joseph K. 7 Pinckney
Anderson H. M. 143 Lincoln

Dietrich Otto, 1098 Tremont
Doherty Neil, 5 Lincoln, Br.
Doherty Patrick, 108 Prince
Doherty Patrick, 30 Cooper
Doherty William, 207 W. Eighth
Dolan John, 1446 Tremont
Dolan Patrick, rear 20 Avery
Donahoe William, rear 326 Main
Donovan Richard, 169 W. Fourth
Dooley James, 18 Fruit
Downey Martin, 21 Prentiss
Driscoll Michael, 108 Ruggles
Driscoll Michael, Lenox, n. Tremont
Drouin Fred, 197 Ruggles
Dunstan Thomas, 188 Hampden
Durham Frank G. 156½ Summer
EDWARDS H. C. DR. 131 Tremont (see page 1942)

The above image is a sample page of the business directory section of the *Boston Directories*. For each volume (published yearly), the business directory selects all the business units from the main directory and sorts them according to product category.

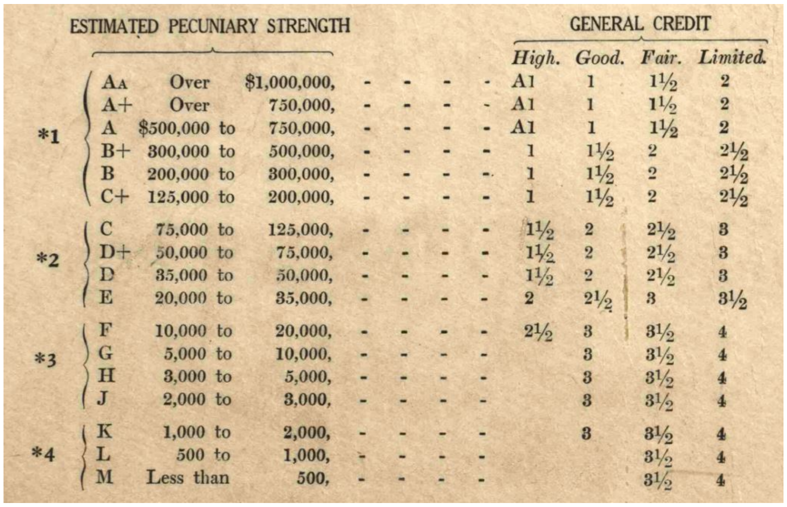
Figure 1.18: A Sample Page of the *Boston Directory* 1890, Business Directory Section



	Ayer G. A.........	Restaurant & Liq.	M	
┐	Ayer J. F. (Charlestown)...	Lumber.	M	4
T	Ayer M. S. & Co...........	Whol. Gro.	B	1
┐	Ayers A. A. (Jam. Plain)..	Carpenter & Builder.		4
	Ayers Melvin D. (Roslindale)...	Carpenter.	G	3
	B			
=	Babb & Stevens.............	Printers.	G	3
◆	Babbitt F. C............	Watchmkr. & Jeweler.		
Z	Babcock C. A..................	Painter.	M	
Z	Babcock John & Co...........	Mnfrs. Varnishes, &c.	C	1½
	Babcock John B. & Co.......	Imps. & Com'n.	E	2
II	Ballou I. H. & Co..	Prod., Flour, &c.	D	2
■	Ballou John.............	Furnaces, &c.	G	3½
=	Ballou Joseph E..............	Printer.	M	
	Ballou M. R....................	Broker.		3
T	Bampton Mrs. Olive L. (Roxbury).	Ret. Gro.	G	3½
	Banadini & Funai	Fruit.	L	4
/	Banash H................	Mnfr. Cigars.	L	
⊏	Banchor John F. & Co......	Whol. & Ret. Liquors.	D	2
●	Banchor & Richardson......	Leather.	D	1½
⊃	Bancroft & Dyer...........	Furniture.	G	3½
/	Bancroft James B........	Cigars, &c.	L	4
Z	Bancroft Joseph H..	Paperhangings.	E	2½
⅄	Bancroft S. A..............	Variety & Periodicals.	M	
■	Bangburn E. B...........	Stoves, &c.	M	

The letter on the right of each business indicates the rating of its pecuniary strength, and the numbers next to the letter is the rating of its credit risk. The rating key is illustrated in Figure 1.20.

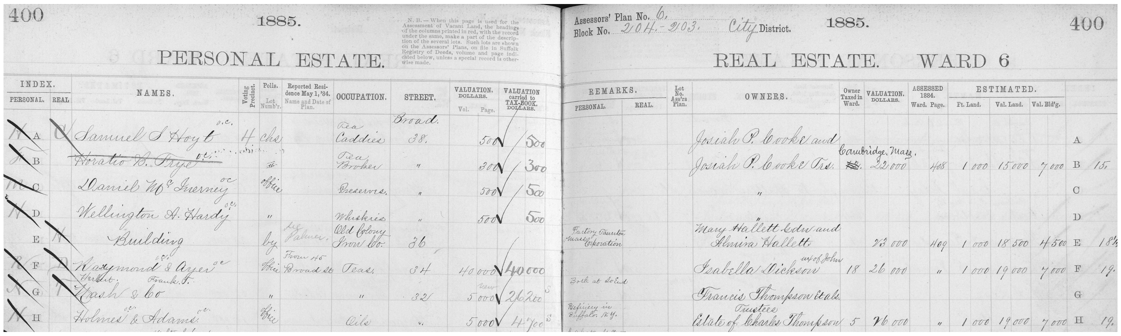
Figure 1.19: A Sample Page of the R.G. Dun & Co's Credit Rating Reference Book, September 1885.



	ESTIMATED PECUNIARY STRENGTH				GENERAL CREDIT			
					High.	*Good.*	*Fair.*	*Limited.*
*1	AA	Over		$1,000,000,	A1	1	1½	2
	A+	Over		750,000,	A1	1	1½	2
	A	$500,000	to	750,000,	A1	1	1½	2
	B+	300,000	to	500,000,	1	1½	2	2½
	B	200,000	to	300,000,	1	1½	2	2½
	C+	125,000	to	200,000,	1	1½	2	2½
*2	C	75,000	to	125,000,	1½	2	2½	3
	D+	50,000	to	75,000,	1½	2	2½	3
	D	35,000	to	50,000,	1½	2	2½	3
	E	20,000	to	35,000,	2	2½	3	3½
*3	F	10,000	to	20,000,	2½	3	3½	4
	G	5,000	to	10,000,		3	3½	4
	H	3,000	to	5,000,		3	3½	4
	J	2,000	to	3,000,		3	3½	4
*4	K	1,000	to	2,000,		3	3½	4
	L	500	to	1,000,			3½	4
	M	Less than		500,			3½	4

Scanned image taken from Sarada and Ziebarth (2015). This rating key applies to the whole study period, 1885-1905.

Figure 1.20: Rating Key of D & B



400 1885.

N. B.—When this page is used for the Assessment of Vacant Land, the headings of the columns printed in red, with the record under the same, make a part of the description of the several lots. Such lots are shown on the Assessors' Plans, on file in Suffolk Registry of Deeds, volume and page indicated below, unless a special record is otherwise made.

PERSONAL ESTATE.

INDEX.		NAMES.	Voting Precinct.	Polls.	Reported Residence May 1, '84.	OCCUPATION.	STREET.	VALUATION. DOLLARS.		VALUATION carried to TAX-BOOK. DOLLARS.
PERSONAL.	REAL			Lot Numb'r.	Name and Date of Plan.			Vol.	Page.	
H A	C	Samuel S Hoyt o.c.	4	chs		Tea Caddies	Broad. 38.		500	500
F B		~~Horatio B. Frye o.c.~~		do		Tea Broker	″		300	300
M C		Daniel Mc Inerney oc		office		Preserves.	″		500	500
H D		Wellington A. Hardy o.c.		″		Whiskies	″		500	500
E	H	Building		by	see Palmer.	Old Colony Iron Co.	36			
R F	D	Raymond o.c. & Ayer o.c		office	From 45 Broad st	Teas.	34		40 000	40 000
N G	T	Herbert. Frank. F. Nash & Co		″		″	32		new 5 000	26 200 S
H H		Holmes o.c. & Adams o.c.		office		Oils	″		5 000	4 700 S

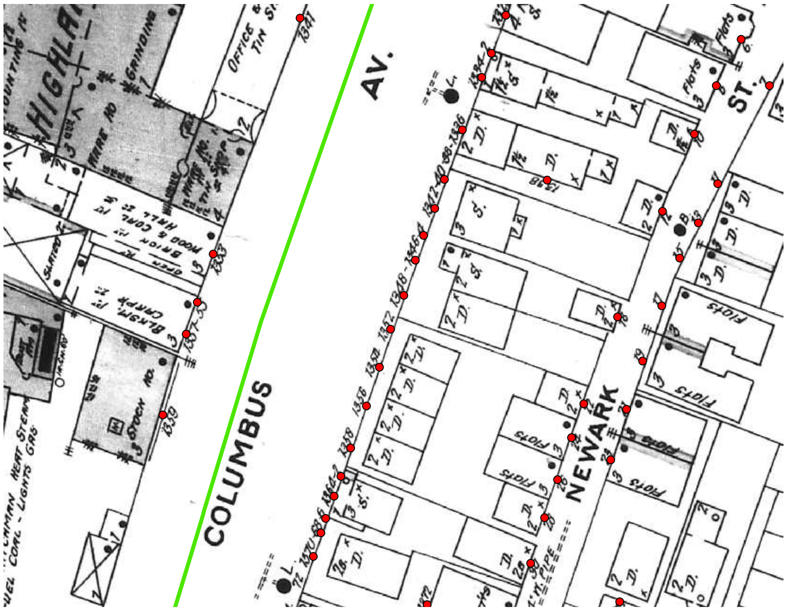
Assessors' Plan No. 6.
Block No. 204.-203. City District. 1885. 400

REAL ESTATE. WARD 6

REMARKS.		Lot No. Ass'rs Plan.	OWNERS.	Owner Taxed in Ward.	VALUATION. DOLLARS.	ASSESSED 1884.		ESTIMATED.				
PERSONAL.	REAL.					Ward.	Page.	Ft. Land.	Val. Land.	Val. Bld'g.		
			Josiah P. Cooke and								A	
			Josiah P. Cooke Trs.	Cambridge, Mass.	22 000		408	1 000	15 000	7 000	B	15.
			″								C	
			″								D	
Factory Taunton Mass. Corporation			Mary Hallett Gdn and Almira Hallett		23 000		409	1 000	18 500	4 500	E	18½
			Isabella Dickson wf of John	18	26 000		″	1 000	19 000	7 000	F	19.
Both at So End			Francis Thompson et als. Trustees								G	
Refinery in Buffalo. N.Y.			Estate of Charles Thompson	5	26 000		″	1 000	19 000	7 000	H	19.

The above image is a sample page of the 1885 Boston Property Tax ledgers. I digitzed a 10% random sample of such pages in the years 1885 and 1898.

Figure 1.21: Boston Property Tax Ledgers



Plot-maps, such as the one above, are georeferenced to the 1930 Boston street centerline map. The red dots correspond to each building/address. The green line represents a portion of the streetcar line.

Figure 1.22: Sanborn Fire Insurance Map with Geo-located Points

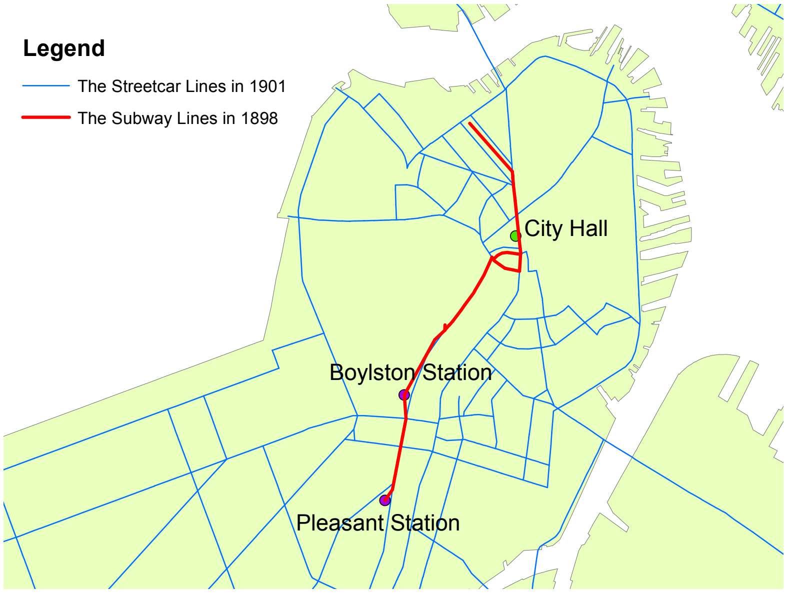


Figure 1.23: The Locations of the Subway Stations in 1898

Chapter 2

Locking in Immigrant Enclaves? Quasi-Experimental Evidence from Late-Nineteenth Century Boston

2.1 Abstract

Immigrants tend to locate in ethnic enclaves within metropolitan areas in the US. Hypotheses and evidence suggest that residence in ethnic enclaves can have either positive or negative consequences for immigrants, implying that the existence of ethnic enclaves in spatial equilibrium may reflect a lack of information on outside communities for enclave residents. This paper examines within-city migration decisions of immigrants in response to a transport shock that Boston quickly electrified its previous horse-drawn streetcar system during 1889-1896. Analyzing new individual-level data merged between city directories and decennial census in Boston from 1885 to 1900, I find that immigrant enclave residents who worked within 25 meters of streetcar rails in 1885 were much more likely to move to a less segregated neighborhood in 1900, compared to enclave residents

who worked between 25 and 50 meters away from streetcar rails. The mechanisms are consistent with interactions in workplace having an impact on choices of residential locations fifteen years after.

2.2 Introduction

Immigrants tend to cluster within their own communities in the US (Cutler, Glaeser, and Vigdor, 2008a). This phenomenon is important because residential segregation of immigrants is commonly believed to hamper integration of immigrants, and has drawn the attention of many policy makers. It then begs the question why immigrants stay in or leave ethnic enclaves.

Evidence and hypotheses suggest that residence in ethnic enclaves can have either positive or negative consequences for immigrants. On the one hand, Edin et al (2003, 2004), Cutler, Glaeser, and Vigdor (2008b), Damm (2009, 2014) show that residence in enclaves has positive effects on immigrants' labor market outcomes in terms of earnings and employment rate. On the other hand, Chiswick (1991) and Lazear (1999) hypothesize that living in an ethnic enclave hampers economic assimilation of immigrants by decreasing the rate of acquisition of host countryspecific human capital (e.g., language) due to reduced social interaction with natives, which can result in adverse labor market outcomes. These ambiguous effects of enclaves suggest that knowing the real effects of moving out of an enclave is very difficult for a particular immigrant. This difficulty is exacerbated by enclaves per se: the spatial concentration of immigrants mechanically results in more physical segregation between immigrants and natives, which increases communication costs between immigrants and natives, and impedes immigrants to learn the potential benefits and costs they might experience if they move out of their enclave. Therefore, immigrants can be "locked-in" enclaves as a spatial equilibrium outcome. It is possible that, with an exogenous shock that significantly reduces the communication costs between neighborhoods, immigrants can learn more information about outside communities, and the ensuing spatial resorting process will result in a new spatial equilibrium with less segregation. The possibility of this “lock-in” effect for ethnic enclaves, however, has not been formally tested in literature.

In this paper, I test the “lock-in” effect for ethnic enclaves, examining the within-city migration decisions of immigrants in response to a sharp upgrade in public transportation infrastructure in late nineteenth century Boston. During 1889-1896, Boston quickly electrified its previous horse-drawn streetcar system. This upgrade increased the speed of the best means of intra-city transportation from 4-5 mph to 8-10 mph, tripled transportation capacity, and enabled services to be provided at lower fares (Warner, 1962). This context is ideal for my study purpose. First, prior to this first generation of modern transit mode, Boston relied on very primitive transport modes: horse trolleys and walking. Contemporary observers describe neighborhoods in Boston as being highly self-contained (For example, Warner, 1962). [1] The electrification of the streetcar system is a major event that integrated neighborhoods in the Boston history. Second, Boston in this period featured highly diverse populations. In 1900, 50% of the city population were first generation immigrants, who came from 38 different foreign countries, and 72.61% were first or second generation immigrants. Moreover, different immigrant groups clustered in different neighborhoods in Boston, as shown in Figure 2.4 in Appendix. The semi-autarkic nature together with the geographic segregation of immigrant neighborhoods imply that information barriers can be very high for immigrants at this time period. Last but not least, this streetcar electrification event was unexpected at least before 1885. By 1885, there still existed a number of decentralized horse-drawn streetcar

[1] Warner (1962) discussed three examples - Dorchester, Roxbury, and West Roxbury - as previously semi-closed communities, which soon became integrated with central Boston after the advances in streetcar technology.

service companies in Boston. A major innovation in streetcar technology was infeasible until a monopolist company - the West End Street Railway Company - consolidated all the other companies in 1888. Moreover, the research and development progressed rapidly and solved major technical problems by 1891, which, from available sources, can hardly be predicted by Bostonians in 1885. Therefore, we avoid the possibility that residents might migrate across neighborhoods in anticipation of future transport infrastructure events. A detailed description of the historical background of this event can be found in Section 1.3 in Chapter 1.

To evaluate the impact of this event, I assemble a novel individual-level data set, merged from the decennial censuses and the Boston city directories, which allows me to trace the exact places of work and places of residence for a random sample of around 2,000 individuals in this city over 1885-1900. I also digitize streetcar rails maps, which allow me to calculate the distance between each individal and the nearest streetcar route to measure the intensity of shocks associated with the streetcar electrification.

The outcome variable of interest is changes in segregation index of neighborhood of residence between 1885 and 1900 for these individuals. A decrease in this index suggests that this individual moved to a less segregated neighborhood. To measure the intensity of the transport infrastructure shock, I construct a treatment dummy based on proximity to existing streetcar rails. In particular, I compares within-city migration outcomes of people whose place of work/residence was within 25m of the streetcar rails [2] (treatment group) to those between 25m and 50m away from the rails (control group). [3] There are good reasons to believe that such a short distance threshold can make a big difference in access to outside communities. First, the areas within 25m of the streetcar rails experienced a much faster increase in employment density than areas between 25m

[2] along the rail-connected streets

[3] Because the residents and employment were highly gravitated toward the streetcar rails, this narrow bandwidth keep as many as 88%/45% (calculated using place of work/residence) observations.

and 50m away from the rails, as shown in Figure 1.5. Second, in my job market paper, I show that the same event caused replacement of sole proprietorships by relatively large companies and partnerships, and this effect decay extremely quickly as one moves away from the streetcar rails, where 25m is the kink.

My baseline result shows that improved transit access has very heterogeneous effects on different individuals. There is no treatment effect for individuals who initially resided in well-integrated communities (segregation index=0); there is no treatment effect for natives either; there is also no treatment effect for treatment defined on residence proximity to the streetcar rails. However, there is strong and significant treatment effect for immigrants who initially resided in "enclaves" and *worked* near the rails: the estimated coefficient suggests that for immigrants whose neighbors were entirely their conationals in 1885, after the streetcar electrification shock in their workplace, those in the treatment group moved to a new location in 1900 where they were exposed to 14 percent point less conationals than those in the control group.

I then examine the heterogeneity of the treatment effect for these initial immigrant "enclave" residents. I find that the effect is stronger for: the first generation immigrants than for the second generations; immigrants who engaged in low-income occupations and low-education occupations; renters and boarders than house owners. These results are consistent with these disadvantaged individuals being the least informed, and therefore the most responsive to transport shocks.

Further analysis of mechanisms reveals that more interactions with people from non-local communities matter. For each workplace, I construct a measure of openness, which is proportional to residential areas covered by all the workers who worked there. A higher value of this openness measure indicates that the employment at this workplace come from diverse communities, and therefore everyone potentially has better access to information on different communities. I find that the treatment effect is larger for immigrant enclave residents who worked at more “open” places than those who worked at less “open” places. To further shed light on this mechanism, I examine whether collegiality relationship (in terms of proximity in workplace) in 1885 predicts co-residence in 1900, using individual pair information. I find that for immigrant pairs (both first and second generation) who worked together in 1885 (within 200m of each other), they were 3 percent point more likely to live together (within 500m of each other) in 1900. This effect is statistically significant and economically large - it doubles the likelihood of living together in 1900 compared to a random pair of immigrants. This effect does not exist among the natives.

This paper contributes to three bodies of literature. First, it adds to a greater understanding of the causes of immigrant enclaves. Edin et al. (2003) and Damm (2009) provide quasi-experimental evidence on the return to living in an ethnic enclave, exploiting spatial dispersal policies on refugees in Sweden and Denmark, respectively. Cutler, Glaeser, and Vigdor (2008b) use instrumental variable approach to estimate the impact of ethnic concentration on immigrant outcomes. All these three papers find a positive effect of enclave size on labor market outcomes. The estimated positive effects, on the one hand, can explain why immigrants form residential enclaves; on the other hand, imply that similar spatial dispersal policies on immigrants might be at the costs of immigrants’ welfare. The results of my paper paint another picture of immigrant enclaves, where immigrant residents lack information about outside communities, and their status quo choice of residence might not be optimal. Spatial integration of neighborhoods can significantly induce voluntary out-migration of enclave residents. The policy implication is that the government can reduce concentration of immigrants without hampering economic benefits of immigrants. Improvement in transport infrastructure is one possible solution.

Second, this paper is related to the literature on immigrant assimilation in the age of mass migration in the US. Previous papers have studied the assimilation process in this period focusing on wage convergence between immigrants and natives (Hanes 1996; Hatton, 1997; Minns, 2000; Abramitzky, Boustan, and Eriksson, 2012, 2014). Recently, Abramitzky, Boustan, and Eriksson (2016) and Carneiro, Lee, and Reis (2016) study cultural assimilation using adoption of an American name as a proxy. My paper complements this strand of literature by looking at a different indicator of assimilation - out-migration from immigrant enclaves - using individual panel data with detail geographic information. The unique historical event I exploit - the electrification of streetcars - helps us identify spatial integration as an important cause of out-migration of immigrants from enclaves at the individual level. The results need to be interpreted with caution though, because this paper does not take into account general equilibrium effects of this event, and therefore cannot determine whether or not this event caused an increase or decrease in the *overall* segregation of immigrants.

Finally, the results of this paper are related to the finding in Cutler, Glaeser, and Vigdor (2008a), that metropolitan areas where public transit is a viable commuting option have witnessed higher increases in segregation of immigrants over the latter half of the 20th century in the US. The reasons why this is true are still not well-understood. While the study context of my paper is very different from today, my results suggest that immigrants as a group are particularly responsive to public infrastructure shocks compared to natives. Therefore, studying the impact of a public transportation infrastructure construction/upgrade event on immigrants in a modern context could be a fruitful direction for future research.

