ESSAYS ON DATA, INFORMATION, AND INVESTMENT DECISIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Feng Chi

December 2025

ESSAYS ON DATA, INFORMATION, AND INVESTMENT DECISIONS

Feng Chi, Ph.D.

Cornell University 2025

Information has always been central to economic decision-making, yet measuring it empirically remains a significant challenge. This dissertation explores innovative ways to capture and quantify the impact of information on decision-making for firms and investors. This thesis is composed of three essays.

The first chapter, "Information Waves and Firm Investment," examines how information quality affects firm investment. This research utilizes the predetermined decennial schedule of the U.S. Census as an exogenous source of variation in information quality available to firms when making entry decisions. As the decade progresses, census data becomes gradually outdated and fails to capture ongoing demographic changes in local markets. Using establishment-level data from the retail and restaurant industries, the analysis shows that deteriorating information quality leads to worse investment decisions, significantly increasing establishment failure rates. This effect is more pronounced in areas with large demographic changes, in industries reliant on precise local data, and for smaller firms, underscoring the role of information in shaping firm outcomes.

The second chapter, "The Use and Usefulness of Big Data in Finance: Evidence from Financial Analysts," co-authored with Byoung-Hyoun Hwang and Yaping Zheng, investigates the influence of non-traditional data such as web traffic, credit card transactions, and satellite imagery on financial markets. This study focuses on sell-side analysts as key information intermediaries. Anal-

ysis of analyst reports reveals an increasing incorporation of alternative data, which is linked to more accurate earnings forecasts and higher trading commissions for brokerages. Furthermore, the use of alternative data is found to help narrow the performance gap between hedge funds and traditional institutional investors.

The third chapter, "A Tax-Shaped Retail Landscape," co-authored with Limin Fang, Mengwei Lin, and Nathan Yang, examines how state business tax policies affect market structure. Using data on retail establishments, the analysis shows that large retail chains are disproportionately more responsive to tax changes. A model of strategic entry demonstrates how uniform tax policies can unintentionally amplify market dominance, leading to asymmetric competitive effects.

BIOGRAPHICAL SKETCH

Feng Chi received her bachelor's degree in Finance from Renmin University of China and her master's degree in Applied Economics and Management from Cornell University.

To my daughter Yvonne

ACKNOWLEDGEMENTS

I am deeply grateful to my advisors for their guidance and support throughout this journey. I would like to thank my committee chair Justin Murfin for his thoughtful feedback and guidance. I also thank David Ng for his generous mentorship and interest in my work. I am especially grateful to Shanjun Li, whose insight, encouragement, and thoughtful guidance were instrumental in shaping my job market paper. I owe special thanks to Byoung-Hyoun Hwang. His mentorship, kindness, and encouragement have profoundly influenced both this dissertation and my development as a researcher.

I am also grateful to the many faculty members at Cornell I have had the privilege to learn from. In particular, I thank Ryan Chahrour, Lawrence Jin and Ben Layden for their advice and continued support, which have greatly contributed to both my research and my professional development.

I also deeply appreciate the support and encouragement of my friends, both at and beyond the Cornell community. I thank Yaping Zheng, my coauthor and close friend, for our countless discussions and the many late nights we spent working together. Her insight, perseverance, and friendship have made this journey more meaningful.

Finally, I thank my family for their love and unwavering support. I am especially grateful to my husband, Nathan, for his constant support and faith in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INFORMATION WAVES AND FIRM INVESTMENT

1.1 Introduction

Information is a critical input for firm decisions, yet empirically assessing its impact on economic outcomes presents significant challenges. Firms make endogenous decisions to acquire information, and the quality of this information is often difficult to observe. To overcome these challenges and quantify the link between information and firm investment outcomes, this paper uses the predetermined release schedule of the U.S. Census data as an exogenous source of variation in information quality.

Despite its public-sector origins, census data has become a vital resource for private business decisions.[1] A 1990 lead article from *The Washington Post* describes the census as "the private sector's most comprehensive planning and marketing tool," highlighting its "unmatched breath and depth as a roadmap to who and where consumers are" (Farhi, 1990). Firms rely on these demographic patterns to evaluate market potential, forecast demand, and inform investment decisions. While some firms access census data directly, many use demographic information supplied by intermediaries, such as data analytics and market research providers, often without realizing that the underlying data comes from the census.[2]

[1] See National Research Council (1995) for examples of business uses of census data across a variety of industries, including retail and restaurant, banking and other financial services, media and advertising, insurance, utilities, health care, and others. One case study illustrates how census data was used to predict potential revenue, thereby informing retail expansion decisions.

[2] Armas (2001) and Thau (2014) describe intermediaries such as Claritas and Esri, which integrate census demographic data into software systems used for market segmentation, mapping,

However, since the decennial census is conducted only once every ten years, its information gradually becomes outdated, especially in small geographic areas sensitive to demographic changes from migration. As local demographics evolve over time, census data remains fixed and no longer reflects these changes. For firms that rely on census data to inform investment decisions, the quality of available information steadily declines, until the next decennial census brings in fresh information.[3] These periodic fluctuations in the quality of demographic information are what I term *information waves*. I hypothesize that as information quality deteriorates, firms are more likely to make misinformed choices, resulting in worse investment outcomes.[4]

To test this hypothesis, I examine failure patterns of establishments in the retail and restaurant industries. Firms in these industries are organized by geographic location and serve localized markets, thereby making site selection decisions essential to the success of their investment (Mian and Sufi, 2014; Adelino, Ma and Robinson, 2017). More importantly, this setting provides a unique opportunity to observe investment decisions (*entry*) and their associated outcomes (*exit*) at an investment-by-investment level, which can be difficult in other industries. Such granularity is crucial for determining the quality of information available to firms at the time of their decisions, and to directly link specific investments to their future outcomes—a connection often obscured in typical fi-

and site selection analyses.

[3]During periods with outdated census data, firms have few alternative data sources of comparable quality in granularity and scope. According to Adair (1991) "Telephone surveys by market research companies often provide the same information, but the census is usually more accurate because the bureau tries to find everyone in the nation. That, in turn, provides amazing detail about every city block."

[4]While positive surprises can help reduce failure rates, as demonstrated in the heterogeneity tests in subsection 1.6.1, firms are more inclined to enter markets that appear favorable based on the available data. As a result, they are more often exposed to conditions that turn out to be less advantageous than initially expected. Consequently, the overall effect of relying on outdated information is a net increase in failure rates.

nancial disclosures that aggregate performance data.

This level of analysis also helps control for time-varying confounding factors by focusing on excess failure rates, which compare the failure rates of new entrants to those of existing establishments within the same geographic market and time period. This comparison allows me to isolate the impact of information quality on new entrants from broader shocks that influence all businesses operating in the same market and year. Moreover, geographic variation provides an additional dimension of changes in information quality, helping to address the challenge of limited observable cycles resulting from infrequent census updates.

Using the sample of Retail Trade (NAICS 44-45) and Accommodation and Food Services (NAICS 72) businesses located within the state of New York from 1985 to 2014, I find that excess failure rates across entry-year cohorts follow the wave pattern that I hypothesized. For each additional year since the most recent census, the failure rate among new entrants increases by a statistically significant 1.6 percentage points. A 10-year gap between two decennial censuses would result in a 16 percentage point increase in failure rate due to outdated information. To put this in perspective, given that the average 5-year failure rate in my sample is 50 percentage points, using census data as old as ten years could raise a firm's baseline failure rate by 32%. This effect is notably stronger before 2000, which likely reflects improvements in the data environment after the early 2000s, including the introduction of the American Community Survey and the growing availability of alternative data sources.

To assess whether the observed pattern could have arisen by chance, I conduct a placebo test that creates hypothetical census schedules by randomly assigning a year in each decade as the census year. Out of the 10,000 randomly

generated schedules, only 193 produce estimates greater than or equal to the original estimate using the true census schedule. The observed p-value of 1.9% suggests that the effect is unlikely to be driven by random timing alone.

Next, I examine heterogeneity in the effect of outdated census data across geographic areas, industries, and firm size. If the main effect is driven by outdated census demographic data, it should be more pronounced in settings where such data plays a more critical role in investment decisions. These cross-sectional results help confirm that the main effect operates through the information channel.

First, the effect of outdated census data is more pronounced in geographic areas that have experienced substantial demographic shifts between censuses. In these areas, the discrepancy between recorded and actual conditions is more pronounced. By contrast, the impact is negligible in areas with stable demographics, where past census data continues to reflect current market conditions.

Second, the effect of outdated census data is stronger in industries with smaller trade areas, such as restaurants and grocery stores, where firms rely more heavily on local demographic information. In contrast, industries with broader geographic reach, such as motor vehicle dealers and furniture stores, are less affected by fluctuations in local conditions. Non-store retailers who do not depend on a physical storefront, such as e-commerce businesses, show no measurable response, consistent with their limited reliance on local demographics.

Third, the effect of outdated census data is concentrated among smaller firms. These firms rely more heavily on public demographic information, mak-

ing them more vulnerable to declines in information quality. In contrast, large firms have been less affected, especially in the latter half of the sample period, as they increasingly rely on alternative data sources that smaller firms often cannot afford.

I conduct additional analyses to address potential alternative explanations. First, I demonstrate that my findings are not driven by business cycles. In the cross-industry analysis, both the restaurant and grocery sectors respond similarly to outdated census data, despite stark differences in their sensitivity to macroeconomic conditions. Moreover, the effect is comparable across local markets that were more or less affected by the early 1990s and early 2000s recessions. Second, I examine whether the findings could be explained by government programs that allocate funding based on census data. For such policies to explain the results, they would need to disproportionately impact new entrants relative to existing establishments. To evaluate this possibility, I focus on place-based economic development programs that specifically target new businesses. Using program eligibility criteria, I show that the results are robust in markets ineligible for these initiatives.

Lastly, I examine whether firms in the retail and restaurant sectors strategically delay entry in anticipation of new census data. I find no significant evidence of such behavior, likely due to the high investment reversibility and localized competition in these industries, which reduce the benefits of waiting for updated information. Additionally, firms may not be fully aware of the staleness of the data they use, further limiting their ability to respond to variation in information quality over time. These findings suggest that the observed effects on failure are unlikely to reflect strategic timing of entry.

This paper makes several contributions to the literature. First, it introduces a novel empirical strategy to measure how variation in information quality affects firm decision-making under uncertainty. The investment under uncertainty literature emphasizes how firms delay investment in response to uncertainty about future conditions (Dixit and Pindyck, 1994; Bloom, Bond and Van Reenen, 2007; Bloom, 2009; Julio and Yook, 2012; Kellogg, 2014; Baker, Bloom and Davis, 2016; Baker, Bloom and Terry, 2024). I provide a new perspective by focusing on a distinct but related issue: uncertainty about current market conditions due to limitations in available data. To quantify this form of informational frictions, the empirical strategy exploits the predetermined, recurring release schedule of U.S. Census data as an exogenous source of variation. This design allows for a clean assessment of how deteriorating information quality influences firm investment.

Second, this paper provides new evidence on the economic value of government-provided data, with a specific focus on the U.S. Census. While the census is primarily designed for political apportionment and federal funding allocation (Serrato and Wingender, 2016), it also plays a critical role in private-sector decisions. The U.S. is unusual in making detailed census data broadly accessible, which has fostered an ecosystem of commercial data users (Donnelly, 2019). As Hughes-Cromwick and Coronado (2019) emphasize, despite descriptive evidence of widespread use by businesses, the economic value of government data remains difficult to measure. This paper addresses the gap by quantifying how the timeliness of census data shapes firm investment outcomes, contributing to a broader literature on the value of public data infrastructure for private-sector decisions (Craft, 1998; Gao and Huang, 2020; Nagaraj, 2022).

Finally, this paper contributes to the literature on firm turnover by introducing a novel information-specific mechanism that explains disparities in failure rates between small and large firms. While previous studies have focused on cross-sectional differences in firm turnover rates (Jovanovic, 1982; Dunne, Roberts and Samuelson, 1989; Hopenhayn, 1992; Asplund and Nocke, 2006), this paper emphasizes the role of information quality in shaping these outcomes. The findings align with Collard-Wexler (2013), who demonstrates the significant impact of uncertainty reduction on firm turnover. The results have broader implications for market structure, given that census data is open access and acts as a public good (Jones and Tonetti, 2020), potentially benefiting all users including small establishments. The accessibility of census data stands in contrast to the exclusivity of private (big) data, which can reinforce the competitive advantage of large firms (Davis et al., 2006; Decker et al., 2016; Begenau, Farboodi and Veldkamp, 2018; Farboodi et al., 2019).

1.2 Institutional Background

1.2.1 The Decennial Census

The decennial censuses from 1970 to 2000 use fairly consistent data collection methods and variable definitions. The short form collects basic demographic variables from 100% of the population, while the long form captures a wide range of socioeconomic and housing variables from a 1-in-6 sample. The 1-in-6 sample is large enough that the estimates are often treated as if they were exact counts (Donnelly, 2019).

As summarized in Table 1.1, tabulated census data are released on a flow basis following the enumeration.[5] Basic population counts are typically published by March of the year after the census, followed by more detailed tabulations of population, housing, and socioeconomic characteristics. The datasets most relevant for business applications, such as Summary File 1 and Summary File 3, are generally available by the middle of the second year after the census.

1.2.2 Decreased Accuracy In-between Censuses

Although census data is highly accurate at the time of collection, its reliability decliness as local conditions change over the course of a decade. Small areas, in particular, can experience rapid population and economic changes during intercensal years.

Various efforts have been made to impute demographic information between decennial censuses. The Census Bureau provides annual population estimates using administrative records,[6] although these estimates are only available down to the county level.[7] Commercial data vendors also make their own estimates, with varying degrees of success (Cropper et al., 2012).

[5]Since 1930 "Census Day" has been April 1 in the first year of each decade (ending in zero). The Census aims to capture a snapshot of the population on this specific reference date, while actual census-taking begins before this date and extends for months thereafter. See "United States Census," *Wikipedia*, `https://en.wikipedia.org/wiki/United_States_census` (accessed July 27, 2022).

[6]For example, birth and death statistics are from health departments; domestic migration data are from the IRS and the Centers for Medicare and Medicaid.

[7]The 2000s estimates are fairly accurate at the county level, as the records were collected at the county level; the 1990s estimates are relatively poor because results are imputed from the state level numbers. A large portion of the error of closure is concentrated in the 5-34 age group. See U.S. Census Bureau, "Intercensal Estimates Methodology" `https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/intercensal/intercensal-nat-meth.pdf` (accessed July 27, 2022).

To produce more timely estimates, in 2005 the Census Bureau began publishing the American Community Surveys based on a rolling sample methodology that surveys a much smaller number of households but more frequently, with the goal of achieving the same level of precision as the data from the decennial long-form sample. However, the margins of error can be substantial for small geographic areas (National Research Council, 2015; Donnelly, 2019).

1.2.3 Use and Impact of Census Data on Business Decision-Making

Census data informs many business decisions, influencing firms through various channels, often without their explicit awareness of its source. This section explores how different types of businesses access and utilize census data, both directly and indirectly, to illustrate its widespread impact on their operations.

1.2.3.1 Direct Access to Census Data

Before the internet era, census data products were distributed in print, on computer tape, or on CD-ROM through a network of affiliated organizations including state executive departments, chambers of commerce, councils of governments, and university research departments. Businesses have been significant users of this data.[8] According to the South Carolina Census State Data Center,

[8]For example, Eckerd drugstore used census data to manage inventory and product offerings tailored to different demographic characteristics in various markets. Similarly, Volvo North America utilized census information to identify optimal locations for its dealerships, and Winn-Dixie grocery wouldn't build a new store unless census data indicated sufficient households to support it (Adair, 1991). Such usage is not exclusive to large firms. According to a report by the Council of Economic Advisers, "Numerous small businesses responded to a request for exam-

35 percent of annual requests for census data came from businesses (National Research Council, 1995).

Local libraries and chambers of commerce play an essential role in helping small businesses access demographic data. They offer access to reference books, statistical yearbooks, and printed reports from the U.S. Census Bureau. Many of these libraries operate business centers staffed by knowledgeable librarians who help entrepreneurs retrieve relevant demographic information.[9] These services are especially valuable to small businesses, as financial institutions often require a detailed business plan when evaluating loan applications.[10] A key part of the business plan is market research. It involves identifying target customers and estimating market size, both of which depend heavily on demographic data.[11]

With the advent of the internet era, access to census data has become more convenient. The Census Bureau's website, launched in the mid-1990s, now provides comprehensive digital access to demographic data. The ease and afford-

ples of business uses of census data" (The Council of Economic Advisers, 2000). For example, a cemetery owner recently asked the Census Bureau to help him determine the number of people of Italian ancestry living near him in order to anticipate the demand for crypts (Farhi, 1990).

[9]See Rhonda Kleiman and Heather Sharpe, *Duke Street Business Center at Lancaster Public Library*, "Business Start-Up Toolkit," `https://cityoflancasterpa.com/wp-content/uploads/2013/10/Business-Startup-Toolkit-April-2013.pdf` (accessed June 25, 2023) for an example of a local library's business center providing guidance to local entrepreneurs on how demographic data from the Census Bureau's website can be used to conduct market research.

[10]For an overview, see *Forbes Advisor*, "How to Write a Successful Business Plan for a Loan," `https://www.forbes.com/advisor/business-loans/how-to-write-a-successful-business-plan-for-a-loan/` (accessed July 17, 2024). For specific examples, see *Bank of America*, "How to Write an Effective Small Business Plan," `https://business.bankofamerica.com/resources/how-to-write-effective-small-business-plan.html` (accessed July 17, 2024); and *Chase*, "5 Reasons You Need a Business Plan," `https://www.chase.com/business/knowledge-center/start/reasons-for-business-plan` (accessed July 17, 2024).

[11]U.S. Small Business Administration, "Market Research and Competitive Analysis," `https://www.sba.gov/business-guide/plan-your-business/market-research-competitive-analysis` (accessed July 17, 2024). The SBA cites the U.S. Census Bureau as a source for demographic data.

ability of online access are especially beneficial for small businesses.[12]

1.2.3.2 Indirect Influence through Intermediaries

Although direct access to census data has become more common in the internet era, many businesses continue to rely heavily on intermediaries for detailed analysis. Popular data analytics companies, market research firms, and location intelligence providers like Claritas, Esri, and SafeGraph, offer analytics tools for customer segmentation, market analysis, and strategic planning. These intermediaries incorporate census demographic data into their platforms, often enriching it with other proprietary data sources and mapping tools (Armas, 2001; Thau, 2014). As a result, businesses using these tools for their investment decisions are influenced by census data, even if they are unaware of the original source.

Another channel is through the real estate sector. Commercial real estate agents are trained to help potential buyers and tenants find the right location for their business needs, with demographic data of the local trade area being a crucial component of their knowledge base. Figure A.1 shows an example of the databases used by commercial real estate brokers, featuring demographic characteristics of various neighborhoods for easy comparison.[13] The underlying data in such databases often comes directly from census sources. This demographic data also influences commercial real estate prices, including rents, which affect the operational costs businesses face at different locations. When

[12] According to Matthew Cunningham, manager of the Texas Business and Industry Data Center "More small businesses would use it just because it's free, especially now that people are in the information age" (Armas, 2001).

[13] See "The Best of Commercial Real Estate Data Sources," *Apto Blog*, `https://www.apto.com/blog/the-best-of-commercial-real-estate-data-sources-demographics-broker-databases/` (accessed July 27, 2022) for other examples.

data becomes outdated, mispricing can cause businesses to overpay for rent or face unexpected competition, ultimately squeezing profit margins and increasing the risk of failure.

1.3 Data and Summary Statistics

1.3.1 Establishment Data

1.3.1.1 Data Source

Establishment data comes from Dun & Bradstreet's National Establishment Time Series (NETS) database, a longitudinal dataset derived from business records that track U.S. establishments over time. Dun & Bradstreet collects information from various sources, including business registrations, trade references, credit inquiries, public records, and direct communication with businesses. The NETS database provides annual snapshots of all U.S. business establishments, capturing key details such as their location, industry classification, size, and operational status. Each establishment is tracked by a unique identifier over time, allowing me to infer entry and exit on an annual basis.

I obtain a sample of the NETS data for Retail Trade (NAICS 44-45) and Accommodation and Food Services (NAICS 72) businesses located within the state of New York covering 30 years from 1985 to 2014.[14] The sample contains 6,459,013 establishment-year observations across 857,792 unique establish-

[14] A large proportion of census tracts are located in the densely populated New York City. To verify that my results are not driven by the tracts in New York City, I report in Table A.6 results using the subsample of census tracts located within (outside) New York City. The effects are very similar in terms of magnitude across the two subsamples.

ments.

Focusing on the retail and restaurant industries offers two key advantages. Firms in these industries are organized by geographic locations and primarily serve a local market, making local demographic statistics crucial to their entry decisions (Mian and Sufi, 2014; Adelino, Ma and Robinson, 2017). More importantly, this setting offers an opportunity to observe establishment-level investment decisions (entry) and subsequent outcomes (exit), which is often difficult in other industries. This granularity allows for a direct link between firms' decisions and the information available to them at the time.

The NETS data offers establishment-level detail with broad industry and geographic coverage, making it a widely used alternative to Census Bureau microdata, which require special access and secure handling.[15] However, differences in data collection methods compared to Census datasets have raised concerns about potential discrepancies. To address these concerns, I compare NETS with publicly available County Business Patterns (CBP) data in subsection A.4.1. The comparison confirms that NETS data aligns well with CBP for the retail and restaurant industries, supporting its use in analyzing establishment entry and exit patterns. In addition, I validate that the key empirical results remain robust to potential sampling differences.

[15]NETS data have been widely used in empirical research across economics, finance, and public policy for studying topics that require establishment-level information. For instance, see Addoum, Ng and Ortiz-Bobea (2020); Currie et al. (2010); Kolko (2012); Levine, Toffel and Johnson (2012); Neumark, Wall and Zhang (2011); Schuetz, Kolko and Meltzer (2012); Tsui et al. (2020).

1.3.1.2 Entry, Exit and Failure Rate

The observed establishment entry (exit) is identified as the year its unique identifier first appears (disappears) in the sample. I make the assumption that the site selection decision is made during the same year as the observed entry, given that it only takes a few months to open a retail business,[16] and that the establishment data are December snapshots. This assumption about entry timing is also consistent with the time it takes for an entrant to become active in theoretical models of industry dynamics, which is often referred to as the "time-to-build" assumption (Ericson and Pakes, 1995) and used in empirical studies about retail dynamics (Arcidiacono et al., 2016; Fang and Yang, 2024; Hollenbeck, 2017; Igami and Yang, 2016; Maican and Orth, 2018; Suzuki, 2013).

To address right censoring in the later years of the sample, I measure investment outcomes using the failure rate within the first 5 years of entry. The retail and restaurant industries have relatively high turnover: half of the establishments in my sample fail within their first 5 years. Long-term survival is less likely to be driven by the initial entry decision.

Five-year failure rates can only be reliably measured for firms with clearly observed entry and exit windows. I define entry cohorts as firms that first appear between 1986 and 2009. Firms already active in the 1985 snapshot have unknown entry dates, and those entering in 2010 or later are not observed long

[16] It generally takes 2-6 months to open a retail business after the site has been selected. See for example, Jean Murray, "How Long Does It Take to Start a Business?" *The Balance*, https://www.thebalancesmb.com/how-long-does-it-take-to-start-a-business-3974594 (accessed July 27, 2022); Jeff Campbell, "How Long Does It Take to Build a Grocery Store?" *The Grocery Store Guy*, https://thegrocerystoreguy.com/how-long-does-it-take-to-build-a-grocery-store/ (accessed July 27, 2022); Katherine Boyarsky, "How Long Does It Take to Open a Restaurant?" *Toast*, https://pos.toasttab.com/blog/on-the-line/how-long-does-it-take-to-open-up-a-restaurant (accessed July 27, 2022).

enough to evaluate five-year outcomes. As a result, the analysis focuses on 24 entry cohorts with complete five-year outcome windows.

1.3.1.3 Market Definition

Success in the retail and restaurant industries depends heavily on location, in large part because establishments in these sectors tend to serve small trade areas.[17] I define markets at the census-tract level and assign each establishment to a tract based on its geographic coordinates, using 2010 TIGER/Line Shapefiles. This approach ensures that market boundaries remain consistent throughout the sample and can be readily linked to census demographic data.[18]

1.3.1.4 Summary Statistics

New York State not only ranks as the third-largest economy in the United States but also embodies a broad mix of urban, suburban, and rural environments. Its diverse industrial base, which ranges from the financial and service hubs of New York City to the manufacturing and agricultural sectors upstate, makes it a microcosm of the national economy.[19] Table A.1 reports establishment counts by NAICS 4-digit sub-industry over the sample period. Table 1.2 summarizes

[17] For a salient illustration of how small these trade areas might be, please refer to this example by Thomadsen (2005) about the fast food industry.

[18] Among the census geographies, county and census tract have relatively stable definitions and ID codes across the relevant censuses in my sample. In contrast, zipcode tabulations are only available since the 2000 census and county subdivisions since the 1990 census; block and block groups may be completely renumbered in the next census; places only cover concentrated settlements; and metropolitan statistical areas update delineations couple of times per decade (Donnelly, 2019).

[19] To verify that my results are not driven by New York City specific dynamics, I report in Table A.6 results using the subsample of census tracts located within (outside) New York City. The effects are very similar in terms of magnitude across the two subsamples.

key market dynamics, including the number of establishments, new entrants, and failure rates at the market-year level.

1.3.2 Demographic Data

1.3.2.1 Data Source

Demographic data comes from the U.S. Decennial Census (1980, 1990, 2000, and 2010) and the American Community Survey (ACS 2008-2012 5-year survey).[20] To identify areas that experience large demographic shifts, I calculate changes in demographic variables between two consecutive decennial censuses at the census-tract level.

1.3.2.2 Demographic Shift

Census tracts are sometimes divided or combined every 10 years to maintain the optimal threshold for population size. To make consistent comparisons, census tract data from 1980, 1990, and 2000 are all mapped to 2010 geography using the LTDB crosswalk developed by Logan, Xu and Stults (2014). In addition, dollar values are inflation adjusted based on the Consumer Price Index research series using current methods (CPI-U-RS) from St Louis Fed. Table A.3 provides details on the construction of each demographic variable.

Table 1.3 summarizes the distribution of demographic changes between two decennial censuses. These census-tract level changes exhibit relatively large

[20] I obtain data on 2010 education, employment, income, and housing value from the ACS 2008-2012 5-year survey, since these variables are not available in the 2010 decennial census.

variation, highlighting that small geographic areas can experience large and sometimes unexpected demographic shifts over a ten-year period.

1.4 Empirical Strategy

To measure the impact of census data on investment outcomes, I compare cohorts of establishments based on their entry years. Different entry cohorts face varying levels of census data quality when they make their entry decisions. Those entering early in the decade benefit from recently released census data that closely reflects market conditions. In contrast, those entering later rely on increasingly outdated data, placing them at a potential disadvantage. As a result, the timing of an establishment's entry can be viewed as a measure of the quality of census data at its disposal.

However, time-varying factors such as macroeconomic conditions can also influence failure rates, independent of information quality. For example, a store that opened just before the financial crisis may have failed due to the broader economic downturn, rather than a poor location decision.

To account for time-varying shocks, I construct a baseline failure rate using all existing establishments in the same local market and calendar year.[21] This baseline failure rate helps eliminate potential confounding effects, assuming time-varying factors affect new and existing establishments similarly within a local market and year.[22] Under this assumption, subtracting the baseline al-

[21] Intuitively, subtracting the average failure rate is analogous to taking out the calendar year fixed effect, which ultimately allows me to control for external factors that are common to all establishments in that year.

[22] I test the robustness of this assumption in subsection 1.7.1

lows me to disentangle the economic impact of information quality at entry (which affects only new entrants) from changes in underlying economic conditions (which affects all existing firms in the market).

The excess failure rate for an entry cohort in a given calendar year is thus defined as the actual failure rate for the entry cohort subtracted by the baseline failure rate across all existing establishments:

$$\Delta f_{imt} = f_{imt} - F_{mt} = \frac{\text{Exit}_{imt}}{B_{imt}} - \frac{\sum_i \text{Exit}_{imt}}{\sum_i B_{imt}}, \tag{1.1}$$

where B_{imt} is the number of establishments that enter at year i into a geographic market m and still exist at the beginning of year t. Here, f_{imt} is the actual failure rate for entry cohort i in market m from calendar year t, and F_{mt} is the average failure rate in calendar year t across all existing establishments in market m.

The excess failure rate of an entry-year cohort i from market m within the first 5 years of entry is calculated by taking a summation of the excess failure rate for the entry-year cohort in each of the 5 years:

$$\text{Excess failure rate}_{im} = \sum_{t=i+1}^{i+5} \frac{\Delta f_{imt} * B_{imt}}{B_{im}}. \tag{1.2}$$

A potential challenge to this approach is whether external shocks systematically impact new entrants differently from incumbents. In section 1.7, I examine potential alternative explanations, such as business cycles or government funding linked to census data, and find no evidence that these channels account for the observed pattern of failures.

1.5 Main Results

1.5.1 Failure Rate Pattern

My central hypothesis is that failure rates across entry-year cohorts follows a wave pattern, characterized by worse investment outcome over time as census data become outdated and a reversal when the new census data is released.

To test this hypothesis, I first examine the pattern of excess failure rates across entry-year cohorts using a flexible specification:

$$\text{Excess failure rate}_{im} = \sum_{i=1986}^{2009} \beta_i \times I(\text{Entry-Year} = i) + \varepsilon_{im}, \tag{1.3}$$

where i is an index for each entry-year cohort and m is an index for market (i.e., census tract). The outcome variable Excess failure rate$_{im}$ is the 5-year failure rate for establishments that enter market m in year i, relative to the average failure rate of establishments in the same market m. Standard errors are clustered at the market level to flexibly account for potential serial correlations in the error term.

Figure 1.1 visualizes the coefficient estimates from Equation 1.3 across entry cohorts. These estimates reveal a wave-like pattern that aligns with the census data release schedule.[23] Following the release of the 1990 census data, the failure rate drops substantially. For the 1991 and 1992 cohorts, as additional census datasets roll out—indicated by the shaded regions—the failure rate declines. This immediate decline in the failure rate suggests that many firms draw

[23]For a detailed schedule, please refer to Table 1.1.

insights from demographic variables in the early release. By the time the 1993 cohort enters, all major census data files have become available. In the following years, failure rates of the subsequent cohorts continue to rise, as the 1990 census data becomes outdated. After the 2000 census, failure rates do not decline immediately, suggesting firms now rely more heavily on detailed socioeconomic data, which are released later in the cycle. Post 2003, once all the new data is assimilated, failure rates move back up again.

In 2005, the Census Bureau introduces the American Community Survey (ACS), which is designed to provide more timely but potentially less precise estimates than the original decennial census.[24] The dotted line marks this transition. Following this point, several ACS datasets become available and the failure rate experiences a more gradual increase.[25]

1.5.2 Failure Rate and Distance to Census Data Release

Building on the descriptive pattern in Figure 1.1, I formalize the relationship between census data quality and firm failure using a parametric specification. Each information wave cycle consists of two distinct phases. The first phase includes entry cohorts from the two years immediately following a census release, when new data files are still being rolled out. During this phase, firms

[24]The American Community Surveys collects socioeconomic characteristics from a significantly smaller sample of the population on a rolling basis. In larger regions or densely populated areas, annual estimates are derived from a full year's worth of data, whereas in smaller areas, samples are aggregated over 5 years to improve accuracy. While the smaller sample size of the American Community Survey makes it more cost-effective and timely, it is not as precise as the decennial census. In some cases, particularly for smaller geographies or specific data breakdowns, the margin of error can be greater than the actual estimate itself (Donnelly, 2019).