The remainder of this chapter is organized as follows. Section 2.3 describes the data. Section 2.4 provides summary statistics. Section 2.5 introduces my empirical strategy and presents the benchmark regression results. Section 2.6 examines heterogeneity of the treatment effect and mechanisms underlying the treatment effect. Lastly, in Section 2.7, I conclude.

2.3 Data

In this section, I provide details on the data used, describe its sources, and construction of key variables.

2.3.1 Measurement of Immigrant Concentration and Census Enumeration District-Level Data

I construct a measure of spatial concentration/segregation for each immigrant group at the neighborhood level, and then use this neighborhood-level measure to construct a individual-level measure of changes in neighborhood environment. Let j index neighborhoods, k index immigrant groups. The over concentration of an immigrant group k in j is measured by

$$E_j^k \equiv Share_j^k - Share^k \tag{2.1}$$

Here, $Share_j^k$ is the share of group k immigrants in neighborhood j population, and $Share^k$ is the share of group k immigrants in city population. The difference between these two values corresponds to the over-representativeness of immigrant group k in neighborhood j. If the share of group k immigrants in neighborhood j is identical to its share in city population, this measure is then 0. For a small immigrant group whose share in city population is close to 0, this measure is bounded above at 1, indicating that neighborhood j consists of immigrants only from this group. This measure of immigrant concentration is in spirit the same as Bayer, McMillan, and Rueben (2004)'s measure of racial segregation.

I use

$$\Delta E_i^k \equiv E^k_{i,j_{t1}} - E^k_{i,j_{t0}} \tag{2.2}$$

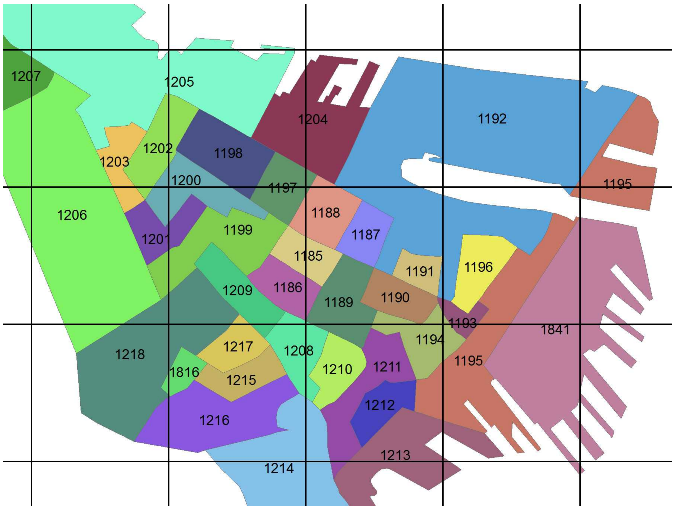
to measure changes in neighborhood environment for an individual i of group k, where $j_{t0}(j_{t1})$ indicates the neighborhood of residence of individual i at the beginning (end) of the study period. A decrease in this measure reflects that individual i moved to a place where she was exposed to a smaller share of conationals between t_0 and t_1 (adjusted by changes in the overall share of group k immigrants).

To construct E_j^k and ΔE_i^k, we need individual-level census data with detailed neighborhood identifier. By the U.S. government's "72-Year Rule," personally identifiable information about an individual can be accessed by researchers 72 years after it was collected for the decennial census. This means that much richer information is available to public for the decennial censuses prior to 1940. The finest geographic identifier for these years of censuses is the "Enumeration District" (ED). While for most cities the boundaries of ED's can only be found on maps in paper format, and therefore are not usable for analysis, Shertzer, Walsh, and Logan (2015) have recently digitized the boundaries of ED's for 10 cities, which allow us to geolocate individuals at a much higher accuracy for these 10 cities. I use the Boston ED shapefiles in 1880 and 1900 in my analysis. [4][5] In my study sample (Boston in 1880 and 1900), a median ED contains 1,700 persons, and is $280m \times 280m$ in size. In contrast, in today's public census data, the finest geographic identifier is "PUMA," which, on average, contains 100,000 persons.

A major difficulty in making use of mapped data over time is that the boundaries of EDs shift from decade to decade. To address this problem, I harmonize ED data to temporally invariant geographically defined areas ("synthetic neighborhoods"), using the same method as in Banzhaf and Walsh (2008) and Shertzer, Walsh, and Logan (2015).

[4] I thank these three authors for kindly sharing these GIS boundary files of ED's with me.

[5] The 1890 Census data was unfortunately destroyed in a fire.



Location: Charlestown, Boston. Grid size: $500m \times 500m$

Figure 2.1: Illustration of a 500m-grid layer, showing the relationship between these grids and enumeration districts in 1900

I utilize 500m grids as "synthetic neighborhoods" in the main empirical analysis, and also report results using 1km grids. The demographic composition of these synthetic neighborhoods is then imputed as the spatially weighted average of the underlying ED level data from each Census. These areas are spatially invariant over time. This method is illustrated in Figure 2.1, using the 1900 ED's in Charlestown as an example.

2.3.2 Boston Directory

I digitize data from the *Boston Directories* published by the Sampson, Murdock, & Company, and printed annually. Each of these volumes lists the names of the inhabitants and firms, their occupations/products, and the places of the business and dwelling houses. Generally, the inhabitants in the directories were in the labor force. I describe the *Boston Directories* in more details in my job market paper.

The advantages of these data over the Censuses include: 1. They contain both residential addresses and workplace addresses. The interactions with neighbors in these two types of places could be very different in nature, so the treatments defined on proximity to place of work and to place of residence could yield very different results. The Census data do not contain workplace information. 2. The addresses can be used to precisely geolocate each individual, while the Censuses do not have such detailed information either. 3. The records are available annually and allows for temporal linking. The 15-year linking rate using the Boston Directories is 50%, while the rate using decennial census records is typically below 30% over 20 years.

I obtain a random sample 3,300 individuals in Boston in the 1900 census data (5%, male, household heads, aged between 30 and 50 in 1900). I then perform a manual match between the census and the 1900 Boston Directory by individual name, occupation, and residence. The matching rate is 65%. Then, from these matched 2,100 individuals in 1900, I trace them back to the 1885 Boston Directory, obtaining their residence and workplace in 1885. The overtime matching rate is around 50%. I choose 1885 as the beginning year because at that time, the electrification of the streetcar system was almost unpredictable, and therefore we can avoid the possibility that people sorted in anticipation of this event. A more detailed description of the linking procedure is in Appendix.

2.3.3 Streetcar Routes

I obtain digital city maps of Boston in 1888 and 1901 from the online David Rumsey Historical Map Collection, which contain streetcar routes and legible street names. I then georeference the two maps such that the points of each of the two city maps are geographically aligned with a common 1930 street centerline shapefile, which I retrieve from the Historical Urban Ecological data set, created by the Center for Population Economics. By overlaying the street centerline shapefile with the georeferenced city maps, I extract the portions of the streets that coincide with the streetcar routes on the city maps and digitize them into new shapefiles. The routes of the two years are shown in Figure 1.3. I then use the streetcar route shapefiles to calculate the distance between each individual's place of work/residence and the nearest streetcar line.

2.3.4 Plot-Level City Maps

The key to combining the digitized streetcar routes data and the individual-level data is to geocode the addresses of on the *Boston Directories* using contemporaneous city maps. I georeference 1,660 plot-level *Sanborn Fire Insurance Maps* of Boston published during the period 1895-1900, which, altogether, cover the entire Boston area. I then manually extract the street name and number of every building on the maps to a GIS shapefile, generating a point shapefile of 100,743 buildings (Figure 1.22 in Appendix shows a sample map). The geographic coordinates of each building are calculated in ArcGIS and then matched to the addresses in the *Boston Directories* by street name and number. [6] For all of the addresses in this study, 90% of them can be geocoded.

The empirical analysis benefits in two ways from geocoding the addresses in the *Boston Directories*. First, I can calculate the distance between each individual's place of work/residence and the nearest streetcar rails at very high accuracy. This distance allowed me to define whether an individual is treated or not based on very narrow thresholds. Second, I can locate where each individual lived, and find the corresponding concentration/segregation index of her neighborhood of residence, and track her relocation patterns over time.

[6] For special addresses, such as "Street A corner Street B," I manually located them on the georeferenced maps.

2.4 Summary Statistics

2.4.1 Distribution of Residential Population and Employment

How did the distribution of the residential population and employment change in the study period? Using the coordinates of both the residential places and the commercial places for the 1% representative sample of the individuals who commuted in the *Boston Directories*,[7] which covered all occupations and industries, I plot the distributions of both residential population and employment by distance to City Hall and their distance to the streetcar rails in 1885 and 1900, seen in Figure 1.5. We find that the spatial patterns of employment growth and residential population growth are quite different. From the first row of Figure 1.5, the majority of employment growth took place in the city center and near the streetcar rails. In contrast, we see from the last row of Figure 1.5 that the residential population density increased primarily in the periphery of the city and further away from the streetcar rails. This contrast suggests that it is important to distinguish between treatment defined on proximity between residence and the rails and between workplace and the rails.

Table 1.3 reports the evolution of the commuting distances of these individuals; the commuting distances are calculated as the distance between their residential addresses and their commercial addresses. The median commuting distance increased from 2.2 km in 1885 to 4 km in 1900, indicating that after the electrification of the streetcar, the population was more mobile. These facts provide quantitative evidence for Warner's (1962) observation that Boston saw the emergence of "streetcar suburbs" in this period, in the sense that more and more people began to live in the suburbs and commute to their

[7] One half of the individuals in the *Boston Directories* had a residential address but no commercial address, and thus, not a commuter. These people probably either worked from home (e.g. as a grocer) or did not have a fixed workplace (e.g., day laborers, peddlers). The fraction of commuters is stable over the study period.

workplace in the Central Business District (CBD, defined as the areas within 1 km of City Hall in Boston).

The facts documented in this section suggests that this event is very suitable for my study purpose: the employment density increased by particularly more near the streetcar rails, and the population became much more mobile than before in my study period. It is a major neighborhood integration event in the Boston history.

2.4.2 Key Explanatory Variables near Treatment Threshold

While the transport shock is sudden and plausibly unexpected, an important assumption is that there is no significant difference in unobservable characteristics between individuals in the treatment and control that drive differential migration decisions between them. While I cannot directly test this assumption, I perform a test whether the values of observables change continuously across the distance cutoff (25m away from the rails, i.e. whether or not along the rails). Passing this test serves as crossing a bar the validity of this assumption.

Figure 2.2 shows that across all relevant observables, we do not see significant differences around the treatment cutoff. The demographic statistics, economic statuses, and initial concentration indexes are all similar. I include all these observables in the main regression analysis.

2.5 Empirical Strategy and Estimation Results

The regression I estimate is

$$\Delta E_i^k = \beta_1 E^k_{i,j_{1885}} + \beta_2 Treatment_{i,j_{1885}} + \beta_3 Treatment_{i,j_{1885}} * E^k_{i,j_{1885}} + \delta_k + X_{i_{1885}} + \varepsilon_i \tag{2.3}$$

Here, i indexes individuals, j indexes neighborhoods, and k indexes immigrant groups (including natives as a group). $E^k_{i,j_{1885}}$ is the segregation index of neighborhood j where individual i of group k lived in 1885, as defined in equation (2.1) in Section 2.3.1. $\Delta E^k_i \equiv E^k_{i,j_{1900}} - E^k_{i,j_{1885}}$ is changes in the segregation index of the neighborhoods where individual i of group k lived between 1885 and 1900, as defined in equation (2.2) in Section 2.3.1. $Treatment_{i,j_1885}$ indicates whether or not individual i lived/worked in a rail-connected location in 1885, where $Treatment_{i,j_1885} = 1$ if the distance to the nearest streetcar rails is 0-25m, $Treatment_{i,j_1885} = 0$ if the distance is 25-50m. δ_k's are immigrant group fixed effects. $X_{i_{1885}}$ is a set of individual characteristics in 1885.

The coefficient β_1 measures whether or not there is mean reversion in the neighborhood environments where individuals lived in. $\beta_1 < 0$ indicates that people who initially lived in a highly segregated neighborhood tended to moved to a less segregated neighborhood, and vice verse. β_2 measures for individuals who initially resided in a well-integrated neighborhood (segregation index $E^k_{i,j_{1885}} = 0$), whether or not there is treatment effect of the streetcar upgrade on the migration outcomes of the treated group. $\beta_2 < 0$ demonstrates that people who initially lived/worked near the rails migrated to a less segregated neighborhood after the transport shock compared to those further away from the rails. β_3 is the main coefficient of interest, which measures for individuals who initially resided in highly segregated neighborhoods (with a high $E^k_{i,j_{1885}}$, called enclave residents thereafter), whether or not those who were close to the rails were more likely to move to a less segregated neighborhood after the treatment. $\beta_3 < 0$ suggests that the streetcar upgrade event induced previous enclave residents to move out of enclaves.

I estimate equation (2.3) for immigrants and natives separately, and for treatment groups defined on both residence proximity and workplace proximity to the rails.

Table 2.1 reports estimation results of equation (2.3) with different controls using different samples. In all the columns, $treatment = 1$ if the distance between *workplace*

Table 2.1: Regression Results: Benchmark

	(1)	(2)	(3)	(4)	(5)
Outcome variable:	Changes in Segregation Index of Residence Neighborhood Between 1885 and 1900				
Samples:	Immigrants	Immigrants	Immigrants	Immigrants	Natives
	1st and 2nd gen.	1st and 2nd gen.	1st and 2nd gen.	1st gen.	
Segregation index 1885	-0.629***	-0.618***	-0.641***	-0.634***	-0.846***
	(0.086)	(0.087)	(0.077)	(0.083)	(0.076)
Treatment	0.015	0.014	0.023*	0.049**	0.008
	(0.013)	(0.013)	(0.013)	(0.019)	(0.020)
Seg index 1885*Treatment	-0.138*	-0.142*	-0.145*	-0.178*	-0.073
	(0.078)	(0.079)	(0.083)	(0.109)	(0.069)
Distance to CBD (residence)		0.005	0.008**	-0.005	0.004
		(0.004)	(0.004)	(0.006)	(0.003)
Distance to CBD (workplace)		-0.001	-0.002	-0.025***	-0.010*
		(0.007)	(0.007)	(0.008)	(0.006)
Married			0.052***	-0.008	-0.026
			(0.014)	(0.014)	(0.023)
Number of Children			0.007*	0.011**	0.006
			(0.004)	(0.005)	(0.005)
Age			-0.003	0.000	-0.001
			(0.007)	(0.010)	(0.008)
Age^2			0.000	-0.000	0.000
			(0.000)	(0.000)	(0.000)
Years in US			-0.003***	-0.002	
			(0.001)	(0.001)	
Group FE	YES	YES	YES	YES	YES
Observations	416	416	416	224	320
R-squared	0.286	0.289	0.345	0.337	0.600

Notes: Standard errors clustered by 500-grid neighborhoods are reported in parentheses: *** indicates statistical significance at the 1% level, ** at the 5% level and * at the 10% level.

and rails is $0-25m$, and $treatment = 0$ if the distance is $25-50m$. Columns (1) - (3) use all the first generation and second generation immigrants in the regression sample, column (4) uses only the first generation immigrants, and column (5) uses only the natives. From columns (1) to (3), we can see that the coefficients for both $Treatment$ and *Segregation index 1885*Treatment* are robust to the addition of control variables. My preferred results are column (3), which include all the controls. The coefficient before $Treatment$ is 0.023, significant at %10 level, suggesting that for individuals who previously lived in a well-integrated neighborhood (segregation index $E^k_{i,j_{1885}} = 0$), the individuals in the treatment group moved to a place where they were exposed to 2.3-percentage-point more conationals 15 years after. On the other hand, the coefficient before *Segregation index 1885*Treatment* is -0.145, also significant at %10 level, suggesting that for individuals who previously lived in a neighborhood consisting only of their conationals (segregation index $E^k_{i,j_{1885}} = 1$), the treated individual moved to a neighborhood with 14.5-percentage-point less conationals. Therefore, the transport shock has an economically large effect on the migration decisions of the enclave residents, but a much smaller effect for people who previously lived outside enclaves. In column (4) of Table 2.1, we can see that these effects are stronger for the first generation immigrants, suggesting that they are the group of people who were subject to more informational constraints. On the other hand, column (5) shows that these effects are much smaller and statistically insignificant for the natives, which is consistent with the interpretation that they were already well-integrated and therefore not responsive to the transport shock.

The baseline regression measures changes by difference in levels. This measure places same weight on going from an immigration concentration share of, say, 80% to 85%, as from 10% to 15%. An alternative measure is log changes, $log(\frac{E^{abs}_{1900}}{E^{abs}_{1885}})$, or changes in proportional terms, $\frac{E^{abs}_{1900}-E^{abs}_{1885}}{0.5(E^{abs}_{1900}+E^{abs}_{1885})}$. These two measures place higher values for changes from 10% to 15% than from 80% to 85%. Another property of the baseline specification

Table 2.2: Robustness Checks: Different Outcome Measures

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable:	$E^{over}_{1900} - E^{over}_{1885}$		$log(\frac{E^{abs}_{1900}}{E^{abs}_{1885}})$		$\frac{E^{abs}_{1900}-E^{abs}_{1885}}{0.5(E^{abs}_{1900}+E^{abs}_{1885})}$	
Exposure index 1885	-0.641***		-2.128***		-2.603***	
	(0.077)		(0.317)		(0.356)	
Exposure index 1885		-0.478***		-2.282***		-2.541***
($E^{over} > 0$, or $E^{abs} > \bar{E}^{abs}$)		(0.121)		(0.248)		(0.233)
Exposure index 1885		-0.836***		-3.727***		-4.468***
($E^{over} < 0$, or $E^{abs} < \bar{E}^{abs}$)		(0.087)		(0.553)		(0.550)
Treatment	0.023*	0.023*	0.238**	0.305***	0.255***	0.284***
	(0.013)	(0.014)	(0.092)	(0.105)	(0.076)	(0.101)
Expo. idx. 1885*Treatment	-0.145*		-0.688**		-0.420*	
	(0.083)		(0.345)		(0.247)	
Expo. idx. 1885*Treatment		-0.205*		-0.756**		-0.397*
($E^{over} > 0$, or $E^{abs} > \bar{E}^{abs}$)		(0.118)		(0.349)		(0.227)
Expo. idx. 1885*Treatment		-0.187		-2.116**		-1.626**
($E^{over} < 0$, or $E^{abs} < \bar{E}^{abs}$)		(0.157)		(0.821)		(0.716)
Observations	416	416	382	382	416	416
R-squared	0.345	0.360	0.210	0.524	0.276	0.583

Notes: Standard errors clustered by 500-grid neighborhoods are reported in parentheses: *** indicates statistical significance at the 1% level, ** at the 5% level and * at the 10% level.
E^{over} indicates the over-exposure index, E^{abs} indicates the exposure index, i.e. without normalization. I use E^{abs} instead of E^{over} in column (3) because E^{over} can be negative, which cannot be taken logarithm.

(equation (2.3)) is that it imposes an restriction that the tendency of regression from above mean to mean is the same as the tendency of regression from below mean to mean. Therefore, a more flexible specification is to allow for asymmetry in these two directions of change:

$$\begin{aligned}\Delta E_i^k =& \beta_1 E^k_{i,j_{1885}} * I\{E_{i,j_{1885}} > \bar{E}\} + \beta_2 E^k_{i,j_{1885}} * I\{E_{i,j_{1885}} < \bar{E}\} + \beta_3 Treatment_{i,j_{1885}} + \\ & \beta_4 Treatment_{i,j_{1885}} * E^k_{i,j_{1885}} * I\{E_{i,j_{1885}} > \bar{E}\} + \\ & \beta_5 Treatment_{i,j_{1885}} * E^k_{i,j_{1885}} * I\{E_{i,j_{1885}} < \bar{E}\} + \delta_k + X_{i_{1885}} + \varepsilon_i\end{aligned} \tag{2.4}$$

where ΔE_i^k can be any of the three definitions of changes, and $\bar{E}$ is 0 if E is the normalized concentration index (equation (2.1)), and is the mean if E is the absolution exposure index, i.e. $Share^k_j$.

Table 2.2 report estimation results using equation (2.3) and (2.4) for each of the three definitions of the outcome. Column (1) replicates the column (3) of Table 2.1. In column (2), we can see that allowing for asymmetry magnifies of the treatment effect (-0.205 and -0.187, compared to -0.145), with similar magnitude in each direction. The results using log differences and proportional changes are reported in columns (3) - (6). From column (3) and (5), we can see that under these different definitions of changes, the treatment effect still exists, and are even more significant. Column (4) and (6) indicate that is asymmetry in the treatment effect: the effect is much large for individuals who were initially in a below average segregated neighborhood than for those who were initially in an above average segregated neighborhood. This is because for the same amount of level changes, these two definitions place higher values for changes starting from a lower base, implicitly assuming that going from an immigration concentration share from 10% to 15% (15% to 10%) is a more significant move than going from 80% to 85% (85% to 80%). In this paper, I focus on results using simple differences as the outcome, not only because they are easier to interpret, but also because the interest of this paper is more about the migration out of previously highly concentrated immigrant enclave.

Table 2.3: Robustness Checks: Different Treatment Definitions

	(1)	(2)	(3)	(4)	(5)
Outcome variable:			$\Delta E_i^k \equiv E^k_{i,j_{1900}} - E^k_{i,j_{1885}}$		
Neighborhood Size:	1km grid	500m grid	500m grid	500m grid	500m grid
Seg index 1885	-0.602***	-0.641***	-0.666***	-0.672***	-0.766***
	(0.083)	(0.077)	(0.078)	(0.093)	(0.057)
Treatment (workplace)	0.001	0.023*	0.017	0.006	
	(0.010)	(0.013)	(0.014)	(0.014)	
Seg. idx. 1885*Treatment (workplace)	-0.071	-0.145*	-0.140*	-0.095	
	(0.084)	(0.083)	(0.082)	(0.092)	
Treatment (residence)					0.009
					(0.016)
Seg. idx. 1885*Treatment (residence)					0.084
					(0.076)
Observations	416	416	432	432	439
R-squared	0.426	0.345	0.336	0.334	0.337

Notes: Standard errors clustered by 500-grid neighborhoods are reported in parentheses: *** indicates statistical significance at the 1% level, ** at the 5% level and * at the 10% level.

Table 2.3 reestimate the baseline specification, using different treatment definitions and outcome variable definitions. Column (2) here replicates the column (3) of Table 2.1, where the outcome variable is calculated using 500m grid neighborhoods. In column (1) of Table 2.3, I recalculate the outcome variable using 1km grid neighborhoods. I find that the signs of the coefficients in column (1) are still the same, but the magnitudes become smaller and not significantly different from 0. This comparison indicates that the effective neighbors with whom immigrants interacted could be highly localized. The baseline regression (column (2)) defines the treatment and control group using the distance thresholds to workplace of $0-25m$ and $25-50m$, respectively. Column (3) uses thresholds of $0-25m$ and $25-100m$, and column (4) uses thresholds of $0-40m$ and $40-100m$. We can see that the coefficient before the interaction term is not sensitive to the control group definitions, but sensitive to the treatment group definitions. This result is consistent with my job market paper, implying that the effect of the transport shock is highly localized near the transit lines.