[25]From 2005, 1-year estimates are provided for areas with a population of at least 65,000; from 2007, 3-year estimates for areas with over 20,000 people; and from 2009, 5-year estimates for all geographies.

may incorporate different components of the census at different times, making its effect on failure rates less predictable. The second phase begins after the first two years, once all major census files have been released. In the second phase, after all major files have been released, the data gradually become outdated, making their effect on failure easier to isolate.[26]

To measure the degree of information staleness, I define a distance variable that captures how long it has been since the initial release of the most recent census data. This variable serves as a proxy for the discrepancy between real-time market conditions and the demographic snapshot available to firms at the time of entry. A larger gap implies greater reliance on outdated information, which increases the likelihood of failure. Specifically, for my sample period, I define S_i as:

$$S_i = \begin{cases} i - 1981, & \text{if } 1981 \leq i < 1991 \\ i - 1991, & \text{if } 1991 \leq i < 2001 \\ i - 2001, & \text{if } 2001 \leq i < 2010 \end{cases} \tag{1.4}$$

To quantify the effect of outdated census data on firm failure, I incorporate this distance variable into a piecewise linear specification. The following model captures this relationship::

[26]Focusing on this phase also makes my results less sensitive to the time to build assumption, as firms will have had more time to incorporate updated census data into their investment decisions. For example, an establishment that enters by the end of 2003 has between 1.5 years (since the release of Summary File 3) and 2.75 years (since the release of Summary File 1) to access census data and make investment decisions, depending on which demographic variables are important to these decisions.

$$\text{Excess failure rate}_{im} = \alpha_1 I(\text{S}_i < 2) + \alpha_2 I(\text{S}_i \geq 2) + \beta_1(\text{S}_i - 2)I(\text{S}_i < 2) + \beta_2(\text{S}_i - 2)I(\text{S}_i \geq 2) + \varepsilon_{im} \quad (1.5)$$

where i is an index for the entry-year cohort and m is an index for the market. Standard errors are clustered at the market (census-tract) level to to flexibly account for potential serial correlations in the error term.

The coefficients α_1 and β_1 represent the intercept and slope of the first phase, when data are still being released. Firms may incorporate different components of the census data as they become available, leading to heterogeneous responses that make this phase less stable for inference. In addition, because this segment covers only two data points per census cycle, the estimate of β_1 is highly sensitive to small variations. By contrast, the coefficients α_2 and β_2 correspond to the second phase, when all major census data files have been made available. Because firms must rely on increasingly outdated information in this phase, it reflects a more stable relationship between census data quality and firm outcomes. Therefore, My primary focus is on interpreting β_2.

Table 1.4 reports the coefficient estimates. The β_2 coefficient estimates are positive and statistically significant, indicating that a longer gap between entry and the most recent census release is associated with higher failure rates. The full sample estimate of 0.016 represents an annual increase of 1.6 percentage points in failure rate. Based on this estimate, a 10-year gap between two decennial censuses would result in a 16 percentage points increase in failure rate when firms have to rely on the most outdated information. Considering that the average 5-year failure rate in my sample is 50 percentage points, outdated

census data increases a firm's baseline failure rate by 32% over a full decade.

I next split the sample into pre-2000 and post-2000 entry cohorts. In both periods, the β_2 coefficient estimates remain positive and statistically significant. However, the effect is notably larger in the pre-2000 period. One reason for the attenuation in later years is the introduction of the American Community Survey, which provided more frequent demographic updates to supplement the decennial census. This transition occurred within the broader context of the information age, as new technologies expanded firms' access to alternative data sources.[27] Overall, the evidence suggests that the impact of census data on firms is influenced by the presence of alternative information sources.

1.5.2.1 Robustness

To assess robustness, I conduct several additional analyses. Full results are reported in the Online Appendix. First, I test whether results are sensitive to how the two linear segments in the main specification are modeled. Specifically, I re-estimate a model that constrains the two segments to connect at the breakpoint rather than allowing separate intercepts. As shown in Table A.7, the results remain similar. Second, I account for potential spatial correlation across markets within the same entry cohort by using alternative clustering methods to compute standard errors. Table A.8 shows that the estimates remain statistically significant under these adjustments. Third, I test whether the benchmark failure rate influences the results. Table A.9 presents estimates excluding the entry cohort from the failure rate calculation to ensure it does not mechanically affect the benchmark. Table A.10 uses an alternative benchmark based only on

[27] Many consumer activities can be tracked by their web browsing history and mobile phone usage, thanks to the growing popularity of the internet and smart phones.

establishments that entered within the past five years to address the concern that newer firms may be more vulnerable to market conditions than more established ones. Both yield similar estimates.

1.5.3 Placebo Test

To assess whether the pattern in Figure 1.1 could arise by chance, I implement a placebo test that randomizes the timing of census releases. Specifically, I randomly shuffle the census years and re-estimate the main specification in Equation 1.5 under each permutation. In each permutation, I assign one year from each decade in the 1970s, 1980s, 1990s, and 2000s as the census year, resulting in 10,000 alternative schedules.[28] For each hypothetical schedule, I recalculate the distance to the most recent census for each entry-year cohorts and then re-estimate Equation 1.5.

Out of the 10,000 simulated schedules, 193 produce β_2 estimates greater than or equal to the original estimate using the true census schedule. The probability of generating by chance a coefficient estimate as large as the original estimate is thus 1.9%. Since this p-value does not rely on standard errors at the geographic market level, it helps verify that spatial correlation is unlikely to drive the main pattern.

[28]For example, one hypothetical set of census schedule could be 1975, 1988, 1992, and 2005. Under this assignment, an establishment entering in 1986 would rely on data from the 1975 census, with a distance of 10 years. A 1970s census year is required in this case, since the 1980s census occurs after the sample begins.

1.6 Heterogeneity in Failure Rate Patterns

This section examines heterogeneity in the impact of outdated census data on failure rates. I test whether the main effect documented in the previous section is more pronounced in (1) geographic areas that experience substantial demographic shifts, (2) industries that depend on localized information in small trade areas, and (3) small firms that lack alternative sources of information. The results show that outdated census data has a stronger impact where having up-to-date demographic information is especially important. I also conduct placebo tests in settings where census data is unlikely to matter and find no discernible effect.

1.6.1 Failure Rate and Shifts in Demographics

1.6.1.1 Incremental Effect

This section explores how the effect of outdated census data varies across various geographic areas. Geographic variation introduces cross-sectional differences in the usefulness of census data beyond entry timing. In areas with stable demographics, census data collected from the past continues to provide an accurate representation of market conditions. However, in areas undergoing rapid demographic changes, the informational value of census data deteriorates more quickly. As a result, the impact of outdated census data on failure rates should be more pronounced in these areas.

To test this prediction empirically, I examine whether the effect of outdated census data is amplified in areas with substantial demographic shifts. I inter-

act the slope terms from Equation 1.5 with an indicator for high demographic change areas and estimate the following regression:

$$\begin{aligned}\text{Excess failure rate}_{im} = {} & \alpha_1 I(\text{S}_i < 2) + \alpha_2 I(\text{S}_i \geq 2) \\ & + \beta_1(\text{S}_i - 2)I(\text{S}_i < 2) + \beta_2(\text{S}_i - 2)I(\text{S}_i \geq 2) + \beta_3 \widetilde{\Delta X}_m \\ & + \gamma_1 \widetilde{\Delta X}_m(\text{S}_i - 2)I(\text{S}_i < 2) + \gamma_2 \widetilde{\Delta X}_m(\text{S}_i - 2)I(\text{S}_i \geq 2) + \varepsilon_{im}\end{aligned} \tag{1.6}$$

where $\widetilde{\Delta X_{im}}$ is an indicator variable equal to 1 if the value of a demographic variable X in market m changes above a threshold between the two decennial censuses surrounding entry-year i. Because different demographic variables may influence business success differently, I estimate Equation 1.6 separately for each demographic variable X, distinguishing between positive and negative changes. Table A.3 provides detailed definition of each demographic variable and Table 1.3 tabulates changes in demographic variables at the census-tract level. Other variables are defined in Equation 1.5.

In areas with substantial demographic change, the total effect of outdated census data on failure rates is given by $\beta_2 + \gamma_2$. The main coefficient of interest γ_2 measures the incremental effect of outdated census data in areas experiencing substantial demographic change. Figure 1.2 plots the coefficient estimates for γ_2 and the associated 95% confidence intervals where changes in demographic variables exceed the $\pm$10% threshold.[29]

Among the demographic variables I analyze, I find statistically significant

[29]To assess the sensitivity of this threshold, I also consider specifications using alternative cutoffs at $\pm$15% and $\pm$20% in Table A.11. The coefficient estimates are similar in magnitude to results from the 10% cutoff.

incremental effects for age, education, income, and housing value. For example, in areas where the share of young adults (*%Young 18–34*) is much lower than firms would have inferred from the previous census, failure rates increase more sharply, as indicated by the positive γ_2 estimate in Figure 1.9. Correspondingly, a favorable surprise moderates the increase in failure rate, as reflected in the negative γ_2 estimate in Figure 1.9. Similar effects are observed for higher education (*%College degree*) and wealth *Median house value*, where an increase in these demographic variables is associated with favorable firm outcomes. In contrast, unexpected declines in the percentage of kids (*%Kids*) and *%Median income* reduce failure rates.

Businesses have heterogeneous preferences regarding the ideal demographic profile.[30] Because of this variation, the estimates reported above reflect average effects across a wide range of firms targeting different segments of the market. Irrespective of the direction of their impact, these large demographic shifts highlight the gap between outdated census data and current market condition, adding another layer to the effect of outdated data on firm failure. In terms of magnitude, the absolute value of the coefficient estimates for the incremental effect γ_2 ranges from 3 to 12 percentage points in absolute terms, equivalent to 19% to 75% of the main effect β_2.

On the other hand, I find no consistent evidence of incremental effects related to demographic variables such as population and unemployment rate. Unlike other demographic variables, local population and unemployment statistics are available at annual and even monthly frequencies.[31] Since busi-

[30]For example, while fast food restaurants and dollar stores might shy away from high-income neighborhoods, luxury boutiques or gourmet restaurants may target them specifically. Similarly, while bars might find neighborhoods with young children less appealing, toy stores or family entertainment centers would find them ideal

[31]At the county level, annual population estimates can be inferred from birth and death

nesses can access this information regularly without waiting for the census data release, population and unemployment data in the decennial census have a limited impact on their failure rate.

Large demographic changes could also reflect local market volatility rather than informational frictions. The construction of the excess failure rate helps address this concern by subtracting the average failure rate of existing establishments in the same market and year. Nevertheless, new entrants might be more vulnerable to volatile market conditions than incumbents. Population and income are key demographic variables that capture the economic condition of a local market. The fact that these variables are not associated with incremental effects further supports the interpretation that the observed patterns operate through the information channel, rather than reflecting exposure to volatility.

1.6.1.2 Placebo Test

I conduct a placebo test using areas that experience little changes in demographics between two censuses. In these areas, entry timing is unlikely to generate meaningful variation in census data quality. Specifically, I construct a subsample of census tracts that have less than 10% absolute change in *%Young*, *%College degree*, *Median house value*, *%Kids* and *%Median income*, the demographic variables shown in the previous section to significantly affect firm failure.[32] I replicate Table 1.4 on this "no surprise" sub-sample and confirm that distance to census data release has no effect on failure rate when demographic conditions are plausibly stable over time. As reported in Table 1.5, the coeffi-

records via National Vital Statistics System (NVSS) and migration data from the IRS, while monthly unemployment statistics are published by Bureau of Labor Statistics.

[32]Imposing the 10% restriction on all demographic variables leaves no observations in the sample.

cient estimate on β_2 is effectively zero (-.0000901).

1.6.2 Failure Rate and Industry

To examine heterogeneity across industries, I estimate Equation 1.5 separately for each NAICS 4-digit industry. In this context, the benchmark failure rate is specific to establishments within the same industry and market.

Figure 1.3 reports the coefficient estimates of β_2 across 33 NAICS 4-digit industries. Notably, restaurants and grocery stores exhibit the strongest sensitivity to outdated census data. These two industries also have the largest number of establishments in my sample, which underscores the importance of proximity to their customer base.

The relative ranking of industries in Figure 1.3 suggests a broader relationship between trade area size and reliance on census data. Businesses that offer highly localized products and services rely more heavily on up-to-date demographic data from their immediate surroundings, as small areas can experience rapid demographic changes. Conversely, retailers selling specialty goods like motor vehicles and furniture serve a broader customer base. Their consumers are willing to travel longer distances to search and compare, which expands their trade areas. As a result, localized demographic fluctuations within these wider trade areas tend to average out, reducing firms' reliance on frequent census updates.

To proxy for trade area size, I categorize retail industries into durable and non-durable goods sectors, following the Bureau of Economic Analysis' classi-

fication of manufacturers of durable and non-durable goods.[33] Durable goods retailers are expected to draw customers from broader trade areas, while non-durable goods retailers typically serve more localized demand. As reported in Figure 1.4, the impact of outdated census data is notably more pronounced for firms in the non-durable goods sector (point estimate 0.0097) than those in the durable goods sector (point estimate 0.0045), which supports my earlier hypothesis regarding trade area size.

As a further test, I examine the NAICS 454 Non-store Retailers category, which includes Electronic Shopping and Mail-Order Houses, Vending Machine Operators, and Direct Selling Establishments. Unlike other retailers, these businesses do not require a physical store front close to their customers, making them much less reliant on local demographic information. As reported in Figure 1.3, the effect of census data on failure rate is statistically insignificant for all three industries in this group.

[33]The Bureau of Economic Analysis (BEA) defines durable goods as those that have a useful life of more than three years. Under this definition, the industries classified into the durable goods sectors include 4411 (Automobile Dealers), 4412 (Other Motor Vehicle Dealers), 4413 (Automotive Parts, Accessories, and Tire Stores), 4421 (Furniture Stores), 4422 (Home Furnishings Stores), 4431 (Electronics and Appliance Stores), 4441 (Building Material and Supplies Dealers), and 4442 (Lawn and Garden Equipment and Supplies Stores). The industries categorized under non-durable goods sectors comprise 4451 (Grocery Stores), 4452 (Specialty Food Stores), 4453 (Beer, Wine, and Liquor Stores), 4461 (Health and Personal Care Stores), 4471 (Gasoline Stations), 4481 (Clothing Stores), 4482 (Shoe Stores), 4483 (Jewelry, Luggage, and Leather Goods Stores), 4511 (Sporting Goods, Hobby, and Musical Instrument Stores), 4512 (Book Stores and News Dealers), 4531 (Florists), and 4532 (Office Supplies, Stationery, and Gift Stores). I exclude 4522 (Department Stores), 4523 (General Merchandise Stores, including Warehouse Clubs and Supercenters), 4533 (Used Merchandise Stores), and 4539 (Other Miscellaneous Store Retailers) were omitted due to their ambiguous nature in strictly categorizing as durable or non-durable goods sectors.

1.6.3 Failure Rate and Firm Size

This section examines whether the impact of outdated census data varies by firm size. One notable difference between large and small firms relates to their ability to obtain information. Large firms can collect data from existing customers, conduct market surveys, and purchase proprietary datasets from commercial vendors. Although census data forms the foundation of their market research, large firms can readily supplement it with alternative sources when it becomes outdated.

Small firms, on the other hand, often lack these resources. In fact, they may not even be aware that some crucial inputs to their decision-making process originate from the census, and that this data may be outdated. For example, entrepreneurs choosing a business location often rely on leasing agents, who use census-derived metrics to help assess whether local demographics fit the proposed business.[34] Both the entrepreneur and the agent might accept this data at face value. As a result, small firms might be more vulnerable to the effects of outdated census data than their larger counterparts.

In the retail sector, the size of a firm can be gauged by the number of its locations. To examine the heterogeneity across firms based on their size, I estimate Equation 1.5 separately for large chains, small chains, and local independents.[35]

[34] An important factor commercial real estate agents consider when they make recommendations to clients is whether the trade area's demographic profile is suitable to the proposed business. Much like corporate real estate planning departments, agents can narrow down the choices using filters based on demographic information, or use such information to justify the location's value to clients. Figure A.1 highlights an example of such databases being directly used by commercial real estate brokers. See "The Best of Commercial Real Estate Data Sources," *Apto Blog*, `https://www.apto.com/blog/the-best-of-commercial-real-estate-data-sources-demographics-broker-databases/` (accessed July 27, 2022) for other examples.

[35] The D&B data groups establishments by ownership. An establishment belonging to a chain might be classified as being owned by a franchisee, yet the parent company (i.e., franchisor)

Large chains are defined as firms with more than 20 locations.[36] This threshold is consistent with existing regulations related to chains,[37] and results are similar under alternative thresholds.[38] The benchmark failure rates are specific to establishments within each size group and census tract, with standard errors are clustered at the census-tract level.

Table 1.6 reports the full set of coefficient estimates across firm sizes, where the β_2 coefficients capture the impact of outdated census information on firm failure rates. Panel A presents these estimates for the full sample, and Panel B breaks them down further into sub-samples by entry cohorts. In the full sample, the β_2 coefficients are positive and statistically significant for small chains and independents, suggesting a higher vulnerability to outdated census data for smaller firms. The corresponding rescaled β_2 coefficients, which adjust for differences in baseline failure rates among firm types, reveal comparable relative effects for both groups. Large chains show a muted response, consistent with their greater access to alternative information sources.

The comparison in Panel B reveals a distinct shift in the pre- and post-2000 periods. Before 2000, the effect of census information on failure rate is positive

is the one that ultimately makes the site selection decisions. See Rick Bisio, "What Franchise Owners Should Know About the Site Selection Process," *Forbes*, https://www.forbes.com/sites/forbescoachescouncil/2021/06/23/what-franchise-owners-should-know-about-the-site-selection-process/?sh=7e97544732baf (accessed July 27, 2022). To avoid misclassifying a chain location owned by a franchisee as an independent, I identify chains based on their trade style names in the data set. See Table A.2 for details.

[36] Table A.1 lists the top 30 chains ranked by the total number of affiliated establishments in the sample.

[37] For an example of FDA menu labeling requirements related to chain establishments, see U.S. Department of Agriculture, "New National Menu Labeling Provides Information Consumers Can Use to Help Manage Their Calorie Intake," *Amber Waves*, https://www.ers.usda.gov/amber-waves/2018/october/new-national-menu-labeling-provides-information-consumers-can-use-to-help-manage-their-calorie-intake/ (accessed July 27, 2022).

[38] Please see Table A.12 for results setting the threshold at 10 establishments, and Table A.13 for results setting the threshold at 50 establishments

and significant across all firm types, including large chains. After 2000, the effect disappears for large and small chains, and although it persists for independents, its magnitude is considerably smaller. This shift coincides with the rise of digital data collection and commercial data vendors that provide firms with more timely demographic insights.[39] The growing availability of alternative sources likely reduced firms' dependence on the decennial census. However, the extent of this transition appears uneven across firm types. The continued sensitivity of independents to outdated data highlights persistent informational disadvantages for smaller firms.

1.7 Alternative Explanations

1.7.1 Differential Response to Business Cycles

A key assumption of my empirical strategy is that new and existing establishments are influenced by time-varying factors in a similar way, so that the average failure rate of establishments within the same local market can serve as a benchmark to remove potential confounding effects. One might be concerned that the entry cohorts behave differently from existing establishments during

[39]One notable example is Synergos Technologies Inc., founded in 2001, which uses postal address data released quarterly, combined with consumer survey data to provide timely and granular demographic data to national and regional companies making strategic location decisions. See Synergos Technologies Inc., `https://www.synergos-tech.com/industries/` (accessed July 27, 2022). Its clients include Kroger (grocery), CVS (pharmacy), Chipotle (quick service restaurant), Family Dollar (dollar store), and Simon (real-estate). Mukherjee, Panayotov and Shon (2021) provide a more general example of private data sources substituting less frequently released government macro data, and Chi, Hwang and Zheng (2024) provide additional examples of alternative data usage.

different phases of the business cycle.[40] In particular, the early 90s and early 00s recessions coincide with outdated census data in terms of timing, which may drive the pattern. However, the industry breakdown in Figure 1.3 shows that the strongest effects occur in restaurants and groceries, despite their notably different sensitivities to the business cycle.

To directly test the impact of the early 90s and early 00s recessions, I split the sample based on the severity of the recession in the local area, measured by changes in the county-level unemployment rate before and after the recession.[41] As reported in Table 1.7, I find similar effects on establishment failure rate (β_2) across sub-samples of local areas that are more and less impacted by the recessions. The absence of differential effects indicates that the observed patterns are unlikely to be explained by recession timing.

1.7.2 Government Policies

This section examines whether government spending programs tied to census data could explain the observed excess failure rates. As reviewed in US Government Accountability Office (2009), the ten largest federal assistance programs

[40] As the literature (Fort et al., 2013; Pugsley and Şahin, 2019; Sedláček and Sterk, 2017) has documented, employment fluctuations at startups and young firms are pro-cyclical: young firms co-vary more with the overall economy than mature firms. Alternatively, it is also conceivable that startups might be counter-cyclical due to selection: better financing conditions during economic boom might allow firms of lower quality to enter the market, which then subsequently fail. However, this financing story seems inconsistent with the cross-industry analysis finding stronger effects for firms with smaller trade areas. These establishments have smaller footprints and sell non-durable goods with lower inventory cost, which requires less upfront financing. If financing conditions or credit cycles were driving the results, we would expect to see stronger effects for capital-intensive retailers like auto dealers or furniture stores.

[41] Based the NBER U.S. recession dates (July 1990 to March 1991 for the early 90s recession and March 2001 to November 2001 for the early 00s recession), I measure the change (percentage change) in same-month unemployment rate before and after the recession (June 1990 to June 1991 for the early 90s recession and February 2001 to February 2002).

that rely at least in part on the decennial census data fall into two broad categories: those that directly increase consumer spending through transfers or subsidies, and those that fund public goods such as education and infrastructure.[42] For the first category, while state-level allocations are based on census data, within-state benefits are distributed according to individual eligibility criteria. As a result, funding misallocation is unlikely to drive the observed differences in failure rates across local markets. The second category includes programs that fund public schools, transportation networks, and local development. While these programs may use more local level data, they primarily serve broader community needs rather than directly influencing retail business conditions.

Ultimately, to the extent that increases in consumer spending and improvements in public infrastructure affect all businesses within the same neighborhood similarly, these government spending channels are unlikely to explain the differential failure rates observed between new and established establishments. In contrast, place-based policies specifically target new business entrants through subsidies and tax benefits, making them particularly relevant to my empirical context (Slattery and Zidar, 2020). One such policy prominent in the New York State is the Economic Development Zones (Empire Zones) Program.[43] To be eligible, a census tract needs to satisfy the following criteria: 1) Poverty rate at or above 20%, 2) Unemployment rate at least 125% of State average, and 3) Population at or above 2,000.[44]

[42]The first category includes programs that provide direct transfers (e.g., Temporary Aid for Needy Families, TANF) or cost subsidies that free up disposable income (e.g., Medicaid, the Children's Health Insurance Program, CHIP, and Section 8 Housing Choice Vouchers). The second category consists of programs funding public education (e.g., Title I Grants, IDEA Part B, Head Start), infrastructure (e.g., Highway Planning and Construction, Federal Transit Formula Grants), and local development (e.g., Community Development Block Grants).

[43]Good Jobs First. 1976–2019. "Subsidy Tracker." https://subsidytracker.goodjobsfirst.org/

[44]These criteria were stipulated by legislature in 1986 when the program was created. Over

Using these institutional features, I split the sample based on a census tract's eligibility for the Economic Development Zones program and estimate Equation 1.5 separately for these two groups. Table 1.8 reports the results. The coefficient estimate of β_2 for census tracts not eligible for Empire Zones funding closely mirror the main result. If anything, the estimated effect is smaller in eligible census tracts. If the state directs subsidies to areas that are no longer in economic distress, firm exits would likely decrease.[45] Thus, misallocated government funding based on census data is unlikely to explain the increase in establishment failure as census data becomes more stale.

1.8 Entry Patterns

The release of census data reduces uncertainty about market conditions and may influence the timing of firm entry. The investment under uncertainty literature (Dixit and Pindyck, 1994) predicts that firms delay irreversible investments when uncertainty is high, waiting until better information becomes available (Bloom, Bond and Van Reenen, 2007; Bloom, 2009; Julio and Yook, 2012; Kellogg, 2014; Baker, Bloom and Davis, 2016). Given the fixed schedule and importance of census data, firms might strategically time their entry to follow the release of fresh data. This sections explores the entry patterns surrounding census data

time, conditions were relaxed to the point that almost any area is eligible. The program was shut down in 2010, due to wide criticisms that the zones no longer correspond to distressed areas and that there is a lack of oversight (New York State Office of the State Comptroller, 2004). If the zones are chosen in ways unrelated to census data, the policy should not affect failure rates of new entrants as census data quality varies over time.

[45] I obtain similar results using eligibility for Opportunity Zones. To qualify for Opportunity Zones in New York State, a census tract should have individual poverty rate of at least 20%, and the median family income no more than 80% of the state median. The first Opportunity Zones were designated in 2018, after my sample period. However, the requirements reflect what could be used in other state and local policies if they are based on census variables.

releases, to better understand how the availability of updated information influences firm behavior under uncertainty, particularly in the context of retail and restaurant sectors.

Investment under uncertainty has been most extensively studied in industries with high irreversibility and long-term commitment, such as manufacturing, mining, and energy. Retail and restaurant sectors differ from these settings in two important respects. First, the retail and restaurant sectors exhibit relatively high investment reversibility compared to other industries, ranking in the 70th percentile for asset redeployability.[46] This flexibility allows firms to repurpose or sell assets with minimal loss, reducing the benefit of delaying. Second, competition in these sectors is highly localized due to small trade areas, which limits the number of firms that can enter and succeed. In such an environment, the strategic advantages of early entry, such as securing prime locations, establishing brand presence, and capturing market share, may outweigh the benefits of waiting for improved information.[47]

$$\begin{aligned}\text{Entry}_{mt} = \beta_0 &+ \beta_1 \text{Pre-census}_t + \beta_2 \text{Census}_t + \beta_3 \text{Post-census}_t \\ &+ \beta_4 \text{GDP growth}_{t-1} + \beta_5 \text{Expected GDP growth}_{t-1} + \beta_6 \text{VXO}_{t-1} \\ &+ \beta_7 \text{EPU}_{t-1} + \beta_8 \text{Forecast dispersion}_{t-1} + \alpha_m + \epsilon_{mt}\end{aligned} \tag{1.7}$$

To examine how entry patterns change around census data releases, I regress market-level entry each year on indicators for census timing and various controls for economic conditions and uncertainty, as specified in Equation 1.7. The

[46] Estimates are based on the redeployability measure from Kim and Kung (2017), averaged over the sample period.

[47] Under the context of retail chain competition, Igami and Yang (2016) demonstrate that these strategic incentives can accelerate entry and expansion.

indicator variables *Pre-census*, *Census*, and *Post-census* correspond to the two years before, during, and after census data release, respectively. The coefficients on these indicators represent the excess number of establishment entries during these periods.

Unlike in the failure rates regressions, where existing establishments provide a natural benchmark for entry cohorts, the entry analysis lack a comparable control group. Including year fixed effects would absorb the variations of interest. Instead, I control for confounding factors using proxies for macroeconomic conditions and uncertainty that may influence entry decisions.

To account for macroeconomic conditions, I include real GDP growth based on data from the St. Louis Federal Reserve as an indicator of overall economic activity, along with the one-year-ahead GDP forecast from the Livingston Survey, which provides a forward-looking measure of anticipated economic growth. To capture uncertainty, I employ three measures: (1) the VXO index from the Chicago Board Options Exchange, which measures the implied volatility of the stock market; (2) the composite Economic Policy Uncertainty (EPU) index for New York state, measuring policy-related economic uncertainty (Baker, Bloom and Davis, 2016); and (3) forecast dispersion, calculated as the standard deviation of GDP forecasts from the Livingston Survey, reflecting disagreement among professional forecasters.

All regression includes market fixed effects to account for market heterogeneity. Standard errors are two-way clustered at the market and year level, and bootstrapped p-values are reported for the key parameters to adjust for the small number of temporal clusters, given the limited number of years in the data.

The results presented in Table 1.9 do not support that firms alter their investment timing around the release of census data. Specifically, the coefficients on the *Pre-census*, *Census*, and *Post-census* indicators are statistically insignificant across all firm types. This pattern is consistent with the earlier discussion on the unique characteristics of the retail and restaurant sectors that reduce the benefits of waiting for updated information.

Another possible explanation is that firms may not be aware that the data they are using is stale. As detailed in subsubsection 1.2.3.1, businesses frequently rely on intermediaries such as data analytics companies, market research firms, and commercial real estate agents for demographic insights. Although these intermediaries rely on census data, end users rarely informed about how frequently the underlying data is updated. Even when firms access the data directly, they may not fully recognize the implications of using potentially outdated information. As a result, firms may base entry decisions on what they perceive to be current market conditions, leading to no observable delays around census data releases.

1.9 Conclusion

This paper empirically measures the impact of information quality on firm investment decisions by examining how outdated census data, which provides key demographic information for market entry, affects firm failure. Using establishment-level entry and exit data on retail firms, I find that outdated information raises the 5-year failure rate by 1.6 percentage points per year. Over the decade-long gap between decennial censuses, this effect accumulates to 16

percentage points, equivalent to a 32% rise over the baseline failure rate.

These findings underscore the value of timely and accurate census data, contributing to ongoing policy debate on the costs and benefits of comprehensive census data collection. When Canada replaced its mandatory long-form census with a voluntary survey in 2011, data gaps disrupted the availability of reliable demographic statistics for public policy decisions, prompting widespread concerns.[48] In the U.S., new confidentiality measures in the 2020 census compromised the reliability of block-group-level data.[49] More recently, disruptions in government data availability have reignited concerns about access to high-quality public statistics.[50] Although census data is primarily designed for political representation and public planning, this paper highlights an additional channel through which lack of data can also distort private market decisions, particularly for small establishments that depend on public data sources. Future research and policy discussions can take this perspective into account when evaluating the broader welfare implications of census data policies.

Beyond the retail and restaurant sectors, many industries rely on demographic data to guide critical business decisions. For instance, manufacturing firms use census data to evaluate labor markets and plan production locations; logistics companies analyze population density to optimize distribution networks; and advertising agencies rely on demographic trends for consumer seg-

[48]See Aarian Marshall, "The Tragedy of Canada's Census," *Bloomberg*, August 28, 2015, https://www.bloomberg.com/news/articles/2015-08-28/the-tragedy-of-canada-s-census (accessed August 15, 2024).

[49]See Michael Wines, "The 2020 Census Suggests That People Live Underwater. There's a Reason," *The New York Times*, April 19, 2022, https://www.nytimes.com/2022/04/19/us/census-privacy.html (accessed April 22, 2022).

[50]For example, in early 2025, federal agencies removed thousands of datasets and webpages, affecting demographic, health, and economic data availability. See Robert Cyran, "Disappearing US Data Dims Economic Outlook," *Reuters*, February 7, 2025, https://www.reuters.com/breakingviews/disappearing-us-data-dims-economic-outlook-2025-02-07/ (accessed April 23, 2025).

mentation and targeted marketing. While this study focuses on sectors where site selection decisions are observable at the establishment level, future research could explore whether similar patterns of information-driven misallocation occur in other industries.

From a methodological perspective, the presence of information waves via census data releases might serve as helpful exogenous drivers of firm exit for research about firm productivity (De Loecker and Syverson, 2021) that relies on production function estimation (Olley and Pakes, 1996). One prevalent identification issue in this stream of research is selection bias, stemming from non-random exit (i.e., observed production levels of firms are conditional that they are active). Census data timing might help augment existing selection correction methods, especially for multi-country research about productivity dispersion (Asker, Collard-Wexler and De Loecker, 2014), whereby firms in different countries experience different release schedules from their respective censuses.

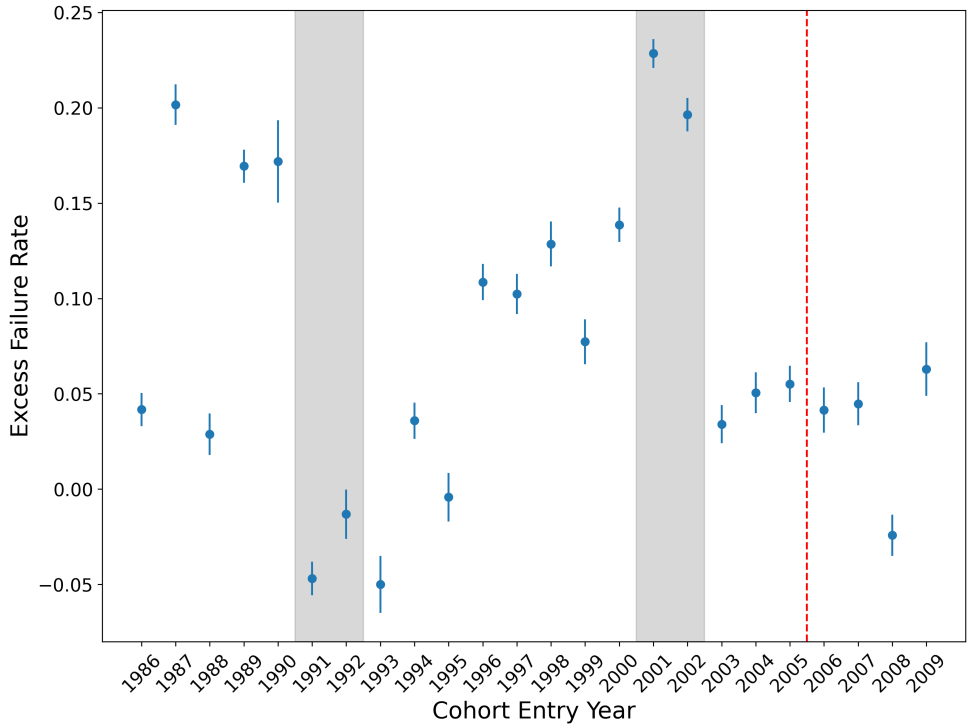
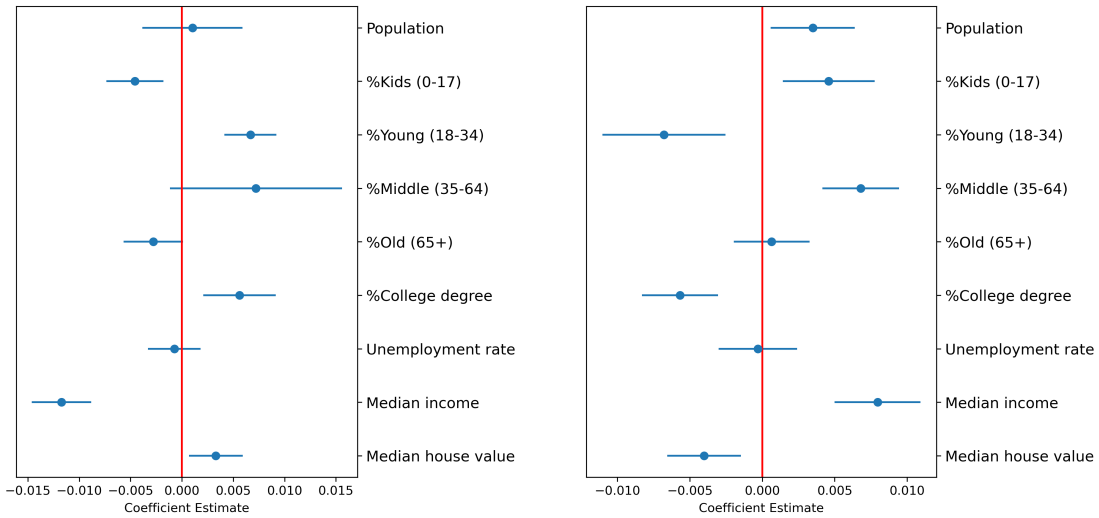


Figure 1.1: Excess Failure Rate by Entry Cohorts

Notes: The figure plots coefficients and associated 95% confidence intervals of the entry-year indicator variables in Equation 1.3. The dependent variable is excess failure rate. Standard errors are clustered at the census-tract level. The shaded areas indicate census data release intervals, while the dotted line denotes the introduction of the American Community Survey.



Negative Shift Positive Shift

Figure 1.2: Coefficient Estimates (γ_2) by Demographic Variable

Notes: This figure reports coefficient estimates γ_2 and associated 95% confidence intervals from the regression Equation 1.6 for each demographic variable. A negative (positive) shift is a decrease (increase) in the demographic variable larger than 10%. Standard errors are clustered at the census-tract level. The vertical line indicates where the coefficient estimate is 0.

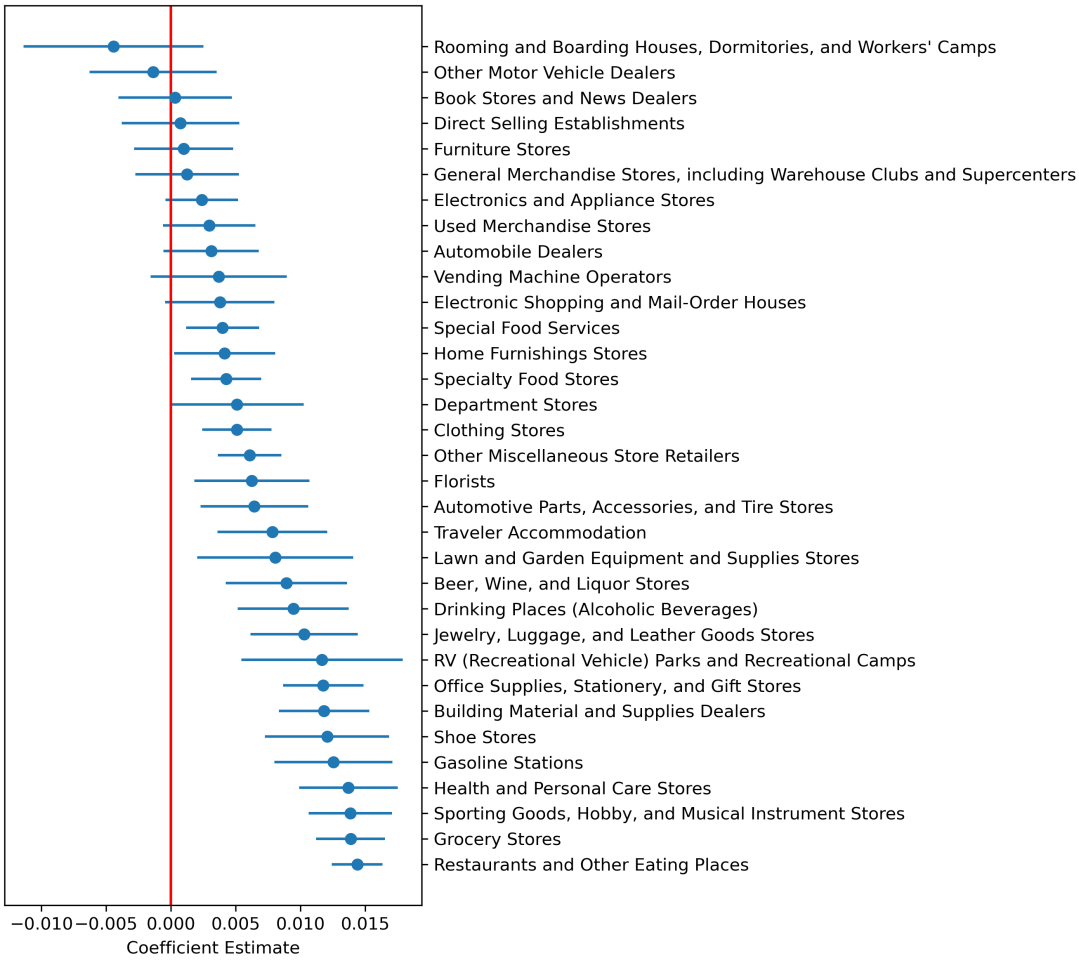


Figure 1.3: Coefficient Estimates (β_2) by NAICS 4-digit Industry

Notes: This figure plots β_2 coefficients estimated from Equation 1.5 and the associated 95% confidence intervals for each NAICS 4-digit industry. Standard errors are clustered at the census-tract level. The vertical line indicates where the coefficient estimate is 0.

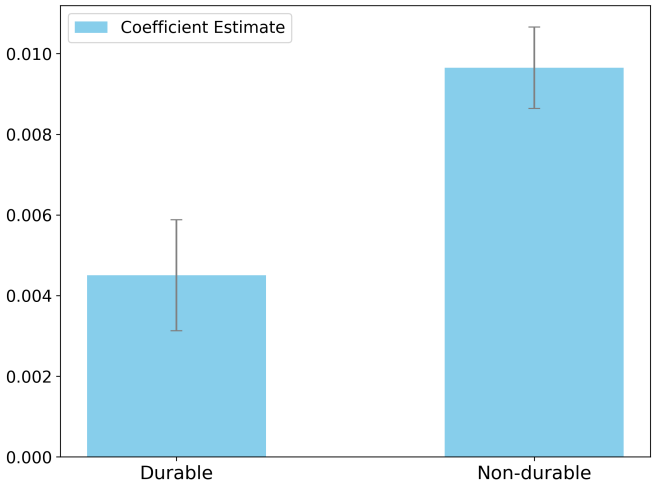


Figure 1.4: Comparison of Coefficient Estimates (β_2) between Durable and Non-durable Goods Retailers

Notes: This figure plots β_2 coefficient estimates from Equation 1.5 and associated 95% confidence intervals across durable and non-durable goods retailers. Standard errors are clustered at the census-tract level.

Table 1.1: Decennial Census Data Release Timeline

Data File	Content	1990 Census	2000 Census
Redistricting Summary File	Population counts used for redistricting	March 1991	March 2001
Summary File 1	Population and housing characteristics	March 1991	June 2001
Summary File 2	Cross tabulations of SF1 by racial groups	August 1991	September 2001
Summary File 3	Detailed socioeconomic characteristics	May 1992	June 2002
Summary File 4	Cross tabulations of SF3 by racial groups	March 1993	October 2002
PUMS	Samples of individual responses	July 1993	April 2003

Sources: https://www.census.gov/programs-surveys/decennial-census/technical-documentation/complete-technical-documents.1990.html

https://www.ibrc.indiana.edu/ibr/2001/spring01/05.pdf

Table 1.2: Market-Level Establishment Counts and Turnover

	p10	p25	p50	p75	p90	N
Number of Establishments	8	16	31	55	91	144912
Number of New Entrants	0	1	3	7	13	144912
5-year Failure Rate for Entry Cohorts	0.00	0.25	0.50	0.67	1.00	100433

Notes: This table presents summary statistics on establishment data across market-years. 5-year Failure Rate for Entry Cohorts is computed for establishments entering up to 2009, while other statistics include all market-years.

Table 1.3: Changes in Demographic Variables

Demographic Variable	p10	p25	p50	p75	p90
Population	-0.09	-0.04	0.02	0.10	0.21
%Kids (0-17)	-0.20	-0.13	-0.04	0.06	0.16
%Young (18-34)	-0.28	-0.18	-0.07	0.03	0.11
%Middle (35-64)	-0.03	0.02	0.07	0.13	0.20
%Old (65+)	-0.24	-0.11	0.03	0.20	0.40
%College degree	-0.33	-0.07	0.14	0.35	0.66
Unemployment rate	-0.55	-0.36	-0.08	0.30	0.86
Median income	-0.21	-0.11	-0.01	0.12	0.28
Median house value	-0.25	-0.14	0.06	0.55	1.21
Observations	89691				

Notes: This table describes the distribution of changes in local demographics between two decennial censuses at the census-tract level. Table A.3 provides details on the construction of each demographic variable.

Table 1.4: Excess Failure Rate and Distance to Census Data Release

	Full Sample	1986-1999 Cohorts	2000-2009 Cohorts
	(1)	(2)	(3)
β_1	0.006	0.034***	-0.032***
	(0.005)	(0.008)	(0.006)
β_2	0.016***	0.025***	0.007***
	(0.001)	(0.001)	(0.001)
α_1	0.105***	0.021	0.164***
	(0.009)	(0.014)	(0.010)
α_2	0.010***	-0.012**	0.027***
	(0.002)	(0.004)	(0.003)
Observations	100,433	56,299	44,134
R^2	0.008	0.027	0.036

Notes: This table reports results from estimating the relationship between excess failure rate and an entry cohort's distance to census data release, using piece-wise linear regressions with the break-point at two years after initial release. The coefficients α_1 and β_1 represent the intercept and the slope of the first phase; the coefficients α_2 and β_2 represent the intercept and the slope of the second phase, respectively. The main coefficient of interest β_2 captures on average how much failure rate increases when entry moves one year further away from when the census snapshot was taken. Standard errors clustered at the census-tract level. *** p<0.001, ** p<0.01, * p<0.05.

Table 1.5: Census Tracts with Little Change in Demographic Variables

	(1)
β_1	0.010
	(0.089)
β_2	-0.000
	(0.012)
α_1	0.099
	(0.150)
α_2	0.079
	(0.043)
Constant	
Observations	276
R^2	0.000

Notes: This table replicates the main specifications using the subsample of census tracts that have less than 10% absolute change in value for *%Young*, *%College degree*, *Median house value*, *%Kids* and *%Median income*. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table 1.6: Chain vs Independent

Panel A: Full Sample

	Large Chain	Small Chain	Independent
	(1)	(2)	(3)
β_1	-0.017	-0.016	0.005
	(0.013)	(0.014)	(0.005)
β_2	0.001	0.011***	0.017***
	(0.001)	(0.002)	(0.001)
Adjusted β_2	0.005	0.037	0.033
α_1	-0.008	-0.014	0.100***
	(0.022)	(0.024)	(0.009)
α_2	-0.021***	-0.025***	0.009***
	(0.005)	(0.006)	(0.003)
Observations	19045	17912	99081
R^2	0.002	0.003	0.008

Panel B: Sub-samples

	1986-1999 Entry Cohorts			2000-2009 Entry Cohorts		
	Large Chain	Small Chain	Independent	Large Chain	Small Chain	Independent
	(1)	(2)	(3)	(4)	(5)	(6)
β_1	-0.023	-0.036*	0.037***	-0.012	0.034	-0.040***
	(0.018)	(0.016)	(0.009)	(0.018)	(0.025)	(0.006)
β_2	0.005**	0.016***	0.025***	-0.003	0.002	0.008***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.003)	(0.001)
Adjusted β_2	0.026	0.053	0.049	-0.015	0.007	0.016
α_1	-0.020	-0.061*	0.020	0.005	0.105*	0.151***
	(0.031)	(0.028)	(0.015)	(0.031)	(0.044)	(0.010)
α_2	-0.046***	-0.041***	-0.004	0.009	0.006	0.018***
	(0.007)	(0.008)	(0.004)	(0.007)	(0.013)	(0.003)
Observations	11529	14143	55239	7516	3769	43842
R^2	0.004	0.006	0.028	0.001	0.002	0.036

Notes: This table summarizes the tests for the relationship between excess failure rate and an entry cohort's distance to census data release separately for large chains (more than 20 outlets), small chains (between 2-20 outlets), and independent establishments. Panel A uses the full sample, while Panel B compares the pre-2000 and post-2000 entry cohorts. The adjusted β_2 coefficients are calculated by dividing the original estimates by the average failure rate for each firm type within the corresponding period. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table 1.7: Effects of Census Data Quality on Establishment Failure Rate by Local Recession Severity

	Early 90s Recession		Early 00s Recession	
	Below Median	Above Median	Below Median	Above Median
	(1)	(2)	(3)	(4)
β_1	0.023**	-0.018*	0.008	0.004
	(0.007)	(0.008)	(0.008)	(0.008)
β_2	0.016***	0.016***	0.015***	0.019***
	(0.001)	(0.001)	(0.001)	(0.001)
α_1	0.119***	0.083***	0.103***	0.107***
	(0.012)	(0.013)	(0.012)	(0.013)
α_2	-0.002	0.026***	0.020***	-0.002
	(0.003)	(0.004)	(0.004)	(0.004)
Observations	58594	41839	54805	45628
R^2	0.008	0.009	0.006	0.011

Notes: This table summarizes replication results of the main specifications on subsamples of census tracts with different impact of recessions. The severity of each recession is measured by the percentage change in county-level same-month unemployment rate before and after the recession (June 1990 to June 1991 for the early 90s recession and February 2001 to February 2002). The county-level unemployment rate series are from the Bureau of Labor Statistics (`https://www.bls.gov/lau/`). The recession dates are from NBER (`https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions`). Column (1) - (2) are based on the early 1990s recession. Columns (3) - (4) are based on the 2000s recession. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table 1.8: Effects of Census Data Quality on Establishment Failure Rate by Census-Based Subsidy Eligibility

	Eligible for Empire Zones		Eligible for Opportunity Zones	
	No	Yes	No	Yes
	(1)	(2)	(3)	(4)
β_1	-0.001	0.045***	-0.006	0.052***
	(0.006)	(0.013)	(0.006)	(0.011)
β_2	0.017***	0.014***	0.017***	0.013***
	(0.001)	(0.002)	(0.001)	(0.001)
α_1	0.098***	0.140***	0.089***	0.163***
	(0.009)	(0.021)	(0.010)	(0.018)
α_2	0.010***	0.010	0.010***	0.010
	(0.003)	(0.006)	(0.003)	(0.005)
Observations	84139	16294	80183	20250
R^2	0.009	0.006	0.009	0.006

Notes: This table summarizes replication results of the main specifications on subsamples of census tracts with respect to eligibility for government subsidy. Column (1) - (2) are based on eligibility for Empire Zones. Columns (3) - (4) are based on eligibility for Opportunity Zones. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table 1.9: Entry Patterns Surrounding Census Data Release

	(1)	(2)	(3)	(4)
	All firms	Large chain	Small chain	Indepedent
Pre-census	-0.021	-0.034	0.203*	-0.190
	(1.611)	(0.065)	(0.075)	(1.490)
Bootstrap p-value	0.99	0.70	0.21	0.91
Census	2.781*	0.049	0.159	2.573*
	(1.065)	(0.056)	(0.102)	(0.975)
Bootstrap p-value	0.11	0.51	0.71	0.12
Post-census	0.792	0.004	0.018	0.770
	(0.862)	(0.045)	(0.062)	(0.777)
Bootstrap p-value	0.52	0.96	0.81	0.48
GDP growth	0.076	0.009	0.079***	-0.013
	(0.234)	(0.011)	(0.018)	(0.212)
Expected GDP growth	-0.293	0.008	-0.054**	-0.247
	(0.337)	(0.014)	(0.017)	(0.313)
VXO	-0.018	-0.003	-0.008*	-0.007
	(0.053)	(0.003)	(0.004)	(0.050)
EPU	-0.006	-0.000	0.000	-0.006
	(0.006)	(0.000)	(0.000)	(0.006)
Forecast dispersion	-0.378	-0.000	-0.029	-0.349
	(0.893)	(0.037)	(0.037)	(0.820)
Constant	6.671***	0.304***	0.224**	6.143***
	(0.976)	(0.059)	(0.073)	(0.897)
Observations	112539	112539	112539	112539
R^2	0.667	0.374	0.336	0.658

Notes: This table reports coefficients estimated from Equation 1.7. Column (1) uses the full sample, while Columns (2) - (4) report sub-sample results for large chain, small chain and independent establishments respectively. Parentheses contain standard errors clustered at the census-tract and year level, and bootstrapped p-valuess are reported for the key parameters to adjust for the small number of clusters. Significance: *** p<0.001, ** p<0.01, * p<0.05.

CHAPTER 2

THE USE AND USEFULNESS OF BIG DATA IN FINANCE: EVIDENCE FROM FINANCIAL ANALYSTS

2.1 Introduction

Sell-side analysts have long played a pivotal role in financial markets (Bradley, 2023). They gather information on publicly traded companies and share their analyses and opinions with investors. To the degree that analysts pass on unique and value-relevant insights, they can contribute significantly to enhancing market efficiency.

Like many roles in knowledge-driven sectors, the analyst profession is being challenged by technological advances: With the advent of modern information technologies and advances in data analytics, we can increasingly track individuals' and businesses' activities through the digital footprints they leave behind. In the realm of investments, an expanding body of research finds that these digital footprints, commonly referred to as "big data" or "alternative data," are useful in predicting company performance.[1]

Alternative data—with its capacity to generate insights in a highly timely, comprehensive, and accurate manner—holds the potential to greatly diminish the usefulness of analysts to the investor community, eventually rendering the analyst profession obsolete.

[1]Examples include Froot et al. (2017), who consider consumer activity estimated in real-time from mobile phone data; Huang (2018) and Zhu (2019), who consider online consumer activity; and Zhu (2019), Kang, Stice-Lawrence and Wong (2021), Katona et al. (2023), and Gerken and Painter (2023), who consider satellite images of retail store parking lots.

Yet, there exists a contrasting perspective. Suppose alternative data can provide unique and value-relevant insights and that investors yearn to incorporate these insights into their decision-making. One challenge that may prevent investors from doing so is the substantial cost associated with subscribing to alternative data. Relatedly, extracting meaningful insights from alternative data requires skill and experience. Instead of each individual investor's bearing the cost of subscribing to and learning about alternative data, it is more economical for a few analysts to incur these expenses and subsequently share their insights with investors. Analysts—who have the resources, readiness, and skill to study alternative data—could, therefore, utilize alternative data to maintain (or increase) their usefulness to investors rather than being made obsolete.

Our paper examines this assertion. Our empirical analysis builds on the following idea: Analysts routinely publish written reports detailing their perspectives, along with the data and analyses underpinning these views. We posit that if analysts accessed alternative data and if the consideration of alternative data meaningfully altered their beliefs, they would discuss such influence in the corresponding reports. By parsing analysts' written reports and checking whether they explicitly reference the use of alternative data, we can thus gauge how often they draw from alternative data and in what contexts. We can also study investors' reactions to the issuance of such reports. To provide more information on our parsing approach, we start with a comprehensive list of in-house data-science teams and external alternative-data vendors. We search for the names of these teams and vendors in analysts' written reports. We then, following prior literature (Hoberg and Moon, 2017, 2019), conduct an iterative keyword search. Among the reports that contain the name of a team or vendor, we extract a list of keywords that analysts use to describe the alternative data.

We then use these keywords to search for additional reports that discuss the use of alternative data. At the end of the process, we manually read the relevant passages of each captured report to verify that the report indeed mentions the use of alternative data.

Given the labor-intensive nature of our identification process, we limit our primary analysis to companies listed in the Dow Jones Industrial Average index (DJI) from June 2009 to May 2019. The DJI includes 30 large publicly traded companies. As the composition of the DJI changes over time, our final sample comprises 35 firms. We search for analyst reports on these 35 firms within the Investext database and integrate annual earnings forecast data from IBES. Ultimately, our dataset encompasses 64,018 written reports and their corresponding annual earnings forecasts, issued by 1,002 distinct analysts employed by 55 brokerage firms.[2]

Our analysis reveals that, by 2009/2010, 11% of the analysts in our sample explicitly reference the use of alternative data in at least one of their reports. By 2018/2019, the corresponding fraction is 28%. Our analysis differentiates eight alternative data categories: app usage, sentiment, employee, geospatial, point of sale, satellite image, web traffic, and others. We find explicit references from all eight categories within the first year of our sample period.

To gauge whether analysts are able to distill unique and value-relevant insights from alternative data, we test whether the annual earnings forecasts from reports that mention the use of alternative data are more accurate than those from reports without alternative data references. We observe within

[2]As we discuss in the main body of the paper, in additional analyses we draw a random sample of 200 companies from the lower half of the size distribution within the CRSP database. We then examine the degree to which our main findings, observed among the largest firms, extend to these smaller firms.

a difference-in-differences specification that annual-earnings-forecast accuracy and alternative-data adoption are positively correlated with each other. Our estimates suggest that the performance improvement accompanying the consideration of alternative data is equivalent to having covered the corresponding firm for 3.6 additional years. The foundation of the analyst business model rests on soft dollar arrangements, in which research services are funded through trading commissions. As investors perceive greater value in an analyst's research, they increase the number of trades routed through the analyst's brokerage and, as a result, pay higher commissions to the corresponding brokerage (Bradley, 2023). The perceived value of an analyst's research and the magnitude of trading commissions are thus directly linked with each other. To ascertain if analysts' adoption of alternative data enhances their value to investors, we therefore examine whether discussions of alternative data positively correlate with the magnitude of commissions received. To implement our test, we obtain institutional-investor trade data from ANcerno, a firm that provides transaction-cost analysis to institutional clients. The ANcerno dataset contains specific information for each institutional trade, including the commissions paid to the broker. For each written report, we consider all trades on the corresponding stock within three months after the report was issued and aggregate the total dollar value of the commissions paid to the corresponding broker. We observe within a difference-in-differences specification that institutional investors pay more in trading commissions when an analyst references alternative data. Our finding aligns with our supposition that analysts can enhance their value to investors by adopting alternative data. Our final test examines the ramifications for investors and financial markets. We hypothesize that barriers to adopting alternative data are more pertinent for smaller or traditional institutional investors,

such as mutual funds, than for hedge funds. Consistent with this conjecture, industry reports document that hedge funds are the primary consumers of alternative data.[3] By distilling insights from alternative data and sharing them with non-hedge-fund investors, analysts could play a crucial role in leveling the playing field for hedge funds and non-hedge-fund investors.

To assess this proposition, we again utilize our institutional investors' trade data. Existing research indicates that hedge funds place more profitable trades than other institutional investors (Stulz, 2007). Our paper examines whether this performance gap narrows when analysts incorporate alternative data and share their insights with investors, including non-hedge-fund investors.

To assess the performance of institutional investors, we construct transaction-based calendar time portfolios following the approach of Seasholes and Zhu (2010) and Ben-David, Birru and Rossi (2019). As we are interested in trades affected by the dissemination of alternative-data insights, we focus on transactions occurring within three months after the issuance of an analyst report that involves the stock mentioned in the report. We employ the methodology outlined by Jame (2018) to distinguish trades by hedge funds from those by other institutional investors.

First, we find that the stocks that hedge funds buy substantially outperform those they sell over the ensuing three months; the outperformance is 10.79% when annualized. The 10.79% figure serves as our benchmark against which we compare the performance of non-hedge-fund institutional investors. Our initial comparison considers cases where analysts do not reference alternative data. In these cases, we find that the stocks that non-hedge funds buy outper-

[3] E.g., https://www.institutionalinvestor.com/article/2bsxas8ahvzgesgpnq0hs/portfolio/its-time-to-cash-in-on-big-data

form those they sell by 3.85%. The results are noticeably different when analysts reference alternative data. In such cases, the outperformance is 9.85%. The similarity in outperformance for hedge funds and non-hedge funds when analysts share their alternative-data insights is consistent with the notion that analysts' use of alternative data helps level the playing field for hedge funds and non-hedge-fund institutional investors. Our paper contributes to several research areas. First, our paper relates to the literature on sell-side analysts. Analysts historically served as vital information intermediaries in financial markets and they have been the subject of extensive research. Bradley (2023) observes that the term "analyst" and its variations have appeared in the titles and abstracts of nearly 2,500 papers in the top finance and accounting journals over the last two decades. An emerging body of work shows that the analyst field is suffering from a rise in passive investing and regulatory changes, which have led to a decline in funding for analyst research (Bradley, 2023). The analyst profession is further threatened by technological advancements. There has been a surge in alternative-data vendors (Grennan and Michaely, 2020). These data appear to predict financial performance above and beyond what is forecasted by human analysts (Froot et al., 2017; Huang, 2018). Relatedly, Jame et al. (2016) show that online platforms, which crowdsource investors' earnings forecasts and make the consensus forecasts publicly available, produce predictions that are more accurate than those provided by human analysts. Coleman, Merkley and Pacelli (2022) study the emergence of "robo-analysts," which are "research firms that focus on the use of technology to mass-produce recommendations with limited human involvement" (page 12). The authors find that the recommendations of Robo-Analysts are superior to those of human analysts. Cao et al. (2023) develop an "AI analyst" by leveraging a range of machine-learning toolkits. Like

Coleman, Merkley and Pacelli (2022), the authors find that their AI Analyst outperforms human analysts. Collectively, the above studies point to a paradigm shift in the way investors can obtain information, with a trend toward advanced data analysis and big data. Our study adds to the analyst literature by providing evidence that some analysts respond to this trend by adopting alternative data themselves. Our results suggest that—by doing so—these analysts have maintained (or enhanced) their relevance to investors. In that vein, our findings shed light on the broader question of how recent technological advances impact the labor market, particularly in knowledge-intensive industries (Agrawal, Gans and Goldfarb, 2019; Frank et al., 2019). Our observations indicate a continued need for human labor, provided that individuals adapt to the evolving technological landscape. One possible reason is that humans can use their intuition to distill insights from machine-generated outputs that, as of now, are lost to machines. As a result, there is value in combining human and machine capabilities, a proposition that aligns with the findings of Cao et al. (2023). Cao et al. (2023) find that while their AI Analyst outperforms human analysts, the most accurate predictions are theoretically achievable by integrating inputs from both AI and human analysts. Another possible reason for the sustained need for human labor is the presence of barriers, which prevent many humans from adopting the technology. As long as these barriers exist, there will be opportunities for certain humans to specialize in these technologies and share their insights with those who have not yet adopted them. Finally, our study contributes to the body of research that examines the consequences of alternative data for financial markets. The existing literature primarily studies the predictive value of alternative data for companies' future performances.[4] There is comparatively little

[4]Additionally, some studies suggest that econometricians can use alternative data, such as satellite images of retail store parking lots, to create proxies for local store performance Kang, Stice-Lawrence and Wong (2021); Gerken and Painter (2023). Researchers can then use these

research investigating the actual usage of alternative data by financial-market participants. Current studies in this area typically take the initial offering of an alternative dataset as a positive shock to alternative-data availability and compare financial-market outcome variables from before and after the shock (Zhu, 2019). Other studies estimate financial-market participants' use of alternative data for a specific stock by measuring how heavily the stock is covered in an alternative dataset (Dessaint, Foucault and Frésard, 2023). In contrast, our research considers situations where analysts explicitly describe their use of alternative data. One limitation of our method is that it might overlook situations where analysts consider alternative data but do not mention their use in their reports. The advantage of our approach is that we can document precisely what kinds of alternative data analysts actually use, for which firms, and in what situations. The descriptive evidence we present may provide a valuable reference source for future work on this topic.

2.2 Analysts' Use of Alternative Data

We begin the main body of our paper with a definition of alternative data. We then describe how we capture analyst reports that reference the use of alternative data.

proxies to study the degree to which local analysts or local investors rely on, or perhaps overemphasize, local information.

2.2.1 Alternative Data and Historical Perspective

Alternative data trace the footprints individuals and firms leave behind through their day-to-day activities. These footprints are also known as "exhaust data." The use of exhaust data is not new to the financial sector. In the past, investment firms dispatched their junior analysts to retail stores to sample the foot traffic; other firms directed their analysts to manufacturing plants to count the number of trucks moving in and out (McMahon and Chu, 2012; Wigglesworth, 2016). What distinguishes alternative data from the previous exhaust data is that, with the advent of modern information technologies and the rise in computing power, we can now source exhaust data instantaneously, comprehensively, and from a large variety of sources. That is, rather than counting foot traffic for select branches manually over several days, we can now comprehensively track how many consumers visit a merchant's website. There are broadly eight alternative data categories: (1) app-usage data, which track the number of active mobile app users and the amount of time they spend on the apps; (2) sentiment data, which include product ratings posted on the Internet and social-media feeds regarding a company's products and services; (3) employee data, which include online job postings, employee opinions, and manager statements; (4) geospatial data, which contain information about the locations in which a company operates branches; (5) point-of-sale data, which include merchant-level transaction data, product-level purchase data, and pricing data; (6) satellite-image data, which include satellite images of parking lots, manufacturing plants, and construction sites; (7) web-traffic data, which track what terms users search for on the Internet and how frequently and for how long users visit a merchant's website; and (8) other, which include data that do not fit cleanly into any of the other seven categories (e.g., data on senders and recipients of barrels loaded

onto vessels).