Finally, in column (5), I estimate the same specification using distance to place of residence as the treatment definition. The coefficient before the interaction term now exhibits the opposite sign, and is insignificant. The contrast in the results based on place of work and place of residence either implies that interactions in these two types of places are very different, or that ownership of houses matters - since 75% of the residents were house owners in my study sample, if the houses appreciated by different degrees between the treatment and control locations, the treatment effect found in column (5) could be due to changes in financial status between the treatment and control. I return to this issue in Section 2.6.

2.6 Heterogeneity and Mechanisms

2.6.1 Heterogeneity

In this section, I consider how the treatment effect varies with the economic statuses of individuals. I construct three dummies: Low Income, Low Education, and No House, which takes the value of 1 if the individual's occupation earnings were below median, occupation average years of education were below median, and did not own a house, respectively, and takes the value of 0 otherwise. I interact each economic status dummy with the 1885 Segregation Index, as well as the interaction term between the 1885 Segregation Index and the treatment dummy. The results are reported in Table 2.4.

Column (1) of Table 2.4 replicates the results in Column (3) of Table 2.1. From the second row of column (2) and the third row of column (3) in Table 2.4, we can see that the interaction between economic status dummy and the 1885 segregation index exhibit positive and significant coefficients. This result indicates that, while there is mean reversion in segregation index, such that 1885 enclave residents tended to move to a less segregated neighborhood in 1900, such a tendency is much weaker for people with low income and low education. Hence, in the absence of a neighborhood integration shock, these disadvantaged people tended to be locked in enclaves.

The coefficients before the triple difference terms of Table 2.4 (i.e. *Segregation Index 1885*Treatment*Economic Status Dummies*) reveal very interesting results. These coefficients in columns (2) to (4) are much larger in magnitude than the coefficient before *Segregation Index 1885*Treatment* in column (1) (-0.145). Take the column (3) result as an example: the coefficient before *Segregation Index 1885*Treatment*Low Education* is -0.511, meaning that for low eduction immigrant enclave residents, those who worked within 25m of the streetcar rails in 1885 moved to a neighborhood in 1900 where there were 51.1 percent point less conationals, compared to low eduction immigrant enclave residents whose workplace were 25-50m away from the rails in 1885! The huge magnitudes of the treatment effect on these disadvantaged immigrant groups indicate that they were among the least informed, and therefore a transport shock that connected them better with outside communities helped them to migrate out of enclaves.

In column (4), the coefficient before *Segregation Index 1885*Treatment*No House* is also very large. While ownership of real estate property is an indicator of financial status, and real estate owners probably became better off financially than renters and boarders after the transport upgrade shock, it also imposes a fixed cost on migration, which can explain why house owners were much less responsive to the transport shock. Therefore, the results in Table 2.4 do not support the idea that better financial resources facilitate the migration out of enclaves; instead, it favors the story that disadvantaged individuals were less informed, and more interactions with people from outside communities help them overcome these information barriers. In the next subsection, I return to testing this mechanism.

2.6.2 Mechanisms

To test whether more interactions with people from non-local communities matter, I construct a measure of openness for each workplace. This measure is proportional to the residential areas covered by all the workers who worked there. Figure 2.3 illustrates this measure. For workplace A, the employment comes from three different residential communities, and therefore has a value of 3 by this openness measure. For workplace B, the workers come from two different residential communities, which corresponds to a value of 2. A higher value of this openness measure shows that the employment at this workplace come from diverse communities, and therefore everyone potentially has better access to information on different communities.

Table 2.4: Regression Results: Heterogeneity

	(1)	(2)	(3)	(4)
Outcome Variable:	Changes in Segregation Index, 1885-1900			
Segregation index 1885	-0.641***	-0.778***	-0.847***	-0.628***
	(0.077)	(0.105)	(0.064)	(0.113)
Segregation index 1885*Low Income		0.197*		
		(0.117)		
Segregation index 1885*Low Educ			0.525***	
			(0.149)	
Segregation index 1885*No House				-0.020
				(0.107)
Treatment	0.023*	0.027**	0.029**	0.017
	(0.013)	(0.013)	(0.013)	(0.014)
Segregation index 1885*Treatment	-0.145*	-0.012	0.028	-0.127
	(0.083)	(0.104)	(0.119)	(0.093)
Seg Idx*Treatment*Low Inc		-0.204*		
		(0.119)		
Seg Idx*Treatment*Low Educ			-0.511**	
			(0.192)	
Seg Idx*Treatment*No House				-0.570**
				(0.243)
Low Income		0.019		
		(0.014)		
Low Educ			0.032**	
			(0.015)	
No House				-0.031
				(0.019)
Observations	416	416	416	416
R-squared	0.345	0.350	0.369	0.355

Notes: Standard errors clustered by 500-grid neighborhoods are reported in parentheses: *** indicates statistical significance at the 1% level, ** at the 5% level and * at the 10% level.

I split the regression sample into immigrants whose workplace was below and above the median openness measure, and use two choices of neighborhood size - 500m and 200m grids - for constructing the openness measure. I report these results in Table 2.5. From column (2) and column (4), we can see that the coefficients before *Segregation Index 1885*Treatment* are larger in magnitude than the benchmark coefficient, suggesting that immigrant enclave residents who worked at more "open" places were more responsive to the transport shock than those who worked at less "open" places.

Table 2.5: Regression Results by Openness Measure in Workplace

	(1)	(2)	(3)	(4)	(5)
	Benchmark	Openess (200m)	Openess (200m)	Openess (500m)	Openess (500m)
VARIABLES		above median	below median	above median	below median
Segregation index 1885	-0.641***	-0.539**	-0.657***	-0.569***	-0.691***
	(0.077)	(0.213)	(0.064)	(0.210)	(0.056)
Treatment	0.023*	-0.005	0.042**	-0.006	0.047**
	(0.013)	(0.016)	(0.020)	(0.015)	(0.022)
Seg. idx. 1885*Treatment	-0.145*	-0.252*	-0.081	-0.221*	-0.146
	(0.083)	(0.137)	(0.095)	(0.126)	(0.093)
Observations	416	202	214	218	198
R-squared	0.345	0.398	0.343	0.338	0.422

Notes: Standard errors clustered by 500-grid neighborhoods are reported in parentheses: *** indicates statistical significance at the 1% level, ** at the 5% level and * at the 10% level.

To further shed light on the result that the treatment effect is stronger for immigrants who worked in more "open" places, I examine whether collegiality relationship (in terms of proximity in workplace) in 1885 predicts co-residence (within a certain distance threshold) in 1900 for individual pairs. If there exist a positive relationship, we can then infer that people did have learned useful information at workplace, and such information affected their residential location choices 15 years later.

I estimate the specification below to test this conjecture:

$$I\{Dist_{ij,1900}^{residence} < 500m/1km\} = \beta_1 I\{Dist_{ij,1885}^{workplace} < 200m\} + \beta_2 X_{ij,1885} + \varepsilon_{ij} \quad (2.5)$$

where ij indicate any pair of individuals, $I\{Dist_{ij,1885}^{workplace} < \bar{D}\}$ is a dummy which takes the value of 1 if the 1885 places of work of i and j are within 200 meters of each other. $X_{ij,1885}$ is a set of individual characteristics that affect the tendency of two individuals to colocate. The coefficient of interest is β_1, with $\beta_1 > 0$ supporting the idea that interactions in workplace affected residential location choices 15 years later. I use pairwise bootstrapping standard errors to address the interdependence of error terms in equation (2.5).

Estimation results of equation (2.5) are reported in Table 2.6. In columns (1) to (3), I use a threshold of 500m as the definition of coresidence, while in columns (4) to (6), I use a threshold of 1km. For each outcome definition, I estimate equation (2.5) using three samples: individual pairs both of whom are natives; are first or second generation immigrants; and are first generation immigrants.

In column (2) of Table 2.6, I find that for immigrant pairs (both first and second generation) who worked together in 1885 (within 200m of each other), they were 3 percent point more likely to live together (within 500m of each other) in 1900. This effect is not only statistically significant, but also economically large - it doubles the likelihood of living together in 1900 compared to a random pair of immigrants. From column (1) and (3), we can see that this effect does not exist among the natives, and is stronger among the first generation immigrants. Columns (4) to (6) confirm this finding under a different definition of co-residence.

That collegiality relationship (in terms of proximity in workplace) predicts co-residence after 15 years only for immigrants but not for natives is striking. It is an indicator that workplace interactions had a powerful impact on immigrants' residential location decisions. This channel, together with the fact that employment density increased substantially faster near the streetcar rails, and that people commuted significantly longer distances during this period, imply that immigrants who initially worked near the streetcar rails suddenly had much better opportunities to contact with people with whom they never met. These interaction opportunities are highly likely the catalysts of the immigrants' decisions to migrate out of ethnic enclaves.

Table 2.6: Regression Results: Mechanism

	(1)	(2)	(3)	(4)	(5)	(6)
Outcome Variable:	Within 500m in 1900 residence			Within 1km in 1900 residence		
Sample:	Natives	1st + 2nd immig	1st immig	Natives	1st + 2nd immig	1st immig
Within 200m in 1885 workplace	0.010	0.030***	0.047***	0.007	0.039***	0.069***
	(0.007)	(0.006)	(0.015)	(0.009)	(0.009)	(0.020)
Observations	15,400	31,375	9,045	15,400	31,375	9,045

Notes: Standard errors are estimated by pairwise bootstraps: *** indicates statistical significance at the 1% level, ** at the 5% level and * at the 10% level.

2.7 Conclusion

Immigrants tend to locate in ethnic enclaves within metropolitan areas in the US. Hypotheses and evidence suggest that residence in ethnic enclaves can have either positive or negative consequences for immigrants, implying that the existence of ethnic enclaves in spatial equilibrium may reflect a lack of information on outside communities for enclave residents. This paper examines within-city migration decisions of immigrants in response to a transport shock that Boston quickly electrified its previous horse-drawn streetcar system during 1889-1896. Analyzing new individual-level data merged between city directories and decennial census in Boston from 1885 to 1900, I find that immigrant enclave residents who worked within 25 meters of streetcar rails in 1885 were much more likely to move to a less segregated neighborhood in 1900, compared to enclave residents who worked between 25 and 50 meters away from streetcar rails. The mechanisms are consistent with interactions in workplace having an impact on choices of residential locations fifteen years after.

Chapter 2, in full, is currently being prepared for submission for publication of the material. The dissertation author, Wei You, was the sole author of this paper.

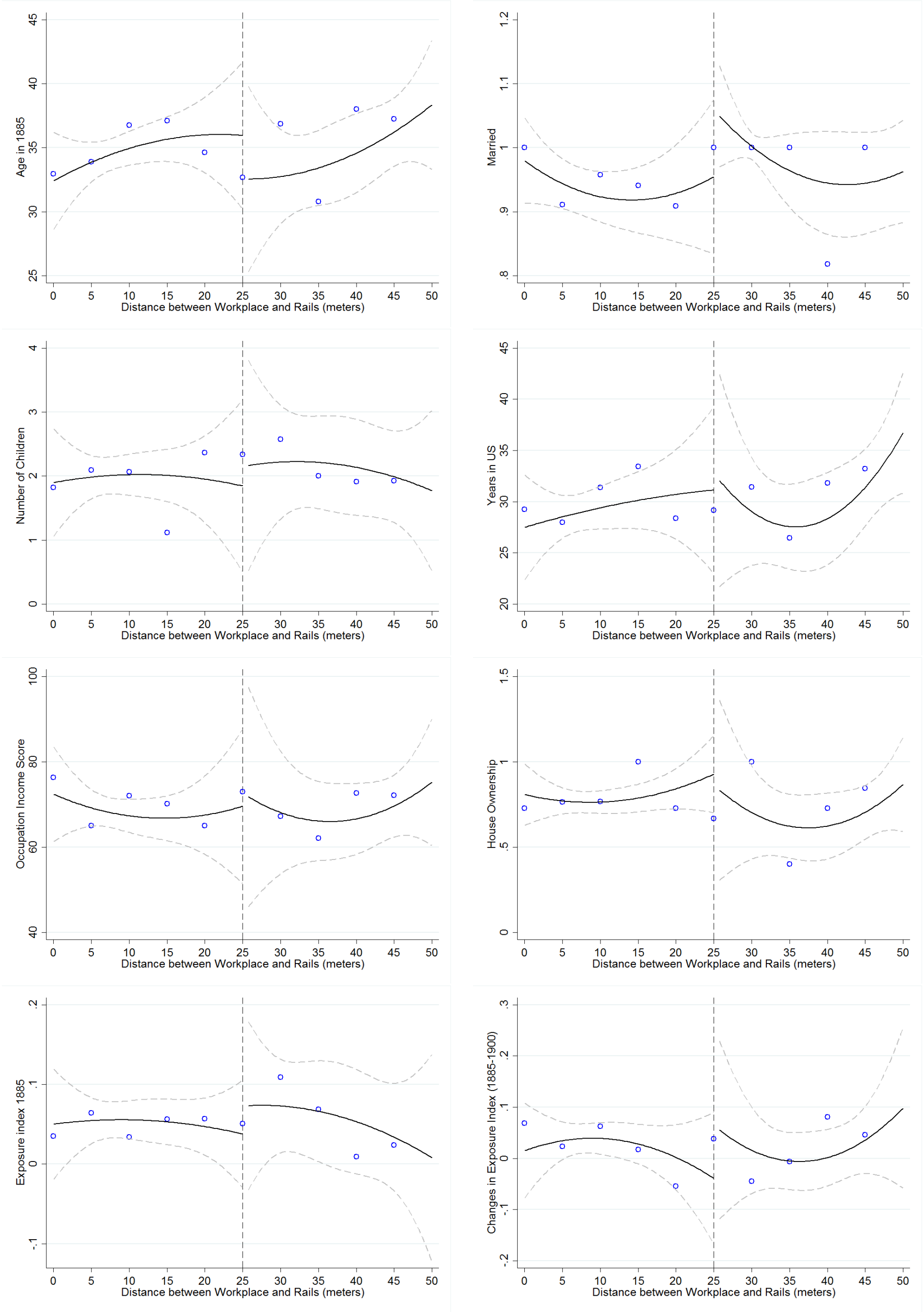
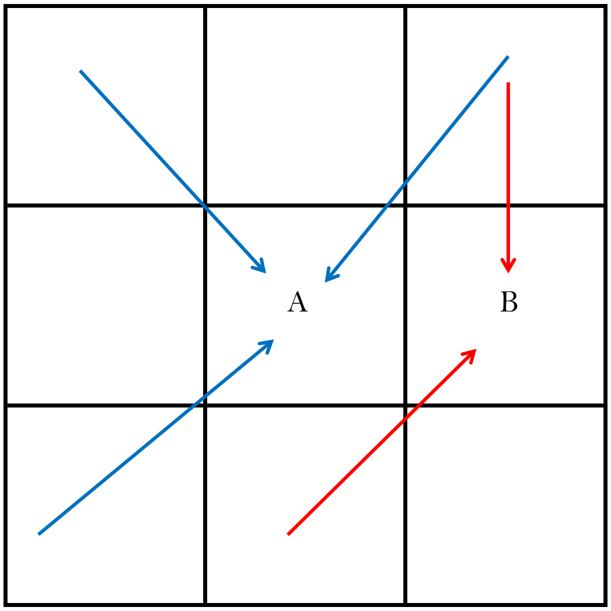


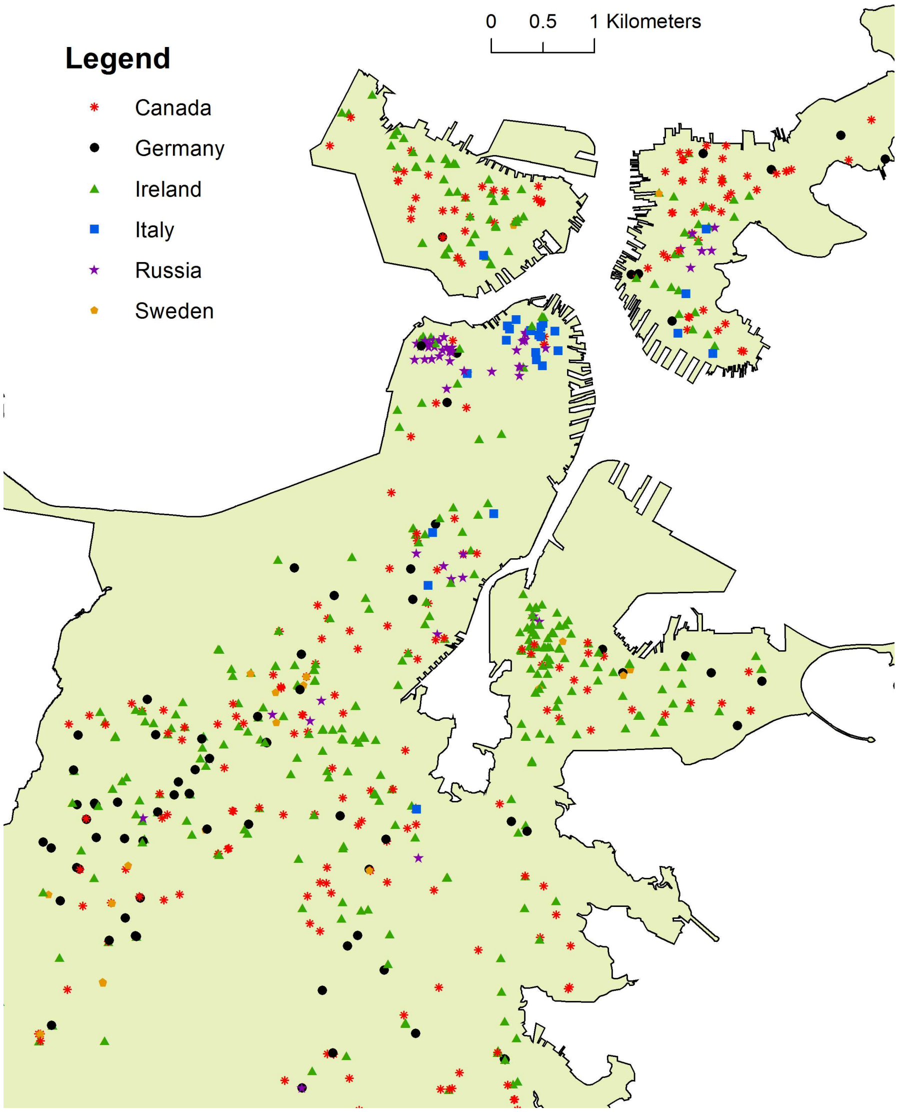
Figure 2.2: Summary Statistics around the Treatment Cutoff



Each arrow indicates a commuting from residence to workplace.
Point A: Openess=3. Point B: Openness=2.

Figure 2.3: Illustration of the openness measure

2.8 Appendix: Supplementary Figures



Source: Geocoded Individual Data Merged from Census to the Boston Directory.

Figure 2.4: Distribution of Immigrants in Boston (1885 locations)

Chapter 3

Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine

3.1 Abstract

Urbanization often occurs in an unplanned and uneven manner, resulting in profound changes in patterns of land cover and land use. Understanding these changes is fundamental for devising environmentally responsible approaches to economic development in the rapidly urbanizing countries of the emerging world. One indicator of urbanization is built-up land cover that can be detected and quantified at scale using satellite imagery and cloud-based computational platforms. This process requires reliable and comprehensive ground-truth data for supervised classification and for validation of classification products. We present a new dataset for India, consisting of 21,030 polygons

from across the country that were manually classified as "built-up" or "not built-up," which we use for supervised image classification and detection of urban areas. As a large and geographically diverse country that has been undergoing an urban transition, India represents an ideal context to develop and test approaches for the detection of features related to urbanization. We perform the analysis in Google Earth Engine (GEE) using three types of classifiers, based on imagery from Landsat 7 and Landsat 8 as inputs. The methodology produces high-quality maps of built-up areas across space and time. Although the dataset can facilitate supervised image classification in any platform, we highlight its potential use in GEE for temporal large-scale analysis of the urbanization process. Our methodology can easily be applied to other countries and regions.

3.2 Introduction

Over the past century, many countries, especially in the developing world, have experienced rapid urbanization (Buhaug and Urdal, 2013; Glaeser, 2014). Between 1950 and 2014, the share of the global population living in urban areas increased from 30% to 54%, and by 2050 it is projected to expand by an additional 2.5 billion urban dwellers, primarily in Asia and Africa (United Nations, 2015). Urbanization also entails an increase in the land area incorporated in cities, which over the next 15 years is projected to grow by 1.2 million km^2 (Seto et al., 2012). The process of urbanization profoundly influences economic (Wu et al., 2013) and social development (Buhaug and Urdal, 2013), and has direct consequences for biodiversity, resource conservation, and environmental degradation (Seto et al., 2012; McKinney, 2002; Pugh, 1996).

Previous literature measures the extent of urban areas using household-survey-based socio-economic data, nighttime lights, and mobile-phone records. With the increasing availability of satellite imagery at ever-improving spatial and temporal resolutions, urban research is shifting towards the use of digital, multispectral images and towards the development of remote-sensing image classification designed to capture urban land features (Taubenbock et al., 2012; Dewan and Yamaguchi, 2009; Bhatta, 2009). The availability of earth-observation data, acquired primarily by Landsat and MODIS satellites, has triggered the development of several classification maps of urban areas (Gaughan et al., 2013; Potere et al., 2009; Schneider, 2010), including multi-class land-cover maps, binary maps that indicate the presence/absence of urban land cover, and maps of variables associated with urban areas, such as impervious surfaces and nighttime light generation (Potere and Schneider, 2009).

In parallel, cloud-based computational platforms have become increasingly accessible and allow one to scale analysis across space and time. One such platform is Google Earth Engine (GEE). GEE leverages cloud-computational services for planetary-scale analysis and consists of petabytes of geospatial and tabular data, including a full archive of Landsat scenes, together with a JavaScript, Python based API (GEE API), and algorithms for supervised image classification.

By definition, supervised classification requires ground-truth labeled data. Several datasets have been proposed to serve as ground-truth for urban research. These include gazetteer datasets of city locations; datasets of sites and boundaries, which are digitized, rated, and assessed by expert analysts; medium-resolution Landsat-based urban maps (Potere et al., 2009); and census-based population databases (Stevens et al., 2015). Crowd-sourced datasets, such as OpenStreetMap (OSM) can also be used to map urban areas (Belgiu and Dragut, 2014; Estima and Painho, 2015), especially when they are combined with remotely-sensed settlement and land cover data (Gaughan et al., 2013). OSM is a valuable source for ground-truth data, primarily because of its vast extent and free availability. However, the completeness of OSM and its suitability for urban research is subject to the number and reliability of OSM contributors (Schlesinger, 2015). The use of OSM for supervised image classification remains challenging due to the risk of imbalanced distribution of class labels (including their spatial coverage), the presence of errors or missing class assignments ("class-noise"), and inaccurate polygon boundary delineations (Johnson and Iizuka, 2016).