2.2.2 Measuring Reliance on Alternative Data

To gauge analysts' use of alternative data, we conduct textual analysis of their written reports. We use the Investext database to download analyst reports for constituents of the DJI from June 1, 2009, through May 31, 2019. The Investext database contains active and historical research reports from brokerages, investment banks, and independent research firms around the globe. At any point in time, constituents of the DJI represent 30 large publicly traded firms. Because the DJI constituent list varies over time, our final sample comprises 35 firms.[5] For each report, we extract the ticker symbol, company name, report date, analyst names, broker name, report title, and full text. The average report in our sample contains 2,152 words, which is the equivalent of roughly five pages. We merge our Investext data with annual-earnings-forecast data from the IBES database. We merge these two datasets based on ticker symbols, company names, broker names, analyst names, and dates of forecast issuances. Of our initial sample of 70,353 Investext reports, we successfully match 65,009 Investext reports (65,009/70,353 = 92.4%). After merging these matches with financial-market data from CRSP and financial-statement data from Compustat, our final sample comprises 64,018 reports and earnings forecasts issued by 1,002 analysts from 55 brokers. Just as some brokerages are missing from

[5]The mean and median market capitalization of firms in our sample (as of 2019) are $250 billion and $222 billion, respectively. To put these numbers in perspective, the 99th market capitalization percentile among firms in the CRSP/Compustat universe (also as of 2019) is $144 billion. DJI constituents are thus substantially larger than most firms in the CRSP/Compustat universe. Online Appendix Table A1 compares the industry distribution of the firms in our sample with that of the firms in the CRSP/Compustat universe. Compared with the CRSP/Compustat universe, our sample overweights the Consumer Staples sector and underweights the Health Care sector.

the IBES dataset (e.g., UBS), others are not included in the Investext database (e.g., Goldman Sachs). The non-comprehensive brokerage coverage raises questions regarding the generalizability of our findings. Although we cannot dismiss the potential for non-representative sampling, when we consider analyst reports and earnings forecasts on DJI constituents in the Investext and the IBES databases, respectively, we find no notable differences in underlying brokerage characteristics. The similarities in brokerage characteristics are evident in several key metrics: the average number of firms each analyst covers (10.07 vs. 8.34), the average number of forecasts issued per month (4.84 vs. 4.15), and the average forecast accuracy (-0.54 vs. -0.52). At least based on these observables, there does not seem to be systematic selection bias.

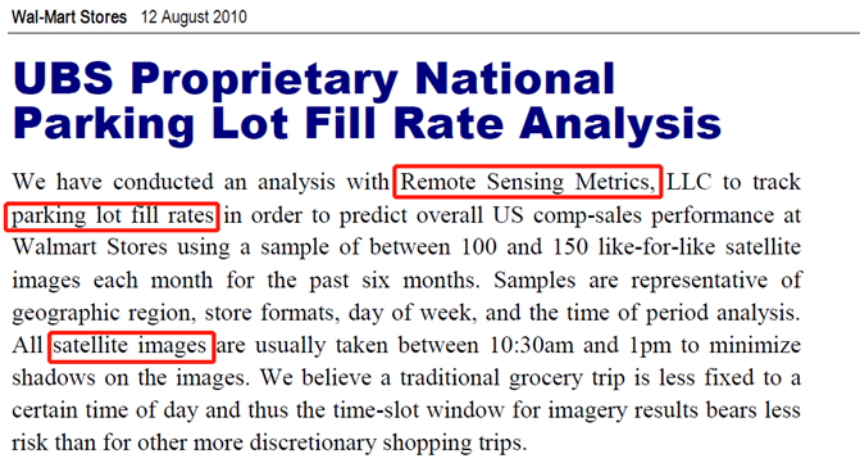
We proceed as follows: We compile a list of in-house data-science teams and external alternative-data vendors from the vendor lists in the J.P. Morgan 2019 Alternative Data Handbook and AlternativeData.org, a platform that connects users to alternative-data providers.[6] To facilitate research on this topic, we report the list of 520 teams and vendors in Online Appendix Figure A1. We use the list of full and abbreviated names as our initial keywords and search for them in analysts' written reports.[7] For each report identified by these initial keywords, we read the passages surrounding the initial keywords to verify that the report indeed adopts alternative data. Some analysts explicitly reference the use of alternative data without disclosing their source. To capture these reports, we follow prior literature (Hoberg and Moon, 2017, 2019) and conduct an iterative keyword-search process. In particular, within our first set of analyst reports that reference an in-house data-science team or external alternative-data

[6]In-house data-science teams specialize in collecting and analyzing large unstructured data, which analysts can use in their valuation efforts.

[7]We convert all names and all text in the reports to lowercase characters.

vendor, we extract a list of keywords that analysts use to describe the alternative data. We then use these new keywords to search for additional reports that incorporate alternative data (but do not reference their sources) and continue expanding our keywords list. Using this iterative process, we arrive at our final set of keywords, which we use to identify reports that reference the use of alternative data. We report our final set of keywords in Appendix 1. In our last step, we (again) read all reports flagged as using alternative data to verify that the analysts indeed discuss the use of alternative data in their analyses.[8]

To illustrate our process by example, one of the alternative data vendors in our sample is "Remote Sensing Metrics," also referred to as "RS Metrics." We first search for reports containing the terms "Remote Sensing Metrics" or "RS Metrics." We find 47 reports that contain these two keywords. The figure below is an excerpt from one such report:[9]



Wal-Mart Stores 12 August 2010

UBS Proprietary National Parking Lot Fill Rate Analysis

We have conducted an analysis with Remote Sensing Metrics, LLC to track parking lot fill rates in order to predict overall US comp-sales performance at Walmart Stores using a sample of between 100 and 150 like-for-like satellite images each month for the past six months. Samples are representative of geographic region, store formats, day of week, and the time of period analysis. All satellite images are usually taken between 10:30am and 1pm to minimize shadows on the images. We believe a traditional grocery trip is less fixed to a certain time of day and thus the time-slot window for imagery results bears less risk than for other more discretionary shopping trips.

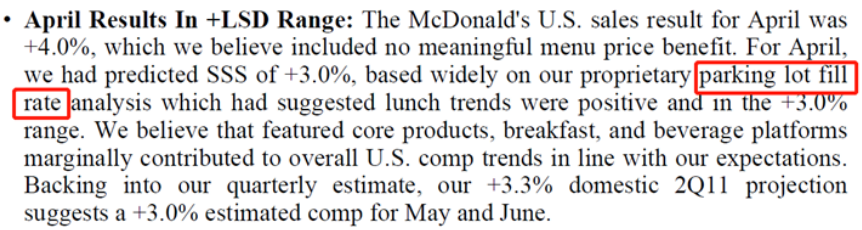
Reading the text surrounding the keyword "Remote Sensing Metrics," we

[8]To researchers interested in further studying the use of alternative data, we caution that analysts' use of the keywords presented in Appendix 1 is a necessary but not a sufficient condition. We found our final step of carefully re-reading all reports to eliminate false positives crucial for cleanly separating reports that adopt alternative data from those that do not.

[9]UBS; Neil Currie, Krista Zuber, and David Eads; Walmart Inc; August 12, 2010.

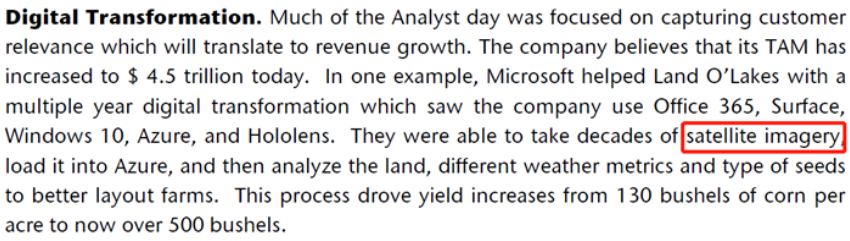
identify two additional keywords related to alternative data: "parking lot fill rates" and "satellite image." We use these new keywords to search for more reports that adopt alternative data but do not reference "Remote Sensing Metrics" or "RS Metrics."

The figure below is an excerpt from one such report:[10]



- **April Results In +LSD Range:** The McDonald's U.S. sales result for April was +4.0%, which we believe included no meaningful menu price benefit. For April, we had predicted SSS of +3.0%, based widely on our proprietary parking lot fill rate analysis which had suggested lunch trends were positive and in the +3.0% range. We believe that featured core products, breakfast, and beverage platforms marginally contributed to overall U.S. comp trends in line with our expectations. Backing into our quarterly estimate, our +3.3% domestic 2Q11 projection suggests a +3.0% estimated comp for May and June.

In our final step, we read all reports that our procedure flags as discussing the use of alternative data to verify that the analysts indeed incorporate alternative data into their reports. For instance, some firms in our sample provide satellite-related products or employ satellite imagery in their business processes (e.g., oil and gas exploration). We exclude such cases. The figure below is an example of a false positive:[11]



Digital Transformation. Much of the Analyst day was focused on capturing customer relevance which will translate to revenue growth. The company believes that its TAM has increased to $ 4.5 trillion today. In one example, Microsoft helped Land O'Lakes with a multiple year digital transformation which saw the company use Office 365, Surface, Windows 10, Azure, and Hololens. They were able to take decades of satellite imagery, load it into Azure, and then analyze the land, different weather metrics and type of seeds to better layout farms. This process drove yield increases from 130 bushels of corn per acre to now over 500 bushels.

Our primary variable, $I(Alternative\ Data_{i,f,t})$, equals one if analyst i's forecast for the annual earnings of firm f at time t is accompanied by a written

[10] Piper Jaffary; Nicole Miller Regan and Joshua C. Long; McDonald's Corporation; May 9, 2011.

[11] Jefferies Group. John DiFucci, Joseph Gallo, and Howard Ma; Microsoft Corporation; May 11, 2017.

report that explicitly references the use of alternative data and zero otherwise.

For $I(Alternative\ Data)$ to equal one, the following three conditions must be met: (a) Analysts have access to alternative data. (b) Analysts believe the alternative data contain clear signals regarding a company's fundamentals and they incorporate those signals into their forecasts and recommendations. (c) Finally, analysts disclose their reliance in their written reports. In Subsection 2.4.1, we discuss the implications of measuring this joint effect for the interpretation of our results.

Our variable may also equal one if an analyst pretends to have studied alternative data. While we cannot rule out this possibility, such fabricated use would expose the analyst to serious career and litigation risks. Online Appendix Figure A2 provides an example of an analyst report that draws from alternative data. As depicted in the figure, discussions of alternative data often include visuals depicting its application, such as satellite images of parking lots and their relation to revenues. After excluding standard boilerplate illustrations, the average number of figures included in analysts' reports is 0.86. When analysts make reference to the use of alternative data, this number rises to 4.16. The high level of detail typically provided by analysts when discussing their use of alternative data suggests further that it is improbable that analysts falsify their use of such data.

2.2.3 Descriptive Evidence Regarding Analysts' Discussion of Alternative Data

Our first test examines how frequently analysts report using alternative data and, if so, in what manner. Panel A of Table 2.1 displays summary statistics for $I(Alternative\ Data)$ across years. Since we have only partial data for 2009 and 2019, we combine the observations in 2009 with those in 2010 and the observations in 2019 with those in 2018. Panel A reveals that, in 2009/2010, 6% of the analyst forecasts are couched in reports that discuss the use of alternative data. By 2018/2019, the corresponding number is 10%. The fraction of analysts discussing the use of alternative data for at least one of the firms they cover is naturally greater than the fraction of reports discussing the use of alternative data. In particular, we find that, as of 2009/2010, 11% of the analysts in our sample report using alternative data for at least one of the firms they cover. This fraction increases to 28% by 2018/2019.

In Panel B we report summary statistics for $I(Alternative\ Data)$ across industry sectors. Analysts most frequently discuss the use of alternative data for firms operating in the Information Technology sector: the average $I(Alternative\ Data)$ is 16%. Alternative data use is also widely discussed for firms operating in Consumer Discretionary (10%), Consumer Staples (10%), Communication Services (9%), Health Care (8%), and Industrials (6%). Analysts infrequently adopt alternative data for firms in the Energy (2%), Financials (1%), and Materials (1%) sectors.

In our study, we manually assign reports mentioning the use of alternative data into the following eight categories: (1) app-usage data, (2) sentiment data,

(3) employee data, (4) geospatial data, (5) point-of-sale data, (6) satellite-image data, (7) web-traffic data, and (8) other types of alternative data. Some reports are assigned to more than one category as analysts occasionally reference multiple alternative-data categories. Appendix 1 details how we allocate the reports across the above eight categories. The results reported in Table 2 indicate that, of the 5,639 forecasts from reports that discuss the use of alternative data, 1,944 (34%) are based on web-traffic data. The next most popular categories are other (23%), followed by point of sale (19%), sentiment (19%), employee (10%), and app usage (8%). The least popular categories are geospatial (5%) and satellite image (3%).[12]

Figure 1 displays two timelines. The first timeline indicates when—for our sample—we observe the first analyst report that explicitly references the use of alternative data from a given alternative-data category. The second timeline indicates when we observe the first one hundred such reports. The sequence for when we observe the first analyst report is as follows: sentiment (June 11, 2009), web traffic (June 12, 2009), point of sale (August 6, 2009), employee (January 5, 2010), geospatial (January 22, 2010), satellite image (May 3, 2010), other (June 12, 2009) and app usage (July 27, 2010). In other words, within essentially the first year of our sample period, we find that analysts explicitly reference the use of alternative data from all eight categories. The sequence for when we observe the first one hundred analyst reports is as follows: web traffic (January 19, 2010), point of sale (March 28, 2011), other (August 11, 2011), sentiment (October 10, 2011), satellite image (May 21, 2012), geospatial (June 5, 2012), app

[12]Online Appendix Figure A3 shows how frequently alternative data from a particular category are discussed across industry sectors. Among others results, the figure reveals that web-traffic data are discussed frequently for firms operating in the Information Technology sector, whereas point-of-sale data are commonly referenced for firms operating in the Consumer Discretionary sector.

usage (September 8, 2014) and employee (August 18, 2015). In other words, by mid-2012, analysts draw extensively from alternative data from six of the eight categories. Only app usage- and employee-level data are not widely adopted until mid-2015. Overall, our evidence shows that analysts frequently incorporate alternative data into their reports, at least for the largest and economically most meaningful firms.

2.3 The Usefulness of Alternative Data

In this section, we examine whether institutional investors, the primary clientele of analysts, value analysts' use of alternative data. We also consider the wider implications of analysts' adoption of alternative data on the quality of institutional investors' decision-making and the competitive balance among investors.

2.3.1 Analysts' Use of Alternative Data and Forecast Accuracy

The hypothesis that institutional investors appreciate analysts' use of alternative data rests on the premise that analysts can derive unique and value-relevant insights from such data. To examine the validity of this assumption, we test whether earnings forecasts from reports that mention the use of alternative data are more precise:

$$Acc_{i,f,t} = \alpha_i + \theta_t + \beta I(Alternative\ Data_{i,f,t}) + \gamma' Controls + \epsilon_{i,f,t} \tag{2.1}$$

The observations are at the analyst/firm/forecast-date level. The construc-

tion of $Acc_{i,f,t}$ follows prior literature (Clement, 1999; Bradley, Gokkaya and Liu, 2017; Green et al., 2014; Harford et al., 2019). We first compute $AFE_{i,f,t}$ as the absolute value of the difference between analyst i's annual earnings forecast for firm f at time t and the corresponding actual reported annual earnings. We then construct $PMAFE_{i,f,t}$ as the difference between $AFE_{i,f,t}$ and $Avg(AFE)_{f,t}$, scaled by $Avg(AFE)_{f,t}$ to reduce heteroskedasticity. $Avg(AFE)_{f,t}$ is the average absolute forecast error across all analysts covering firm f as of the corresponding forecast period, excluding analyst i and other analysts, who also report drawing from alternative data in their coverage of firm f as of time t. $PMAFE_{i,f,t}$ thus measures analyst i's forecast accuracy relative to the forecast accuracies of all analysts who cover the same firm at the same time but do not report drawing from alternative data. Negative values, or smaller forecast errors, indicate above-average performance. Positive values, or larger forecast errors, indicate below-average performance. To facilitate interpretation, $Acc_{i,f,t}$ equals $PMAFE_{i,f,t} \times (-1)$.[13] We provide descriptive statistics regarding Acc and all other variables we use in this paper in Online Appendix Table A3.

We include both analyst-firm, $\eta_{i,f}$, and firm-year, $\theta_{f,t}$, fixed effects. Angrist and Pischke (2008) show that our two-way fixed-effects specification is equivalent to the basic difference-in-differences specification. The estimate of $I(Alternative\ Data)$ thus indicates how much more accurate an analyst becomes in the post-adoption period relative to the pre-adoption period compared with analysts covering the same firm over the same period that do not reference alternative data.

Controls include the following analyst characteristics: *Forecast Age*, *Ana-*

[13] In robustness checks, we base our analysis directly on $AFE_{i,f,t}$ (Bradley, Gokkaya, and Liu, 2017). As shown in Online Appendix Table A2, the results based on the absolute forecast error are similar to those based on $Acc_{i,f,t}$.

lyst/Firm Experience, *Analyst Experience*, *#Firms Covered*, *Forecast Frequency*, and *Broker Size*.[14] We do not control for firm characteristics as our fixed effects subsume them. Since our final sample comprises 64,018 written reports and earnings forecasts, the number of observations on which we estimate regression equation (1) is 64,018. We double-cluster our standard errors at the analyst- and year-month levels.

Difference-in-differences specifications often produce causal evidence. Our setting does not lend itself to causal inferences. That is, while the estimate of $I(Alternative\ Data)$ indicates how much more accurate an analyst becomes in the post-adoption period relative to the pre-adoption period compared with her non-adopting counterparts, it cannot tell us how much of the abnormal performance improvement is truly caused by alternative data because alternative-data adoption is endogenous. For instance, alternative-data adoption may coincide with an analyst's decision to exert greater effort covering the corresponding firm, which, in turn, leads to improved forecast accuracy ("increased effort channel").

We try to gauge the relevance of the increased-effort channel by constructing various measures of analyst effort used in the literature, including the timeliness of forecasts, the number of forecast revisions, and analyst activity during earnings conference calls (Merkley, Michaely and Pacelli, 2017; Hwang, Liberti and Sturgess, 2019; Grennan and Michaely, 2020). We then explore whether adopting alternative data associates with greater effort. As detailed and tabulated in Online Appendix Table A4, alternative-data adoption correlates neither with the timeliness of forecasts nor with the number of forecast revisions. The adoption of alternative data also is not associated with the number of questions asked

[14]We detail the construction of these variables in Appendix 2.

during earnings conference calls, the number of words spoken, or the types of questions asked; it correlates marginally positively with the number of conference calls attended.

While we generally fail to find empirical support for the increased-effort channel, our tests may lack power. Our point estimate of how much an analyst's forecast accuracy improves after she adopts alternative data should be interpreted with this caveat in mind.

We present our regression results in Table 2.3. The results reported in column (1) show that the coefficient estimate of $I(Alternative\ Data)$ is 0.214 (t-statistic = 6.30). To illustrate the economic significance of this estimate, a 0.214 improvement would move an analyst who is at the median in terms of forecast accuracy to the 62nd percentile.

Another way to gauge the economic significance is to compare the estimate of $I(Alternative\ Data)$ with those of our control variables. For instance, the results reported in column (1) show that forecast accuracy increases significantly with the number of years an analyst has been covering a particular firm: The estimate of *Analyst/Firm Experience* is 0.060 (t-statistic = 2.72). Comparing the estimate of $I(Alternative\ Data)$ with that of *Analyst/Firm Experience* suggests that the performance improvement accompanying the adoption of alternative data is equivalent to having covered the corresponding firm for 3.6 additional years.

Overall, while we cannot provide strong causal evidence, our results are at least consistent with the premise that analysts can derive distinct and value-relevant insights from alternative data.

2.3.2 Analysts' Use of Alternative Data and Trading Commissions

The analyst business model is rooted in soft dollar agreements. When institutional investors find greater value in an analyst's research, they allocate a higher volume of trades through the brokerage that employs the analyst and, consequently, pay a greater amount in trading commissions to the corresponding brokerage. A simple test to determine if institutional investors place value on analysts' adoption of alternative data is thus to correlate the reported use of alternative data with the amount in trading commissions received.

We obtain institutional investor trade data from ANcerno, a firm that provides transaction-cost analysis to institutional clients, including investment managers, like Fidelity Investments, and plan sponsors, like CalPERS (Hu et al., 2018). While the ANcerno dataset covers a wide set of funds, the coverage is not comprehensive. Hu et al. (2018) estimate that the institutions in the ANcerno dataset account for 15% of all institutional trading volume. As discussed by Hu et al. (2018) and other studies utilizing the ANcerno dataset (Puckett and Yan, 2011; Jame, 2018), the dataset does not seem to suffer from sample selection bias.

For each institution covered by ANcerno, the dataset contains specific information for each of their trades, such as ticker symbol, date, trade direction, number of shares traded, and—crucially—identifiers for the investment manager, the broker executing the trade, and the commissions paid to the broker.

The ANcerno data contain records on all 35 DJI constituents. While the ANcerno data span the period running from 1997 through 2015, as of 2011 ANcerno

no longer reports the identifier, which would have allowed us to distinguish trades by individual institutions. Given that our sample of earnings forecasts and analyst reports commences in 2009, our brokerage-commissions analysis thus encompasses the years 2009 and 2010. This period includes a total of 7,634 analyst reports. Of these, 2,877 are issued by brokerages that are not included in the ANcerno database. Our sample for this part of our paper thus comprises 4,757 analyst reports.

To examine how the distribution of commissions is influenced by the dissemination of alternative-data insights, we consider all transactions completed within the first three months following a report's issuance that involve the stock covered in the report. We then total the commissions earned by the brokerage issuing the report. The median commissions earned by a brokerage in the first three months following a report's publication is $11,578.40. To the degree that *ANCerno* covers 15% of all institutional trades and the coverage is representative, we estimate that the total median commissions earned by a brokerage in the first three months following a report's publication is $77,189.33.

We re-estimate regression equation (1) but replace the earnings-forecast-accuracy variable with our new trading-commissions variable. The estimated coefficient β now indicates how much more in trading commissions the analyst's brokerage receives in the post-adoption period relative to the pre-adoption period compared with analysts covering the same firm over the same period but not adopting alternative data. We again double-cluster our standard errors at the broker- and year-month levels.

As reported in Table 2.4, our estimate suggests that, on average, the adoption of alternative data increases the total commissions paid to the brokerage in

the first three months by an additional $11,858.82 ($t$-statistic = 1.86), essentially doubling the median commission earned by a brokerage. In additional analyses, we experiment with time frames other than a three-month period. The results of these tests are in line with those reported in this study and are available upon request.

Overall, our findings suggest that institutional investors recognize and value analysts' incorporation of alternative data and compensate them accordingly.

2.3.3 Analysts' Alternative Data Use and the Playing Field Among Institutional Investors

The adoption of alternative data by sell-side analysts has the potential to reshape not only the relationship between analysts and investors but also the competitive balance among the investors they serve. As alluded to in the introduction, hedge funds have been at the forefront of utilizing alternative data, while other institutional players, such as mutual funds, exhibit slower adoption rates. Additionally, existing research suggests that hedge funds execute more profitable trades than non-hedge funds. This outperformance may arise in part because hedge funds have superior data.

In this part of the study, we investigate whether the dissemination of alternative data insights through analyst reports has leveled the playing field for hedge funds and non-hedge-fund investors. To assess this possibility, we again combine our analyst data with the ANcerno data and explore whether analysts' adoption of alternative data has narrowed the performance gap usually seen

between hedge funds and non-hedge funds.

As we are interested in trades affected by the dissemination of alternative-data-driven insights, we again focus on transactions that occur within three months after the issuance of an analyst report and that involve the stock discussed in the report. Analyses based on alternate windows produce results that are similar to those presented here and are available upon request. Applying Jame (2018)'s methodology, we separate transactions executed by hedge funds from those made by other institutional investors.[15]

We further separate non-hedge-fund trades by whether they are preceded by analyst reports that adopt alternative data or whether they are placed following reports that do not draw from alternative data. To ensure that the investor composition behind the trades associated with alternative data and those without such a linkage is comparable, we restrict our analysis to investors who appear in both subsets of trades.

We assess the performance of investors' trades by creating transaction-based calendar-time portfolios, following the approach of Seasholes and Zhu (2010) and Ben-David, Birru and Rossi (2019). Stocks purchased (sold) on day t are added to the buy (sell) portfolio at the beginning of day $t + 2$. We assume a holding period of three months. We compute DGTW-adjusted returns for each stock/day and construct value-weighted portfolio returns. DGTW-adjusted returns are returns on a stock minus the value-weighted returns on a portfolio of stocks in the same size, book-to-market, and momentum quintiles. We report the time-series average of the buy-minus-sell portfolio returns, whereby each daily portfolio return is weighted by the number of trades contributing to the

[15]For additional details, we refer the reader to Jame (2018).

portfolio on that specific day.

We report the results in Table 2.5. We find that the stocks that hedge funds buy outperform those they sell. The difference in the stocks' ensuing returns is 10.79% when annualized. The statistical significance is modest (t-statistic = 1.54), which largely reflects the high standard errors tied to the comparatively low number of hedge funds and the short sample period.

The performance is noticeably different for non-hedge funds. When analysts do not discuss the use of alternative data, the stocks that non-hedge funds buy outperform those they sell by 3.85% when annualized. The difference in the stocks' ensuing returns of 3.85% for non-hedge funds is *similar to the* numbers reported by Puckett and Yan (2011) and Busse et al. (2019), who also analyze *ANCerno* data.

Interestingly, when analysts incorporate alternative data, the difference in the stocks' ensuing returns for non-hedge funds improves significantly, reaching 9.85% when annualized. This figure closely matches that achieved by hedge funds. Again, the standard error of the mean is very high because of the comparatively low number of analyst reports discussing the use of alternative data and the short sample period. Still, our results are at least consistent with the proposition that when they employ alternative data analysts can equalize the competitive environment and allow the non-hedge funds that pay attention to alternative-data insights to match the performance typically associated with hedge funds.

Another investor group that may benefit from alternative-data adoption by analysts is retail investors. Although they do not comprise the primary audi-

ence for analysts, retail investors can also access analyst reports, depending on their retail broker accounts and the investment-related platforms to which they subscribe.

To study the possible benefits retail investors might derive from analysts' use of alternative data, we utilize Trade and Quote (TAQ) data. We construct a measure of retail order imbalance following the method of Barber et al. (2023). Our findings, presented in Online Appendix Table A5, reveal that retail investors are three times more likely to adjust their trades in line with changes in analyst recommendations when these recommendations come from reports that incorporate alternative data. These results imply that at least some retail investors follow analyst recommendations more closely when they are anchored in alternative data. To gauge the future

performance of these trades, we estimate regressions of one-week, one-month, or three-month future returns on retail investor order imbalances. We distinguish between orders executed following the release of an analyst report that discusses alternative data and those following the release of reports that do not include such discussions. As shown in Online Appendix Table A6, retail investor order imbalances more positively predict future returns when the orders are placed in the presence of alternative data–inclusive analyst reports. This finding is consistent with the notion that using alternative data assists retail investors in executing more profitable trades.

In separate analyses, we assess the broader market's reaction to analysts' use of alternative data. Specifically, we examine whether the market's response to changes in analyst research outputs, including earnings forecasts, price-target forecasts, and overall recommendation levels, becomes more pronounced when

those outputs are couched in alternative data. We detail our empirical design and our results in Online Appendix Table A7. In short, our results again suggest that broad sections of investors more closely follow analyst outputs when they are anchored in alternative data.

2.4 Discussion and Additional Analyses

Before concluding, we discuss possible reasons we do not observe more analyst reports mentioning the use of alternative data. We also examine the extent to which our findings regarding the largest U.S. firms apply to smaller firms. Finally, we explore variations in our key independent variable, $I(Alternative\ Data)$ and our main empirical specification.

2.4.1 Why Do We Not Observe More Analyst Reports Mentioning the Use of Alternative Data?

The seeming benefits accompanying the adoption of alternative data, such as higher trading commissions, raise the question why analysts do not incorporate alternative data into their reports more frequently.

Our key independent variable, $I(Alternative\ Data)$, can equal zero for at least three reasons. First, analysts may not have the resources or willingness to study alternative data, causing I(Alternative Data) to be zero ("resource limitations"). Second, even if analysts access and study alternative data, they may not always uncover clear, unique, or value-relevant insights. Analysts are unlikely to dis-

cuss such "failed" uses of alternative data in the limited space that is available to them in their reports ("intermittent usefulness"). Finally, analysts may identify relevant signals but choose strategically not to detail these insights in their reports ("strategic considerations").

2.4.1.1 The Relevance of Resource Limitations and Intermittent Usefulness

To gauge the relevance of the resource-limitations and intermittent-usefulness perspectives for explaining why $I(Alternative\ Data)$ does not equal one more frequently, we estimate the following probit regression:

$$I(Alternative\ Data_{i,f,t}) = \alpha + \beta' X_{i,f,t} + \delta' Controls_{i,f,t} + \epsilon_{i,f,t} \tag{2.2}$$

The observations are at the analyst/firm/forecast-date level. *Controls* include the following analyst- and firm-level characteristics: *Analyst/Firm Experience*, *Analyst Experience*, *#Firms Covered*, *Forecast Frequency*, *Broker Size*, *Size*, M/B, and *Momentum*. We detail the construction of these variables in Appendix 2. We double-cluster our standard errors at the analyst- and year-month levels.

2.4.1.1.1 Variables Tied to Resource Limitations Our first key independent variable is an indicator of whether an analyst's brokerage has an in-house data-science team. Our second key variable is the number of an analyst's colleagues working in the same city who have already discussed the use of alternative data in at least one of their reports as of the analyst's forecast date. The expenses necessary to access and analyze alternative data are presumably reduced when analysts are supported by an in-house data-science team. Moreover, sup-

pose that individuals in knowledge-based industries owe much of their success to their colleagues (Hwang, Liberti and Sturgess, 2019). In that case, analysts surrounded by peers who already utilize alternative data should find the learning curve less steep. If resource limitations are an important reason that $I(Alternative\ Data)$ does not equal one more frequently, we should observe more frequent alternative-data adoptions when there are fewer resource limitations; we should thus observe positive coefficient estimates for the first two key independent variables.

2.4.1.1.2 Variables Tied to Intermittent Usefulness Similarly, if intermittent usefulness is an important reason that $I(Alternative\ Data)$ does not equal one more frequently, we should observe a positive correlation between the frequency of alternative-data adoptions and measures of its usefulness.

Alternative data offer three key advantages. First, alternative data provide immediate indications of a company's performance, allowing analysts to draw from current rather than stale information. Second, the origin of alternative data from independent third-party vendors mitigates the risk of bias or distortion often associated with data produced by company managers. Third, the detailed nature of alternative data—for instance data based on breaking down information into specific products or branches—provides analysts with a more detailed and nuanced understanding of a company's performance.

Building on these considerations, we propose that alternative data are more useful and, as a result, become more frequently discussed in analyst reports when (1) receiving instantaneous signals regarding a company's performance is critical, (2) concerns of misrepresentation are urgent and traditional data are

ambiguous, and (3) analysts lack access to granular data.

Receiving instantaneous signals regarding a company's performance becomes critical when there are relatively few company announcements and when the uncertainty regarding a company's performance is high. We thus conjecture that the usefulness of alternative data and the likelihood of alternative-data adoption are higher when a firm files relatively few Form 8-Ks, when stock return volatility is high, and when the absolute value of earnings surprises is high. For a description of the precise construction of these and the subsequent key independent variables outlined below, we direct the reader's attention to Appendix 2.

That managers and firms are removed from the "alternative data–generating process" becomes particularly relevant when misrepresentation concerns are urgent and traditional data are ambiguous. As measures for concerns of misrepresentation and the ambiguity of traditional data, we consider whether a company has had to restate its earnings (Wilson, 2008) and the absolute value of discretionary accruals (Bhattacharya, Desai and Venkataraman, 2013).

Finally, private meetings with management are one of the key channels through which analysts can obtain a more nuanced perspective on a company's performance (Green et al., 2014; Soltes, 2014; Brown et al., 2015; Bengtzen, 2017). Not all analysts are granted private meetings with management, however, putting them at a significant disadvantage.

To measure whether analyst i has preferential access to the management of firm f, we consider whether analyst i works for a broker that hosts an investor conference in which firm f participates. Green et al. (2014) argue that broker-

hosted investor conferences, which provide selected investors opportunities to interact with senior corporate managers, offer insights into whether a particular analyst has preferential access to the firms participating in the conference.

2.4.1.1.3 Results Regarding The Relevance of Resource Limitations and Intermittent Usefulness

We report the results of the analyses of resource limitations and intermittent usefulness in Table 2.6. We find that analysts incorporate alternative data into their reports less frequently when their brokerages do not deploy in-house data-science teams and when analysts do not have colleagues that already draw from alternative data. These results are consistent with the resource-limitations perspective.

Consistent with the intermittent-usefulness idea, we find that analysts adopt alternative data more frequently when a company provides relatively few announcements, when stock-return volatility and the absolute value of earnings surprises are high, when the firm has had to restate its earnings and the absolute value of discretionary accruals are high, and when analysts lack preferential access to management through private meetings.[16]

Overall, our results suggest that resource limitations and intermittent usefulness are important determinants of $I(Alternative\ Data)$ and help to explain why analysts do not discuss alternative data in their reports more frequently.

[16]In Online Appendix Table A8, we describe analyses showing that the incremental improvements in earnings-forecast accuracy associated with discussions of alternative data also strengthen when the absolute value of earnings surprises are high and when the firm has had to restate its earnings.

2.4.1.2 The Relevance of Strategic Considerations

Analysts may choose not to disclose their reliance on alternative data for two strategic reasons. First, some analysts may strive to continuously provide fresh perspectives. An analyst reporting the use of alternative data in one year may thus avoid mentioning her continued reliance in the following years to avoid repetition.

To investigate this possibility, we consider cases where analysts report using alternative data for a particular firm in a particular year. We find that these analysts discuss their reliance on the same alternative-data category in the subsequent year 55% of the time.

It appears likely that an analyst who incorporated alternative data into her report in one year could easily obtain an updated version of the data in the ensuing year. The fact that in 45% of the cases analysts do not reference the same alternative data category in the subsequent year suggests either that the perceived value of alternative data decreases rapidly ("intermittent usefulness") or that analysts avoid repeating their continued reliance ("strategic considerations").

To shed light on the relevance of these possibilities, we modify our earnings-forecast-accuracy regression. Our adjusted independent variable equals one if an analyst does not discuss the use of alternative data in year t but did so in year $t-1$. Suppose analysts no longer mention the use of alternative data because they find the data no longer useful. Suppose further that analysts' assessments of the diminished usefulness are accurate. In that case, our revised independent variable should no longer be significantly associated with earnings-forecast ac-

curacy.

Our analysis yields a coefficient estimate of 0.070 (t-statistic = 2.57). This robust estimate is consistent with the idea that analysts continue to benefit from alternative data insights but, to avoid repetition, opt not to reiterate their reliance. At the same time, the estimate of 0.070 is markedly lower than that of our main specification (coefficient estimate = 0.214, t-statistic = 6.30), suggesting that, in many cases, the utility of alternative data does diminish over time. This result again points to intermittent usefulness as an important reason that $I(Alternative\ Data)$ does not equal one more frequently.

Another strategic consideration that may dissuade analysts from disclosing their use of alternative data is that they fear that disclosing their methodologies and data sources would make it easier for competing analysts to imitate their outputs. This could result in the erosion of the competitive advantage the original analyst initially held. We find limited evidence for the relevance of this particular consideration. In our sample, we detect only 92 cases where the initial adoption of alternative data is replicated by another analyst in the concurrent year for the identical company. In these situations, we find that the original analyst receives, on average, $20,059.55 in commissions (for trades in the corresponding stock occurring within the first three months of the report publication). Subsequent to the imitation, the total three-month average commission of the original analyst stands at $25,537.25, suggesting that the financial repercussions from imitation are limited.[17]

[17]Concerns over imitation might still affect analysts' reporting behavior if they are unaware that imitations are rare and do not affect an original analyst's trading commissions.

2.4.1.2.1 Variables Tied to Resource Limitations Our first key independent variable is an indicator of whether an analyst's brokerage has an in-house data-science team. Our second key variable is the number of an analyst's colleagues working in the same city who have already discussed the use of alternative data in at least one of their reports as of the analyst's forecast date. The expenses necessary to access and analyze alternative data are presumably reduced when analysts are supported by an in-house data-science team. Moreover, suppose that individuals in knowledge-based industries owe much of their success to their colleagues (Hwang, Liberti and Sturgess, 2019). In that case, analysts surrounded by peers who already utilize alternative data should find the learning curve less steep. If resource limitations are an important reason that $I(Alternative\ Data)$ does not equal one more frequently, we should observe more frequent alternative-data adoptions when there are fewer resource limitations; we should thus observe positive coefficient estimates for the first two key independent variables.

2.4.2 The Use and Usefulness of Alternative Data Among Small Firms

Our tests so far have involved constituents of the DJI. To explore the prevalence and impact of alternative data within the small-cap sector, we analyze a random sample of 200 companies drawn from the lower half of the size distribution in the CRSP database. The sample period is 2009–2019. Of these companies, 143 have earnings forecasts and analyst reports data in IBES and Investext. The median market capitalization of our small firms is $1.013 billion. In comparison,

the median market capitalization of DJI constituents is $11.854 billion. We retrieve a total of 13,123 analyst reports for our 143 small firms over the 2009–2019 period, compared with 64,018 reports for the 35 firms in the DJI. Our small firms thus receive substantially less extensive coverage, which is no surprise given that analysts' key clientele, institutional investors, invest primarily in larger, more liquid stocks.

Following the same procedure described in Subsection 2.2.2, we identify the fraction of reports that discuss the use of alternative data. We find that for our small firms, analysts discuss alternative data in only 2% of their reports, compared with 9% for the DJI constituents. It thus appears that analysts use alternative data significantly less frequently among small firms. One factor that might explain this result is that analysts are reluctant to bear the financial and learning costs associated with alternative-data adoption for firms that attract limited attention from institutional investors. It is also possible that the quality of alternative data is lower for smaller firms.

We find that references to alternative data are accompanied by more accurate earnings forecasts even among our small firms. As reported in Online Appendix Table A9, the coefficient estimate is 0.197, suggesting that an analyst at the median accuracy level improves to the 60th percentile. The magnitude of the association is similar to that for DJI constituents.

We also re-run our analyses of the trading commissions received by analysts' brokerages and the performance differential between hedge funds and non-hedge-fund institutional investors. Since institutions trade significantly less extensively in smaller firms, the sample sizes for these additional analyses are substantially smaller. As shown in Online Appendix Table A10, we find

evidence that institutional investors continue to reward analysts who discuss alternative data by directing more trades through their brokerages. The coefficient estimate is 2,818.23 (t-statistic = 3.38). Compared with the estimate of 11,858.82 for DJI constituents, this estimate is substantially smaller. The weaker economic significance suggests that there is a weaker incentive for analysts to adopt alternative data for smaller firms, which could explain why the fraction of reports that reference alternative data is much lower for smaller firms than for DJI constituents.

When we repeat our transaction-based calendar-time portfolio analysis for the smaller firms, we find that analysts' discussions of alternative data coincide with a narrower performance gap between institutional investors and hedge funds. However, the standard errors are so large that none of the average returns are statistically different from zero.

2.4.3 Differences in Usefulness by Alternative Data Types

2.4.3.1 Variation Across Alternative Data Categories

While our results suggest that alternative data contain unique and value-relevant insights, the usefulness of the data may vary by type. To explore this possibility, we first replace $I(Alternative\ Data)$ with eight indicator variables, each denoting whether an analyst uses alternative data from a particular alternative-data category. We then re-estimate our earnings-forecast-accuracy regression (1). The results reported in Table 2.7 show that the adoption of alternative data from six of the eight categories is associated with statistically significant performance improvements. Within those six, the ranking in descending

order based on the magnitude of the coefficient estimates is as follows: (1) app usage, (2) sentiment, (3) employee, (4) other, (5) point of sale, and (6) web traffic.

Unlike the adoption of alternative data from the above six categories, our results show that (7) geospatial data and (8) satellite image data are not associated with more accurate earnings forecasts. As shown in Table 2.2, these are also the two categories that analysts report using least frequently. In 2017, Ernst & Young Global Limited surveyed hedge funds and asked which datasets, in their experience, have been the *least* accurate and *least* insightful.[18] The two datasets that are by far the most frequently mentioned are "geolocation" and "satellite." While the survey conducted by Ernst & Young Global Limited represents a one-time snapshot of investors' opinions, we nevertheless find the overlap between the survey results and our regression results revealing.[19]

In another test, we replace $I(Alternative\ Data)$ with the number of distinct alternative data categories discussed in an analyst's report. We then re-estimate our regressions within the subset of cases where an analyst reports the use of alternative data from at least one category. In short, we detect a strong positive correlation between the number of alternative data categories and the accuracy of earnings forecasts. This finding implies that, when analysts' views are supported by a wider range of data types, their predictions are particularly accurate (Table 2.7).

[18]The survey results are viewable at https://alternativedata.org/stats/.

[19]In separate tests, we also gauge the usefulness of alternative data categories across industries. Specifically, we re-estimate our regressions separately for each Global Industry Classification Standard Sector. We present the results in Online Appendix Figure A4. Among other findings, our results suggest that app-usage data are particularly beneficial for forecasting the performance of firms that operate in the Information Technology sector, while point-of-sale data is particularly advantageous for predicting the performance of firms that operate in the Consumer Staples sector.

2.4.3.2 Variation Across More or Less Proprietary Alternative Data

Among our eight alternative-data categories, four categories—(1) sentiment data, (2) employee data, (3) geospatial data, and (4) web-traffic data—are comparatively more accessible to the public and less dependent on external data vendors ("accessible data"). For example, investors can acquire variants of sentiment data through social media platforms, catch a glimpse of employee-related information through public websites (e.g., Glassdoor.com), gather geospatial data through a combination of Google Maps and publicly available demographic data, and retrieve web-traffic data using tools such as Google Trends. In contrast, data from the remaining four categories—(5) app-usage data, (6) point-of-sale data, (7) satellite image data, and (8) other unspecified types of data—are generally not accessible to the general public ("proprietary data").

In separate tests, we gauge whether accessible data are as useful to analysts for predicting a company's performance as proprietary data are. We find that, among analyst reports that discuss alternative data, 64.2% base their analysis on accessible data, while 49.3% rely on proprietary data. Proprietary data are adopted more frequently when analysts work for larger brokerages or brokerages with internal data-science teams; they are also adopted more frequently when more colleagues in the same locale lean on alternative data in their reports.

Regarding the implications for forecast accuracy, results reported in Table 2.7 show that references to proprietary data (coefficient estimate = 0.243, t-statistic = 6.82) and accessible data (coefficient estimate = 0.188, t-statistic = 4.07) come with similar increases in the precision of earnings forecasts; the two coefficient

estimates are not statistically different from each other (F-statistic = 0.99, p-value = 0.32).

2.4.4 Alternative Data Use and Earnings-Forecast Accuracy: Instrumental Variable- and Matching Analyses

To study the relationship between alternative-data use and earnings-forecast accuracy, our primary analysis adopts a standard difference-in-differences method and assesses how much more accurate an analyst becomes in the post-adoption period relative to the pre-adoption period compared with analysts covering the same firm over the same period that do not reference alternative data. As alluded to in Subsection 2.3.1, an analyst's decision to adopt alternative data is not exogenous and, instead, could reflect the decision to exert greater effort. To address this concern, we experiment with an instrumental-variable approach. The basic premise behind our instrumental-variable approach is that an individual analyst's ability to use alternative data depends on her firm's infrastructure and technology investment ("relevance"). At the same time, certain brokerage-level decisions, which allow analysts to access alternative data, are made independently of a single analyst's preferences or actions ("exclusion restriction").

We consider two instruments. The first instrument, *First Time Use*, is an indicator variable that equals one when an analyst's colleague, working in the same city, adopts alternative data for the first time. The second, *Software Budget*, refers to the allocated budget for software purchases at the broker-year level, sourced from Aberdeen's Computer Intelligence Technology Database. We assess the

validity of these instruments through the Hansen J-test for overidentifying restrictions and confirm their efficacy with underidentification tests and F-tests in the first-stage regression. Our regression results pass all diagnostic tests, implying that our instrumenting strategy is valid.

We report the results from the instrumental-variable approach in Online Appendix Table A11. In column (1) we report the results from the first-stage test. The coefficient estimates of *First Time Use* and *Software Budget* are positive and significant at the 1% level, suggesting that our instruments are highly correlated with the endogenous variable. The results reported in column (2) show that the second-stage coefficient estimates of the instrumented $I(Alternative\ Data)$ are positive and significant. The estimates are similar in magnitude to those obtained from the original regression specification.

We also experiment with a matching-sample analysis. We pair each alternative-data report within our sample with a non-alternative-data report written by analysts covering the same firm during the same forecast period. We ensure that the analysts with the matched reports work for brokerages that fall into the same size quintile (based on the number of analysts employed) and the same firm-specific experience quintile (based on the number of years of having covered the respective firm). If multiple reports satisfy the above criteria, we select the report with the forecast horizon that is most similar to that of the alternative data report. We then repeat our forecast accuracy analysis within this matching sample.

The results obtained with this matching-sample approach are presented in Online Appendix Table A12. The coefficient estimate for $I(Alternative\ Data)$ is 0.204, with a t-statistic of 3.83. Again, this estimate is very similar to that

obtained in our primary analysis.

2.4.5 Other Key Performance Indicators

Our final test considers revenue-forecast accuracy as another crucial analyst output. Revenue forecasts are sourced directly from the IBES database and the accuracy of these forecasts is measured in the same manner as earnings-forecast accuracy. Not all analysts have revenue forecasts in the IBES database, though, so the sample size drops to 27,661.

In addition to revenue-forecast accuracy, one might also consider the accuracy of cost forecasts. Unfortunately, the IBES database does not contain cost forecasts. Instead, we compute "residual forecasts" by taking the difference between revenue-per-share forecasts and earnings-per-share forecasts and constructing a measure of accuracy based on these residual forecasts. The resulting measure may be seen as capturing any improvement in earnings-forecast accuracy that cannot be tied to more accurate revenue forecasts.

We report our results in Online Appendix Table A13. In column (1) and column (2) we present results based on revenue-forecast accuracy and residual-forecast accuracy, respectively. The results reported in column (1) reveal that discussing alternative data is associated with significantly more accurate revenue forecasts. Conversely, the findings reported in column (2) suggest that the use of alternative data does not lead to more accurate residual forecasts, given that the estimate of $I(AlternativeData)$ is close to zero. Put differently, our results indicate that, once any improvement tied to more accurate revenue forecasts is removed from the analysis, the adoption of alternative data no longer

leads to more accurate earnings forecasts. Such a suggestion seems reasonable, considering that most alternative data pertain to a company's sales, not profits.

2.5 Conclusion

Our study documents the dynamic role of sell-side analysts. Rapid advancements in information technology and the emergence of alternative-data sources are creating a wealth of information which, from an investor's point of view, might dominate those provided by sell-side analysts, eventually rendering the analyst profession obsolete.

Our paper points to a more nuanced picture. We propose that the complexity and costs associated with accessing and interpreting alternative data represent a significant challenge for many investors. This challenge opens a window of opportunity for analysts to serve as essential conduits of information. Instead of investors each bearing the financial and learning costs associated with studying alternative data, it is more efficient for a few analysts to absorb these expenses and become proficient in parsing alternative data. Analysts can then share their insights with investors, who, in turn, can integrate these signals with other information to determine a stock's value.

Our research findings are in line with this perspective. We find that analysts have begun to adopt alternative data and that investors value analysts' alternative-data-driven insights and reward them accordingly. Our study thus highlights the enduring importance of human expertise in financial analysis, even in an era of increasing reliance on big data and technology. Simultaneously, our findings indicate that maintaining this relevance requires a transformation

in the function and approach of human analysts.

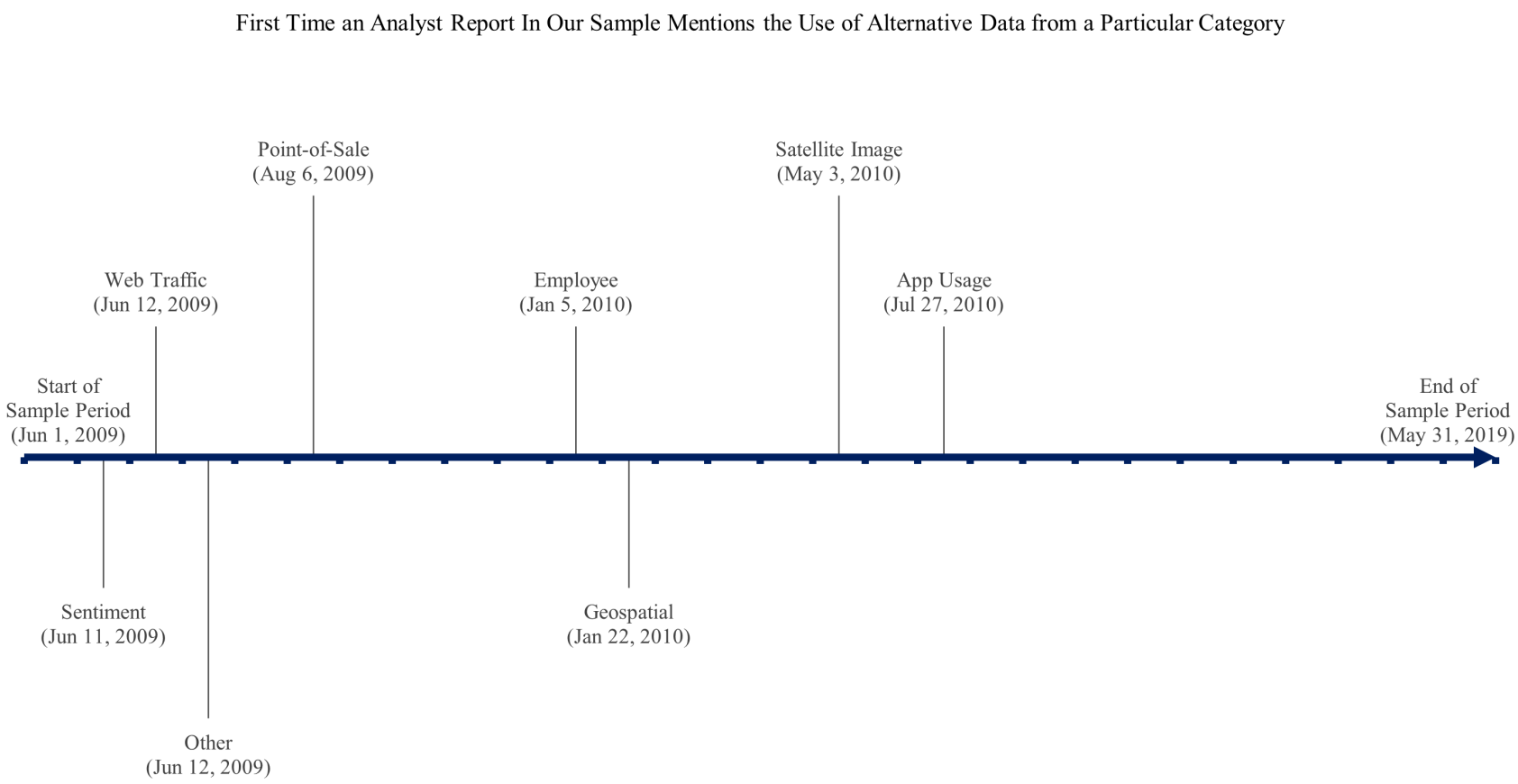


Figure 2.1: This figure displays two timelines, indicating when—for our sample of firms in the Dow Jones Industrial Average Index—we observe the first analyst report, or, the first one hundred analyst reports, explicitly referencing the use of alternative data from a particular alternative-data category. We describe our alternative data categories in Subsection 2.2.3.

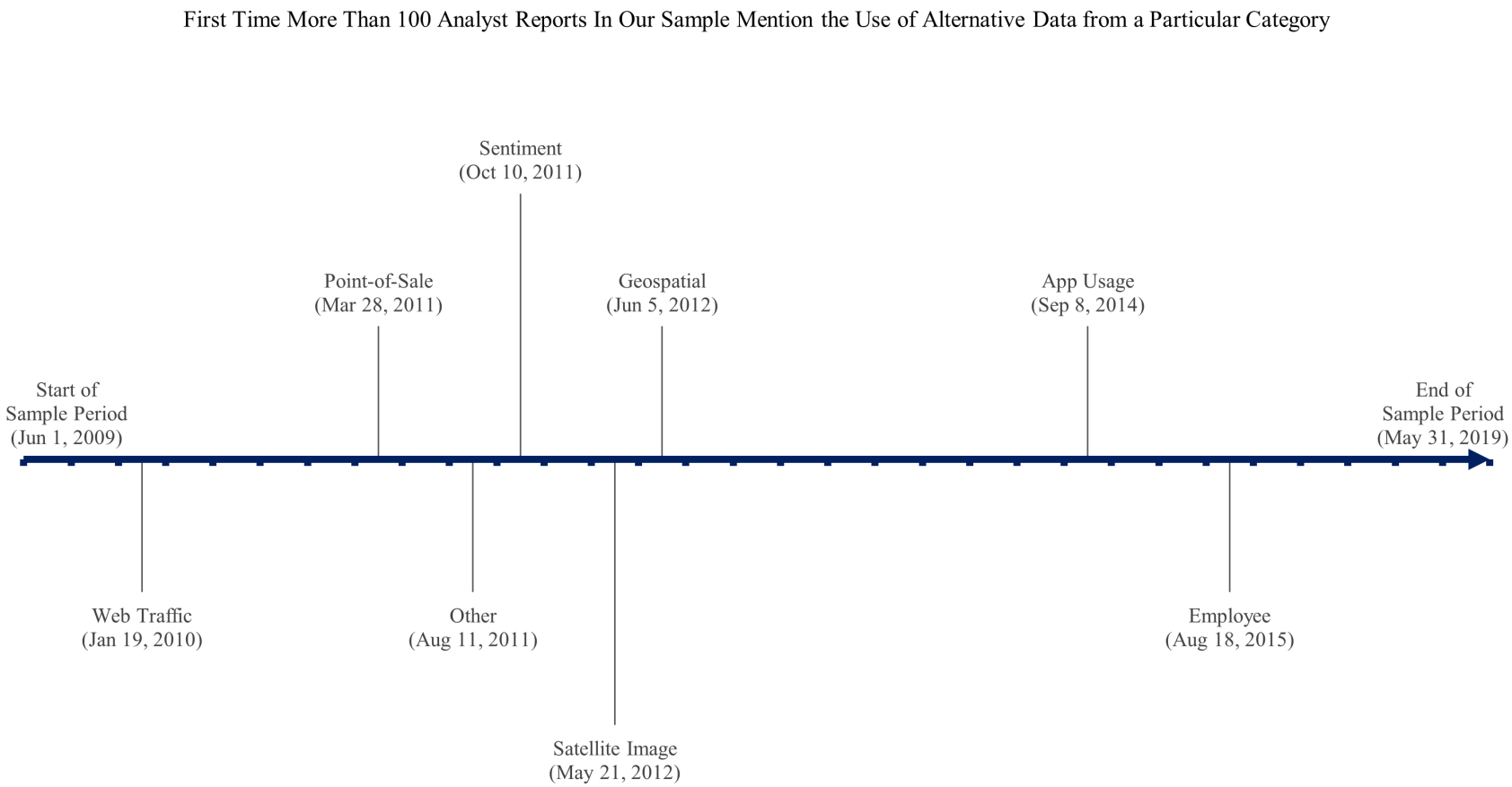


Figure 2.1: Continued.

Table 2.1: Numbers and Fractions of Analyst Forecasts Explicitly Supported by Alternative Data

	Number of Forecasts		Fraction of Forecasts
	Explicitly Supported by Alternative Data	**Not Explicitly Supported by Alternative Data**	**Explicitly Supported by Alternative Data**
Panel A: By Year			
2009/2010	515	7,616	6%
2011	615	6,239	9%
2012	488	6,769	7%
2013	490	6,348	7%
2014	497	6,058	8%
2015	694	5,998	10%
2016	729	5,691	11%
2017	659	5,444	11%
2018/2019	952	8,216	10%
2009–2019	**5,639**	**58,379**	**9%**
Panel B: By Industry Sector			
Energy	40	2,512	2%
Materials	29	2,596	1%
Industrials	580	8,398	6%
Consumer Discretionary	443	4,194	10%
Consumer Staples	841	7,199	10%
Health Care	661	7,487	8%
Financials	103	8,082	1%
Information Technology	2,513	13,597	16%
Communication Services	429	4,314	9%

Notes: This table presents the numbers and fractions of analyst forecasts explicitly supported (not explicitly supported) by alternative data. Our sample contains all Dow Jones Industrial Average Index firms from June 1, 2009, through May 31, 2019. We combine the years 2009 and 2010 and the years 2018 and 2019 as we have only partial data for those years.

Table 2.2: Numbers of Analyst Forecasts Explicitly Supported by Data from a Particular Category

Alternative Data Category	**Number [Fractions] of Forecasts Explicitly Supported by Alternative Data**
App Usage	476 [8%]
Employee	543 [10%]
Geospatial	257 [5%]
Point of Sale	1,080 [19%]
Satellite Image	171 [3%]
Sentiment	1,062 [19%]
Web Traffic	1,944 [34%]
Other	1,322 [23%]

Notes: This table presents the numbers and fractions of analyst forecasts explicitly supported (not explicitly supported) by alternative data. Our sample contains all Dow Jones Industrial Average Index firms from June 1, 2009, through May 31, 2019. We combine the years 2009 and 2010 and the years 2018 and 2019 as we have only partial data for those years.

Table 2.3: Alternative Data and Forecast Accuracy

	(1)
I(Alternative Data)	0.214***
	(6.30)
Forecast Age	-0.251***
	(-12.34)
Analyst/Firm Experience	0.060***
	(2.72)
Analyst Experience	0.056
	(1.00)
#Firms Covered	0.029
	(0.52)
Forecast Frequency	0.033
	(1.11)
Broker Size	-0.000
	(-0.85)
Analyst–Firm Fixed Effects	Yes
Firm–Year Fixed Effects	Yes
N	64,018
Adjusted R^2	0.238

Notes: This table reports coefficient estimates derived from regressions of forecast accuracy on whether an analyst explicitly references the use of alternative data in her corresponding written report. The dependent variable, *Acc*, is constructed as $PMAFE \times (-1)$, where $PMAFE$ is the difference between the absolute forecast error of an analyst and the average absolute forecast error across all analysts not referencing alternative data, scaled by the average absolute forecast error. Negative (positive) $PMAFE$ values indicate above (below) average performance. $I(\textit{Alternative Data})$ equals one if the analyst's earnings forecast is explicitly supported by alternative data and zero otherwise. All other variables are defined in Appendix 2. We report *t*-statistics in parentheses. Standard errors are double-clustered at the analyst and year-month levels. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 2.4: Alternative Data and Trading Commissions

	(1)
I(Alternative Data)	11,858.82*
	(1.86)
Forecast Age	-3,393.30
	(-1.15)
Analyst/Firm Experience	1,416.47
	(0.63)
Analyst Experience	1,233.77*
	(1.88)
#Firms Covered	5,091.09
	(0.22)
Forecast Frequency	-2,909.48
	(-0.33)
Broker Size	-720.16***
	(-4.23)
Broker-Firm Fixed Effects	Yes
Firm-Year Fixed Effects	Yes
N	4,757
Adjusted R^2	0.623

Notes: This table reports coefficient estimates from regressions of trading commissions paid to a brokerage on whether an analyst explicitly references the use of alternative data in her written report. Observations are at the brokerage/firm/forecast-date level. We consider trades and associated trading commissions in the ANcerno database. Focusing on how alternative-data insights affect the allocation of trades and commissions, we restrict to transactions within three months after the issuance of an analyst report that involve the stock discussed in the report. For each analyst report, we compute the total commissions earned by the brokerage issuing the report and use this as the dependent variable. $I(\textit{Alternative Data})$ equals one if the corresponding report explicitly describes the use of alternative data as in Subsection 2.2.2 and zero otherwise. All remaining variables are defined in Appendix 2. We report t-statistics in parentheses. Standard errors are double-clustered at the broker and year-month levels. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 2.5: Alternative Data and Portfolio Returns

	Buy – Sell [Daily]	**Buy – Sell [Annual]**
	(1)	(2)
Hedge funds	0.04%	10.79%
	(1.54)	
Non-Hedge Funds without Alternative Data	0.02%**	3.85%
	(2.19)	
Non-Hedge Funds with Alternative Data	0.04%	9.85%
	(1.27)	

Notes: This table reports the performance of transaction-based buy-and-sell calendar-time portfolios. We consider trades in the *ANcerno* database. As we are interested in trades affected by the dissemination of alternative-data insights, we focus on transactions that occur within three months after the issuance of an analyst report and that involve the stock discussed in the report. Following James (2018), we separate trades into those made by hedge funds and those made by non-hedge funds. We further separate non-hedge-fund trades by whether they are preceded by an analyst report that adopts alternative data or by reports that do not use alternative data. We assess trade performance using transaction-based calendar-time portfolios, following Seasholes and Zhu (2010) and Ben-David, Birru, and Rossi (2019). Stocks purchased (sold) on day t are added to the buy (sell) portfolio at the beginning of day $t+2$, with a holding period of three months. We compute DGTW-adjusted returns for each stock/day (the return on a stock minus the value-weighted return on a portfolio of stocks in the same size, book-to-market, and momentum quintiles) and construct value-weighted portfolio returns. Each daily portfolio return is weighted by the number of trades contributing to the portfolio on that day. We also report the annualized difference. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 2.6: Variation in the Use of Alternative Data

	(1)	(2)
$I(In-House\ Data\ Science\ Team)$	0.514***	0.104***
	(4.42)	(3.55)
Σ*Colleagues Alternative Data*	0.068**	0.008
	(2.15)	(1.48)
Rank(Number of 8-Ks)	-0.100***	-0.012**
	(-3.30)	(-2.20)
Rank(Return Volatility)	0.113***	0.011**
	(4.02)	(2.54)
Rank(Earnings Surprise)	0.057***	0.009***
	(2.76)	(3.04)
$I(Earnings\ Restatement)$	0.300***	0.072***
	(3.98)	(3.55)
Rank(Discretionary Accruals)	0.044*	0.012***
	(1.72)	(2.85)
$I(Lack\ of\ Preferential\ Access\ to\ Management)$	0.160	0.028*
	(1.61)	(1.93)
Analyst/Firm Experience	-0.006	-0.001
	(-0.81)	(-0.91)
Analyst Experience	0.004	0.001
	(0.60)	(0.98)
#Firms Covered	0.135	-0.005
	(1.00)	(-0.29)

Table 2.6 Continued

	(1)	(2)
Forecast Frequency	-0.024	0.003
	(-0.57)	(0.58)
Broker Size	0.001	0.000
	(0.99)	(1.47)
Size	0.210***	0.046***
	(3.43)	(3.57)
M/B	0.012	0.001
	(1.29)	(0.54)
Momentum	0.175	0.027
	(1.25)	(1.24)
Analyst-Firm Fixed Effects	No	No
Firm-Year Fixed Effects	No	No
Industry Fixed Effects	No	Yes
N	64,018	64,018
Pseudo R^2	0.118	0.122

Notes: In this table, we report coefficient estimates derived from regressions of alternative-data adoption on various analyst- and firm-level characteristics. The observations are at the analyst/firm/forecast-date level. In column (1) we report the results derived from a probit model, while in column (2) we report the results derived from a linear probability model to allow for consistent parameter estimation while including fixed effects. The dependent variable equals one if the analyst's earnings forecast is explicitly supported by alternative data and zero otherwise. $I(In-House\ Data\ Science\ Team)$ equals one if the analyst's brokerage has an in-house data-science team. Σ *Colleagues Alternative Data* is the number of colleagues working in the same city as the corresponding analyst and drawing from alternative data for one of the

firms they cover in the same forecast period. *Number of 8-Ks* is the total number of Form 8-Ks filed during the previous annual forecast period. *Return Volatility* is the standard deviation of daily stock returns during the previous annual forecast period. *Earnings Surprise* is measured as in Livnat and Mendenhall (2006). $I(Earnings\ Restatement)$ equals one if the corresponding firm has had to restate its financial accounts. We compute *Discretionary Accruals* as in Kothari, Leone and Wasley (2005). $I(Lack\ of\ Preferential\ Access\ to\ Management)$ equals one if the corresponding firm did not participate in a conference hosted by the corresponding analyst's broker over the previous year. To facilitate a comparison of the coefficient estimates, we convert Σ *Colleagues Alternative Data*, *Number of 8-Ks*, *Return Volatility*, *Earnings Surprise*, and *Discretionary Accruals* into quintile-rank variables, ranging from one if the corresponding realization is in the bottom quintile of its distribution to five if the corresponding realization is in the top quintile. We define all remaining variables in Appendix 2. We report z-statistics in parentheses. We double-cluster our standard errors at the analyst- and the year-month levels. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 2.7: Differences in the Usefulness by Alternative Data Types

	(1)	(2)	(3)	(4)
$I(Category = App\,Usage)$	0.310***			
	(4.04)			
$I(Category = Sentiment)$	0.230***			
	(2.88)			
$I(Category = Employee)$	0.212***			
	(3.09)			
$I(Category = Geospatial)$	-0.033			
	(-0.31)			
$I(Category = Point\ of\ Sale)$	0.183***			
	(3.44)			
$I(Category = Satellite\ Image)$	0.053			
	(0.53)			
$I(Category = Web\,Traffic)$	0.137**			
	(2.09)			
$I(Category = Others)$	0.187***			
	(3.88)			
$\Sigma\ Categories$		0.175*		
		(1.75)		
$I(Source = Proprietary\ Data)$			0.243***	
			(6.82)	
$I(Source = Accessible\ Data)$				0.188***
				(4.07)

Table 2.7 Continued.