Despite the significant progress in the field of machine learning and the increasing availability of satellite imagery, there is still a scarcity of ground-truth labeled datasets that have been developed specifically to detect urban areas (Miyazaki et al., 2011). In this study, we aim to fill the need for this valuable data, and to provide, for the first time, reliable and comprehensive open-source ground-truth data for supervised classification that delineates urban areas in one country. We present a new dataset consisting of 21,030 polygons in India that were manually labeled as "built-up" or not built-up and use these data for supervised image classification and detection of urban areas. As a large and geographically diverse country that has been undergoing an urban transition, India represents an ideal context to illustrate the applicability of our approach for mapping urbanization. The results demonstrate the potential for integrating high-resolution satellite imagery, cloud-based computational platforms and ground-truth data to measure and to analyze the urbanization process. Although our study focuses on India, the methodology we develop can easily be applied to other parts of the world.

This study differs from previous efforts to map urban areas in four respects. First, we construct a large-scale and comprehensive georeferenced dataset that is designed for the express purpose of mapping urban areas. We make it available and accessible for the use and validation of existing classification products. As noted above, validated ground truth datasets are in short supply and many of those that do exist are small in size or spatial extent. Second, we validate this dataset and demonstrate its applicability for mapping urban areas at the national level for India. We present, in one study, an assessment of alternative classifiers and examine the effect of various inputs and class combinations on the performance of the classifiers. Third, we propose a methodology that is designed to evaluate the spatial generalizability of the classifiers. We use a spatial k-fold cross-validation procedure, which enables us to evaluate the performance of the classifiers in a large and geographically heterogeneous context. Finally, we leverage the computational power of GEE and its full Landsat archive to introduce a practicable and adaptable procedure for temporal analysis of urban areas at scale.

To summarize, the objectives of this study are: (1) to present a large-scale dataset for supervised image classification of built-up areas; (2) to integrate this dataset into the GEE platform; and (3) to compare different types of classifiers and inputs in GEE. The dataset can be downloaded as a Google fusion or KML file format. [1]

The remainder of this article is organized as follows. In Sub-Section 3.2.1, we discuss the literature on urbanization and remote-sensing methods for urban research. In Section 3.3, we describe the study area and the methodology used to construct and to assess the dataset. In Sections 3.4 and 3.5, we present and evaluate the results. In Section 3.6, we offer a concluding discussion.

3.2.1 Measuring Urbanization by Means of Remote Sensing

Urbanization occurs as rural areas are incorporated into cities, typically through sprawl radiating out from the city center or linearly along major transportation corridors (Sudhira et al., 2004; Baum-Snow, 2007). The growth of cities, which often occurs in unplanned and uneven patterns (Sudhira and Ramachandra, 2007), changes the spatial distribution of population sub-groups (Rahman et al., 2011; Barnes et al., 2001), and affects land cover and land use (LC/LU) (Bhatta, 2009; Schneider, 2012) through the

[1] The dataset can be accessed online as a Google Fusion Table at: $https : //www.google.com/fusiontables/DataSource?docid = 1fWY4IyYiV - BA5HsAKi2V9LdoQgsbFtKK2BoQib0rows : id = 1$ (Note: class “1” = “BU”, class “2” = “NBU”).

construction of built-up structures and impervious surfaces (Sudhira et al., 2004; Bhatta et al., 2010; Jat et al., 2008).

Previous literature characterizes urbanization, alternatively, as an increase in the share of the population living in cities, the level of non-agricultural employment or production, the pace of resource consumption, or the presence of traffic congestion (Frenkel and Ashkenazi, 2008). Spatial metrics of urbanization include urban land area, population density, spatial geometry, accessibility, and building types, as well as various features of land use (Yue et al., 2013). However, the dichotomy between "urban" and "rural" is not universal (Dahly and Adair, 2007). Urban areas are often defined according to social or administrative indicators derived from census-based sources, which, by their nature, vary in their availability, consistency, and spatial and temporal resolutions.

Given the spatial dimensions of urbanization, remote-sensing analysis of satellite images is valuable for mapping urban areas, and analyzing and modeling urban growth and land-use change (Herold et al., 2003). Many features associated with urbanization can be detected in satellite images and used to delineate the boundaries of urban areas, including nighttime lights, LC and LU. However, the delineation of urban areas often differs according to the nature of input data (Schneider et al., 2010), which may capture different dimensions of urbanization, such as population distribution, national income levels, or the distribution of physical structures. For example, it is common to see disparities between the extent of lighted areas and other spatial measures of urban extent (Small et al., 2005), due in considerable part to the relatively coarse spatial resolution of these datasets (Elvidge et al., 2004).

In this paper, we use satellite images to define urban extents according to built-up land cover, which can be observed in satellite images (Sudhira et al., 2004; Jat et al., 2008) and that is closely related to urbanization (Sudhira et al., 2004; Bhatta et al., 2010). Detection of LC/LU using remote sensing can be performed at the level of a pixel (pixel-based), or at the level of an object (object-based), where pixels are grouped together to provide contextual information, such as image texture, pixel proximity, and salient geometric attributes of features. While several studies suggest that object-based classifiers outperform pixel-based classifiers in LC/LU classifications tasks (Whiteside and Ahmad, 2005; Myint et al., 2011; Whiteside et al., 2011; Bhaskaran et al., 2010), other studies suggest that pixel-based and object-based classifiers perform similarly when utilizing common machine-learning algorithms (Duro et al., 2012). In addition, object-based classification requires significantly more computational power than pixel-based classification and there is no universally accepted method to determine an optimal scale level for image segmentation (Myint et al., 2011), especially when analyzing large-scale geographically diverse regions. Thus, object-based classification is typically conducted when the unit of analysis is relatively small, such as a city (Myint et al., 2011; Bhaskaran et al., 2010), or a region of a country (Whiteside and Ahmad, 2005; Whiteside et al., 2011; Duro et al., 2012; Robertson and King, 2011).

In this study, we adopt a pixel-based classification approach to detect built-up areas in India that utilizes the full spectral imagery available in Landsat, as well as NDVI (Normalized Difference Vegetation Index) and NDBI (Normalized Difference Built-up Index) indices. We apply three types of classifiers that are integrated into GEE: Classification and Regression Tree (CART) (Breiman et al., 1984), Random Forest (Breiman, 2001), and Support Vector Machines (SVM) (Cortes and Vapnik, 1995).

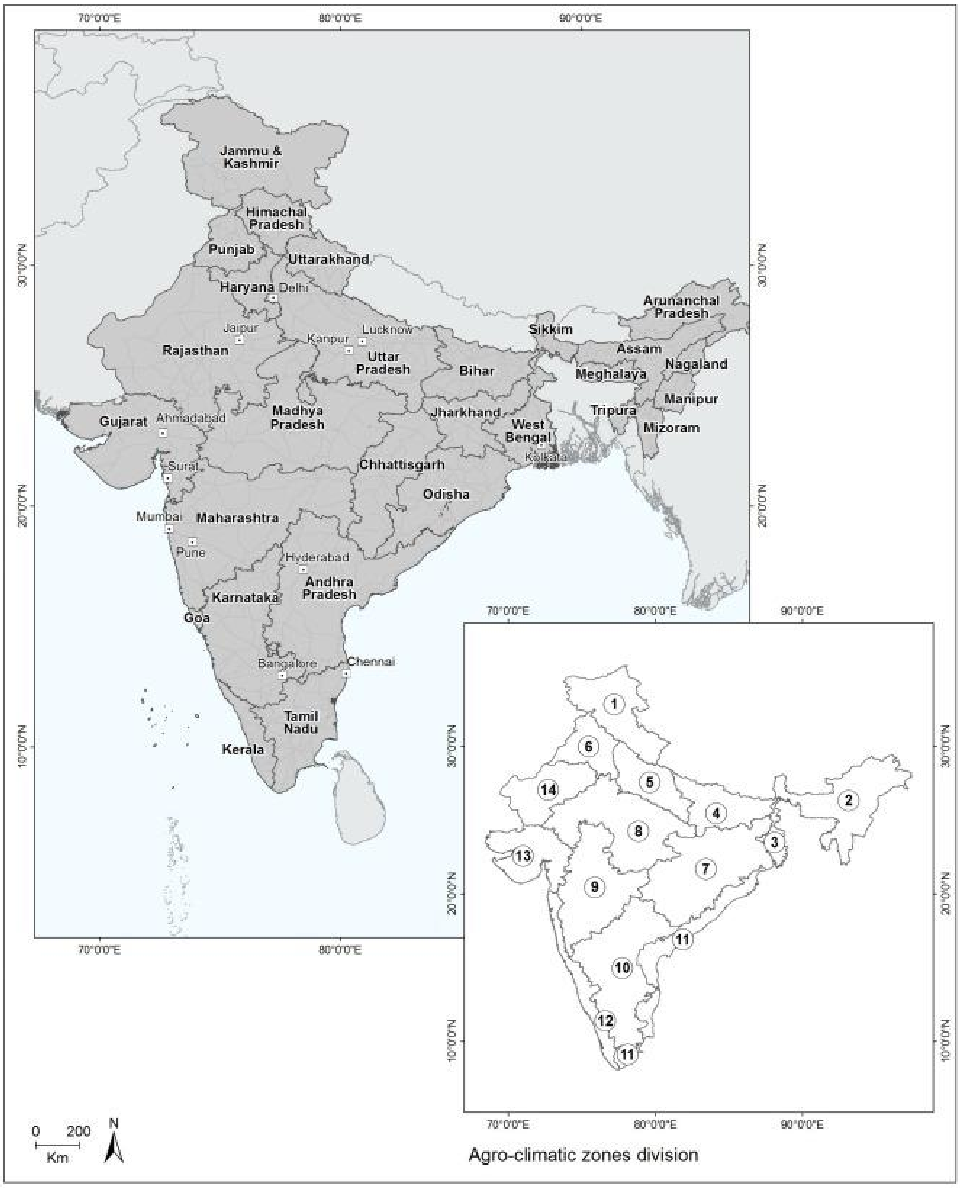
3.3 Materials and Methods

3.3.1 Study Area

India is one of the largest (3.287 million km^2 in size) and most populated countries in the world. In 2014, 1.295 billion people resided in the nation's 29 states which are distributed across 15 geographical regions (see Figure 3.1 (Indian Agriculture Statistics Research Institute, 2006)), 32.4% of whom lived in urban areas (World Bank, 2006). The country is urbanizing rapidly. In the last decade, the growth of its urban population outpaced the growth of its rural population by 31.80% to 12.18% (Census of India, 2011), due primarily to natural urban population growth and secondarily to rural-to-urban migration (Chen and Raveendran, 2011). This trend is expected to continue (Sudhira and Gururaja, 2012). By 2050, half of India's population is projected to be urban (United Nations, 2015).

The urbanization of India is also reflected in the rapid expansion of built-up areas (Sudhira, et al., 2004; Prakasam, 2010; Moghadam and Helbich, 2013) and low-density sprawl (Chen and Raveendran, 2011), together with a decline of other types of land cover, including open land, agriculture land, and bodies of water (Ramachandra et al., 2012; Chadchan and Shankar, 2012; Sharma and Joshi, 2013). By capturing the distinct spectral profile of built-up areas, by means of earth observation, it is possible to map and to quantify the extent of urbanization and the pace of urban growth.

India contains 15 distinct agro-climatic zones. These zones are geographical regions characterized by relatively homogenous environmental-physical characteristics, such as soil type, rainfall, temperature, and water resources (Singh, 2006). India's unusually large number of climatic zones reflects the country's latitudinal expanse and widely varying elevation and rainfall. Previous studies have shown that these zones vary in their agriculture growth, rural poverty and population density (Palmer-Jones and Sen, 2003). By randomly sampling areas within India for our analysis, the countrys geographic diversity allows us to create a training set that would incorporate agro-climatic zones found in the large majority of developing countries. This feature makes our training set of potential value for analysis throughout the tropics, as well as in sub-tropical regions.



The spatial extent of the agro-climatic zones was digitized according to the National Portal on Mechanization and Technology.

Figure 3.1: India's states (top); and agro-climatic zones (without zone 15 - the islands region) (bottom).

3.3.2 Dataset Construction

We define the boundaries of urban areas according to one property of urbanization: built-up land cover (i.e., the boundaries between built-up (BU) and not built-up (NBU) areas). We define BU areas as polygons where the majority of space (more than 50%) is paved or covered by human-made surfaces and used for residential, industrial, commercial, institutional, transportation, or other non-agricultural purposes. All other land cover is defined as NBU. Similar definitions for urban areas are proposed by (Potere et al., 2009; Schneider et al., 2010) who characterize a pixel as "urban" when the built environment spans the majority (50% or greater) of the sub-pixel space.

Our classification utilizes a dataset consisting of 21,030 polygons, 30 m × 30 m in size, that are randomly distributed throughout India and manually labeled as BU or as NBU (the methodology is described in Figure 3.2). To construct this dataset, we begin withWorldPop, a per-pixel population estimation dataset (Gaughan et al., 2013; Stevens et al., 2015), and create an initial random stratified sample of BU and NBU areas. WorldPop depicts a grid of per-pixel estimates of population densities, in a spatial resolution of approximately 100 m (we use India's population dataset for 2010, available at: www.worldpop.org.uk). The maximum value of a pixel is 1523 (i.e., 1523 people per hectare). A visual comparison between WorldPop dataset and Google Earth satellite imagery shows that a threshold of 40 persons per hectare (pixel) closely matches the extent of India's settlements and populated areas. We thus set a threshold of 40 persons per pixel as an initial indicator to identify highly populated areas. These areas constitute 0.41% of the country's land area and account for 19.2% of the country's population.

We define populated areas as clusters of neighboring pixels whose values are higher than, or equal to, 40 (i.e., 40 persons per pixel). We convert these clusters to polygons (a vector format), where the polygons represent the boundaries of highly populated areas. We define the adjacent periphery to these highly populated areas by calculating the width of each polygon's enclosing rectangle (W_i), where W_i is the length of the shorter side of a given polygon's enclosing rectangle. To capture peripheral rural areas around cities, we create a buffer around each polygon, which is twice the size of its enclosing rectangle ($2W_i$). We sample our NBU examples from these peripheral areas (Figure 3.3), such that in the classification we will consider pixels from established urban areas and immediate surrounding areas that have yet to experience urbanization. We focus on rural areas adjacent to urban areas because our BU/NBU classification targets the boundaries of cities and is therefore designed to characterize the process of urban sprawl. From this universe of high-population-density cores and surrounding peripheral rural areas in India, we randomly sample 20,151 polygons, 40% of which are from the core and 60% of which are from the periphery (7928 and 12,223 polygons, respectively), where the number of sampled polygons in each of Indian state is proportional to the state's total population. We oversample polygons from the periphery to account for heterogeneity in the types of land cover found in NBU regions. In order to have sample representation of rural areas that are distant from urban zones, we randomly sample an additional 879 polygons from outside of the core and periphery areas of cities, for a total of 21,030 polygons.

We overlay the polygons with the Google Earth high-resolution base map and manually classify each polygon as BU or as NBU using a visual interpretation method. The polygons are manually labeled by two graduate students who were provided with extensive training and supervised by the researchers. We provided each student an equal proportion of samples. The students labeled each polygon either as BU or as NBU in Google Earth by a visual interpretation of the most recent available satellite image (typically from 2014 to 2015). The students were instructed to label polygons that have at least 50% of their area covered with built-up land cover (according to the definition above) as BU and otherwise as NBU. The manual labeling resulted in a dataset (a KML

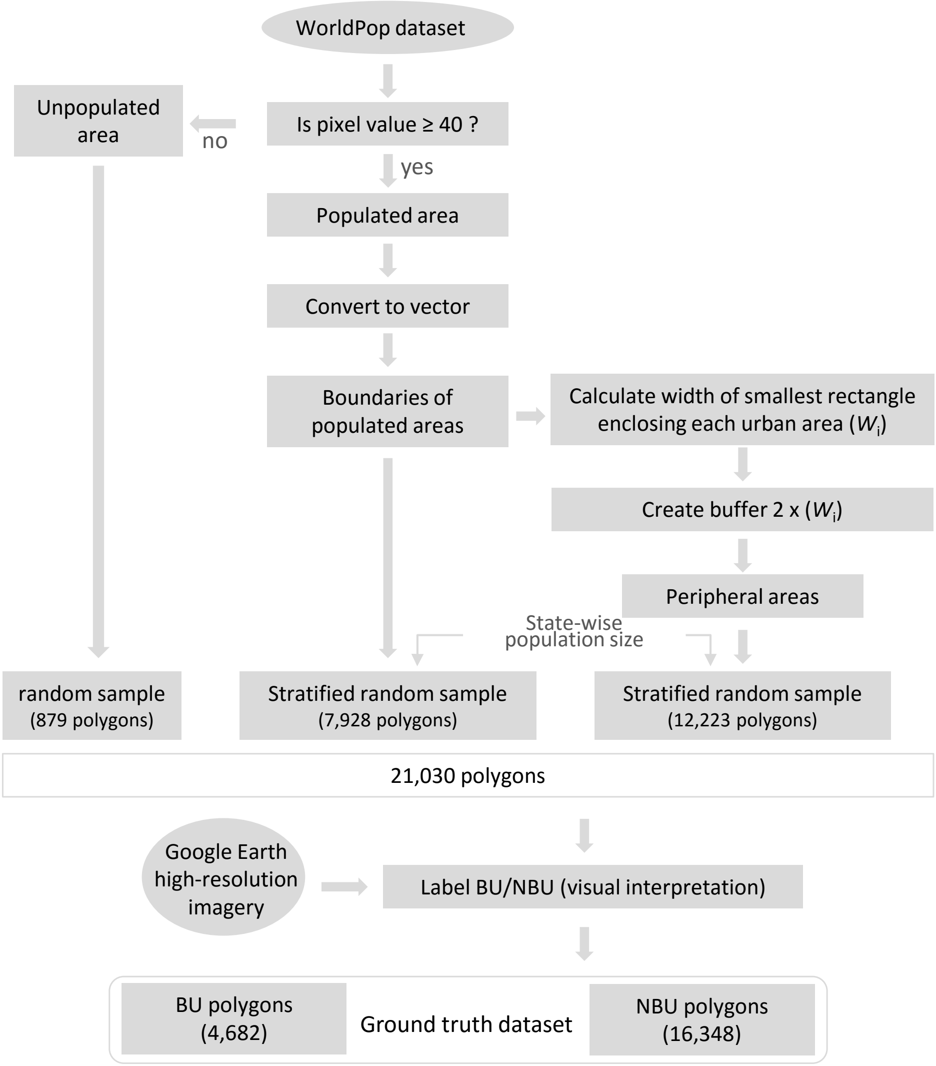

WorldPop dataset
Is pixel value ≥ 40 ?
no
Unpopulated area
yes
Populated area
Convert to vector
Boundaries of populated areas
Calculate width of smallest rectangle enclosing each urban area (W_i)
Create buffer 2 x (W_i)
Peripheral areas
State-wise population size
random sample (879 polygons)
Stratified random sample (7,928 polygons)
Stratified random sample (12,223 polygons)
21,030 polygons
Google Earth high-resolution imagery
Label BU/NBU (visual interpretation)
BU polygons (4,682)
Ground truth dataset
NBU polygons (16,348)

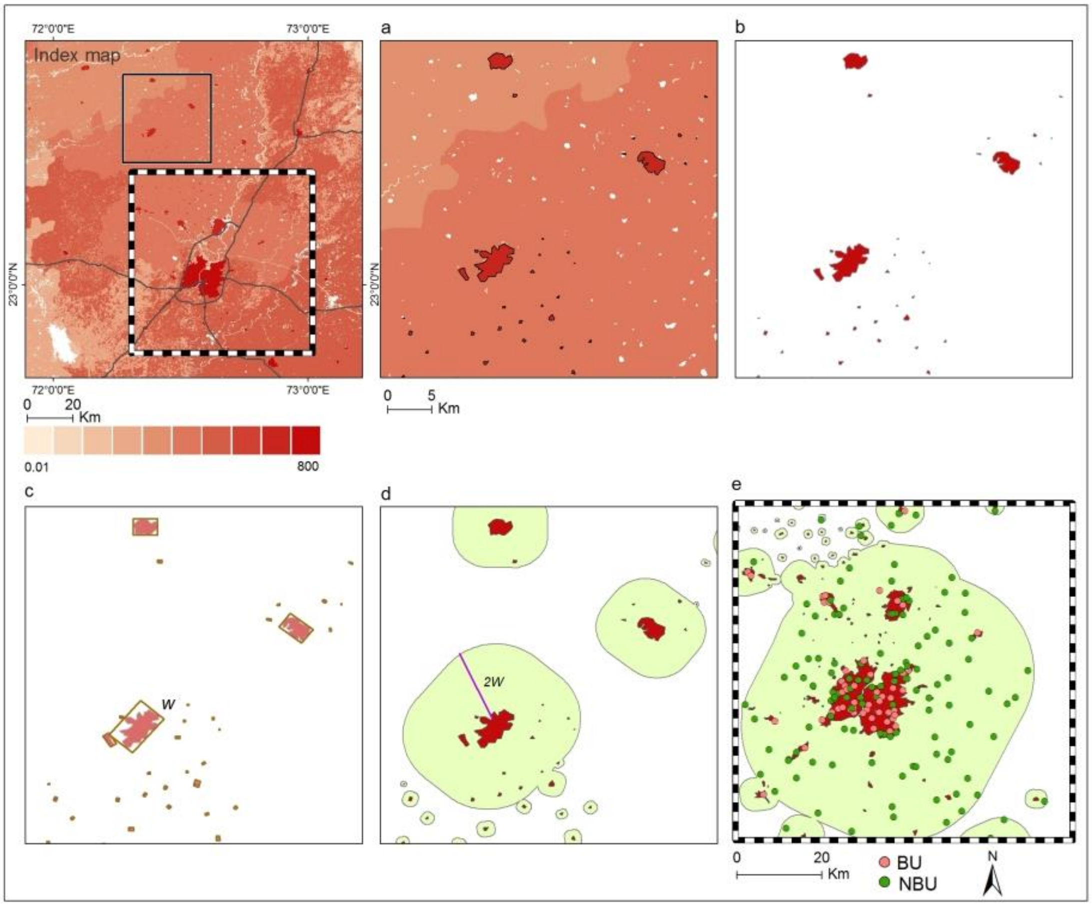
Figure 3.2: The procedure to generate the ground-truth dataset.

file) of 4682 polygons that were labeled as BU and 16,348 polygons that were labeled as NBU (some polygons from the urban core did not contain a majority of built-up pixels, leading us to label them as NBU, whereas some polygons from peripheral areas surrounding cities did have a majority of built-up pixels, leading us to label them as BU). The KML file is then converted to a Google fusion table, which is used for supervised classification in GEE.