	(1)	(2)	(3)	(4)
Forecast Age	-0.250***	-0.181***	-0.251***	-0.252***
	(-12.36)	(-3.67)	(-12.19)	(-12.31)
Analyst/Firm Experience	0.060***	-0.017	0.059***	0.061***
	(2.68)	(-1.24)	(2.60)	(2.79)
Analyst Experience	0.055	0.378	0.059	0.057
	(0.98)	(1.50)	(1.03)	(1.02)
#Firms Covered	0.030	-0.295	0.029	0.034
	(0.54)	(-1.10)	(0.53)	(0.60)
Forecast Frequency	0.033	-0.101	0.033	0.029
	(1.09)	(-0.96)	(1.08)	(0.96)
Broker Size	-0.000	0.001	-0.000	-0.001
	(-0.91)	(1.46)	(-0.84)	(-0.86)
Analyst-Firm Fixed Effects	Yes	Yes	Yes	Yes
Firm-Year Fixed Effects	Yes	Yes	Yes	Yes
N	64,018	5,639	64,018	64,018
Adjusted R^2	0.239	0.406	0.237	0.236

Notes: In this table we report results derived from analyses similar to those associated with Table 3. For column (1), we replace $I(Alternative\ Data)$ with $I(Category = X)$, which equals one if the analyst explicitly references the use of alternative data from alternative-data category X. For column (2), we replace $I(Alternative\ Data)$ with $\Sigma\ Categories$, which is the number of distinct alternative-data categories an analyst references in her report. For column (3), we replace $I(Alternative\ Data)$ with $I(Source = Proprietary\ Data)$ and $I(Source = Accessible\ Data)$, which equals one if the analyst explicitly references the use of proprietary and

accessible alternative data as described in Subsection 2.4.3.2, respectively. We report t-statistics in parentheses. We double-cluster our standard errors at the analyst- and year-month levels. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

CHAPTER 3

A TAX-SHAPED RETAIL LANDSCAPE

3.1 Introduction

As states compete to foster economic growth, tax policies have become a pivotal tool for attracting business investment. However, their impact on market structure remains a subject of debate. By shaping the costs associated with property, equipment, workforce training, investment, and research, these policies influence the sunk costs of market entry. In industries with localized trade areas, such as retail, geographic markets often sustain only a limited number of firms serving nearby customers. This raises an important question: are certain types of firms more responsive to tax policies than others? This question holds particular relevance in policy debates, where critics argue that tax incentives disproportionately benefit large firms. For example, an analysis of New York State's Empire Zone program revealed significant concentration of benefits among the largest retailers:

> *The benefits are highly concentrated. Of the approximately 5,000 firms that claimed credits in 2006, just 500 claimed almost $391 million, or 76 percent, of the $514 million total... Within that group of 500 firms, a handful account for $100 million... Many of the large corporations are "big box" retailers. Of the 25 firms among the Fortune 1000 that received Empire Zone credits, 11 were retail chains. The firms—Wal-Mart, Home Depot, Costco, Target, Walgreen's, Lowe's, Staples, Kohl's, Family Dollar, Radio Shack, and Dick's Sporting Goods—got $30 million in credits.*

This critique suggests that tax policies may disproportionately favor the largest retail chains, reflecting their heightened responsiveness to such incentives. In the context of declining business dynamism and increased state tax subsidies (Shambaugh et al., 2018), evaluating the validity of this claim is crucial to inform policy design. To address this issue, we analyze comprehensive data on retail establishment entry in the United States from 1990 to 2014. Specifically, we investigate whether retail entry patterns vary across states with different tax policies and whether the effects of these policies differ by retailer type (e.g., large vs. small chains).

Our main empirical findings, based on a spatial discontinuity design, reveal that the aggregate relationship between net taxes and retail establishment counts is modest and statistically

imprecise.[1] However, this muted average effect masks heterogeneity, as the negative association between taxes and entry is driven almost entirely by large retail chains, while smaller chains exhibit little systematic responsiveness to tax policy.[2] Taken together, the results highlight that favorable tax policies primarily stimulate expansion by dominant retailers, with far weaker effects among smaller competitors.

Motivated by these empirical patterns, we explore potential economic mechanisms that could be consistent with these patterns. We do so by developing a new theoretical model of entry with taxes to examine a key economic mechanism. Larger retailers with extensive geographic coverage can leverage demand-side (e.g., customer loyalty, brand equity) or cost-based (e.g., scale economies) advantages, making them more likely to benefit from entering new markets. Consequently, a uniform "one-size-fits-all" tax policy that influences sunk costs of entry can lead to asymmetric responses among firms. Our model confirms that differences in firm size-based advantages explain why larger retailers respond more strongly to tax policies than smaller competitors.[3]

3.2 Related Literature

Our study contributes to several key areas of literature. First, we add to the body of work that examines local and national government policies and their impact on businesses. This literature spans a wide range of contexts, including the effects of right-to-work laws on manufacturer entry (Holmes, 1998), land use zoning regulations on retail activity (Bertrand and Kramarz, 2002), information waves triggered by census data releases (Chi, 2024), and business improvement districts targeting retail revitalization (Shoag and Veuger, 2018). The most relevant aspect of our

[1] This result aligns with prior evidence linking business dynamism to tax reductions, such as Sedlácek and Sterk (2019).

[2] Large chains are defined as those with sufficient scale to capture market share advantages, consistent with evidence that physical presence enhances customer value through re-patronage (Zhang, Chang and Neslin, 2022).

[3] In the Appendix, we also explore model extensions additional scenarios, including proportional tax policies as well as firms with asymmetric entry costs. Across these extensions, our central theoretical insight, that tax policies disproportionately impact firms with preexisting advantages, remains robust. Together, these findings underscore the need to account for firm heterogeneity when designing tax policies, as uniform approaches may unintentionally amplify market concentration while shaping local retail landscapes.

work is the growing focus on place-based tax policies and their influence on firm investment, particularly in manufacturing, innovation, and entrepreneurship (Atkinsa et al., 2023; Corinth and Feldman, 2024; Criscuolo et al., 2019; Da Rin, Di Giacomo and Sembenelli, 2011; Devereux, Griffith and Simpson, 2007; Fajgelbaum et al., 2019; Giroud and Rauh, 2019; Kline and Moretti, 2014; Slattery and Zidar, 2020; Suárez Serrato and Zidar, 2016). We complement this body of research by providing new insights into the relationship between localized tax policies and the evolving retail landscape. Our findings show that while these policies can stimulate entry, their benefits are unevenly distributed between firms, leading to asymmetric responses that favor larger players.

Second, we contribute to the literature on the determinants of retail entry in geographic markets[4] Our findings build on this research by showing how government tax policies serve as a critical driver of retail market dynamics, particularly by reinforcing the dominance of large chains. In addition, we extend these findings beyond sector-specific analyses to offer insights that apply broadly across retail sectors, uncovering universal implications for antitrust and urban development. Beyond competition concerns, we also shed light on how such policies might influence the availability of amenities (Couture and Handbury, 2020), local labor market conditions (Basker, 2005), and spillovers to complementary businesses (Shoag and Veuger, 2018).

Large retailers have demonstrated size-based advantages, including resilience to macroeconomic shocks (Basker, Vickers and Ziebarth, 2018), product line variety (Hollenbeck and Giroldo, 2022), productivity and efficiency (Javorcik and Li, 2013), and strategic entry deterrence (Fang and Yang, 2024; Igami and Yang, 2016). Our work complements this line of research by demonstrating how government policies further enhance market dominance for large retailers. To complement this literature, we provide evidence that tax policies amplify these advantages in ways that are broadly relevant across retail sectors, moving beyond the narrow sectoral focus of much of the existing literature.

Finally, we contribute to the literature examining increasing market concentration in non-

[4]This work falls within the broader empirical work that has examined competition in retail markets, with a focus on location selection and market expansion decisions (Arcidiacono et al., 2016; Hollenbeck, 2017; Igami and Yang, 2016). These works highlight the dominant role of large firms in driving retail growth (Haltiwanger, Jarmin and Krizan, 2010) and the persistence of market dominance over time (Geurts and Van Biesebroeck, 2016; Hanner et al., 2015; Jarmin, Klimek and Miranda, 2009).

traded sectors, particularly retail. Studies have documented the rise of "superstar firms" in retail (Autor et al., 2017) and national concentration trends in service industries (Hsieh and Rossi-Hansberg, 2022). This literature suggests that rising fixed costs (e.g., technology investments) and declining marginal costs have tilted competitive advantages toward larger firms, often at the expense of smaller competitors (Gutiérrez and Philippon, 2017).

3.3 Motivating Patterns

This section examines the relationship between state tax policies and retail establishment entry. We begin by describing the empirical context of U.S. retail markets and the data used to measure tax policies and establishment entry patterns. Using this data, we document key empirical relationships, focusing on the potential heterogeneity in how tax policies influence establishment entry. These patterns provide suggestive evidence of asymmetries in the responsiveness of different types of retailers to tax changes, motivating the theoretical and structural analyses that follow.

3.3.1 Empirical Context

Our analysis draws on three key sources of data: retail establishment records, state tax measures, and local market demographics. Together, these datasets provide a comprehensive foundation for examining the relationship between tax policies and retail market dynamics. Below, we provide a detailed description of each data component.

3.3.1.0.1 Retail landscape. To study the entry of retail establishments, we utilize the National Establishment Time-Series (NETS) Database. This dataset provides detailed information on each establishment's exact location, entry and exit years, employee count, and headquarter address. For our analysis, we focus on a subset of establishments classified under Stan-

dard Industrial Classification (SIC) codes corresponding to retail trade from 1990 to 2014.[5] As highlighted by Rossi-Hansberg, Sarte and Trachter (2021), retail trade is one of the industries well-represented in the NETS database.[6]

The dataset includes all relevant establishment locations in the U.S. during the sample period. Given that retail establishments typically sell non-tradable goods targeted toward local markets rather than national or international ones, we define geographic markets using Census Federal Information Processing Standards (FIPS) codes, which uniquely identify counties and their equivalents in the United States.[7] A key feature of our empirical design is that establishment counts are treated as long-run equilibrium outcomes rather than short-run fluctuations. This interpretation is natural in retail markets, where sunk costs of entry and local demand conditions imply that observed establishments reflect relatively stable industry structures. As such, the cross-sectional and border-based variation we exploit is most consistent with static entry frameworks, a perspective we will eventually adopt in our discussion about potential economic mechanisms.

Each establishment's entry decision is thus organized at the county-year level, enabling us to construct the primary outcome variable: establishment counts. These counts represent the total number of active establishments in a given market and year. Figure 3.1 illustrates the variation in retail establishment counts across counties and over time, providing a snapshot of the data's geographic and temporal heterogeneity.

[5]The Appendix Table C.1 provides a detailed breakdown of the retail sectors included in our data.

[6]The NETS database offers several key advantages for analyzing local market dynamics, as highlighted by Rossi-Hansberg, Sarte and Trachter (2021). First, it provides granular establishment-level data, including detailed geographic and industrial classifications, allowing for nuanced analyses of local and national trends. Second, its unique identifiers enable the tracking of establishment entry, exit, and transitions over time, making it particularly valuable for studying market dynamics. Third, NETS has been validated against Census data, showing strong consistency in industries such as Retail Trade and Services. This reliability is critical to understanding geographical concentration and competition patterns. Finally, NETS captures the complete universe of establishments, avoiding confidentiality constraints that limit access to other datasets, while retaining high levels of representativeness for retail markets.

[7]FIPS codes are a widely used standard for geographic classification in economic and demographic studies.

3.3.1.0.2 Retail chain type.

To capture differences in entry patterns across retailer types, we classify establishments as being large or small chains.[8] A retailer is defined as a **large chain** if it operates 10 or more outlets nationwide, a threshold inspired by zoning ordinances used by municipalities (e.g., Provincetown, San Francisco), while a **small chains** are those with 2 to 10 outlets nationwide.[9]

We use chain size (i.e., large vs. small) as a proxy for a retailer's ability to capture market share, given the absence of establishment-level revenue data. This proxy is motivated by the potential advantages physical store presence offers in terms of customer value and loyalty.[10] Chains with a larger physical footprint are more likely to enjoy these benefits, making size an informative indicator of a retailer's competitive capability.

3.3.1.0.3 Business taxes.

For information about business taxes, we utilize state-level business tax and incentive data from the Panel Database on Incentives and Taxes (PDIT), developed by the Upjohn Institute for Employment Research (Bartik, 2017). This dataset employs a "rules-based" approach to estimate tax intensity, collecting data on the rules governing each tax, tax credit, and incentive program in a locality, and predicting the effective tax and incentive levels faced by firms based on these rules (Slattery and Zidar, 2020). The PDIT dataset has been widely used in studies of state tax policies and their effects on firm behavior, including entrepreneurship (Fazio, Guzman and Stern, 2019). According to Bartik (2017), the dataset is well-suited for capturing differences in overall tax and incentive regimes across states, including variations in property taxes and abatements, although it may not fully capture localized nuances within metropolitan areas. Notably, the dataset excludes consumer-facing taxes, such as sales taxes paid by individuals.

The PDIT database records tax rates and business incentives relevant to new and expanding firms for 45 industries (including retail trade) across 47 cities and 33 states from 1990 to 2015. It includes detailed predictions of the tax incentives a firm would receive in a city, based on its hy-

[8] A retailer is defined as a chain if it has 2 or more outlets.

[9] Some municipalities have enacted "chain store bans" using these thresholds. For example, San Francisco prohibits retail firms with 11 or more nationwide outlets from entering specific neighborhoods.

[10] Physical stores enhance customer engagement through product inspections and multi-sensory shopping experiences, particularly important for experience goods (Nelson, 1970).

pothetical balance sheet. For our analysis, we extract state-level taxes and incentives specific to the retail trade sector during our sample period of 1990 to 2014. The tax measures, expressed as a percentage of industry value-added, are normalized to facilitate comparisons across states and industries. This normalization reflects the effective burden of taxes and benefits of incentives relative to the economic output of the sector.

The database defines a state's tax burden as the total taxes net of total incentives, incorporating components such as property taxes, sales taxes, corporate income taxes and tax credits or subsidies for job creation, investment, research and development, property abatements and job training. These components are listed in Table 3.1. Taxes and incentives are adjusted to 2015 dollars using the GDP-implicit price deflator (Bartik, 2017). This construction facilitates the aggregation of tax and incentive components into a single index, which we use as a proxy for state tax policies in the retail trade sector.[11]

To ensure comparability across states, the database standardizes business profiles and accounts for state-specific apportionment formulas, throwback rules, and tax-incentive interactions over a 20-year horizon. National and state averages are weighted by GDP contributions, while specialized taxes (e.g., unemployment insurance, public utility taxes) and location-specific industries (e.g., mining) are excluded to maintain uniformity. These adjustments make the PDIT dataset a robust tool for analyzing state tax regimes and their effects on retail market dynamics.

Bartik (2017) notes that property taxes account for approximately 65% of a state's overall business tax burden, making them the dominant component of the tax index. Other taxes, such as unemployment insurance and industry-specific levies (e.g., severance taxes), are excluded from the index to reflect their limited relevance to the broader business tax regime. While these omissions are unlikely to impact our analysis of the retail trade sector, we acknowledge the potential for measurement error in the constructed tax proxy. To address this potential concern, we conduct additional sensitivity analyses, detailed in the Appendix, to test the robustness of our results to variations in the tax data.

By focusing on retail trade, categorized in PDIT as a "non-export-based industry," our anal-

[11] We provide more details in section C.1 about how the tax measures are constructed in the database.

ysis aligns with the sectoral focus of our empirical framework. The database provides a comprehensive measure of the tax and incentive environment faced by retail firms, enabling us to examine the interplay between state tax policies and retail establishment entry dynamics.

3.3.1.0.4 Market size. Finally, to proxy for market size, we obtain data about population at the county level from the US Census. This market size proxy is consistent with typical research about retail entry (Bresnahan and Reiss, 1991). Note that this information from the Census is unlikely to be perfect, especially for information during years far removed from the decennial Census surveys (Chi, 2024).[12] For this reason, these characteristics are at best proxies, so all of our empirical analysis will necessitate the use of geographic, border, and temporal fixed effects.

Combining all sources, we have a panel data containing the the state tax policies and the number of retail establishments across over 60 retail sectors at the four-digit SIC code level in over 3000 counties in the US from 1990 to 2014. We first highlight some general patterns in state taxes and retail establishment entry over time. Figure 3.2 provides the average net taxes (Figure 3.5), along with the average growth of retail establishments across markets (Figure 3.5) over time.

From these figures, we observe decreasing net taxes at the state level. This pattern is consistent with the introduction of tax reforms (beginning in the late 1980s) that have led to falling corporate tax rates (Chodorow-Reich, Zidar and Zwick, 2024). Furthermore, the growth of retail establishments at the market level appears to oscillate around 10%. Large chains appear to experience faster growth rates relative to small chains for most of the years.

Moreover, Figure 3.3 provides a snapshot of the geographic and temporal variation for the net taxes, where the light to dark blue ranges capture low to high net tax states. As can be seen, the state-level net tax ranges from 2% to 8%, with a lot of variations both across states and time. In particular, there are many bordering states with differential taxes. For example, New Mexico and Texas share a border, but New Mexico's net tax is about 2% higher than that of Texas in 1990. In 2000, the tax gap closes to about 1%, only to widen again in 2014 back to about 2%.

[12]We consider analysis that uses an alternative measure of population in the Online Appendix.

3.3.2 Taxes and the Retail Landscape

To explore the relationship between state tax policies and retail establishment entry, we leverage variation across markets and time. This variation arises from changes in state tax policies during the sample period and differences in market characteristics between geographic regions. Taking advantage of this variation, we estimate the impact of tax decreases on retail establishment counts, distinguishing between large versus small chains.

3.3.2.1 Making Use of Spatial Discontinuities

We proceed by presenting a spatial discontinuity identification approach using borders, similar to Dell (2010) and Holmes (1998). Specifically, we compare counties that border state boundaries. On one side of the border, counties are likely subject to a different set of tax policies than those counties on the other side. This identification strategy relies on the assumption that contiguous border counties serve as adequate controls.[13] Figure 3.4 illustrates the geographic and temporal variation in the number of establishments for the border counties, and Table 3.2 shows the variation in the data that can be exploited to identify causal effects.

In Table 3.2, the total number of establishments and the number of establishments per capita are listed for neighboring counties, whereby one county has relatively higher taxes than the other. This comparison showcases the potential identification power of using the spatial discontinuity in the form of state borders. As shown in the table, the number of establishments per capita is larger on the side of the border with lower taxes. Although the total number of establishments follows the reverse pattern, it may simply be that counties in higher tax states have a bigger population. For this reason, in our analysis we control the population.

With this border identification strategy, we estimate the following semiparametric specification:

[13] That is, these contiguous border counties have similar underlying economic conditions as their neighbors on the other side of the border. In particular, we need sufficient differences in net taxes within cross-state county-pairs. Moreover, counties would ideally be more similar to its cross-state counterpart than to a randomly chosen county.

$$y_{mt} = \beta_0 + \beta_1 \text{tax}_{st} + \beta_2 \text{pop}_{mt} + f(\text{geographic location}_m) + \phi_b + \delta_t + \varepsilon_{mt}. \quad (3.1)$$

The dependent variable, y_{mt}, represents the count of retail establishments of a certain type (e.g., large or small chains) in market m at time t, scaled by their respective national averages. To obtain these counts, we aggregate the number of establishments across all retail sectors to the county and year level. Note that we use scaling as a means to compare specifications for large versus small chains, as large chains, by their definition, will naturally have more establishments in absolute terms.

Moreover, we incorporate retail sector fixed effects to account for difference in retail sector composition across borders. Furthermore, tax_{st} is state-retail net tax in state s in year t, and δ_t represents year fixed effects. pop_{mt} is the population for county m in year t.

The geographic controls are incorporated in the specification using $f(\text{geographic location}_m)$, which is a standard regression discontinuity polynomial. This term controls for geographic heterogeneity via a smooth function of geographic location[13] (i.e., longitude and latitude). For our analysis, we consider linear, quadratic, and cubic functional-form specifications for $f(\text{geographic location}_m)$. For example, if (x, y) represent the latitude and longitude of a county centroid, then a cubic polynomial function can be written as $\alpha_1 x + \alpha_2 y + \alpha_3 xy + \alpha_4 x^2 + \alpha_5 y^2 + \alpha_6 x^3 + \alpha_7 y^3 + \alpha_8 x^2 y + \alpha_9 x y^2$.

Furthermore, ϕ_b is state border fixed effects, which ensure that the comparison is between county m and its neighboring county on the same state border, but not between different state borders.[14]

3.3.2.2 Summary of Empirical Patterns

We begin by examining the impact of state taxes on overall retail entry. Table 3.3 reports the results across a range of specifications. When aggregating all chains, the estimated effect of net tax

[14]Our design implicitly assumes that unobserved determinants of retail entry evolve smoothly across borders, such that discontinuities reflect state tax policy. This assumption is plausible in practice given the similarity of contiguous counties along borders (Dell, 2010; Holmes, 1998). Potential violations, such as correlated shifts in other state-level business policies, are unlikely to coincide precisely with border demarcations.

rates is small and statistically indistinguishable from zero. This motivates a more disaggregated analysis that distinguishes between the entry decisions of large and small chains.

While the estimated coefficients may appear modest in absolute magnitude, they reflect economically meaningful adjustments in retail structure when benchmarked against observed establishment counts. Because the dependent variable is normalized by national averages, a 0.01 increase in a state's net business tax—roughly a 15–20 percent rise from the sample mean and comparable to the cross-border differences observed between states such as New Mexico and Texas—corresponds to about a five-percent lower count of large-chain establishments. Given that the average county in our sample hosts approximately 98 large-chain outlets, this effect translates to roughly five fewer establishments per county. By contrast, the corresponding effect for small chains, which average 47 outlets per county, is statistically indistinguishable from zero. Although moderate in elasticity terms, these differences accumulate across space and time, implying that even relatively small and persistent tax differentials can reshape the spatial allocation of large-chain activity.

The results reveal a noticeable asymmetry. For large chains, the estimated coefficients on net tax are consistently negative and statistically significant across all specifications, indicating that higher taxes reduce the number of establishments. By contrast, the corresponding estimates for small chains are close to zero and statistically insignificant. In other words, the aggregate null effect masks important heterogeneity, as tax policy appears to primarily shape the entry behavior of large chains, while having little measurable influence on small ones.[15]

Large chains, by virtue of their scale and financial flexibility, appear better positioned to translate reductions in tax burdens into additional expansion, while smaller chains may face frictions that limit their responsiveness. As a result, the benefits of tax cuts tend to accrue disproportionately to larger players in the retail sector. This asymmetric responsiveness raises the question of what underlying mechanisms drive the divergence. The theoretical framework that follows takes up this challenge by interpreting firm strength as rooted in either demand-side advantages, such as brand loyalty and customer reach, or cost-side advantages, such as economies of scale. In this way, the model proposes a conceptual bridge between the empirical patterns

[15]Viewed through the lens of entry as investment, this result echoes prior evidence that corporate tax policies often exert weaker effects on small businesses, such as those documented in Finland (Harju, Koivisto and Matikka, 2022).

and their microeconomic mechanisms.

3.3.2.2.1 Market scope. Next, we examine whether the main results are affected by the market scope of the chains considered in the establishment tabulation. For example, one may be concerned that small chains are unresponsive to tax policy changes because their scope of markets for expansion is limited (compared to large chains).

To address this potential concern, we develop a robustness test in which we restrict the tabulation to outlets belonging to chains that maintain a presence on both sides of a state border. This restriction ensures that the construction of establishment counts reflects chains with credible operations on both sides of a border. By focusing on these chains, the test mitigates concerns that the observed asymmetries in entry patterns are driven by small or large chains that operate exclusively in one state and thus never respond to tax policies of neighboring states.

The results in Table 3.4 remain consistent with the baseline findings. When redefining the market scope to include only chains with presence on both sides of a border, the estimated tax coefficients for large chains continue to be negative and statistically significant, whereas those for small chains remain close to zero and imprecisely estimated. This refinement yields smaller baseline counts but a comparable pattern of sensitivity to state taxes. Because the dependent variable remains scaled by national averages, a 0.01 increase in a state's net business tax—approximately a 15–20 percent rise from the sample mean—corresponds to about a five-percent lower count of large-chain establishments. Given that the average county under this restricted definition hosts roughly 66 large-chain outlets, the effect translates to approximately three fewer establishments per county. By contrast, the corresponding effect for small chains, which average four outlets per county, remains negligible. Taken together, these results indicate that the concentration of the tax effect among large chains is not an artifact of market definition but reflects a persistent feature of cross-border retail structure.

3.3.2.2.2 Placebo. To further assess the validity of the border-based empirical strategy, we conduct a placebo analysis that evaluates whether the estimated effects arise in settings where intuitively, *no* cross-border treatment should exist. The basic idea is to replicate the border specification within each state, comparing interior counties to the average of other interior counties

in the same state and year.

We implement the placebo test in three steps. First, we restrict our sample to interior counties located more than 60km from any state border. Second, for each interior county, we compute the difference between its establishment count and the average establishment count across all interior counties in the same state and year. This creates a differenced outcome variable that measures how each interior county deviates from its state's interior average. Third, we estimate the same regression specification used in our main analysis, but now applied to these interior county differences.

Formally, for county m in state s, belonging to state-pair p (its actual border pairing) at time t, we estimate

$$\triangle y_{mst} = \beta \, \text{tax}_{st} + \gamma \, \triangle \text{pop}_{mst} + \alpha_m + \delta_{pt} + \varepsilon_{mst},$$

where

$$\triangle y_{mst} = y_{mst} - \bar{y}_{st}^{\text{interior}}, \qquad \triangle \text{pop}_{mst} = \text{pop}_{mst} - \bar{\text{pop}}_{st}^{\text{interior}}.$$

Here y_{mst} denotes the (scaled) retail-establishment count in interior county m, and $\bar{y}_{st}^{\text{interior}}$ is the mean across other interior counties within state s in year t. The county fixed effects α_m absorb time-invariant county characteristics, while the state-pair-by-year effects δ_{pt} capture contemporaneous shocks common to both sides of an actual border pair.

This specification applies the same spatial discontinuity estimator as in the main analysis but to counties that should not experience cross-border treatment variation. If our research design is sound, we should find statistically insignificant tax coefficients in this interior-only sample. Thus, finding null effects provides a falsification test, confirming that the main results are not artifacts of state-level shocks or within-state spillovers but reflect genuine differences in border counties' responses to state tax changes.

The placebo results reported in Table 3.5 show that the estimated coefficients on the state tax variable are small in magnitude and statistically indistinguishable from zero in all specifications. This pattern is consistent with the construction of the placebo test: because interior counties are compared against their own state-year averages, no systematic treatment variation exists in this setting. The absence of significant coefficients confirms that our empirical design does not spuriously generate tax effects where none should be present.

The logic behind this test hinges on the fact that interior counties face identical state tax environments as their comparison group. Unlike border counties that can respond to neighboring states' tax rates, interior counties' establishment decisions should be insensitive to their own state's tax level when measured relative to other interior counties in the same state. Finding null effects in this setting confirms that our border design correctly isolates causal tax effects rather than capturing broader state-level confounders that would affect both border and interior counties equally.

In summary, this placebo approach serves as a powerful falsification test. If unobserved state-level factors were driving our results, we would expect to see similar tax coefficients in both border and interior samples. Conversely, finding significant effects only at borders—where tax differentials create meaningful variation—but not in interiors, provides strong evidence that we are identifying genuine responses to tax policy rather than spurious correlations.

3.4 Theoretical Analysis of Entry and Taxes

This section introduces an oligopoly entry model to interpret how uniform tax policies interact with heterogeneity between firms. The framework emphasizes preexisting differences in firm strength (e.g., demand-side advantages, cost efficiencies, established market presence) and demonstrates how these asymmetries condition responses to taxes. By allowing for such variation, the model captures the differential burdens that uniform policies impose, illustrating how taxes can potentially amplify structural advantages, alter competitive intensity, and ultimately shape market coverage. In this way, the framework proposes a parsimonious structure for interpreting the empirical results and situating them within broader policy debates about whether uniform incentives disproportionately benefit large incumbents relative to smaller rivals.

3.4.1 Actions and Payoffs

We model this scenario as a game of entry under incomplete information. Two firms, indexed by $i = 1, 2$, independently choose whether to enter the market. We denote firm i's entry decision

as $a_i \in \{0, 1\}$, where $a_i = 1$ indicates market entry and $a_i = 0$ indicates no entry. The payoffs for each firm are as follows. Firm 1's variable profit is 0 if it does not enter, 1 if it enters as a monopolist, and $\rho \in [0, 1]$ if it enters as part of a duopoly (i.e., with firm 2), and firm 2's variable profit is similarly defined, with its duopoly profit being $1 - \rho$.

The market size is normalized to 1, and total demand is fixed. Consequently, the entry of a second firm does not expand the market size. This assumption might be appropriate for mature retail sectors that have reached a steady-state. Moreover, the parameter ρ represents firm 1's relative strength, while $1 - \rho$ represents firm 2's relative strength in the market. We interpret firm strength capturing structural asymmetries that can arise either from the demand side (e.g., consumer loyalty, brand reach) or the cost side (e.g., scale economies, capital access). This dual interpretation ensures the model remains consistent with multiple channels suggested by the empirical findings.

The asymmetries in firm strength may relate to demand-side advantages, such as consumer loyalty or product differentiation, or from cost-side advantages, such as economies of scale. Below, we provide two interpretations of ρ, demonstrating how it may reflect firm size through either demand-based or cost-based mechanisms. These interpretations also connect the theoretical model to empirical analyses where firm size is often observable.

3.4.1.0.1 Demand-based interpretation of ρ and firm size. We now offer a formal interpretation of ρ as a proxy for firm 1's strength through demand-side channels. Specifically, consider a post-entry Cournot competition model in which firms 1 and 2 choose quantities q_1 and q_2, with total market supply given by $Q = q_1 + q_2$. Each firm has a size, denoted s_1 and s_2, which shifts the demand it faces. The inverse demand function is specified as

$$p_i = \alpha s_i - \beta Q,$$

where α reflects the extent to which firm size enhances willingness-to-pay for its products, and β captures the sensitivity of price to total output. For expositional clarity, we assume marginal costs are zero, thereby isolating demand-side effects. Solving the Cournot model yields equilibrium quantities

$$q_i = \frac{\alpha s_i - \beta q_j}{2\beta}, \quad i \neq j,$$

which, under symmetry ($s_1 \simeq s_2 = s$), simplify to

$$q_i = \frac{\alpha s}{3\beta}, \quad p_i = \frac{\alpha s}{3}.$$

Firm i's profit is therefore

$$\Pi_i(s) = p_i q_i = \frac{\alpha^2 s^2}{9\beta}.$$

Differentiating with respect to firm size gives

$$\frac{\partial \Pi_i}{\partial s} = \frac{2\alpha^2}{9\beta}\, s > 0.$$

This derivative shows that equilibrium profits are strictly increasing in firm size, without additional restrictions on parameters. Hence, firm size s can be viewed as a proxy for ρ, capturing how larger firms enjoy greater demand-side advantages. Illustrative examples include Walmart's ability to leverage its store footprint to increase demand-side scale economies (i.e., one-stop shopping via product variety), or Amazon's capacity to use customer loyalty programs to amplify demand-side strength. Both cases are consistent with our interpretation that ρ summarizes such demand-side advantages via firm size.