3.3.3 Pre-Processing and Scene Selection

We use Landsat 7 and Landsat 8 as inputs for image classification (Table 3.1 presents a description of the spectral bands). Although the spectral resolution of Landsat 7 is lower than that of Landsat 8, the former satellite was launched in 1999 (Landsat 8 was launched in 2013) and thus allows for a longer time horizon over which to study urbanization. Since a composite of pre-processed scenes of Landsat 7 is available in GEE, we use a Landsat 7 annual TOA percentile composites (2014) (referred to as Landsat 7). This composite includes Top of Atmosphere (TOA) calibrated Landsat 7 (ETM+) images (filtered to 2014), excluding images with a negative sun elevation. The composite includes pixels with the lowest cloud cover, computed as per-band percentile values and scaled to 8 bits ([0,255]) (bands 15,7) or to units of Kelvin-100 (band 6). For Landsat 8, we apply a standard TOA calibration on USGS Landsat 8 Raw Scenes (filtered to 2014) and assign a cloud score to each pixel. We select the lowest possible range of cloud scores and compute per-band percentile values from the accepted pixels. We scale the values to 8 bits.

To improve the classification when using Landsat 7 as the input, we add two additional indices: the Normalized Difference Vegetation Index (NDVI) (Pettorelli et al., 2005) and the Normalized Difference Built-up Index (NDBI) (Zha et al., 2003).



We begin with WorldPop dataset, a grid of per-pixel estimates of population densities (a). Then, we extract clusters of neighboring pixels whose values are greater than or equal to 40. These clusters represent highly populated areas and are converted into a vector format (polygons) (b). We calculate the width (W_i) of the shorter side of each polygon's enclosing rectangle (c) and create a buffer around each polygon that is twice this width ($2W_i$) (these buffers represent the periphery of the populated areas) (d). Finally, we randomly sample 7,928 and 12,223 polygons from the highly populated and from their periphery, respectively (e).

Figure 3.3: The procedure to generate the stratified random sample

Table 3.1: The Bands that were Used as Features for the Classification

	Spectral Band	**Wavelength**	**Resolution**
		(Micrometers)	**(Meters)**
Landsat 7			
B1	Band 1blue-green	0.450.52	30
B2	Band 2green	0.520.61	30
B3	Band 3red	0.630.69	30
B4	Band 4reflected IR	0.760.90	30
B5	Band 5reflected IR	1.551.75	30
B6	Band 6thermal	10.4012.50	120
B7	Band 7reflected IR	2.082.35	30
NDVI	$\frac{(B4-B3)}{(B4+B3)}$		30
NDBI	$\frac{(B5-B4)}{(B5+B4)}$		30
Landsat 8			
B1	Band 1Ultra blue	0.430.45	30
B2	Band 2Blue	0.450.51	30
B3	Band 3Green	0.530.59	30
B4	Band 4Red	0.640.67	30
B5	Band 5Near Infrared (NIR)	0.850.88	30
B6	Band 6SWIR 1	1.571.65	30
B7	Band 7SWIR 2	2.112.29	30
B8	Band 8Panchromatic	0.500.68	15
B10	Band 10Thermal Infrared (TIRS) 1	10.6011.19	100 (resampled to 30)
B11	Band 11Thermal Infrared (TIRS) 2	11.5012.51	100 (resampled to 30)

- NDVI expresses the relation between red visible light (which is typically absorbed by a plant's chlorophyll) and near-infrared wavelength (which is scattered by the leaf's mesophyll structure). It is computed as:

$$(NIR-RED)/(NIR+RED) \tag{3.1}$$

  where NIR is the near infra-red wavelength and RED is the red wavelength. The values of NDVI range between (-1) and (+1). An average NDVI value in 2014 was calculated for each pixel (with Landsat 7 32-Day NDVI Composite).

- NDBI expresses the relation between the medium infra-red and the near infra-red wavelengths. It is computed as:

$$(MIR-NIR)/(MIR+NIR) \tag{3.2}$$

  where MIR is the medium infra-red and NIR is the near infra-red wavelength. The index assumes a higher reflectance of built-up areas in the medium infra-red wavelength range than in the near infra-red.

3.3.4 Detection of Built-Up Areas

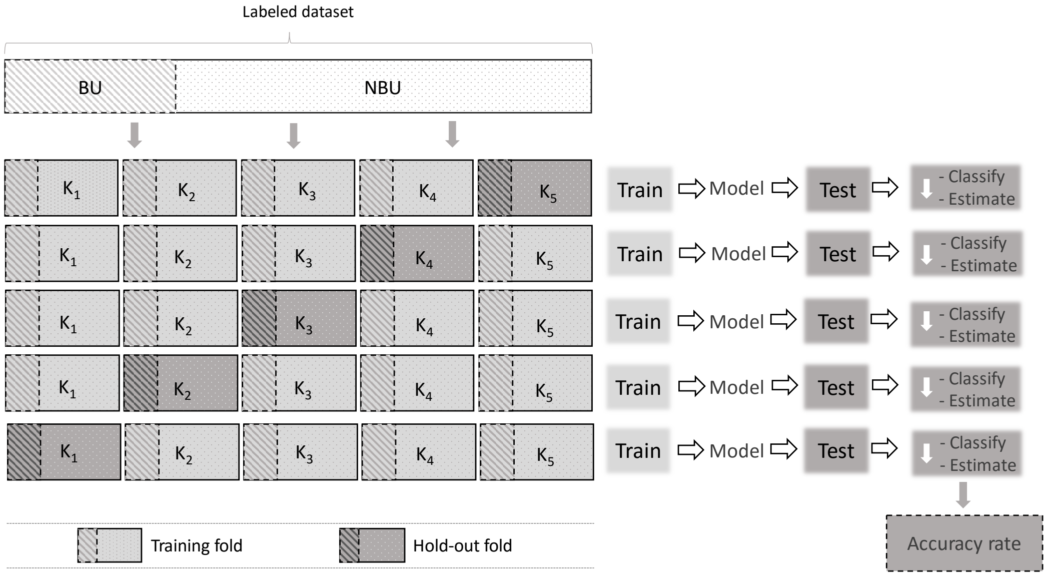
We perform detection of built-up areas in GEE. First, we overlay the labeled polygons on the input. We collect all Landsat pixels within the regions of these polygons (a total of 5,092 BU examples and 17,751 NBU examples), including the reflectance values (per band) and the index values of the examples. Note that the number of the sampled pixels (examples) differs from the number of polygons in the dataset because the polygons do not overlap entirely with Landsat's pixels; these variables are the input for the classifiers (the classifiers' feature space). In addition, each example included an output: a binary class - BU or NBU. We use this set to train, test and evaluate the performance of the classifiers.

3.3.5 Accuracy Assessment

The performance or the accuracy of a classifier refers to the probability that it will correctly classify a random set of examples (Kohavi, 1995). To assure a "fair" assessment of a classifier's generalization, the data used to train the classifier must be separated from the data that is used to assess its accuracy. Thus, labeled data is typically divided into a training set and a test set (a validation set may also be used to "tune" the classifier's parameters). Different data splitting heuristics can be used to assure a separation between the training and test sets (Kohavi, 1995), including the holdout method, in which the data is divided into two mutually exclusive subsets: a training set and a test/holdout set; bootstraping, in which the dataset is sampled uniformly from the data, with replacement; and cross-validation, also known as k-fold cross-validation, in which the data are divided into k subsets (optimally 5 or 10, to allow a less biased estimation (Rodriguez et al., 2010)) with k "experiments". The cross-validation procedure ensures that each example is included exactly once in the test fold and that each example in the test fold is not used to train the classifier. Averaging the overall accuracy across all k partitions yields k accuracy values, or k hold-out estimators, and a variance estimation of the classification error (Salzberg, 1997; Arlot and Celisse, 2010). Though each of these methods can be used to assess the performance of a given classifier, cross-validation is a widely accepted procedure (Refaeilzadeh et al., 2009) that provides a robust estimate of a classifier's generalization error (Blum et al., 1999). When the instances are representative of the underlying population and when sufficient instances are available for training, this procedure results in an unbiased estimate of the accuracy of the classifier over the population (Bradford and Brodley, 2001).

In this study, we adopt a k-fold cross-validation procedure (with k “experiments”) to estimate the accuracy of the classifiers. In each experiment, the examples in one of the data folds is left out for testing and the examples in the remaining k-1 fold are used to train the classifier. The performance quality of the trained classifier is tested on the left-out fold, and the overall performance measure is then averaged over the k folds (over the k experiments) (Figure 3.4).

We first conduct a 5-fold cross validation by dividing the data into five randomly stratified folds (while maintaining a constant proportion of BU and NBU examples per fold). Then, to evaluate the spatial generalization of the classifiers, we conduct a 14-fold cross validation by dividing the data into 14 distinct geographical regions according to India’s agro-climatic zones (Singh, 2006) (see Figure 3.1) (note: we exclude zone number 15, which is the islands region). Each zone includes between 558 and 2695 BU and NBU examples (see Table 3.2).



In each “experiment,” the examples in one of the data folds is left out for testing and the remaining examples in the k-1 fold are used to train the classifier. The performance quality of the trained classifier is tested on the left-out fold (in each “experiment”), and the overall performance measure is then averaged over the k folds (k “experiments”) (This figure is adapted from Refaeilzadeh et al. (2009))

Figure 3.4: k-fold (5-fold) cross validation scheme.

Table 3.2: Built-up (BU) and Not Built-up (NBU) Examples per Agro-Climatic Zone

Zone Number	Number of Examples		BU/NBU Ratio	
	BU	NBU	BU	NBU
1	82	476	14.7%	85.3%
2	169	825	17.0%	83.0%
3	222	837	21.0%	79.0%
4	425	1816	19.0%	81.0%
5	671	2024	24.9%	75.1%
6	382	953	28.6%	71.4%
7	326	1545	17.4%	82.6%
8	333	1066	23.8%	76.2%
9	421	1464	22.3%	77.7%
10	645	1979	24.6%	75.4%
11	391	1197	24.6%	75.4%
12	250	894	21.9%	78.1%
13	262	805	24.6%	75.4%
14	103	467	18.1%	81.9%
Total	4682	16348		

Note: The table indicates the number of built-up (BU) and not built-up (NBU) polygons per agro-climatic zone and the ratio between BU and NBU examples per zone.

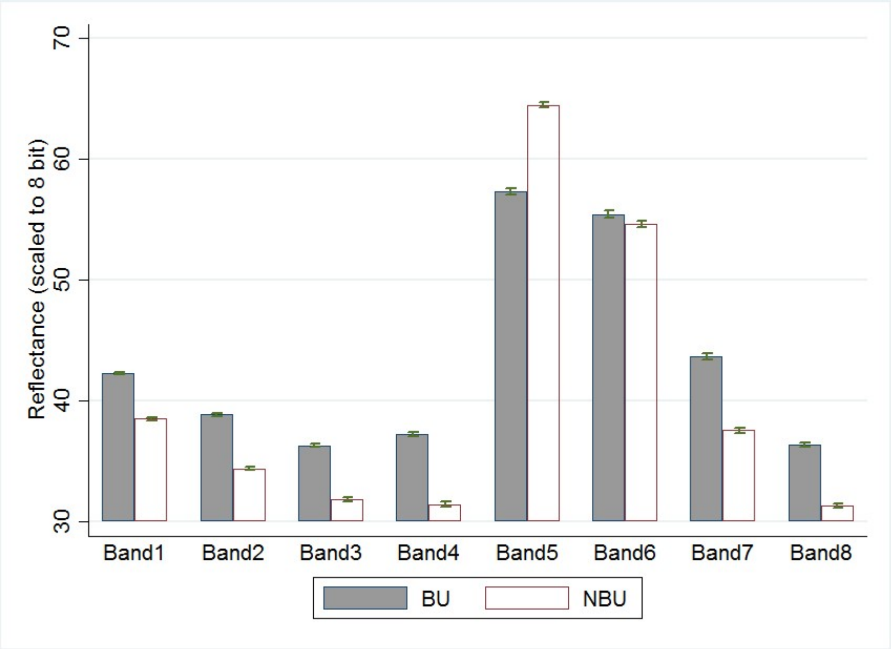
3.4 Results

We now turn to describe the dataset and to present an evaluation of the classification of built-up areas in India using the three classifiers and different combinations of training-set examples and inputs. We assess the performance of the classifiers and map the classified built-up areas. As a preliminary step to validate our BU/NBU dataset's examples, we examine the reflectance profile of the examples, calculated as the average reflectance value of the sampled regions/pixels per band, scaled to 8 bits (Figure 3.5). Consistent with built-up areas containing structures and impervious surfaces that are reflective relative to vegetation and undeveloped land of non-built-up areas, the reflectance of NBU regions is lower than the reflectance of BU regions in all bands except band 5 (the near infra-red range). This anomaly in band 5 is likely due to higher reflectance of vegetation land cover in this wavelength range. A t-test of equal means and a KolmogorovSmirnov test show that BU and NBU regions are characterized by a significantly different ($p < 0.001$, for both tests) reflectance values in all bands (Table 3.3). The BU/NBU distinction is also expressed by significantly different ($p < 0.001$, for both tests) NDVI and NDBI values. As seen in Figure 3.6, the distribution of the NDVI values of NBU regions is to the right of that of BU regions, while the distribution of the NDBI values of NBU regions is a left-skewed and flatter than of BU regions. The standard error bounds of the average reflectance values within BU and NBU regions are relatively small in all bands.

3.4.1 Detection of Built-Up and Not Built-Up Areas

Evaluation of the Classifiers

GEE includes several classifiers for pixel-based image classification. In this study we compare the performance of three prominent ones - SVM, CART and Random Forest



Note: Per-band percentile values were scaled to 8 bits.

Figure 3.5: The mean and 95% confidence intervals of reflectance values of built-up (BU) and not built-up (NBU) regions (Landsat 8 bands)

Table 3.3: Average Reflectance Values of Built-Up (BU) and Not Built-Up (NBU) Regions (Landsat 8 bands).

		B1	B2	B3	B4	B5	B6	B7	B8	NDVI	NDBI
BU	(mean)	42.27	38.84	36.29	37.2	57.29	55.44	43.64	36.36	0.21	-0.02
	(st. err.)	0.048	0.058	0.073	0.099	0.126	0.151	0.137	0.085	0.001	0.001
NBU	(mean)	38.51	34.37	31.82	31.43	64.49	54.65	37.54	31.27	0.35	-0.10
	(st. err.)	0.069	0.075	0.081	0.097	0.104	0.132	0.12	0.087	0.001	0.001
t-tests of	(t-stats)	44.91	47.17	40.95	41.5	-43.99	3.92	33.54	41.92	-85.98	57.02
equal means	(p-value)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
K-S tests*	(p-value)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: Per-band percentile values were scaled to 8 bits.
* KolmogorovSmirnov tests of equality of distributions

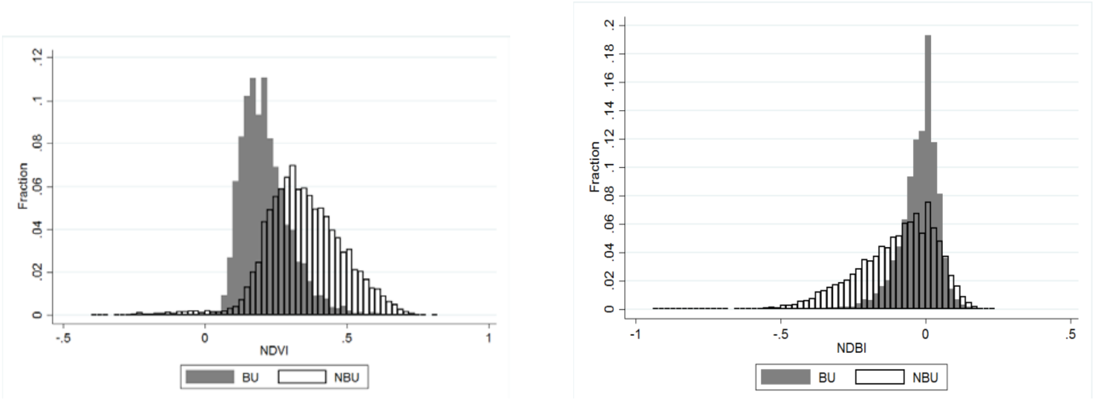


Figure 3.6: Histogram of NDVI (Normalized Difference Vegetation Index) and NDBI (Normalized Difference Built-up Index) values by built-up (BU) and not built-up (NBU) examples

- in detecting BU and NBU areas in India. A five-fold cross-validation test shows that Random Forest (with 100 decision trees) achieves the highest overall accuracy rate - defined as the percentage of examples that were classified correctly - while SVM achieves the lowest. With Landsat 8 as the input, these two classifiers predict correctly 87.1% and 83.1% of the examples, respectively. Previous studies have suggested that the number of decision trees of the Random Forest is generally proportional to the classifier's accuracy (Rodriguez-Galiano et al., 2012). The results show that though the performance of Random Forest improves as the number of trees increase, this pattern holds only up to 10 trees (see Figure 3.7). The classifier's performance remains nearly the same with 50 and with 100 decision trees.

We refer to the class "Built Up" (BU) as positive and to the class "Not Built-Up" (NBU) as negative and evaluate the performance of the classifiers using three additional estimators: (1) True-Positive Rate (TPR) (the percentage of actual BU examples classified correctly as BU); (2) True-Negative Rate (TNR) (the percentage of actual NBU examples classified correctly as NBU); and (3) the average of TPR and TNR (referred to as the balanced accuracy rate).

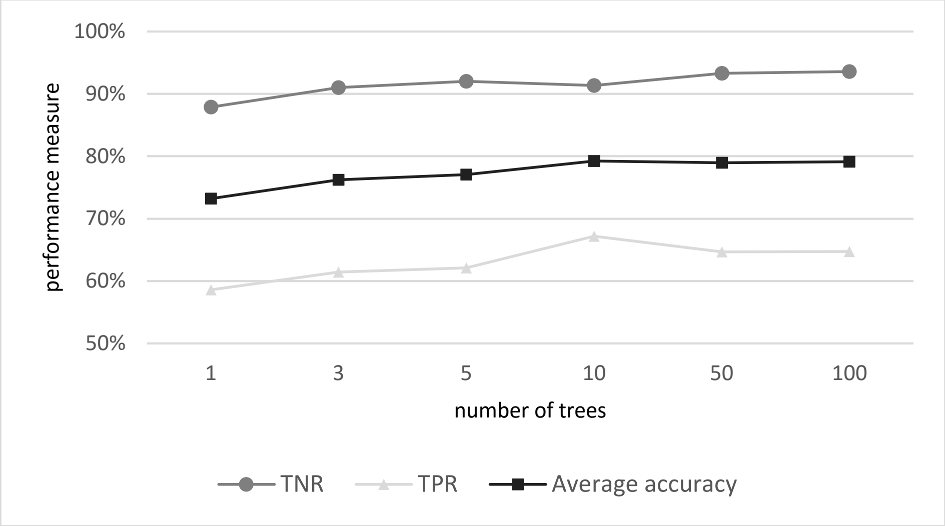
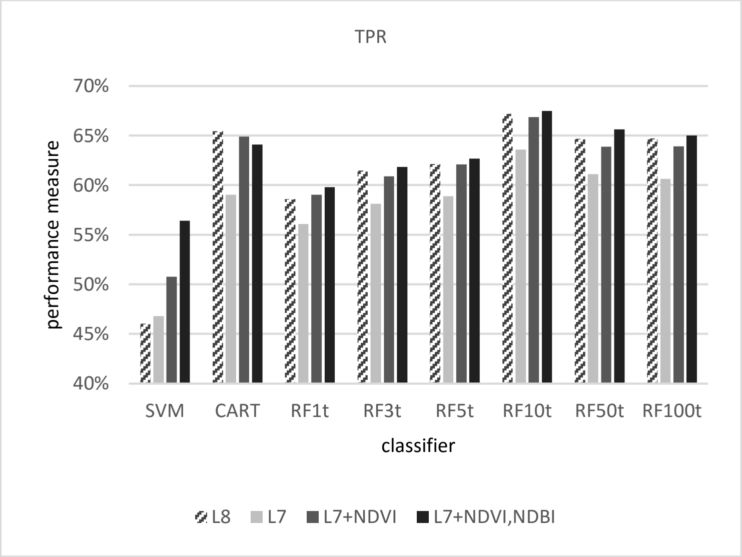
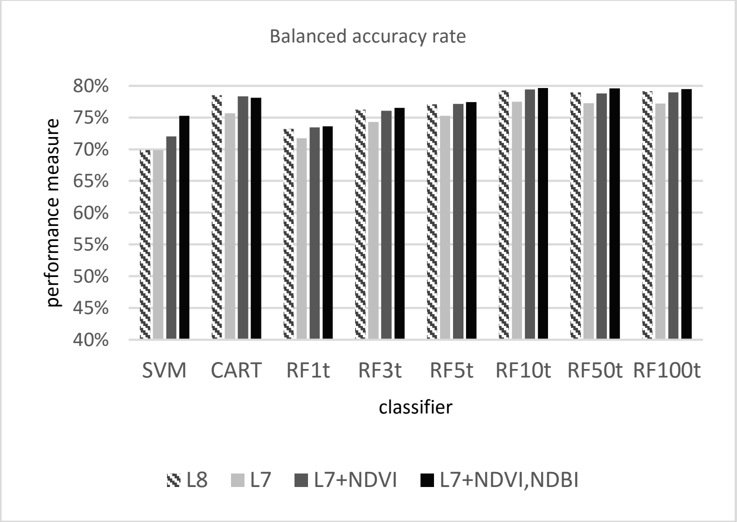


Figure 3.7: The effect of the number of trees of Random Forest on the True-positive rate (TPR), True-negative rate (TNR) and the balanced accuracy rate.

Random Forest (with 10 decision trees) shows the highest balanced accuracy rate (79.7%) while SVM shows the lowest (around 69%) (see Figure 3.8). The classifiers' TPR ranges between 46% (with SVM) and 67% (with Random Forest, 10 trees). As expected, performance with Landsat 8 exceeds that of Landsat 7 likely because of the former's higher resolution relative to the latter. However, when NDVI and NDBI are added to Landsat 7's bands, performance using this input improves. With the exception of SVM, Landsat 7 plus NDVI and NDBI as inputs performs similarly to Landsat 8 as the input. As seen in Figure 3.8, the addition of these two indices primarily improves the balanced accuracy rate of SVM; the classifier's TPR increases from 47% to 56%, and, accordingly, its balanced accuracy rate increases from 70% to 75%. We relate this to the linear kernel that we use with SVM, which is unable to express nonlinear functions from the input variables to the predicted classes.

The performance of the classifiers can also be described in a confusion matrix, where the predicted classes of the examples in the test set are compared with their actual





Examined classifiers: SVM, CART, and Random forest with 1 (RF1t), 3 (RF3t), 5 (RF5t), 10 (RF10t), 50 (RF50t), and 100 (RF100t) trees. Inputs: Landsat 8 (L8), Landsat 7 (L7), Landsat 7 with NDVI (L7 + NDVI) and Landsat 7 with NDVI and NDBI (L7 + NDVI + NDBI).