3.4.1.0.2 Cost-based interpretation of ρ and firm size.

We next offer a cost-side interpretation of ρ through economies of scale. Again, we model post-entry Cournot competition with two firms. The inverse demand is

$$p = \alpha - \beta Q, \quad Q = q_1 + q_2,$$

where α reflects baseline demand and β captures the sensitivity of price to total output. Firm size s_i now enters through marginal cost:

$$c_i = c - \delta s_i,$$

with c denoting base cost and δ measuring the cost-reducing effect of size. Solving the Cournot model yields equilibrium outputs

$$q_i = \frac{\alpha - 2c_i + c_j}{3\beta}, \qquad i \neq j.$$

Under symmetry ($s_1 \simeq s_2 = s$), this simplifies to

$$q_i = \frac{\alpha - c + \delta s}{3\beta}, \qquad p = \frac{\alpha + 2c - 2\delta s}{3}.$$

Firm i's profit is therefore

$$\Pi_i(s) = (p - c_i)q_i = \frac{(\alpha - c + \delta s)^2}{9\beta}.$$

Differentiating with respect to firm size gives

$$\frac{\partial \Pi_i}{\partial s} = \frac{2\delta}{9\beta}(\alpha - c + \delta s).$$

Hence, firm profits are increasing in size whenever

$$s > \frac{c - \alpha}{\delta}.$$

This threshold condition illustrates how size-related cost reductions enhance profitability. Interpreting ρ as a proxy for size therefore captures cost-side advantages, consistent with cases such as Home Depot or Walmart, where scale economies and operational efficiencies translate into lower marginal costs and higher profitability.

Having established the demand-side and cost-side mechanisms by which firm size influences payoffs, we now consider the sunk costs associated with market entry. Specifically, each firm faces an entry cost, denoted $\kappa < 1$, and has private information about its own productivity, ε_i, which is assumed to be independently and identically distributed according to a uniform distribution $\mathrm{U}[0, 1]$. This distribution enables analytical solutions to the entry game, and the productivity shocks ε_i is known only to the firm. The expected payoff for firm 1 when entering the market is given by:

$$\pi_1(a_1 = 1, a_2) = 1 - a_2(1 - \rho) - \kappa - \varepsilon_1,$$

and for firm 2, it is:

$$\pi_2(a_1, a_2 = 1) = 1 - a_1\rho - \kappa - \varepsilon_2.$$

These payoff functions incorporate both the sunk cost of entry and the uncertainty about productivity, allowing us to characterize the equilibrium entry probabilities in closed form.

Our theoretical framework adopts a static entry game rather than a dynamic model. This choice reflects the empirical context of retail markets, where sunk costs and geographic demand imply that observed establishment counts approximate long-run equilibrium structures. Because our empirical analysis exploits cross-sectional and border-based variation in establishment presence, the relevant outcomes correspond to equilibrium entry probabilities rather than transitory fluctuations. Static entry models are especially well-suited for industries such as retail

where local markets are likely mature. Thus, while our model abstracts from dynamic considerations, the static structure provides a tractable and empirically consistent framework for linking tax policy to equilibrium entry patterns.

3.4.2 Strategies, Expectations, and Equilibrium Entry

Given the variable profits, sunk costs, and private information defined in the model, we derive the entry conditions for firms 1 and 2. Entry decisions are made simultaneously, and firms form expectations about the actions of their competitor. Specifically, firm 1 forms an expectation about the probability of firm 2 entering the market, $\sigma_2 = \Pr(a_2 = 1)$, and firm 2 forms an analogous expectation about firm 1, $\sigma_1 = \Pr(a_1 = 1)$. Each firm's entry decision is therefore a function of its competitor's choice probability. Firm 1 will choose to enter if and only if:

$$1 - (1 - \rho)\sigma_2 - \kappa \geq \varepsilon_1.$$

Similarly, firm 2 will enter if and only if:

$$1 - \rho\sigma_1 - \kappa \geq \varepsilon_2.$$

Given that the private productivity shocks, ε_1 and ε_2, are independently and identically distributed according to a uniform distribution $\mathrm{U}[0, 1]$, the choice probabilities for firms 1 and 2 can be expressed as:

$$\sigma_1 = 1 - (1 - \rho)\sigma_2 - \kappa,$$

$$\sigma_2 = 1 - \rho\sigma_1 - \kappa.$$

The strategy profiles, expressed in terms of choice probabilities σ_1 and σ_2, are characterized by a Bayesian Nash Equilibrium (BNE), as defined below:

Definition 1. A strategy profile $\{\sigma_i^*\}_{\forall i}$ constitutes a Bayesian Nash Equilibrium (BNE) if and only if:

$$\pi_i(\sigma_i^*, \sigma_j^*) \geq \pi_i(\sigma_i, \sigma_j^*) \quad \forall\, \sigma_i \in [0, 1].$$

This definition ensures that neither firm has a unilateral incentive to deviate from the equilibrium strategy profile (σ_1^*, σ_2^*). To determine the BNE, we solve the system of equations for σ_1

and σ_2:

$$\sigma_1^* = \frac{(1-\kappa)\rho}{1-\rho(1-\rho)},$$
$$\sigma_2^* = \frac{(1-\kappa)(1-\rho)}{1-\rho(1-\rho)}.$$

The BNE is therefore given by the strategy profile (σ_1^*, σ_2^*). In the following section, we analyze the implications of tax policies that alter the cost of entry, κ.

Our theoretical framework adopts a static entry game rather than a dynamic model. This choice reflects the empirical context of retail markets, where sunk costs and geographic demand imply that observed establishment counts approximate long-run equilibrium structures. Because our empirical analysis exploits cross-sectional and border-based variation in establishment presence, the relevant outcomes correspond to equilibrium entry probabilities rather than transitory fluctuations. Static entry models are especially well-suited for industries such as retail where local markets are likely mature. Thus, while our model abstracts from dynamic considerations, the static structure provides a tractable and empirically consistent framework for linking tax policy to equilibrium entry patterns.

Although the empirical analysis exploits spatial tax differences across state borders, the theoretical framework is intentionally non-spatial. The model captures how heterogeneity in firm characteristics shapes the sensitivity of entry decisions to tax policy *within* a local market, rather than re-allocations of firms across borders. Consequently, the empirical design does not require an explicit location-choice mechanism as it relies solely on cross-state variation in tax rates as an exogenous source of identifying variation, consistent with the model's focus on differential entry-cost responsiveness.

3.4.3 Taxes

Under the Bayesian Nash Equilibrium (BNE), we evaluate the effects of a tax policy that increases the cost of entry by τ. Results for a tax cut would mirror those discussed here. This policy modifies the equilibrium strategies for firms 1 and 2, which we denote as $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$. The

new equilibrium strategies under this tax policy are given by:

$$\tilde{\sigma}_1 = \sigma_1^* - \frac{\tau\rho}{1-\rho(1-\rho)},$$

$$\tilde{\sigma}_2 = \sigma_2^* - \frac{\tau(1-\rho)}{1-\rho(1-\rho)}.$$

By comparing these equilibrium choice probabilities to those without the tax policy, the marginal impact of taxation can be assessed. But before we discuss some of the key insights from this model, we first review a set of assumptions that are relevant for our analysis.

3.4.3.0.1 Asymmetric tax effects on entry.

The effects $(\tilde{\sigma}_1 - \sigma_1^*)/\tau$ and $(\tilde{\sigma}_2 - \sigma_2^*)/\tau$ are both strictly negative for all $\rho \in [0,1]$, indicating that higher taxes reduce firm entry probabilities. Conversely, lower taxes increase entry probabilities, promoting more establishment entry into the industry as a whole. This observation highlights an industry-level benefit associated with favorable tax conditions. However, the tax-induced changes in entry probabilities are asymmetric across firms, depending on their relative strengths. This asymmetry disappears only when $\rho = 0.5$, where firms are equally strong. To formalize this analytical result, we state the following proposition below.

Proposition 1. *Under a uniform tax policy that increases the entry cost by τ, the sensitivity of equilibrium entry probabilities $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$ to the tax policy depends on the relative strength of the firms, ρ. Specifically,*

$$\frac{\partial(\tilde{\sigma}_1 - \tilde{\sigma}_2)}{\partial\tau} = \frac{1-2\rho}{1-\rho(1-\rho)}.$$

This difference in sensitivity is symmetric (i.e., $\frac{\partial(\tilde{\sigma}_1-\tilde{\sigma}_2)}{\partial\tau} = 0$) when $\rho = 0.5$, while for $\rho > 0.5$, the stronger firm (firm 1) reacts more strongly to the tax policy compared to the weaker firm (firm 2). Conversely, when $\rho < 0.5$, the weaker firm exhibits greater sensitivity.

Proof. The Bayesian Nash Equilibrium (BNE) entry probabilities for firms 1 and 2 without taxes are given by:

$$\sigma_1^* = \frac{(1-\kappa)\rho}{1-\rho(1-\rho)}, \quad \sigma_2^* = \frac{(1-\kappa)(1-\rho)}{1-\rho(1-\rho)}.$$

When a uniform tax τ is introduced, the entry cost increases to $\kappa + \tau$, and the equilibrium probabilities become:

$$\tilde{\sigma}_1 = \frac{(1-\kappa-\tau)\rho}{1-\rho(1-\rho)}, \quad \tilde{\sigma}_2 = \frac{(1-\kappa-\tau)(1-\rho)}{1-\rho(1-\rho)}.$$

The marginal impact of taxation on each firm's entry probability is then:

$$\frac{\partial \tilde{\sigma}_1}{\partial \tau} = -\frac{\rho}{1-\rho(1-\rho)}, \quad \frac{\partial \tilde{\sigma}_2}{\partial \tau} = -\frac{1-\rho}{1-\rho(1-\rho)}.$$

The difference in tax sensitivity is therefore:

$$\frac{\partial(\tilde{\sigma}_1 - \tilde{\sigma}_2)}{\partial \tau} = \frac{\partial \tilde{\sigma}_1}{\partial \tau} - \frac{\partial \tilde{\sigma}_2}{\partial \tau}.$$

Substituting the derivatives, we have:

$$\frac{\partial(\tilde{\sigma}_1 - \tilde{\sigma}_2)}{\partial \tau} = -\frac{\rho}{1-\rho(1-\rho)} + \frac{1-\rho}{1-\rho(1-\rho)}.$$

Simplifying the numerator gives:

$$\frac{\partial(\tilde{\sigma}_1 - \tilde{\sigma}_2)}{\partial \tau} = \frac{(1-\rho)-\rho}{1-\rho(1-\rho)} = \frac{1-2\rho}{1-\rho(1-\rho)}.$$

To confirm symmetry when $\rho = 0.5$, substitute $\rho = 0.5$ into the expression:

$$\frac{\partial(\tilde{\sigma}_1 - \tilde{\sigma}_2)}{\partial \tau} = \frac{1-2(0.5)}{1-(0.5)(1-0.5)} = \frac{0}{0.75} = 0.$$

Thus, the tax effect is symmetric at $\rho = 0.5$. For $\rho > 0.5$, $1 - 2\rho < 0$, implying that firm 1 (the stronger firm) is more sensitive to taxes than firm 2. Conversely, when $\rho < 0.5$, $1 - 2\rho > 0$, implying that firm 2 (the weaker firm) is more sensitive. □

This analytical result demonstrates the potential asymmetric sensitivity of firm entry probabilities to uniform tax policies. The degree of asymmetry depends on ρ, which captures the relative strength of the firms. When $\rho = 0.5$, the two firms are of equal strength, resulting in symmetric responses to tax changes. However, for values of ρ greater than 0.5, firm 1 benefits more from entry and consequently exhibits a stronger reaction to tax changes compared to firm 2. In contrast, when ρ is less than 0.5, firm 2's relative weakness amplifies its sensitivity to changes in entry costs.

In summary, the result underscores the role of firm strength in determining the impact of tax policy, as uniform taxes, while designed to treat firms equally, introduce asymmetric distortions by disproportionately affecting firms with different strengths. This asymmetry not only exacerbates market concentration by favoring stronger firms when taxes are reduced but also disproportionately deters weaker firms under higher taxes, thereby reducing market dynamism. A notable consequence of this tax-induced asymmetry is its potential to distort the market landscape,

as uniform tax policies shift entry rates away from levels that minimize market concentration. The degree of market concentration is inversely related to the equality of entry probabilities $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$; hence, as $\tilde{\sigma}_1 - \tilde{\sigma}_2$ increases with τ, market concentration rises, favoring the stronger firm and amplifying the disparities in market outcomes.

Intuitively, stronger firms are more sensitive to taxes because they have more to gain (or lose) from market entry. With their larger market share and ability to command higher payoffs due to scale or brand recognition, their decision to enter is more finely balanced. A tax increase disproportionately raises their effective entry costs relative to the potential payoffs, making them more responsive to these changes. Conversely, weaker firms, with smaller payoffs and less dominance, are less impacted by incremental changes in entry costs since their stakes in the market are already limited.

3.4.3.0.2 Taxes and under-served markets.

An additional implication of the framework concerns the incidence of retail deserts—markets that fail to attract any entry. Uniform taxes can widen disparities in consumer access by increasing the likelihood that weaker markets remain under-served. In this way, the model links firm-level asymmetries not only to competitive intensity but also to broader distributional outcomes across communities.

Without tax policy, the probability that neither firm enters is

$$\Psi = (1 - \sigma_1^*)(1 - \sigma_2^*),$$

and under a uniform tax τ it becomes

$$\tilde{\Psi} = (1 - \tilde{\sigma}_1)(1 - \tilde{\sigma}_2) = \Psi + \frac{\tau\big(1 - 3\rho + 3\rho^2 + 2\kappa\rho(1-\rho) + \tau\rho(1-\rho)\big)}{(1 - \rho(1-\rho))^2}.$$

Minimizing $\tilde{\Psi}$ with respect to τ yields the interior first-order condition

$$\hat{\tau}_{\min} = -\kappa + \frac{3}{2} - \frac{1}{2\rho} - \frac{1}{2(1-\rho)}.$$

Because $\kappa \in [0, 1)$, $\hat{\tau}_{\min}(\rho) < 0$ for all $\rho \in (0, 1)$, so the planner's feasible choice under a non-negativity constraint is

$$\hat{\tau} = \max\{0, \hat{\tau}_{\min}\} = 0.$$

Proposition 2. *If negative taxes are not permitted, the constrained optimal tax that minimizes under-served markets is $\hat{\tau} = 0$. Comparative statics for the interior (unconstrained) solution show that $\hat{\tau}_{\min}$ is*

decreasing in ρ when $\rho > 1/2$ and increasing when $\rho < 1/2$.

Proof. Differentiating the interior solution gives

$$\frac{\partial \hat{\tau}_{\min}}{\partial \rho} = \frac{1 - 2\rho}{2\rho^2(1-\rho)^2},$$

which is negative for $\rho > 1/2$, zero at $\rho = 1/2$, and positive for $\rho < 1/2$. □

The derivative pattern for $\hat{\tau}_{\min}$ illustrates how firm heterogeneity would shape optimal taxation if subsidies were allowed: stronger asymmetry (large ρ) would call for lower or even negative taxes to maintain coverage. Under the realistic non-negativity constraint, however, the planner chooses $\hat{\tau} = 0$ everywhere, implying that any positive uniform tax increases the prevalence of under-served markets. Consequently, in a counterfactual setting with rebates or targeted subsidies, relief would be most valuable in highly asymmetric markets.

In practice, the non-negativity constraint implies a simple policy takeaway. Any positive *uniform* tax increases the incidence of under-served markets, so the constrained optimum is to avoid additional uniform taxation ($\hat{\tau} = 0$). This does not rule out policy interventions altogether. If a jurisdiction can deploy *targeted* instruments (e.g., place-based rebates, entry subsidies, or firm-contingent incentives) then the interior comparative statics for $\hat{\tau}_{\min}$ are informative about where relief would be most valuable (namely, markets with greater asymmetry in strength). Absent such tools, however, raising a uniform tax risks widening coverage gaps, whereas keeping it at zero avoids exacerbating retail deserts.

3.4.4 Policy Implications

In summary, our empirical and theoretical analysis suggests an important tension for policy design. Uniform tax policies, although administratively simple, risk amplifying existing disparities in market power by disproportionately stimulating entry among larger firms. From a welfare perspective, such policies may inadvertently increase concentration while doing little to foster the dynamism of smaller competitors. A more nuanced approach would tailor incentives to account for firm heterogeneity. For example, subsidies targeted toward smaller chains

or caps on incentive eligibility for dominant incumbents. These considerations underscore the importance of aligning tax policy with competitive objectives, rather than treating all firms symmetrically.

3.5 Conclusion

We present new insights into the impact of place-based tax policies on shaping local retail landscapes. Using comprehensive data on retail entry across the United States, we demonstrate that tax policies have asymmetric effects between firms of different sizes. Our empirical analysis confirms that tax effects on establishment entry are disproportionately driven by the largest retail firms. Specifically, policies that *reduce* taxes ultimately *reinforce* the growth of already dominant retail firms. Through a model of entry with taxes, we identify one potential mechanism for these patterns, showing that stronger firms are more sensitive to tax policy changes. Consequently, *favorable* tax policies disproportionately *increase* entry among large retail chains relative to smaller chains.

Although our analysis sheds light on observed effects of tax policy on retail entry and potential economic mechanisms, it abstracts from the role of local government competition. Local governments often compete to attract firm investment through increasingly favorable tax incentives (Lin, 2024; Slattery, 2022). Such competition could exacerbate the asymmetric effects documented here, amplifying the dominance of large retailers at the expense of smaller competitors. Future research could explore the extent to which such bidding magnifies the market power of already dominant firms, leading to broader economic distortions. Our findings thus point to an important direction for future research, namely, how inter-jurisdictional competition for retail investment interacts with firm heterogeneity, and whether such competition systematically reinforces the dominance of large chains.

Although our theoretical framework provides a parsimonious explanation for the heterogeneous responses of large and small chains, it is necessarily stylized. We abstract from heterogeneity within small and large chains and do not model dynamic considerations such as sequential entry, learning, or exit. Likewise, we do not incorporate the possibility of cross-border

substitution directly into the theoretical structure, even though our empirical design and robustness checks explore this potential aspect. Thus, extending the framework to incorporate these additional layers of heterogeneity and dynamics represents another direction for future research.

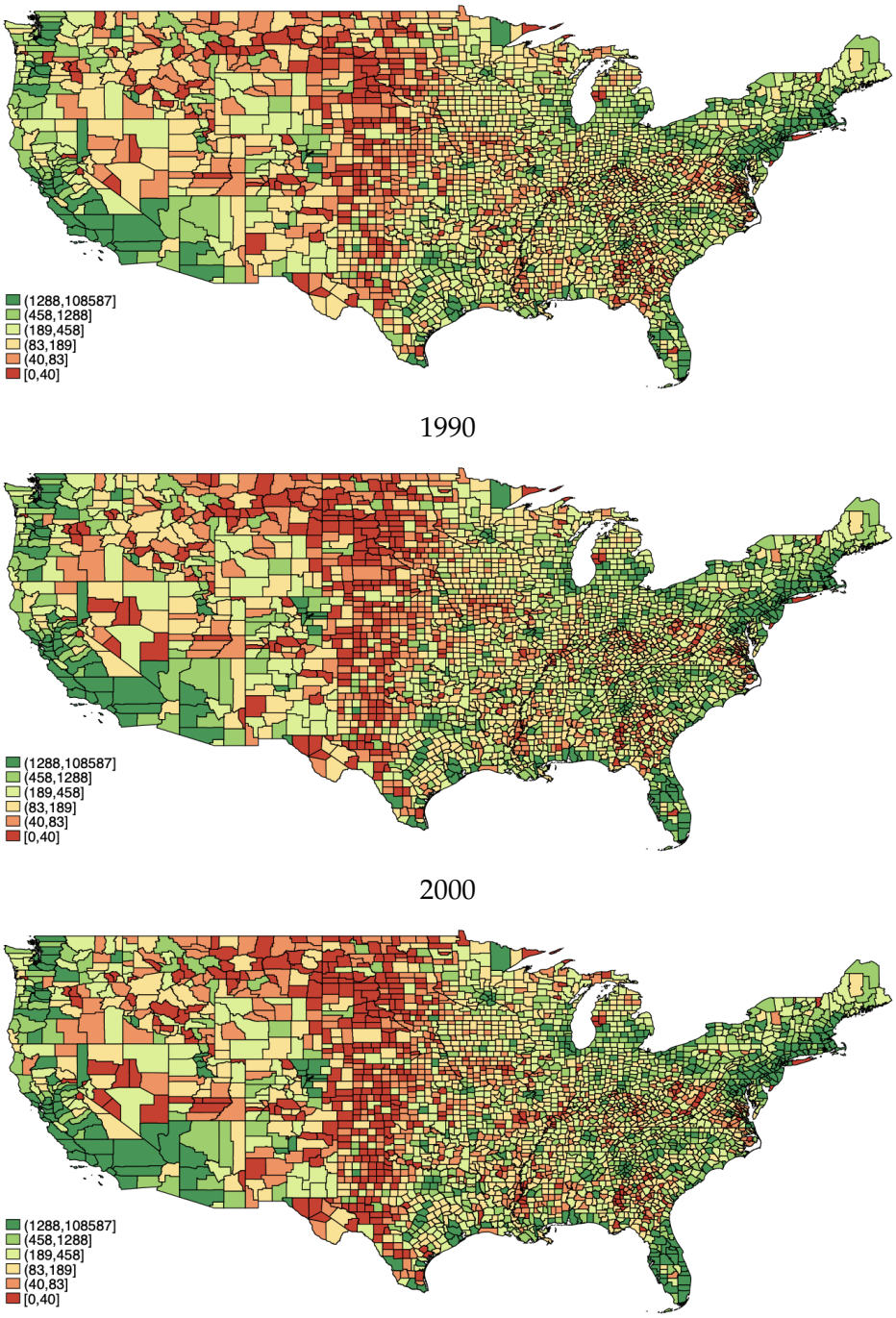
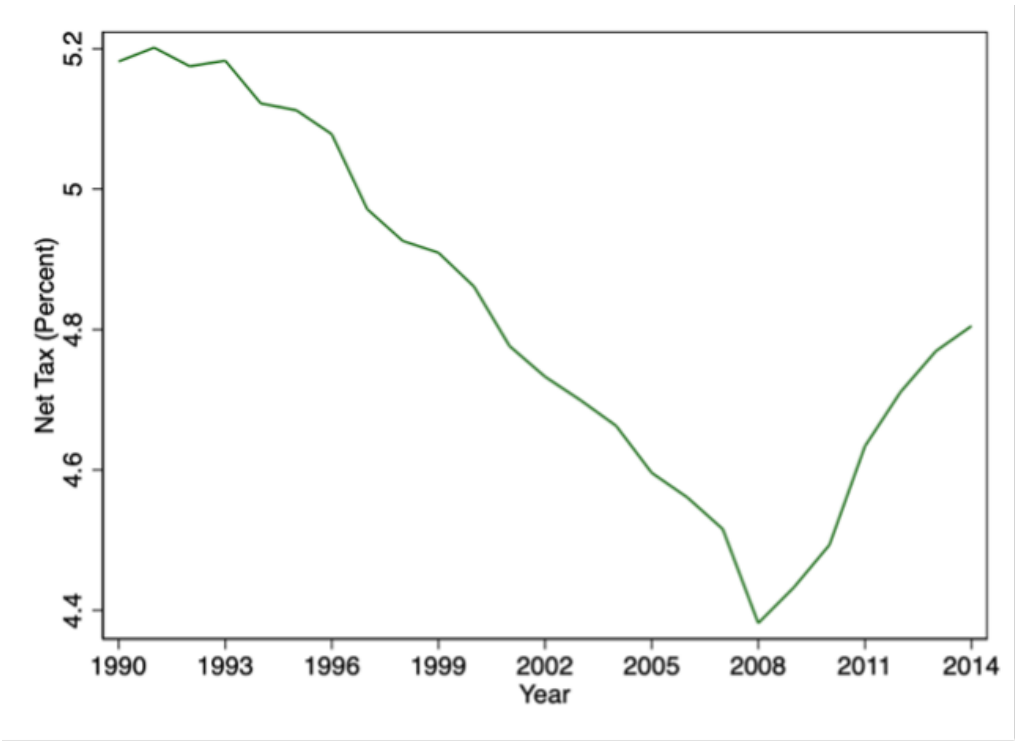
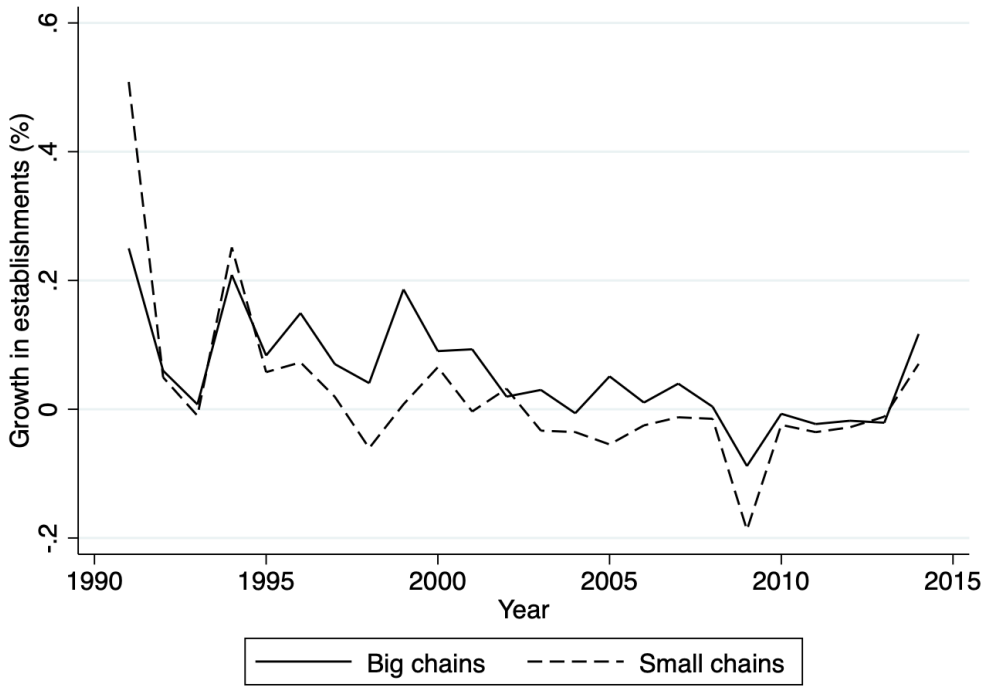


Figure 3.1: Geographic and Temporal Variation in Establishment Counts

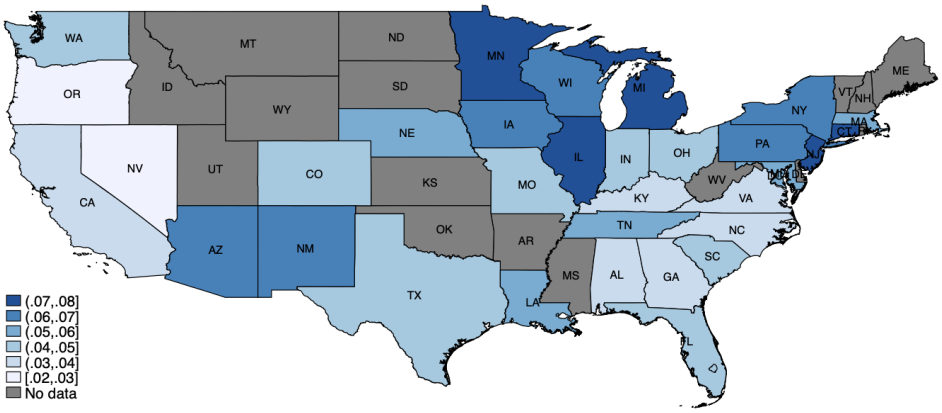


(a) Averaged Net Taxes

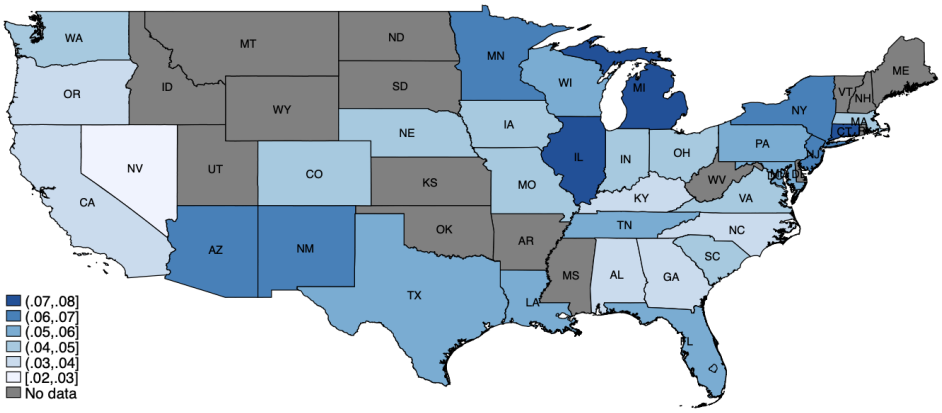


(b) Market-Level Growth Rates

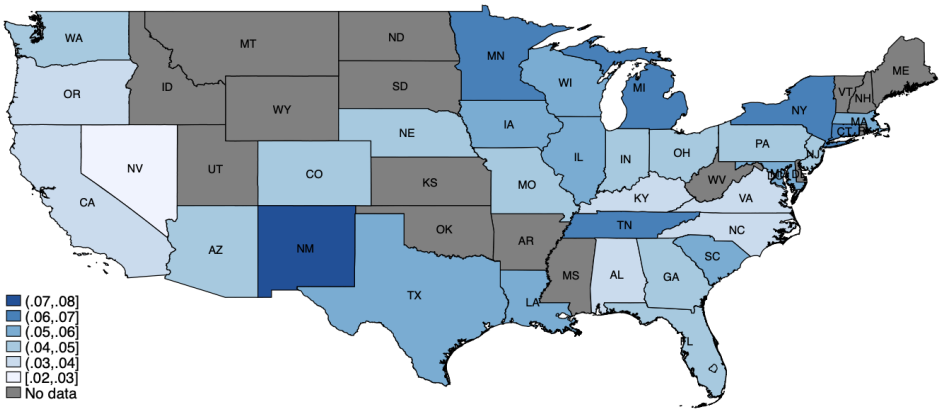
Figure 3.2: Net Tax and Establishment Entry Dynamics



1990

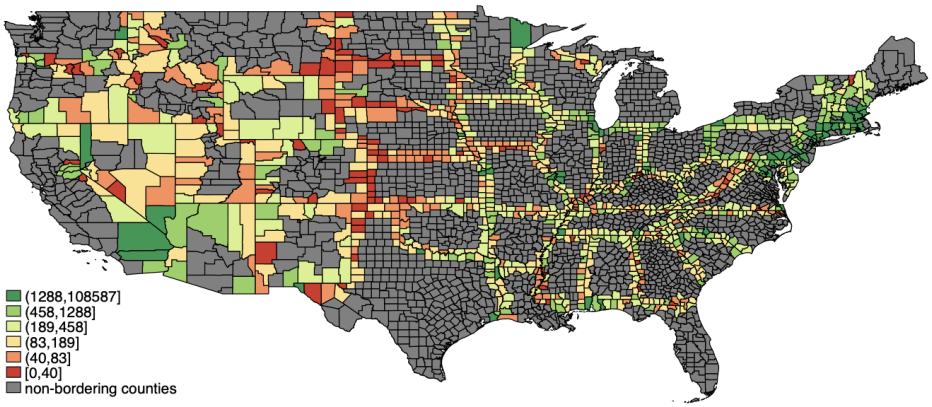


2000

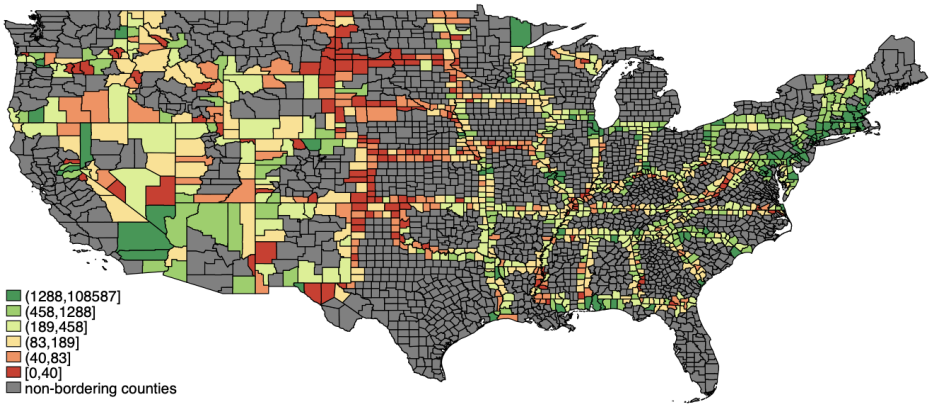


2014

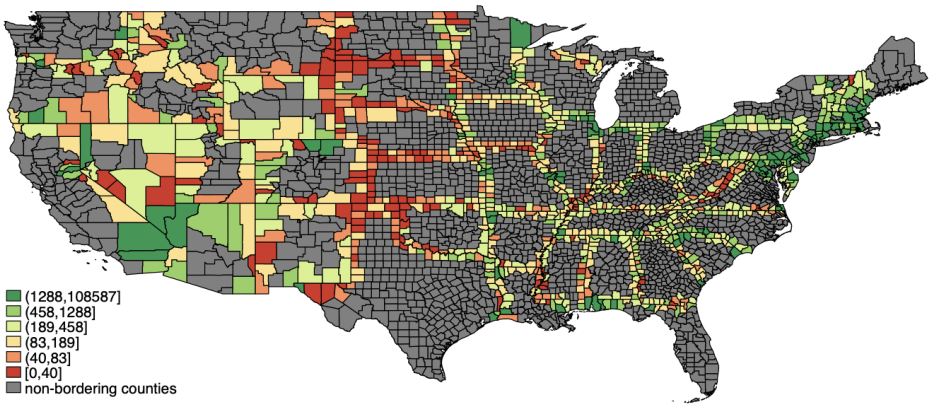
Figure 3.3: Geographic and Temporal Variation in Net Tax



(a) 1990



(b) 2000



(c) 2014

Figure 3.4: Geographic and Temporal Variation in Establishment Counts for Border Counties

Table 3.1: Relevant State Taxes that Impact Operating/Fixed Costs

Component	**Description**
Property taxes	Taxes the firm will pay that relate to the physical properties they own
Sales taxes	Taxes firm will pay on its business inputs (e.g., initial purchase of building materials, initial purchase of machinery)
Corporate income taxes	Taxes firm will pay that is based on their profits
Job creation credit	Tax credit firm receives for each job that is created
Investment credit	Tax credit that allows firms to deduct investment costs from their taxes
Research credit	Tax credit that allows firms to deduct R&D costs from their taxes
Property abatement	Subsidy offered on certain types of real estate or business opportunities
Job training subsidy	Subsidy offered to help cover personnel training costs

Table 3.2: Average Establishment Counts on Both Sides of the State Border

Distance from Border (km)	Establishment Counts	Establishment Counts per 10000 Capita
	Higher Tax Side of Border	
>100	1997	79
75-100	1391	71
50-75	564	71
25-50	674	73
1-25	844	71
0	767	74
	Lower Tax Side of Border	
0	723	77
1-25	689	73
25-50	491	77
50-75	392	75
75-100	636	72
>100	787	78

Note: This table shows the number of establishments in counties located on both sides of the state border. The first column measures the minimum distance from a county to its closest state border.