Figure 3.8: The True-positive rate (TPR) and the balanced accuracy rates by classifier and satellite product.

class (resulting in four possible combinations: TP (True-positive), TN (True-negative), FP (False-positive) and FN (False-negative)). The confusion matrix of the five-fold cross validation (Table 3.4) describes the predicted and the actual class of the tested examples in the five experiments. As noted above, in each experiment, a different subset (fold) is used for the evaluation, and each example - and all examples - are tested exactly once. Several performance estimators can be calculated from this confusion matrix. We present three that are related to the classification of the positive (BU) class: (1) Overall accuracy rate: the portion of instances that were classified correctly (calculated as: $(TP+TN)/(TP+TN+FP+FN)$); (2) Precision rate: the portion of instances that were correctly predicted as positive out of all instances that were predicted as positive (calculated as $TP/(TP+FP)$); and (3) Recall rate: the portion of instances that were correctly predicted as positive out of all actual positive instances (calculated as $TP/(TP+FN)$). Since the best performance is achieved with Landsat 8 as the input, we use Landsat 8 as the input in subsequent analysis.

We also evaluate the classifiers at the geographical level of agro-climatic zones. A k-fold cross-validation test was conducted by dividing the examples into 14 folds according to their geographical location (zone). Similar to the results shown above (where the examples were divided into five random folds), Random Forest (with 10 trees) shows the best performance while SVM shows the worst. When Landsat 8 is used as the input, the TPR and the balanced accuracy rate of Random Forest (with 10 trees) are 66% and 78.7%, respectively, while only 54% and 74%, respectively, with CART. This result provides an additional dimension to the assessment of the classifiers' accuracy, and confirms their generalization as predictors under varying geographical conditions.

Our analysis is based on a large training set relative to past work in remote sensing, with over 20,000 hand-labeled polygons. In many settings, constructing a training set of this magnitude may be infeasible. To provide insight into the importance of training-

Table 3.4: A Confusion Matrix of the Five-Fold Cross Validation Tests

			Predicted								
			L8			L7			L7 + NDVI + NDBI		
			BU	NBU	Total	BU	NBU	Total	BU	NBU	Total
Actual	SVM	BU	2347	2750	5097	2376	2700	5076	2863	2213	5076
		NBU	1122	16727	17849	1261	16659	17920	1044	16876	17920
		Total	3469	19477	22946	3637	19359	22996	3907	19089	22996
			Ac: 0.83 Re: 0.46 Pre: 0.67			Ac: 0.82 Re: 0.46 Pre: 0.65			Ac: 0.85 Re: 0.56 Pre: 0.73		
	CART	BU	3335	1762	5097	2998	2078	5076	3253	1823	5076
		NBU	1500	16349	17849	1380	16540	17920	1413	16507	17920
		Total	4835	18111	22946	4378	18618	22996	4666	18330	22996
			Ac: 0.85 Re: 0.65 Pre: 0.69			Ac: 0.85 Re: 0.59 Pre: 0.68			Ac: 0.85 Re: 0.64 Pre: 0.69		
	RF3t	BU	3133	1964	5097	2951	2125	5076	3140	1936	5076
		NBU	1601	16248	17849	1709	16211	17920	1576	16344	17920
		Total	4734	18212	22946	4660	18336	22996	4716	18280	22996
			Ac: 0.84 Re: 0.61 Pre: 0.66			Ac: 0.83 Re: 0.58 Pre: 0.63			Ac: 0.84 Re: 0.61 Pre: 0.66		
	RF5t	BU	3167	1930	5097	2989	2087	5076	3181	1895	5076
		NBU	1423	16426	17849	1494	16426	17920	1402	16518	17920
		Total	4590	18356	22946	4483	18513	22996	4583	18413	22996
			Ac: 0.85 Re: 0.62 Pre: 0.69			Ac: 0.84 Re: 0.58 Pre: 0.66			Ac: 0.85 Re: 0.62 Pre: 0.69		
	RF10t	BU	3424	1673	5097	3229	1847	5076	3426	1650	5076
		NBU	1543	16306	17849	1539	16381	17920	1471	16449	17920
		Total	4967	17979	22946	4768	18228	22996	4897	18099	22996
			Ac: 0.86 Re: 0.67 Pre: 0.68			Ac: 0.85 Re: 0.63 Pre: 0.67			Ac: 0.86 Re: 0.67 Pre: 0.70		
	RF50t	BU	3297	1800	5097	3102	1974	5076	3332	1744	5076
		NBU	1196	16653	17849	1180	16740	17920	1153	16767	17920
		Total	4493	18453	22946	4282	18714	22996	4485	18511	22996
			Ac: 0.86 Re: 0.64 Pre: 0.73			Ac: 0.86 Re: 0.61 Pre: 0.72			Ac: 0.87 Re: 0.65 Pre: 0.74		
	RF100t	BU	3299	1798	5097	3078	1998	5076	3299	1777	5076
		NBU	1151	16698	17849	1116	16804	17920	1088	16832	17920
		Total	4450	18496	22946	4194	18802	22996	4387	18609	22996
			Ac: 0.87 Re: 0.64 Pre: 0.74			Ac: 0.86 Re: 0.60 Pre: 0.73			Ac: 0.87 Re: 0.65 Pre:0.75		

Note: The confusion matrix is calculated for the five experiments. In each experiment a different fold is used as the test fold and each example is tested exactly once. Key: TP: True-positive; TN: True-negative; FP: False-positive; FN: False-negative; Accuracy rate (Ac): the portion of instances that were classified correctly (calculated as: $(TP+TN)/(TP+TN+FP+FN)$); Recall (Re): the portion of instances correctly predicted as positive out of all actual positive instances (calculated as: $TP/(TP+FN)$); Precision rate (Pre): the portion of instances that were correctly predicted as positive out of all instances predicted as positive (calculated as $TP/(TP+FP)$).

set size for the analysis, we next examine how the prediction rate of the alternative classifiers compares for randomly drawn training sets of different dimensions. We conduct experiments with 800, 1600, 4000, 8000 and 16,000 randomly drawn examples and evaluate each experiment using a five-fold cross-validation test. In each experiment, we use the same test sets (approximately 4500 examples) and a similar proportion between BU and NBU examples (equal to the proportion in the full sample).

The results show strongly improved performance as the size of the training set increases, both, in terms of the TPR and balanced accuracy (see Figure 3.9). CART shows the largest improvement; for example, as the training set size increases from 800 to 16,000 examples, CART's balanced accuracy increases from 74% to 78%. On the other hand, SVM does not show a significant improvement as the training set size increases; the balanced accuracy remains around 73%.

In the experiments described above, we maintain a constant proportion between the BU and the NBU examples in the training set (similar to the proportion in the full dataset). In an additional experiment, we examine the effect of varying the proportion between BU and NBU training examples on the classifiers' performance. In each experiment, we use all BU examples in the dataset as training examples and increase the size of the training set by adding NBU training examples. This allows us to evaluate whether performance improves as the size of the training set increases despite an increased imbalance between BU and NBU examples. We conduct a five-fold cross validation test by increasing the number of NBU examples in the training set and maintaining a constant number of BU examples (4000) to a total of 6000, 8000, 10,000, 14,000 and 16,000 BU and NBU examples in the training set. The size of the test set and the proportion between the BU and NBU examples remain constant (Landsat 8 is used as the input).

Results show a moderate improvement in the classifiers' performance as the training set's size increases, primarily with CART. Although the size of the training

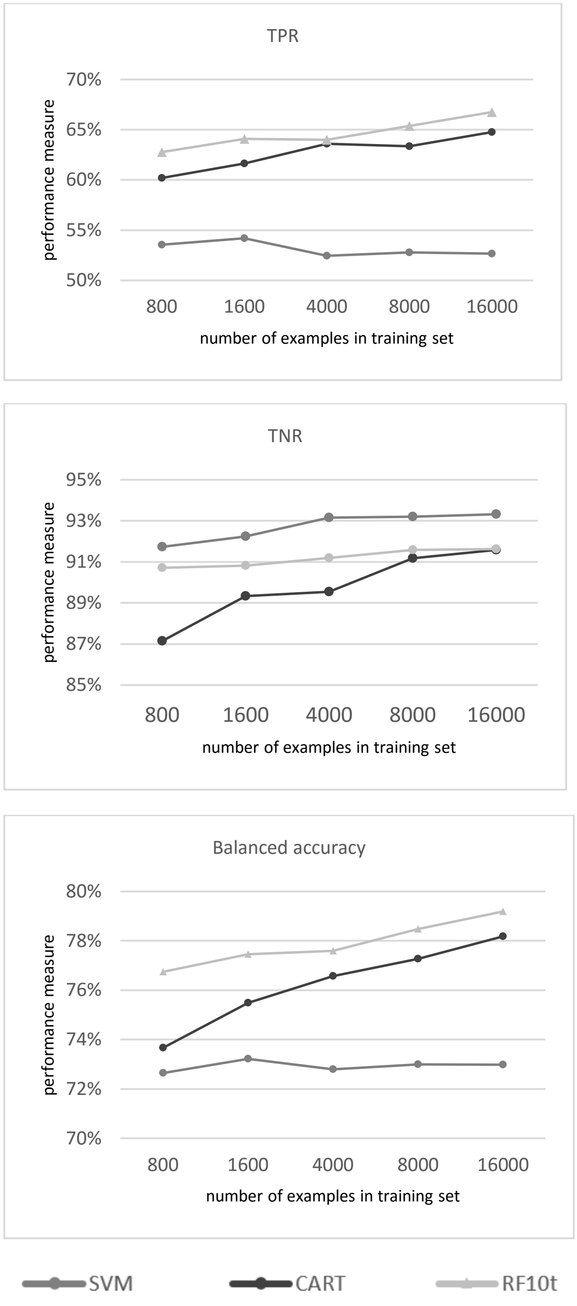


Figure 3.9: The effect of the training set size on the True-positive rate (TPR), True-negative rate (TNR) and balanced accuracy (with Landsat 8 as the input)

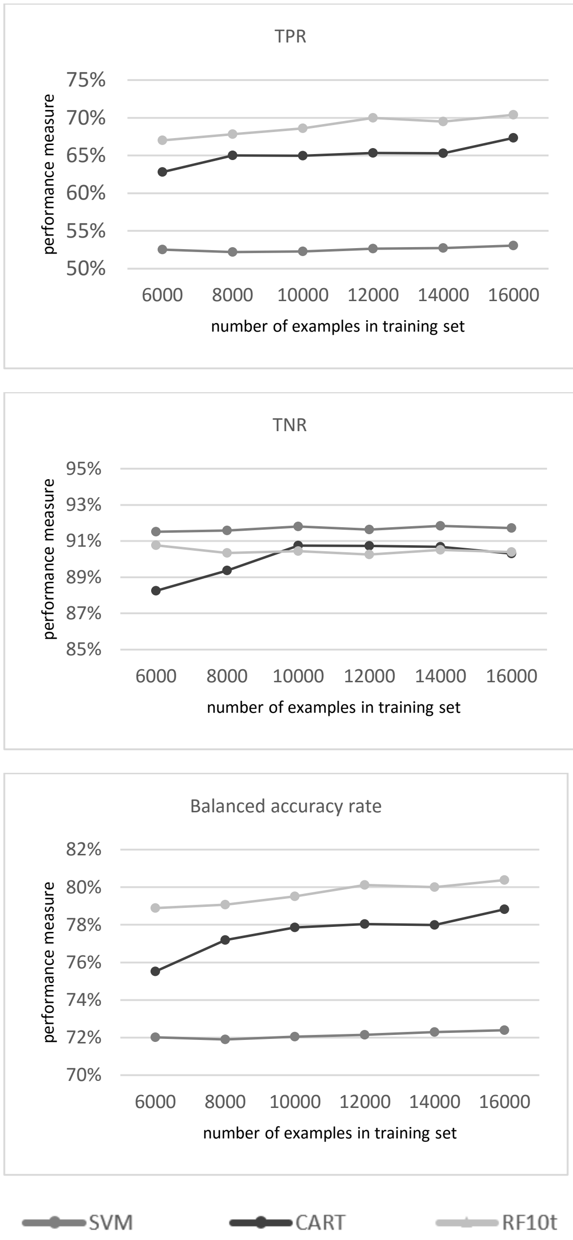
set is increased only by adding NBU examples, the TPR also improves (in addition to TNR). As the size of the training set increases from 6000 to 16,000 examples, CART's TPR increases from 63% to 67% and its TNR increases from 88% to 90% (see Figure 3.10). Thus, although we increase the disproportion between BU and NBU examples, the addition of NBU examples improves the overall performance of the classifiers.

Mapping the Classification

In the final classification process, we use the trained classifier to map built-up and not built-up areas over new examples/pixels. The classified image (in a spatial resolution of 30 m) was post-processed to discard isolated pixels and improve the homogeneity of the classified image.

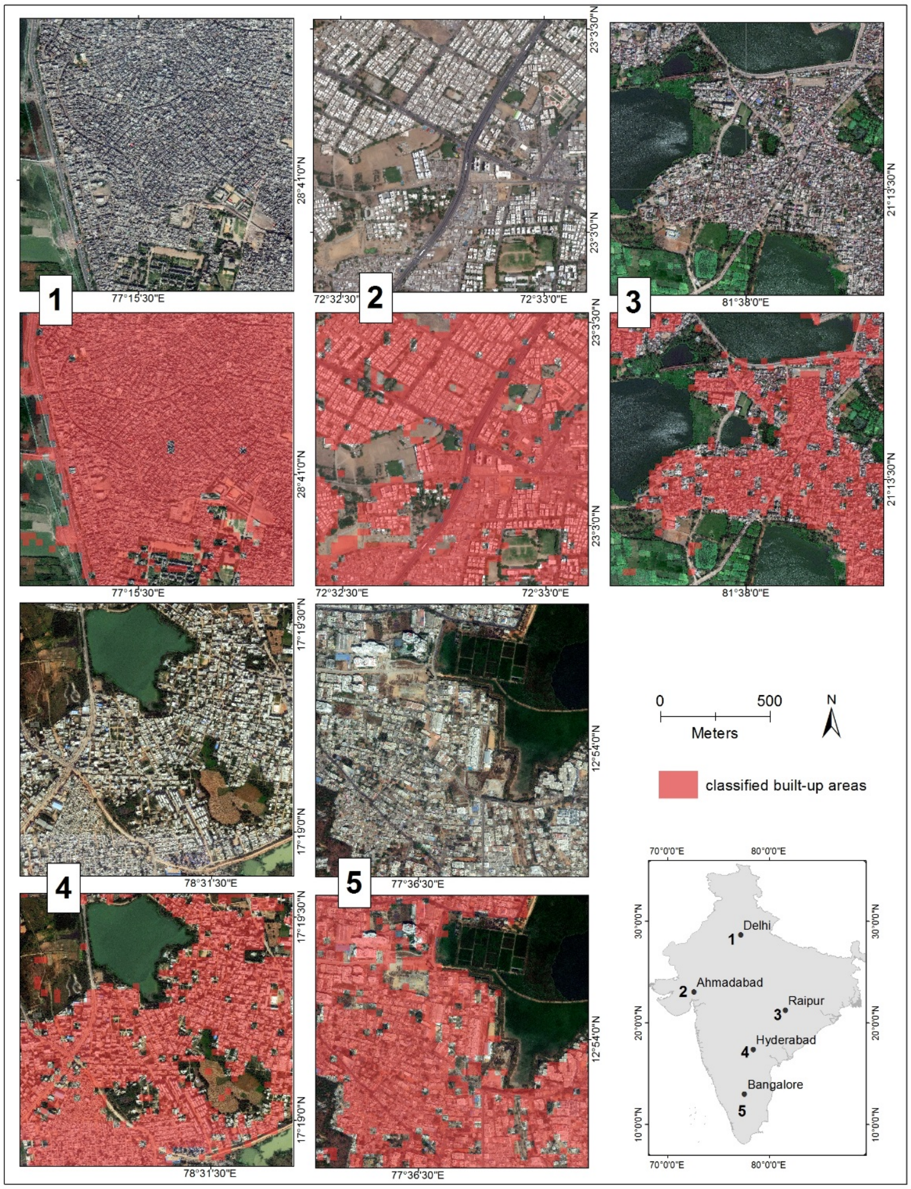
Figure 3.11 presents, as an illustration, a classified image of built-up areas in five regions across India and Figure 3.12 presents examples of this classified image in a finer (higher) resolution (we present a classified image of each site below its corresponding high-resolution satellite image). The classified image captures the fabric of built-up urban areas, as well as the fine boundaries between built-up areas, as well as the fine boundaries between built-up areas and various types of land cover (e.g., vegetation, water bodies and open spaces).

The classifier can also be used to map urban areas across time. Since our ground-truth dataset is collected based on 2014 imagery, using it as a training set with the 2000 imagery may result in "class-noise" (Johnson and Iizuka, 2016) due to mislabeled examples. Thus, we first train the classifier (using Random forest, with 10 trees) with Landsat 7, filtered to 2014 as the input (in addition to per-pixel NDVI and NDBI values). Then, we use the trained classifier to map the extent of urban areas in 2000 (using the same feature space, based on Landsat 7, filtered to 2000) (Figure 3.13 presents examples of the classified image in 2000, and Figure 3.14 presents a comparison between the extent



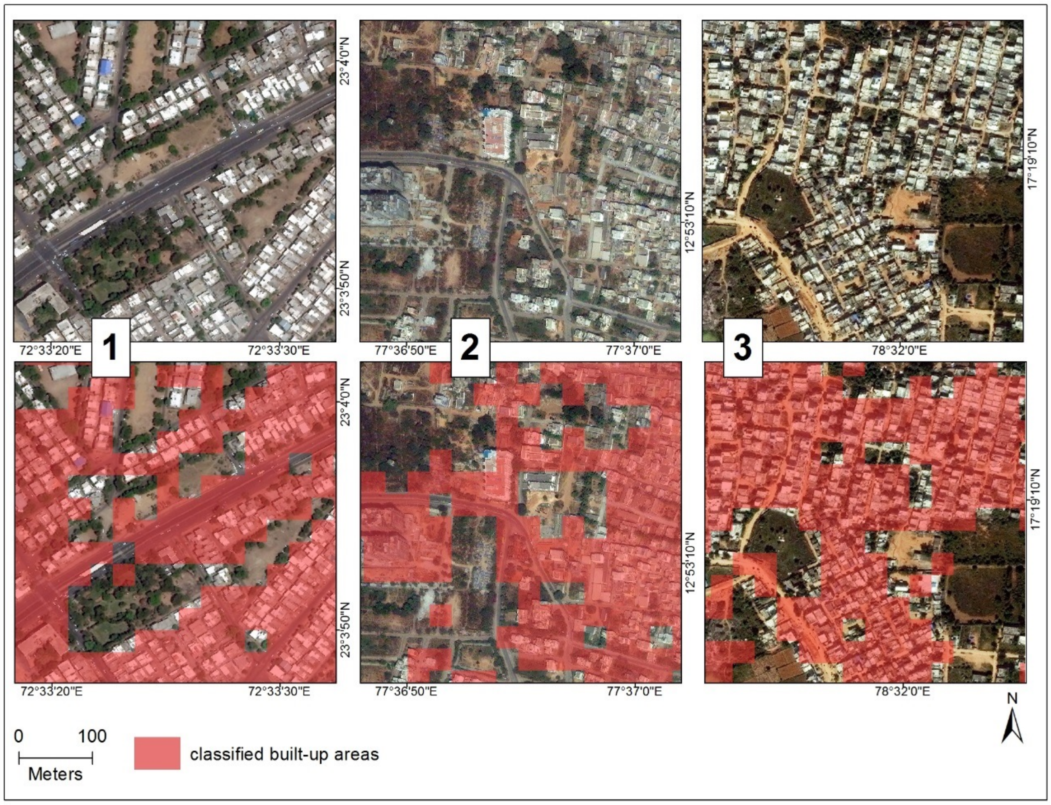
The x-axis represents the total number of examples in the training set (each training set includes 4000 BU examples) (with Landsat 8 as the input)

Figure 3.10: The effect of the training set size on the True-positive rate (TPR), True-negative rate (TNR) and the balanced accuracy.



Classifier: Random forest with 10 trees, with Landsat 8 as the input. Satellite images from DigitalGlobe. Includes copyrighted material of DigitalGlobe, Inc. (Westminster, CO, Canada), All Rights Reserved.

Figure 3.11: Classification of built-up areas (visualized in red) compared to raw satellite images in five regions in India.



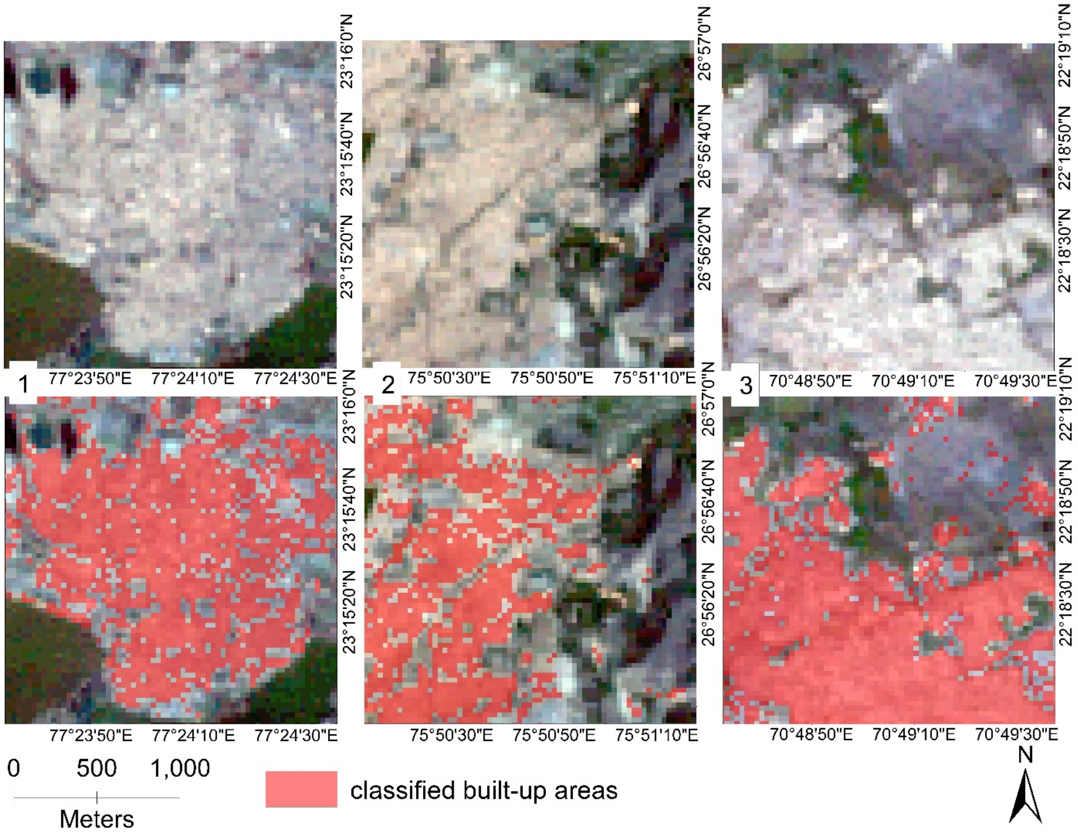
Classifier: Random forest with 10 trees, with Landsat 8 as the input. Satellite images from DigitalGlobe. Includes copyrighted material of DigitalGlobe, Inc. (Westminster, CO, Canada), All Rights Reserved.

Figure 3.12: A detailed examination of the classification of built-up areas (visualized in red) compared to raw satellite images in five regions in India

of urban areas in 2000 and in 2014 in the city of Ahmedabad). As an assessment of this classification, we choose 200 random polygons from our dataset, visually examine them against 2000 Landsat 7 imagery and assign each polygon with a class (BU or NBU). Then, we compare the classified image with this ground-truth dataset. The examination reveals an overall accuracy of 86% and a TPR of 58.6% (Table 3.5 presents a confusion matrix of this test). Finally, Figure 3.15 shows the advantage of using our methodology to map urbanization relative to the WorldPop dataset. The classified image is able to capture various types of LC/LU (e.g., built-up areas, parks and open spaces) that is not possible using estimates of local area populations.

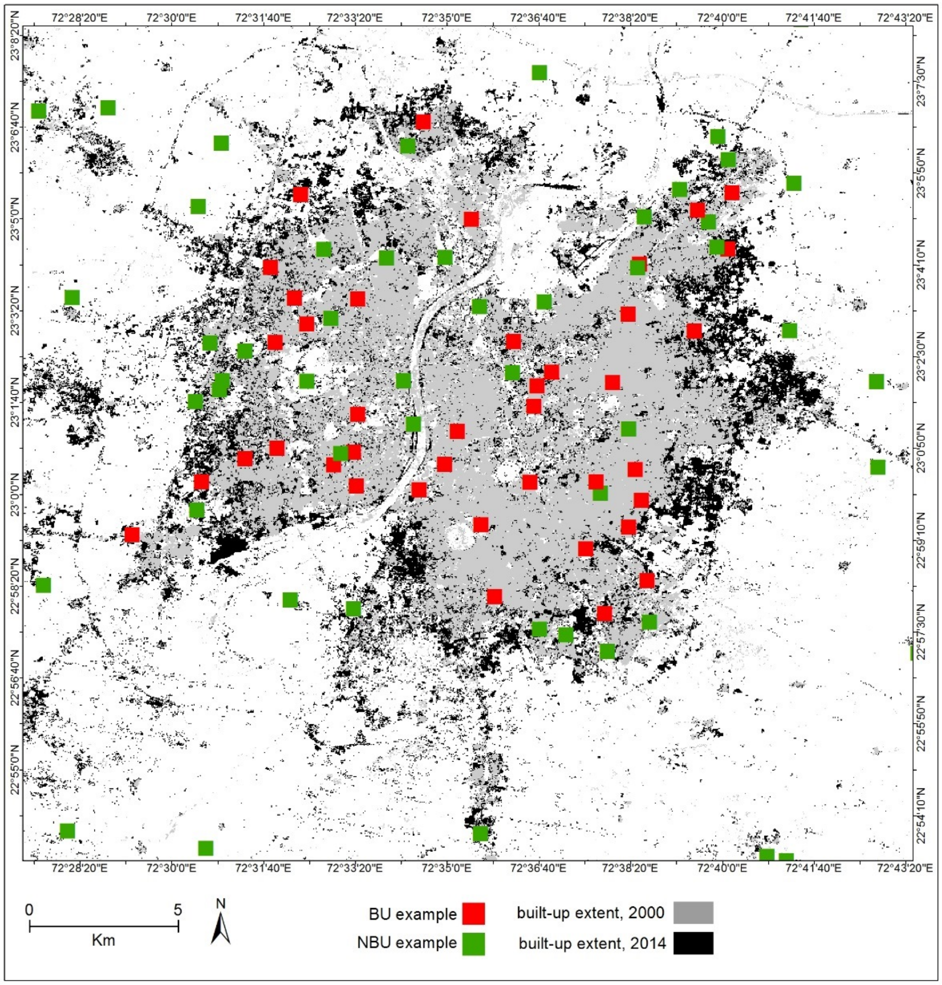
Table 3.5: Confusion Matrix Describing the Classifier's Performance (Detection of Urban Areas in 2000)

		Predicted		
		BU	**NBU**	**Total**
Actual	BU	34	24	58
	NBU	4	138	142
	Total	38	162	200



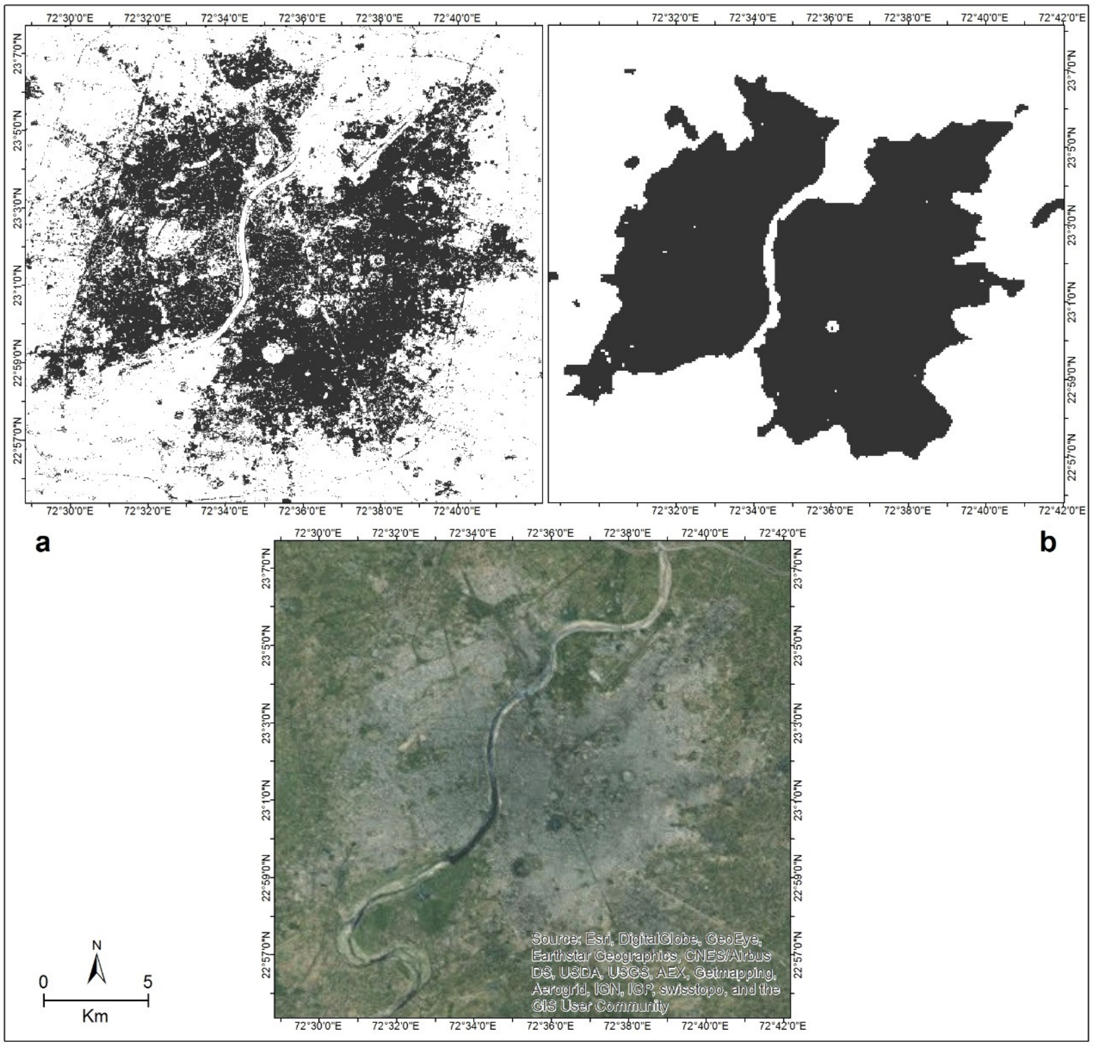
Classifier: Random Forest (10 trees). Input: Landsat 7 (plus NDVI and NDBI). The classifier was trained with Landsat 7 filtered to 2014, and the trained classifier was used to classify Landsat 7 filtered to 2000.

Figure 3.13: Detection of built-up areas in three Indian cities in 2000 - Erode, Visakhapatnam and Nagpur (bottom) - compared to the raw Landsat 7 filtered to 2000 (top).



Classifier: Random Forest (10 trees). Input: Landsat 7 (plus NDVI and NDBI). The classifier was trained with Landsat 7 filtered to 2014; the trained classifier was used to classify Landsat 7 filtered to 2000.

Figure 3.14: Detection of the boundaries of Ahmedabad, India, in 2000 and in 2014, together with built-up (BU) and not built-up (NBU) examples used for the training.



Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Note: population density data from WorldPop: www.worldpop.org.uk. Detection of built-up areas in 2010 was done with Random Forest (10 trees) using Landsat 7 (plus NDVI and NDBI) as the input.

Figure 3.15: Estimation of the boundaries of Ahmedabad, India (in 2010) according to: (a) classification of built-up areas; and (b) population density.

3.5 Discussion

In recent decades, there have been substantial research investments in attempting to understand the social and physical dynamics related to urbanization. Though urbanization is one of the major potential threats to the global environment (Sudhira et al., 2004; Jat et al., 2008), its rate and magnitude have not been quantified with sharp precision at global scale. For many low-income countries, the last significant mapping efforts occurred in the 1960s and 1970s (Tatem et al., 2007). Urban extent can be measured by different means, including population counts, nighttime illumination intensity, and detecting the unique LC/LU characteristics and physical attributes associated with urban areas (Taubenbock, et al., 2012; Dewan and Yamaguchi, 2009; Bhatta, 2009; Orenstein et al., 2011). With the increasing availability of satellite data at ever-improving spatial and temporal resolutions, urban research is rapidly shifting towards the use of image-classification methods designed to extract the “urbanized land” that can be observed and captured in multispectral imagery (Schneider et al., 2010).

Several datasets of urban extent have now been developed to map urban areas at global scale (Gaughan et al., 2013; Potere et al., 2009; Stevens et al., 2015). However, these datasets show considerable disagreement on the location and extent of urban land (Schneider et al., 2010; Miyazaki et al., 2011) and the majority of the existing information provides classified raster images that have limitations across space and time (Bhatta, 2009; Patel et al., 2015; Gislason et al., 2006; Shao and Lunetta, 2012; Wieland and Pittore, 2014) or that use ground-truth data that are limited in size, with no more than several thousand examples (Miyazaki et al., 2011).

With the availability of cloud-based platforms such as GEE, it is now feasible to monitor urbanization in multi-spatial and temporal resolutions and to understand urban dynamics globally. High-resolution ground-truth data are fundamental for any supervised image classification, including classification of built-up land cover. Training data remain scarce, making it difficult to apply modern remote-sensing techniques (Xie et al., 2015). At the current time, ground-truth data lag far behind the ever-growing supplies of satellite imagery and analytical tools for image classification. Though ground-truth labeled data for urban areas can be extracted from several existing datasets-e.g., Landsat-based urban maps and crowd-source-based datasets such as OpenStreetMap (Belgiu and Dragut, 2014)-validated and processed datasets that are designed specifically for mapping urban areas are in scarce supply. This paper aims to fill this gap.

Ground-truth data can be used, in conjunction with high-resolution satellite imagery and cloud-based computational platforms to detect built-up land cover. GEE is a platform with tremendous potential for urban research at scale. With appropriate ground-truth data, GEE can serve as an accessible and feasible platform for image classification and analysis of large and geographically diverse regions. Though GEE has been used in previous studies for various applications, including population (Patel et al., 2015; Trianni et al., 2015) and forest cover (Hansen et al., 2013) mapping, ours is the first to provide comprehensive open-source ground-truth data that can serve as a training set for supervised classification of built-up land cover and for evaluation/validation of existing classifiers and classification products.

Of the three types of classifiers that we examine in GEE (SVM, CART and Random Forest), Random Forest achieves the best performance (a balanced accuracy rate of 80%). This classifier produces high-quality maps of built-up areas across space and time in India. Although the performance of CART and Random Forest are better when Landsat 8 is used as the input than when Landsat 7 is used (due perhaps to the higher spectral resolution of Landsat 8), performance improves substantially when NDVI and NDBI are added to Landsat 7, especially with SVM. Similar to the findings of (Wieland and Pittore, 2014; Qian et al., 2015), performance also improves as the size of the training set increases. Importantly, we find that increasing the size of the training set by expanding the number of NBU examples and holding BU examples fixed leads to marked improvements in accuracy.

We note several limitations of the analysis. First, the dataset was labeled according to 20142015 imagery using a visual-interpretation method, which, by its nature, may be subject to idiosyncratic variation across individuals performing the manual classification. As noted in previous studies (Johnson and Iizuka, 2016), “class-noise” may impact the accuracy of the classification. Second, our analysis is limited to India. Creating manually labeled ground-truth data is expensive and time consuming. However, crowd-sourcing platforms may allow researchers to scaleat low costthe labeling method and to construct larger and more comprehensive ground-truth datasets. Although various methods have been suggested for combining census-based data with satellite imagery (Gaughan et al., 2013) and to extract training data from different sources, such as from nighttime lights (Xie et al., 2015), validation by means of visual interpretation remains inescapable for the maintenance of accurate ground-truth training data. Third, the sampling method was designed to detect the boundaries between built-up areas and their periphery; we primarily sampled examples from highly populated areas and from their adjacent, low-population environs. This approach may create a risk of false-positive detections when classifying distant/remote areas.

3.6 Conclusions

During the past century, many countries, especially in the developing world, have been experiencing rapid urbanization and complex changes in patterns of land cover and land use. Understanding the various ecological, environmental, social and economic impacts of these processes is essential for the preservation of a sustainable human society.

The increasing availability of satellite imagery at different spatial and temporal resolutions has shifted urban research towards the use of digital, multispectral images and the development of remote-sensing image classification methods designed to capture urban land features, such as non-vegetative, human-constructed elements. Though numerous low and medium-scale urban maps have been developed to capture urban land features, these maps are generally limited in their temporal or spatial resolution and cannot be used for analysis of continuous urbanization processes. Moreover, previous studies have generally analyzed urbanization processes over small regions, due in part to computational limitations and the lack of ground-truth data for supervised classification. As parallel computational platforms with much larger storage and capacity become accessible to researchers, it is possible to expand the spatial and temporal units of analysis and to investigate urbanization processes over larger areas and over longer periods of time. Expanding this research frontier creates an urgent need for ground-truth data that can facilitate the development of supervised machine-learning algorithms and enable reliable evaluation and validation.

This paper contributes to this domain by providing ground-truth data that will further efforts to understand the urbanization processes at scale. The dataset we present consists of 21,030 polygons in India that were manually labeled as “built-up” or “not built-up” through a visual interpretation method. Though existing datasets, such as OSM and others, can facilitate supervised classification of urban areas, the majority of these were not developed for this purpose and therefore require further processing and validation. Our large-scale georeferenced dataset was developed to facilitate the detection of urban areas at a national level and to provide a handy and reliable tool for temporal analysis of urban zones and their rural peripheries. Although GEE is steadily evolving as a platform for remote-sensing research at scale, its potential for urban research has not been fully explored. In this study, we highlight the use of GEE for urban research and demonstrate the applicability of our dataset for detection of urban areas in a country with a large population and a diverse land cover. We validate the dataset and show that when used with traditional classifiers available in GEE, the classifiers achieve an overall accuracy rate of around 87%. Our methodology, which is designed to evaluate the spatial generalizability of classifiers, shows that classifier performance is similar when the examples in the training and test sets are sampled from areas with heterogeneous land-cover characteristics. This evaluation procedure is thus suitable for studies that analyze large-scale regions.

Extensions to our approach may improve the classification of urban land cover by modifying the inputs to the classifiers or their dimensions, and by adding additional features to the input's feature space. Incorporating nighttime-light data, socio-economic variables, and physical/geographical characteristics to satellite imagery may offer opportunities to improve the accuracy rate of classifiers. Further extensions to our approach may also include the application of learning algorithms and evaluation with various tuning parameters of the classifiers.

Chapter 3 is coauthored with Ran Goldblatt, Gordon Hanson, and Amit Khandelwal, and in full, is a reprint of the material as it appears in: "Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine", *Remote Sensing*, 2016, 8(8), 634. I thank my coauthors for the permission to use this chapter in my dissertation. The dissertation author, Wei You, was the primary author of this paper.

3.7 Appendix

Description of the classifiers used in this study:

- CART (Classification and Regression Tree) (Elvidge et al., 2004) is a binary decision tree. The classifier recursively examines each examples variables with logical if-then questions in a binary tree structure. Questions are asked at each node of the tree, and each question typically looks at a single input variable. The variables are compared with a predetermined threshold, so that the examples are optimally split into "purer" subsets. The examples are split to an overly large tree until reaching a terminal node (when the nodes have less than a defined number of examples or when further split will result in almost the same outcome). The tree is then pruned back through the creation of a nested sequence of less complex trees. The class is predicted at the terminal node according to the proportion of the classes in the training examples that reached that node.

- SVM (Support Vector Machines) identifies decision boundaries that optimally separate between classes. First, the n input vectors (examples) $S = \{X_1, X_2, \ldots, X_n\}$ are mapped to the output classes by a linear decision function on a (possibly) high-dimensional feature space $F = \{\phi(X_1, X_2, \ldots, X_n)\}$. SVM then optimizes the hyperplane that separates the classes by maximizing the margin between the support vectors of the classes (these are the examples that are closest to the decision surface) (Myint et al., 2011). In this study we used a basic linear SVM.

- Random Forests are tree-based classifiers that include k decision trees (k predictors). When classifying an example, the example variables are run through each of the k tree predictors, and the k predictions are averaged to get a less noisy prediction (by voting on the most popular class). The learning process of the forest involves some level of randomness; each tree is trained over an independently random sample of examples from the training set and each node's binary question in a tree is selected from a randomly sampled subset of the input variables (Tatem et al., 2007).

Bibliography

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson, "Cultural Assimilation during the Age of Mass Migration," *Working Paper*, 2016.

_ , Leah Platt Boustan, and Katherine Eriksson, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, May 2012, *102* (5), 1832–1856.

_ , _ , and _ , "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, June 2014, *122* (3), 467–506.

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf, "The Economics of Density: Evidence From the Berlin Wall," *Econometrica*, November 2015, *83* (6), 2127–2189.

Arlot, Sylvain and Alain Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, 2010, *4*, 40–79.

Atack, Jeremy, "Returns to Scale in Antebellum United States Manufacturing," *Explorations in Economic History*, November 1977, *14* (4), 337–359.

_ , "Industrial Structure and the Emergence of the Modern Industrial Corporation," *Explorations in Economic History*, January 1985, *22* (1), 29–52.

_ , "Firm Size and Industrial Structure in the United States During the Nineteenth Century," *The Journal of Economic History*, June 1986, *46* (02), 463–475.

_ , Fred Bateman, and Robert A. Margo, "Steam power, Establishment Size, and Labor Productivity Growth in Nineteenth Century American Manufacturing," *Explorations in Economic History*, April 2008, *45* (2), 185–198.

_ , Matthew Jaremski, and Peter L. Rousseau, "American Banking and the Transportation Revolution before the Civil War," *The Journal of Economic History*, December 2014, *74* (04), 943–986.

Bartel, Ann P., "Where Do the New U.S. Immigrants Live?," *Journal of Labor Economics*, October 1989, *7* (4), 371–391.

Bauer, Thomas, Gil S. Epstein, and Ira N. Gang, "Enclaves, language, and the location choice of migrants," *Journal of Population Economics*, November 2005, *18* (4), 649–662.

Baum-Snow, Nathaniel, "Did Highways Cause Suburbanization?," *The Quarterly Journal of Economics*, 2007, *122* (2), 775–805.

Bayer, Patrick, Robert McMillan, and Kim S. Rueben, "What drives racial segregation? New evidence using Census microdata," *Journal of Urban Economics*, November 2004, *56* (3), 514–535.

_ **, Stephen L. Ross, and Giorgio Topa**, "Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes," *Journal of Political Economy*, December 2008, *116* (6), 1150–1196.

Beaman, Lori A., "Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S.," *The Review of Economic Studies*, January 2012, *79* (1), 128–161.

Belgiu, Mariana and Lucian Dragu, "Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, October 2014, *96*, 67–75.

Bernard, Andrew B., Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum, "Plants and Productivity in International Trade," *The American Economic Review*, September 2003, *93* (4), 1268–1290.

Bhaskaran, Sunil, Shanka Paramananda, and Maria Ramnarayan, "Per-pixel and object-oriented classification methods for mapping urban features using Ikonos satellite data," *Applied Geography*, December 2010, *30* (4), 650–665.

Bhatta, B., S. Saraswati, and D. Bandyopadhyay, "Urban sprawl measurement from remote sensing data," *Applied Geography*, December 2010, *30* (4), 731–740.

Bhatta, Basu, "Analysis of urban growth pattern using remote sensing and GIS: a case study of Kolkata, India," *International Journal of Remote Sensing*, August 2009, *30* (18), 4733–4746.

Billings, Stephen B., "Estimating the Value of a New Transit Option," *Regional Science and Urban Economics*, November 2011, *41* (6), 525–536.

Bleakley, Hoyt and Aimee Chin, "Age at Arrival, English Proficiency, and Social Assimilation among US Immigrants," *American Economic Journal: Applied Economics*, January 2010, *2* (1), 165–192.

Blum, Avrim, Adam Kalai, and John Langford, “Beating the Hold-out: Bounds for K-fold and Progressive Cross-validation,” in “Proceedings of the Twelfth Annual Conference on Computational Learning Theory” COLT ’99 ACM New York, NY, USA 1999, pp. 203–208.

Blumenberg, Evelyn, “Moving in and moving around: immigrants, travel behavior, and implications for transport policy,” *Transportation Letters*, April 2009, *1* (2), 169–180.

Borjas, George J., “Ethnicity, Neighborhoods, and Human-Capital Externalities,” *The American Economic Review*, 1995, *85* (3), 365–390.

_ , “To Ghetto or Not to Ghetto: Ethnicity and Residential Segregation,” *Journal of Urban Economics*, September 1998, *44* (2), 228–253.

Bowes, David R. and Keith R. Ihlanfeldt, “Identifying the Impacts of Rail Transit Stations on Residential Property Values,” *Journal of Urban Economics*, July 2001, *50* (1), 1–25.

Bradford, Jeffrey P. and Carla E. Brodley, “The Effect of Instance-Space Partition on Significance,” *Machine Learning*, March 2001, *42* (3), 269–286.

Breiman, Leo, *Classification and Regression Trees*, Wadsworth/Thomson Learning, 1984.

_ , “Random Forests,” *Machine Learning*, October 2001, *45* (1), 5–32.

Buhaug, Halvard and Henrik Urdal, “An urbanization bomb? Population growth and social disorder in cities,” *Global Environmental Change*, February 2013, *23* (1), 1–10.

Carneiro, Pedro, Sokbae Lee, and Hugo Reis, “Please Call Me John: Name Choice and the Assimilation of Immigrants in the United States, 1900-1930,” *Working Paper*, 2016.

Cervero, Robert and Michael Duncan, “Transit’s Value-Added Effects: Light and Commuter Rail Services and Commercial Land Values,” *Transportation Research Record: Journal of the Transportation Research Board*, January 2002, *1805*, 8–15.

Chadchan, J. and R. Shankar, “An analysis of urban growth trends in the post-economic reforms period in India,” *International Journal of Sustainable Built Environment*, June 2012, *1* (1), 36–49.

Chandler, Alfred D., *The Visible Hand: The Managerial Revolution in American Business*, Harvard University Press, 1977.

_ **and Takashi Hikino**, *Scale and Scope: The Dynamics of Industrial Capitalism*, Belknap Press, January 1990.

Chiswick, Barry R., "Speaking, Reading, and Earnings among Low-Skilled Immigrants," *Journal of Labor Economics*, April 1991, *9* (2), 149–170.

_ **and Paul W. Miller**, "Do Enclaves Matter in Immigrant Adjustment?," *City & Community*, March 2005, *4* (1), 5–35.

Cortes, Corinna and Vladimir Vapnik, "Support-vector networks," *Machine Learning*, September 1995, *20* (3), 273–297.

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor, "The Rise and Decline of the American Ghetto," *Journal of Political Economy*, June 1999, *107* (3), 455–506.

_ , _ , and _ , "Is the Melting Pot Still Hot? Explaining the Resurgence of Immigrant Segregation," *The Review of Economics and Statistics*, July 2008, *90* (3), 478–497.

_ , _ , and _ , "When are ghettos bad? Lessons from immigrant segregation in the United States," *Journal of Urban Economics*, May 2008, *63* (3), 759–774.

Dahly, Darren L. and Linda S. Adair, "Quantifying the urban environment: A scale measure of urbanicity outperforms the urbanrural dichotomy," *Social Science & Medicine*, April 2007, *64* (7), 1407–1419.

Damm, Anna Piil, "Ethnic Enclaves and Immigrant Labor Market Outcomes: Quasi-Experimental Evidence," *Journal of Labor Economics*, April 2009, *27* (2), 281–314.

_ , "Neighborhood quality and labor market outcomes: Evidence from quasi-random neighborhood assignment of immigrants," *Journal of Urban Economics*, January 2014, *79*, 139–166.

DeSoto, Hernando, *The Other Path: the Invisible Revolution in the Third World*, Harper & Row, 1989.

Dewan, Ashraf M. and Yasushi Yamaguchi, "Land use and land cover change in Greater Dhaka, Bangladesh: Using remote sensing to promote sustainable urbanization," *Applied Geography*, July 2009, *29* (3), 390–401.

Donaldson, Dave and Richard Hornbeck, "Railroads and American Economic Growth: A Market Access Approach," *The Quarterly Journal of Economics*, February 2016.

Duro, Dennis C., Steven E. Franklin, and Monique G. Dub, "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery," *Remote Sensing of Environment*, March 2012, *118*, 259–272.

Edin, Per-Anders, Peter Fredriksson, and Olof slund, "Ethnic Enclaves and the Economic Success of ImmigrantsEvidence from a Natural Experiment," *The Quarterly Journal of Economics*, February 2003, *118* (1), 329–357.

_ , _ , and _ , “Settlement policies and the economic success of immigrants,” *Journal of Population Economics*, February 2004, *17* (1), 133–155.

Elvidge, ChristopherD, Jeffrey Safran, IngridL Nelson, BenjaminT Tuttle, Vinita Ruth Hobson, KimberlyE Baugh, JohnB Dietz, and EdwardH Erwin, “Area and Positional Accuracy of DMSP Nighttime Lights Data,” in “Remote Sensing and GIS Accuracy Assessment,” CRC Press, July 2004, pp. 281–292.

Estima, Jacinto and Marco Painho, “Investigating the Potential of OpenStreetMap for Land Use/Land Cover Production: A Case Study for Continental Portugal,” in Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney, and Marco Helbich, eds., *OpenStreetMap in GIScience*, Lecture Notes in Geoinformation and Cartography, Springer International Publishing, 2015, pp. 273–293.

Frenkel, Amnon and Maya Ashkenazi, “Measuring Urban Sprawl: How Can We Deal with It?,” *Environment and Planning B: Planning and Design*, February 2008, *35* (1), 56–79.

Gamba, Paolo and Martin Herold, *Global Mapping of Human Settlement: Experiences, Datasets, and Prospects*, CRC Press, June 2009.

Gaughan, Andrea E., Forrest R. Stevens, Catherine Linard, Peng Jia, and Andrew J. Tatem, “High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015,” *PLOS ONE*, February 2013, *8* (2), e55882.

Gibbons, Stephen and Stephen Machin, “Valuing Rail Access Using Transport Innovations,” *Journal of Urban Economics*, January 2005, *57* (1), 148–169.

Gislason, Pall Oskar, Jon Atli Benediktsson, and Johannes R. Sveinsson, “Random Forests for land cover classification,” *Pattern Recognition Letters*, March 2006, *27* (4), 294–300.

Glaeser, Edward L., “A World of Cities: The Causes and Consequences of Urbanization in Poorer Countries,” *Journal of the European Economic Association*, October 2014, *12* (5), 1154–1199.

_ , Matthew E. Kahn, and Jordan Rappaport, “Why do the poor live in cities? The role of public transportation,” *Journal of Urban Economics*, January 2008, *63* (1), 1–24.

Gollin, Douglas, “Nobody’s Business but My Own: Self-Employment and Small Enterprise in Economic Development,” *Journal of Monetary Economics*, March 2008, *55* (2), 219–233.

Grnqvist, Hans, “Ethnic Enclaves and the Attainments of Immigrant Children,” *European Sociological Review*, September 2006, *22* (4), 369–382.

Guerra, Erick, Robert Cervero, and Daniel Tischler, "Half-Mile Circle: Does It Best Represent Transit Station Catchments?," *Transportation Research Record: Journal of the Transportation Research Board*, December 2012, *2276*, 101–109.

Gururaja, K. V. and H. S. Sudhira, "Population crunch in India: is it urban or still rural?," *Current Science*, July 2012, *103* (1), 37–40.

Hanes, Christopher, "Immigrants Relative Rate of Wage Growth in the Late 19th Century," *Explorations in Economic History*, January 1996, *33* (1), 35–64.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend, "High-Resolution Global Maps of 21st-Century Forest Cover Change," *Science*, November 2013, *342* (6160), 850–853.

Harris, John R. and Michael P. Todaro, "Migration, Unemployment and Development: A Two-Sector Analysis," *The American Economic Review*, 1970, *60* (1), 126–142.

Hatton, Timothy J., "The Immigrant Assimilation Puzzle in Late Nineteenth-Century America," *The Journal of Economic History*, 1997, *57* (1), 34–62.

Hellerstein, Judith K., Melissa McInerney, and David Neumark, "Neighbors and Coworkers: The Importance of Residential Labor Market Networks," *Journal of Labor Economics*, October 2011, *29* (4), 659–695.

Herold, Martin, Noah C. Goldstein, and Keith C. Clarke, "The spatiotemporal form of urban growth: measurement, analysis and modeling," *Remote Sensing of Environment*, August 2003, *86* (3), 286–302.

Hewitt, Christopher and W. E. Hewitt, "The Effect of Proximity to Urban Rail on Housing Prices in Ottawa," *Journal of Public Transportation*, December 2012, *15* (4).

Holmes, Thomas J. and John J. Stevens, "An Alternative Theory of the Plant Size Distribution, with Geography and Intra- and International Trade," *Journal of Political Economy*, 2014, *122* (2), 369–421.

Hornbeck, Richard, "Barbed Wire: Property Rights and Agricultural Development," *The Quarterly Journal of Economics*, May 2010, *125* (2), 767–810.

_ **and Daniel Keniston**, "Creative Destruction: Barriers to Urban Growth and the Great Boston Fire of 1872," *Working Paper*, 2016.

Hsieh, Chang-Tai and Benjamin A. Olken, "The Missing Missing Middle," *The Journal of Economic Perspectives*, July 2014, *28* (3), 89–108.

_ **and Peter J. Klenow**, "Misallocation and Manufacturing TFP in China and India," *The Quarterly Journal of Economics*, November 2009, *124* (4), 1403–1448.

_ and _, "The Life Cycle of Plants in India and Mexico," *The Quarterly Journal of Economics*, May 2014, p. qju014.

Jat, Mahesh Kumar, P. K. Garg, and Deepak Khare, "Monitoring and modelling of urban sprawl using remote sensing and GIS techniques," *International Journal of Applied Earth Observation and Geoinformation*, February 2008, *10* (1), 26–43.

Jensen, Anders, "Employment Structure and the Rise of the Modern Tax System," *Working Paper*, 2016.

Johnson, Brian A. and Kotaro Iizuka, "Integrating OpenStreetMap crowdsourced data and Landsat time-series imagery for rapid land use/land cover (LULC) mapping: Case study of the Laguna de Bay area of the Philippines," *Applied Geography*, February 2016, *67*, 140–149.

Kahn, Matthew E., "Gentrification Trends in New Transit-Oriented Communities: Evidence from 14 Cities That Expanded and Built Rail Transit Systems," *Real Estate Economics*, June 2007, *35* (2), 155–182.

Ko, Kate and Xinyu Cao, "The Impact of Hiawatha Light Rail on Commercial and Industrial Property Values in Minneapolis," *Journal of Public Transportation*, March 2013, *16* (1).

Krugman, Paul, "Increasing Returns and Economic Geography," *Journal of Political Economy*, June 1991, *99* (3), 483–499.

Lagakos, David, "Explaining Cross-Country Productivity Differences in Retail Trade," *Journal of Political Economy*, March 2016, *124* (2), 579–620.

LaPorta, Rafael and Andrei Shleifer, "The Unofficial Economy and Economic Development," *Brookings Papers on Economic Activity*, 2008, *2008*, 275–352.

_ and _, "Informality and Development," *The Journal of Economic Perspectives*, July 2014, *28* (3), 109–126.

Lazear, Edward P., "Culture and Language," *Journal of Political Economy*, December 1999, *107* (S6), S95–S126.

Levy, Santiago, *Good Intentions, Bad Outcomes: Social Policy, Informality, and Economic Growth in Mexico*, Brookings Institution Press, 2008.

Lewis, W. Arthur, "Economic Development with Unlimited Supplies of Labour," *The Manchester School*, May 1954, *22* (2), 139–191.

Logan, John R., Weiwei Zhang, and Miao David Chunyu, "Emergent Ghettos: Black Neighborhoods in New York and Chicago, 18801940," *American Journal of Sociology*, January 2015, *120* (4), 1055–1094.

Lucas, Robert E., "On the Size Distribution of Business Firms," *The Bell Journal of Economics*, 1978, *9* (2), 508–523.

Margo, Robert A., "Economies of Scale in Nineteenth Century American Manufacturing Revisited: A Resolution of the Entrepreneurial Labor Input Problem," *NBER Working Paper*, 2013.

McCaig, Brian and Nina Pavcnik, "Informal Employment in a Growing and Globalizing Low-Income Country," *The American Economic Review*, May 2015, *105* (5), 545–550.

McKenzie, David, "Identifying and Spurring High-Growth Entrepreneurship: Experimental Evidence from a Business Plan Competition," *Working Paper*, 2016.

McKINNEY, Michael L., "Urbanization, Biodiversity, and Conservation," *BioScience*, October 2002, *52* (10), 883–890.

Melitz, Marc J., "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, November 2003, *71* (6), 1695–1725.

Michaels, Guy, "The Effect of Trade on the Demand for Skill: Evidence from the Interstate Highway System," *Review of Economics and Statistics*, October 2008, *90* (4), 683–701.

Minns, Chris, "Income, Cohort Effects, and Occupational Mobility: A New Look at Immigration to the United States at the Turn of the 20th Century," *Explorations in Economic History*, October 2000, *37* (4), 326–350.

Miyazaki, Hiroyuki, Koki Iwao, and Ryosuke Shibasaki, "Development of a New Ground Truth Database for Global Urban Area Mapping from a Gazetteer," *Remote Sensing*, June 2011, *3* (6), 1177–1187.

Moghadam, Hossein Shafizadeh and Marco Helbich, "Spatiotemporal urbanization processes in the megacity of Mumbai, India: A Markov chains-cellular automata urban growth model," *Applied Geography*, June 2013, *40*, 140–149.

Morris, Eric, "From Horse Power to Horsepower," *Access*, 2007.

Most, Doug, *The Race Underground: Boston, New York, and the Incredible Rivalry That Built America's First Subway*, Macmillan, February 2014.

Munshi, Kaivan, "Networks in the Modern Economy: Mexican Migrants in the U. S. Labor Market," *The Quarterly Journal of Economics*, May 2003, *118* (2), 549–599.

Myint, Soe W., Patricia Gober, Anthony Brazel, Susanne Grossman-Clarke, and Qihao Weng, "Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery," *Remote Sensing of Environment*, May 2011, *115* (5), 1145–1161.

Orenstein, Daniel E., Bethany A. Bradley, Jeff Albert, John F. Mustard, and Steven P. Hamburg, "How much is built? Quantifying and interpreting patterns of built space from different data sources," *International Journal of Remote Sensing*, May 2011, *32* (9), 2621–2644.

Palmer-Jones, Richard and Kunal Sen, "What has luck got to do with it? A regional analysis of poverty and agricultural growth in rural India," *The Journal of Development Studies*, October 2003, *40* (1), 1–31.

Patacchini, Eleonora and Yves Zenou, "Ethnic networks and employment outcomes," *Regional Science and Urban Economics*, November 2012, *42* (6), 938–949.

Patel, Nirav N., Emanuele Angiuli, Paolo Gamba, Andrea Gaughan, Gianni Lisini, Forrest R. Stevens, Andrew J. Tatem, and Giovanna Trianni, "Multitemporal settlement and population mapping from Landsat using Google Earth Engine," *International Journal of Applied Earth Observation and Geoinformation*, March 2015, *35, Part B*, 199–208.

Pettorelli, Nathalie, Jon Olav Vik, Atle Mysterud, Jean-Michel Gaillard, Compton J. Tucker, and Nils Chr. Stenseth, "Using the satellite-derived NDVI to assess ecological responses to environmental change," *Trends in Ecology & Evolution*, September 2005, *20* (9), 503–510.

Potere, David, Annemarie Schneider, Shlomo Angel, and Daniel L. Civco, "Mapping urban areas on a global scale: which of the eight maps now available is more accurate?," *International Journal of Remote Sensing*, November 2009, *30* (24), 6531–6558.

Publications, United Nations, *World Urbanization Prospects 2014: Highlights*, United Nations Environment Programme, August 2014.

Pugh, Cedric, *Sustainability the Environment and Urbanisation*, Routledge, January 2014.

Qian, Yuguo, Weiqi Zhou, Jingli Yan, Weifeng Li, and Lijian Han, "Comparing Machine Learning Classifiers for Object-Based Land Cover Classification Using Very High Resolution Imagery," *Remote Sensing*, December 2014, *7* (1), 153–168.

Rahman, Atiqur, Shiv Prashad Aggarwal, Maik Netzband, and Shahab Fazal, "Monitoring Urban Sprawl Using Remote Sensing and GIS Techniques of a Fast Growing Urban Centre, India," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, March 2011, *4* (1), 56–64.

Ramachandra, TV, Bharath H. Aithal, and Durgappa D. Sanna, "Insights to urban dynamics through landscape spatial pattern analysis," *International Journal of Applied Earth Observation and Geoinformation*, August 2012, *18*, 329–343.

Rauch, James E., “Modelling the Informal Sector Formally,” *Journal of Development Economics*, January 1991, *35* (1), 33–47.

Redding, Stephen J. and Matthew A. Turner, “Transportation Costs and the Spatial Organization of Economic Activity,” *NBER Working Paper*, 2014.

Refaeilzadeh, Payam, Lei Tang, and Huan Liu, “Cross-Validation,” in LING LIU and M. TAMER ZSU, eds., *Encyclopedia of Database Systems*, Springer US, 2009, pp. 532–538.

Robertson, Laura Dingle and Douglas J. King, “Comparison of pixel- and object-based classification in land cover change mapping,” *International Journal of Remote Sensing*, March 2011, *32* (6), 1505–1529.

Rodriguez-Galiano, V. F., B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, “An assessment of the effectiveness of a random forest classifier for land-cover classification,” *ISPRS Journal of Photogrammetry and Remote Sensing*, January 2012, *67*, 93–104.

Rodriguez, Juan D., Aritz Perez, and Jose A. Lozano, “Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, March 2010, *32* (3), 569–575.

Ryan, Sherry, “The Value of Access to Highways and Light Rail Transit: Evidence for Industrial and Office Firms,” *Urban Studies*, April 2005, *42* (4), 751–764.

Salzberg, Steven L., “On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach,” *Data Mining and Knowledge Discovery*, September 1997, *1* (3), 317–328.

Sarada and Nicolas L. Ziebarth, “Distance and the Informativeness of Credit Ratings,” *Working Paper*, 2015.

Schlesinger, Johannes, “Using Crowd-Sourced Data to Quantify the Complex Urban FabricOpenStreetMap and the UrbanRural Index,” in Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney, and Marco Helbich, eds., *OpenStreetMap in GIScience*, Lecture Notes in Geoinformation and Cartography, Springer International Publishing, 2015, pp. 295–315.

Schneider, Annemarie, “Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach,” *Remote Sensing of Environment*, September 2012, *124*, 689–704.

_ , Mark A. Friedl, and David Potere, “Mapping global urban areas using MODIS 500-m data: New methods and datasets based on urban ecoregions,” *Remote Sensing of Environment*, August 2010, *114* (8), 1733–1746.

Sequeira, Sandra, Nathan Nunn, and Nancy Qian, "Migrants and the Making of America: The Short- and Long-Run Effects of Immigration during the Age of Mass Migration," *Working Paper*, 2017.

Seto, Karen C., Burak Gneralp, and Lucy R. Hutyra, "Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools," *Proceedings of the National Academy of Sciences*, October 2012, *109* (40), 16083–16088.

Shao, Yang and Ross S. Lunetta, "Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points," *ISPRS Journal of Photogrammetry and Remote Sensing*, June 2012, *70*, 78–87.

Sharma, Richa and P. K. Joshi, "Monitoring Urban Landscape Dynamics Over Delhi (India) Using Remote Sensing (19982011) Inputs," *Journal of the Indian Society of Remote Sensing*, September 2013, *41* (3), 641–650.

Shertzer, Allison, Randall P. Walsh, and John R. Logan, "Segregation and neighborhood change in northern cities: New historical GIS data from 19001930," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, October 2016, *49* (4), 187–197.

_ , Tate Twinam, and Randall P. Walsh, "Race, Ethnicity, and Discriminatory Zoning," *American Economic Journal: Applied Economics*, July 2016, *8* (3), 217–246.

Siodla, James, "Razing San Francisco: The 1906 Disaster as a Natural Experiment in Urban Redevelopment," *Journal of Urban Economics*, September 2015, *89*, 48–61.

Slemrod, Joel, "Cheating Ourselves: The Economics of Tax Evasion," *Journal of Economic Perspectives*, March 2007, *21* (1), 25–48.

Small, Christopher, Francesca Pozzi, and C. D. Elvidge, "Spatial analysis of global urban extent from DMSP-OLS night lights," *Remote Sensing of Environment*, June 2005, *96* (34), 277–291.

Sokoloff, Kenneth L., "Was the Transition from the Artisanal Shop to the Nonmechanized Factory Associated with Gains in Efficiency?: Evidence from the U.S. Manufacturing Censuses of 1820 and 1850," *Explorations in Economic History*, October 1984, *21* (4), 351–382.

Stevens, Forrest R., Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem, "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data," *PLOS ONE*, February 2015, *10* (2), e0107042.

Sudhira, H. S., T. V. Ramachandra, and K. S. Jagadish, "Urban sprawl: metrics, dynamics and modelling using GIS," *International Journal of Applied Earth Observation and Geoinformation*, February 2004, *5* (1), 29–39.

Tatem, Andrew J., Abdisalan M. Noor, Craig von Hagen, Antonio Di Gregorio, and Simon I. Hay, "High Resolution Population Maps for Low Income Nations: Combining Land Cover and Census in East Africa," *PLOS ONE*, December 2007, *2* (12), e1298.

Taubenbck, Hannes, Thomas Esch, Andreas Felbier, Michael Wiesner, Achim Roth, and Stefan Dech, "Monitoring urbanization in mega cities from space," *Remote Sensing of Environment*, February 2012, *117*, 162–176.

Trianni, G., G. Lisini, E. Angiuli, E. A. Moreno, P. Dondi, A. Gaggia, and P. Gamba, "Scaling up to National/Regional Urban Extent Mapping Using Landsat Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, July 2015, *8* (7), 3710–3719.

Tybout, James R., "Manufacturing Firms in Developing Countries: How Well Do They Do, and Why?," *Journal of Economic Literature*, 2000, *38* (1), 11–44.

Vessali, Kaveh V., "Land Use Impacts of Rapid Transit: A Review of the Empirical Literature," *Berkeley Planning Journal*, January 1996, *11* (1).

Vuchic, Vukan R., *Urban Transit Systems and Technology*, John Wiley & Sons, February 2007.

Warner, Sam B., *Streetcar Suburbs: the Process of Growth in Boston, 1870 - 1900*, Harvard University Press and The M.I.T. Press, 1962.

Warner, Sam Bass, *Streetcar Suburbs: The Process of Growth in Boston, 1870-1960*, Harvard University Press, 1962.

Whiteside, Timothy G., Guy S. Boggs, and Stefan W. Maier, "Comparing object-based and pixel-based classifications for mapping savannas," *International Journal of Applied Earth Observation and Geoinformation*, December 2011, *13* (6), 884–893.

Wieland, Marc and Massimiliano Pittore, "Performance Evaluation of Machine Learning Algorithms for Urban Pattern Recognition from Multi-spectral Satellite Images," *Remote Sensing*, March 2014, *6* (4), 2912–2939.

Xie, Michael, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon, "Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping," *arXiv:1510.00098 [cs]*, September 2015.

ya Wu, Kai, Xin yue Ye, Zhi fang Qi, and Hao Zhang, "Impacts of land use/land cover change and socioeconomic development on regional ecosystem services: The case of fast-growing Hangzhou metropolitan area, China," *Cities*, April 2013, *31*, 276–284.

Yue, Wenze, Yong Liu, and Peilei Fan, "Measuring urban sprawl and its drivers in large Chinese cities: The case of Hangzhou," *Land Use Policy*, March 2013, *31*, 358–370.

Zha, Y., J. Gao, and S. Ni, “Use of normalized difference built-up index in automatically mapping urban areas from TM imagery,” *International Journal of Remote Sensing*, January 2003, *24* (3), 583–594.

Ziv, Oren, “Geography in Reduced Form,” *Working Paper*, 2016.