Table 3.3: Retail Chain Establishment Counts and State Taxes

All chains				
Net Tax (percent)	-0.02	-0.02	-0.02	-0.02
	(0.02)	(0.02)	(0.02)	(0.02)
Population (10,000)	0.08***	0.08***	0.08***	0.08***
	(0.00)	(0.00)	(0.00)	(0.00)
Observations	15794	15794	15794	15794
R^2	0.92	0.92	0.92	0.92
Big chains				
Net Tax (percent)	-0.05**	-0.05**	-0.05**	-0.05**
	(0.02)	(0.02)	(0.02)	(0.02)
Population (10,000)	0.08***	0.08***	0.08***	0.08***
	(0.00)	(0.00)	(0.00)	(0.00)
Observations	15794	15794	15794	15794
R^2	0.90	0.91	0.91	0.91
Small chains				
Net Tax (percent)	0.04	0.04	0.04	0.04
	(0.03)	(0.03)	(0.03)	(0.03)
Population (10,000)	0.07***	0.07***	0.07***	0.07***
	(0.00)	(0.00)	(0.00)	(0.00)
Observations	15794	15794	15794	15794
R^2	0.81	0.81	0.81	0.82
Border FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Geography Control	No	Linear	Quadratic	Cubic

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The dependent variable, y_{mt}, represents the count of retail establishments of a certain type (e.g., large or small chains) in market m at time t, scaled by their respective national averages. Each column pertains to different functional form assumptions for the geographic controls. Column (1) has no geographic controls, Column (2) uses linear, Column (3) uses quadratic, and Column (4) cubic functions to control for geography. Standard errors are clustered by county.

Table 3.4: Retail Chain Establishment Counts and State Taxes with Market Scope Condition

All chains				
Net Tax (percent)	-0.07**	-0.07**	-0.07**	-0.07**
	(0.03)	(0.03)	(0.03)	(0.03)
Population (10,000)	0.08***	0.08***	0.08***	0.08***
	(0.01)	(0.01)	(0.01)	(0.01)
Observations	15719	15719	15719	15719
R^2	0.87	0.87	0.87	0.87
Big chains				
Net Tax (percent)	-0.07**	-0.07**	-0.07**	-0.07**
	(0.03)	(0.03)	(0.03)	(0.03)
Population (10,000)	0.08***	0.08***	0.08***	0.08***
	(0.00)	(0.00)	(0.00)	(0.00)
Observations	15719	15719	15719	15719
R^2	0.87	0.87	0.88	0.88
Small chains				
Net Tax (percent)	-0.01	-0.01	-0.00	0.00
	(0.09)	(0.09)	(0.09)	(0.09)
Population (10,000)	0.08***	0.08***	0.08***	0.08***
	(0.03)	(0.03)	(0.03)	(0.03)
Observations	15719	15719	15719	15719
R^2	0.50	0.50	0.51	0.51
Border FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Geography Control	No	Linear	Quadratic	Cubic

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The dependent variable, y_{mt}, represents the number of retail establishments of a given type (all, large, or small chains) in county m and year t, scaled by national averages. The sample is restricted to state-pair border markets where chains operate on both sides of the border. Each column corresponds to alternative specifications with different geographic control functions. All regressions include border and year fixed effects, with standard errors clustered by county.

Table 3.5: Placebo Test for Retail Chain Establishment Counts and State Taxes

	All chains	Big chains	Small chains
Net Tax (percent)	-0.007	0.012	-0.045
	(0.021)	(0.039)	(0.037)
Differenced Population (10,000)	0.082***	0.160***	-0.078***
	(0.012)	(0.025)	(0.022)
Observations	15362	15362	15362
R^2	0.985	0.982	0.966
County FE	Yes	Yes	Yes
State-Pair $\times$ Year FE	Yes	Yes	Yes

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The dependent variable, Δy_{mit}, is defined as the difference between a county's establishment count and the average establishment count in interior counties of the same state and year, scaled by national averages. Standard errors are two-way clustered by state and state-pair.

APPENDIX A

APPENDIX FOR CHAPTER 1

A.1 Example of Census Data Usage

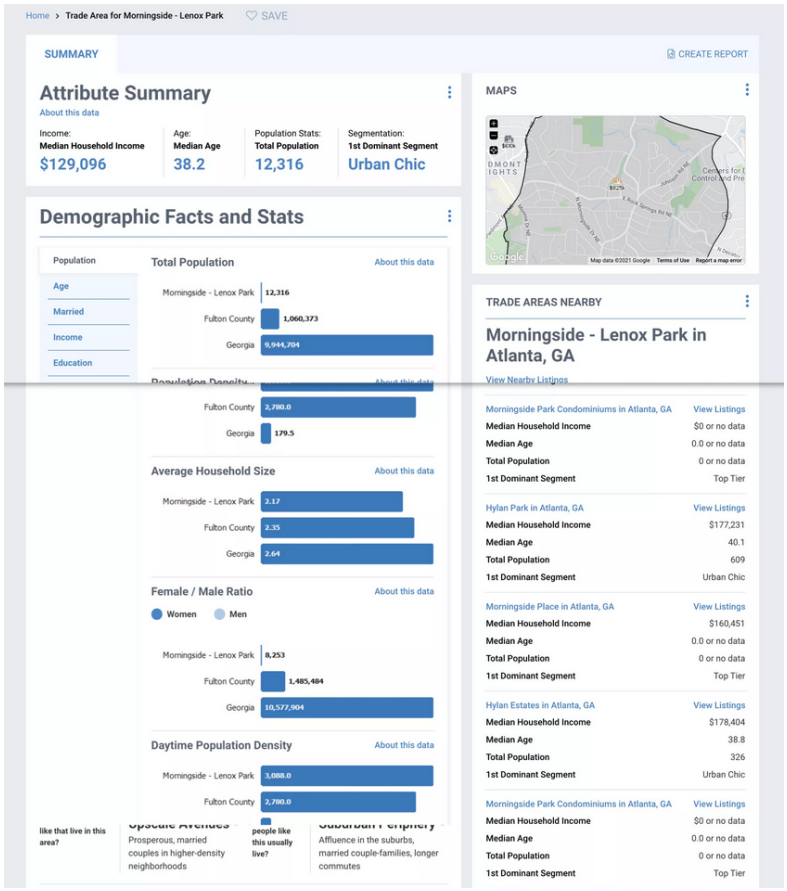


Figure A.1: Example Trade Area Analysis Report Tool For Commercial Real Estate Brokers

Notes: The figure provides an example of the demographic data used by commercial real estate brokers that market commercial properties to potential tenants. This report is an illustration of the data included in the Realtors Property Resource database provided by the National Association of Realtors. Source: `https://blog.narrpr.com/tips/commercial-trade-area-details/`.

A.2 Additional Descriptive Statistics

Table A.1: Number of Establishments by Industry

NAICS4		Counts
4411	Automobile Dealers	22829
4412	Other Motor Vehicle Dealers	7587
4413	Automotive Parts, Accessories, and Tire Stores	13435
4421	Furniture Stores	16749
4422	Home Furnishings Stores	17364
4431	Electronics and Appliance Stores	34200
4441	Building Material and Supplies Dealers	25635
4442	Lawn and Garden Equipment and Supplies Stores	4400
4451	Grocery Stores	69016
4452	Specialty Food Stores	41069
4453	Beer, Wine, and Liquor Stores	10447
4461	Health and Personal Care Stores	18928
4471	Gasoline Stations	17776
4481	Clothing Stores	74856
4482	Shoe Stores	12374
4483	Jewelry, Luggage, and Leather Goods Stores	20193
4511	Sporting Goods, Hobby, and Musical Instrument Stores	32169
4512	Book Stores and News Dealers	8596
4522	Department Stores	5285
4523	General Merchandise Stores, including Warehouse Clubs and Supercenters	12209
4531	Florists	10547
4532	Office Supplies, Stationery, and Gift Stores	33294
4533	Used Merchandise Stores	13536
4539	Other Miscellaneous Store Retailers	68230
4541	Electronic Shopping and Mail-Order Houses	7000
4542	Vending Machine Operators	3833
4543	Direct Selling Establishments	10525
7211	Traveler Accommodation	15868
7212	RV (Recreational Vehicle) Parks and Recreational Camps	2886
7213	Rooming and Boarding Houses, Dormitories, and Workers' Camps	460
7223	Special Food Services	24002
7224	Drinking Places (Alcoholic Beverages)	23310
7225	Restaurants and Other Eating Places	179184

Table A.2: Top 30 Chains by Number of Establishments in the Sample

Chain	Number of establishments
SUBWAY	1586
MCDONALDS	1323
DUNKIN DONUTS	1278
RITE AID	1105
CVS	803
SEVEN ELEVEN	718
BURGER KING	672
RADIO SHACK	633
MOBIL	587
ECKERD	566
SUNOCO	536
STARBUCKS	524
A & P	471
PAYLESS SHOE	430
WENDYS	428
FAMILY DOLLAR	410
STEWARTS	374
BASKINROBBINS	366
KFC	365
DOMINOS PIZZA	344
CARVEL ICE CREAM	342
DOLLAR GENERAL	333
PIZZA HUT	319
GNC	311
TIM HORTONS	308
DUANE READE	295
GETTY	290
WALGREENS	288
EXXON	274
HOLIDAY INN	267

Notes: To identify locations associate with a brand, I first standardize the trade style names by cleaning up the text strings. First, I remove numbers, special characters, and common indicators for a branch such as "STORE", "RESTAURANT", and "REST". Then I manually go through the top 100 brands in terms of the number of locations and use combination of keywords to identify variants of the brand name. KFC, for example, is sometimes recorded under "KENTUCKY FRIED CHICKEN", "K F C", or "KENTUCKY FRD CHICKEN". Finally, I combine locations with various alternative names under the same standardized brand name. This table lists the top 30 brands by number of establishments after cleaning the sample.

A.3 Demographic Variable Definition

Table A.3: Demographic Variable Definition

Variable	Definition
Population	Total population
%Kids (0-17)	Persons under 18 years old as a percentage of Total population
%Young (18-34)	Persons 18 to 34 years old as a percentage of Total population
%Middle (35-64)	Persons 35 to 64 years old as a percentage of Total population
%Old (65+)	Persons 65 years old and over as a percentage of Total population
%College degree	Percentage of Bachelor's degree or more among Persons 25 years and over
Unemployment rate	Percentage of Employed among Civilian Population In Labor Force 16 Years And Over
Median income	Median Household Income in 2010 dollars
Median house value	Median House Value for Specified Owner-Occupied Housing Units in 2010 dollars

A.4 Robustness and Sensitivity Analyses

A.4.1 Validation of NETS Data Using CBP data

The NETS data offers establishment-level data as a more accessible alternative to Census Bureau microdata, such as the Longitudinal Business Database (LBD), which requires special access and secure handling. While the LBD is based on administrative records and survey data,[1] NETS is constructed using various sources. Differences in data collection methods can lead to inconsistencies, underscoring the importance of validating NETS data. Since direct validation against the LBD is not possible due to its restricted access, I compare NETS with the publicly available County Business Patterns (CBP) data. Although CBP cannot distinguish between establishment entry and exit, it provides aggregate measures of establishment counts by industry and geography, offering a useful benchmark for assessing broader patterns.

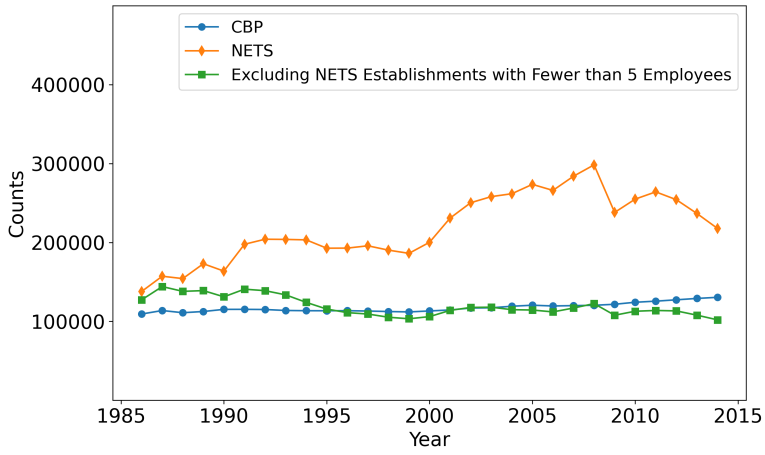
Previous research has examined the alignment between NETS and CBP data, addressing key discrepancies and identifying conditions under which the datasets show consistency. Barnatchez, Crane and Decker (2017) note that NETS includes many nonemployer establishments, which are not covered by CBP and tend to be very small. They also highlight discrepancies in sectors such as Agriculture, Mining, and Construction. Rossi-Hansberg, Sarte and Trachter (2021) further confirm that NETS data trends align closely with CBP after excluding problematic sectors identified by Barnatchez, Crane and Decker (2017) and applying restrictions, such as excluding establishments with fewer than five employees.

Following the approach in Barnatchez, Crane and Decker (2017), I validate my NETS sample for data the retail and restaurant sectors in New York state with corresponding CBP data.

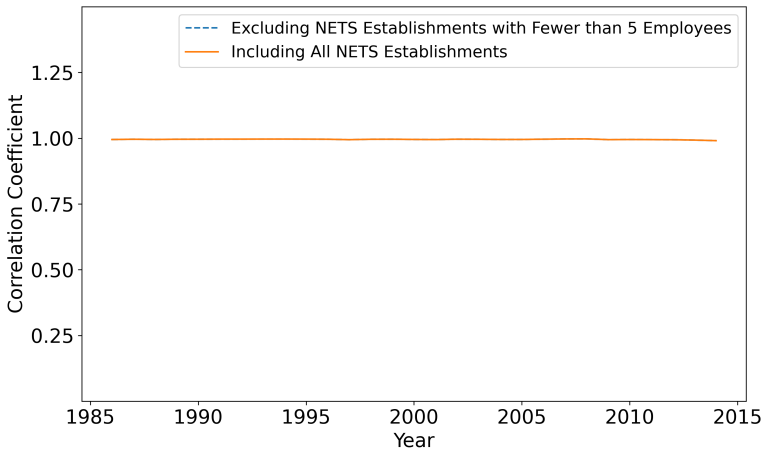
[1]The Longitudinal Business Database (LBD) is constructed using administrative records such as payroll tax filings, and survey data from the Economic Census. While these sources provide a comprehensive foundation for tracking establishments, certain factors can lead to timing discrepancies in the recognition of births and deaths. Small single-establishment firms may not report status changes immediately due to delays in payroll tax filings or EINs remaining active after closures, while multi-unit firms may be recorded as entering or exiting the market because of ownership changes or restructuring, even if their operations remain unchanged. The Economic Census, conducted every five years, serves as a benchmark for systematically updating firm status but also reveals artificial spikes in entry and exit that require adjustments to better reflect true business activity.

Panel A of Figure A.2 compares establishment counts over time across CBP data, unrestricted NETS data, and NETS data excluding establishments with fewer than five employees. Consistent with Barnatchez, Crane and Decker (2017) and Rossi-Hansberg, Sarte and Trachter (2021), the unrestricted NETS data include more establishments than the CBP data, while the restricted NETS data aligns more closely with CBP. Panel B of Figure A.2 reports the cross-sectional correlation across counties in establishment counts between NETS and CBP data. Both the restricted and unrestricted NETS datasets exhibit correlations above 0.99 in every year, consistent with the findings in Rossi-Hansberg, Sarte and Trachter (2021).

As the NETS dataset includes smaller establishments not fully captured in CBP, a key concern is whether differences in sampling coverage influence the results. As a robustness check, Table A.4 replicates Table 1.6 using only establishments with at least 5 employees. The results are consistent with the main findings: β_2 is still positive and statistically significant for small chains and independent establishments but remains insignificant for large chains.



Panel A: Establishment Count



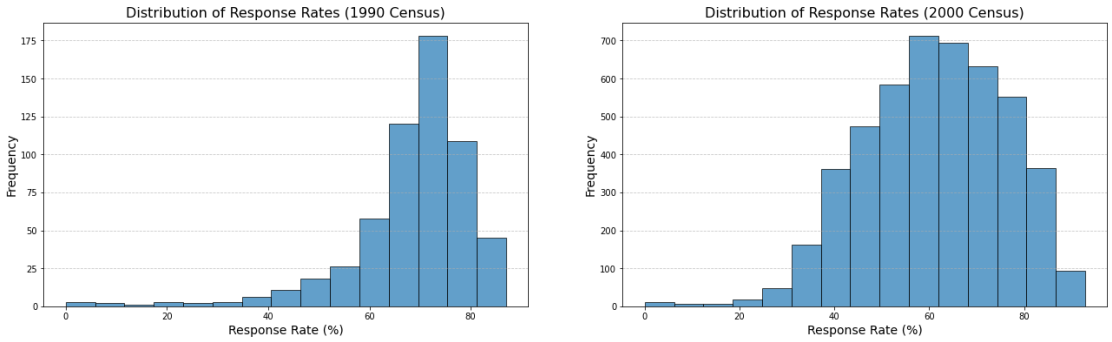
Panel B: County-level Correlation

Figure A.2: NETS and CBP Data Comparison

A.4.2 Census Response Rate

While the U.S. Census aims to count the entire population, achieving complete coverage remains challenging in practice. Declining response rates in recent decades have raised concerns about data quality, particularly in small-area statistics. This section examines whether variations in response rates systematically influence the estimated relationship between establishment failure rates and census data release.

Figure A.3 shows the distribution of response rates across geographic areas for the 1990 and 2000 censuses. The 1990 data only captures initial response rates before follow-up efforts, yet still achieves relatively high participation with a median of 71%. In contrast, the 2000 census reports a median response rate of 62%, reflecting a slight decline. These distributions are not directly comparable, given that the 1990 census only recorded data at a more aggregated place level.



1990 Census Response Rate (Place Level) 2000 Census Response Rate (Tract Level)

Figure A.3: Distribution of Census Response Rates

The 2000 census provides final response rates at the tract level, allowing for precise matching with the local markets in my sample. Table A.5 reports regression coefficients estimated separately for areas with low, medium, and high response rates. Even in the 2000 census, areas with very low response rates are uncommon, with the bottom decile having response rates lower than 43%. The β_2 estimates across response rate groups are consistent with the original results from Table 1.4, particularly for areas above the lowest response threshold.

Table A.4: Excluding Establishments with Fewer than 5 Employees

	Large Chain	Small Chain	Independent
	(1)	(2)	(3)
β_1	-0.022	-0.033*	0.015
	(0.013)	(0.017)	(0.008)
β_2	-0.002	0.008***	0.014***
	(0.001)	(0.002)	(0.001)
Adjusted β_2	-0.012	0.029	0.028
α_1	-0.022	-0.048	0.215***
	(0.022)	(0.029)	(0.013)
α_2	-0.020***	-0.022**	0.058***
	(0.005)	(0.008)	(0.004)
Observations	16525	12648	67740
R^2	0.003	0.002	0.008

Notes: This table replicates Table 1.6 using establishments with at least 5 employees. The adjusted β_2 coefficients are calculated by dividing the original estimates by the average failure rate for each firm type within the corresponding period. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table A.5: Subsample Analysis Based on Census Response Rates

	Low Response Rate ($< 43\%$)	Medium Response Rate (43%–73%)	High Response Rate ($> 73\%$)
	(1)	(2)	(3)
β_1	0.058***	0.015*	-0.031**
	(0.018)	(0.007)	(0.012)
β_2	0.011***	0.016***	0.018***
	(0.002)	(0.001)	(0.001)
α_1	0.164***	0.121***	0.055**
	(0.029)	(0.011)	(0.020)
α_2	0.010	0.010**	0.017**
	(0.009)	(0.003)	(0.005)
Observations	8528	51666	20863
R^2	0.004	0.008	0.010

Notes: This table reports results from the main specification Equation 1.5, estimated separately for areas classified by response rates: high response rate (top quartile, $> 73\%$), low response rate (bottom decile, $< 43\%$), and medium response rate (43%–73%). Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

A.4.3 Excluding New York City

A significant number of census tracts in the dataset are located in the densely populated New York City area. To ensure that the results are not disproportionately influenced by these tracts, I conduct a robustness check by splitting the sample into census tracts within and outside New York City. The results in Table A.6 demonstrate that the effects are consistent in magnitude across both subsamples, indicating that the findings are not driven by New York City-specific dynamics.

Table A.6: Excess Failure Rate and Distance to Census Data Release: NYC vs Others

	(1)	(2)
	NYC Census Tracts	Non-NYC Census Tracts
β_1	0.041***	-0.020***
	(0.007)	(0.007)
β_2	0.017***	0.016***
	(0.001)	(0.001)
α_1	0.142***	0.076***
	(0.012)	(0.012)
α_2	-0.017***	0.029***
	(0.004)	(0.003)
Observations	41743	58690
R^2	0.009	0.008

Notes: This table reports coefficient estimates from Equation 1.5 using the subsamples of census tracts within (outside) of the New York City. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

A.4.4 Alternative Specification

Table A.7: Excess Failure Rate and Distance to Census Data Release (Connected Segments)

	Full Sample	1986-1999 Cohorts	2000-2009 Cohorts
	(1)	(2)	(3)
β_1	-0.046***	0.016***	-0.108***
	(0.002)	(0.003)	(0.003)
β_2	0.015***	0.025***	0.004***
	(0.001)	(0.001)	(0.001)
Constant	0.017***	-0.009*	0.038***
	(0.002)	(0.004)	(0.003)
Observations	100,433	56,299	44,134
R^2	0.007	0.027	0.033

Notes: This table reports results using an alternative specification where the two segments connect at the break-point rather than having separate intercepts. The coefficients β_1 and β_2 represent slopes of the first and second phases, respectively. Standard errors clustered at the census-tract level. *** p<0.001, ** p<0.01, * p<0.05.

A.4.5 Alternative Clustering

In the main specification, standard errors are clustered at the census-tract level to account for temporal correlation across entry cohorts within the same market. For robustness, I consider three alternative clustering methods. First, standard errors are clustered at the county level to address potential spillovers across census tracts within the same county. Second, I implement two-way clustering at the census-tract and county-and-entry-year levels, which allows for spatial correlation within each cohort while constraining it to markets within the same county. This intermediate approach avoids the limitations of clustering solely by cohort, given the small number of entry-year clusters. Third, I cluster at the census-tract and entry-year levels directly to fully account for spatial correlation within cohorts. For this specification, I report p-values from wild bootstrap tests to correct for the small number of entry-year clusters. Table A.8 reports results using these alternative clustering approaches and the results remain statistically significant.

A.4.6 Alternative Benchmark Excluding the Entry Cohort

In the main specification, I calculate the baseline failure rate using all existing establishments within the same market. This approach captures the overall market conditions in a given year, accounting for time-varying factors that affect all businesses in the same location, and corresponds to a time fixed effect.

As a robustness check, I construct the baseline failure rate by excluding establishments in the entry cohort. This alternative ensures that the benchmark is unaffected by the performance of the cohort being analyzed. As shown in Table A.9, the results are qualitatively and quantitatively very similar to the main result from the original baseline specification.

Table A.8: Alternative Clustering Approaches

	(1)	(2)	(3)
β_1	0.006	0.006	0.006
	(0.010)	(0.032)	(0.125)
β_2	0.016***	0.016***	0.016***
	(0.001)	(0.002)	(0.004)
Wild bootstrap p-value			0.013
α_1	0.095***	0.095**	0.095
	(0.019)	(0.044)	(0.181)
α_2	0.010	0.010*	0.010
	(0.008)	(0.006)	(0.016)
Observations	100433	100433	100433
R^2	0.008	0.008	0.008

Notes: This table replicates Table 1.4 using alternative ways of clustering standard errors. In Column (1), standard errors are clustered at the county level. In Column (2), standard errors are two-way clustered at the census-tract and county-and-entry-year level. In Column (3), standard errors are two-way clustered at the census-tract and entry-year level. Given the relatively small number of entry-year clusters, p-values from wild bootstrap tests with two-way clustering are reported for β_2. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table A.9: Alternative Benchmark Excluding the Entry Cohort

	Full Sample	1986-1999 Cohorts	2000-2009 Cohorts
	(1)	(2)	(3)
β_1	0.007	0.053***	-0.051***
	(0.006)	(0.010)	(0.007)
β_2	0.019***	0.028***	0.008***
	(0.001)	(0.001)	(0.001)
α_1	0.118***	0.040**	0.169***
	(0.010)	(0.016)	(0.011)
α_2	0.011***	-0.008**	0.026***
	(0.003)	(0.004)	(0.004)
Observations	100433	56299	44134
R^2	0.008	0.028	0.041

Notes: This table summarizes my sensitivity analysis of the benchmark failure rate using other establishments (excluding the entry cohort of interest) in the same census tract to measure the market failure rate of a given calendar year. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

A.4.7 Alternative Benchmark Based on Establishments that Entered in the Recent 5 Years

My empirical strategy relies on the assumption that time-varying factors affect failure rates similarly across establishments, allowing the baseline failure rate to remove potential confounding effects. However, if younger and older establishments systematically differ in their vulnerability to market conditions, using all existing establishments as the benchmark may introduce bias. To assess the implication of this assumption, in Table A.10, I show that the results remain similar when using a benchmark based only on relatively young establishments—those that entered the same market within the past five years.

Table A.10: Alternative Benchmark Based on Establishments that Entered in the Recent 5 Years

	Full Sample	1986-1999 Cohorts	2000-2009 Cohorts
	(1)	(2)	(3)
β_1	0.015**	0.083***	-0.060***
	(0.007)	(0.011)	(0.009)
β_2	0.021***	0.034***	0.006***
	(0.001)	(0.001)	(0.001)
α_1	0.061***	0.046**	0.059***
	(0.011)	(0.019)	(0.013)
α_2	-0.107***	-0.123***	-0.100***
	(0.003)	(0.004)	(0.004)
Observations	100433	56299	44134
R^2	0.011	0.022	0.044

Notes: This table provides a summary of the sensitivity analysis of the benchmark failure rate using other establishments of similar age (within 5 years) in the same census tract to measure the market failure rate of a given calendar year. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

A.4.8 Alternative Thresholds for Demographic Changes

I define large changes in a demographic variable as those exceeding ±10%. As a sensitivity check, I also consider alternative cutoffs at ±15% and ±20%. Table A.11 shows that the coefficient estimates are similar in magnitude across these alternative cutoffs.

Table A.11: Sensitivity Check for Changes in Demographic Variables Cutoffs

Demographic Variable	Cutoff for Changes in Demographic Variable					
	<-20%	<-15%	<-10%	>10%	>15%	>20%
Population	0.001	0.001	0.001	0.003**	0.003	0.003
	(0.004)	(0.003)	(0.002)	(0.001)	(0.002)	(0.002)
%Kids(0-17)	0.000	-0.002	-0.005***	0.005***	0.004*	0.002
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
%Young (18-34)	0.008***	0.007***	0.007***	-0.007***	-0.006*	-0.006
	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.004)
%Middle (35-64)	-0.002	0.006	0.007*	0.007***	0.009***	0.008***
	(0.008)	(0.006)	(0.004)	(0.001)	(0.002)	(0.002)
%Old (65+)	0.001	-0.002	-0.003*	0.001	0.002	0.002
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)
%College	0.005**	0.005***	0.006***	-0.006***	-0.005***	-0.004***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Unemployment rate	-0.001	-0.001	-0.001	-0.000	0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Median income	-0.014***	-0.013***	-0.012***	0.008***	0.006***	0.007***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Median house value	0.002	0.004**	0.003**	-0.004***	-0.004***	-0.004***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

Notes: This table provides a summary of the sensitivity analysis for the incremental effects of demographic shifts on failure rate using different cutoffs to define a large change. The γ_2 coefficients and associated standard errors are estimated from the regression Equation 1.6 for each demographic variable with various cutoffs. Standard errors are clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

A.4.9 Alternative Thresholds for Chain Size

As a sensitivity check, I redefine large chains using alternative thresholds: more than 10 locations or more than 50 locations, instead of the baseline of 20 locations. The results remain similar across these alternative definitions.

Table A.12: Chain vs Independent using Alternative Definition of Large Chain (10 Locations)

Panel A: Full Sample

	Large Chain	Small Chain	Independent
	(1)	(2)	(3)
β_1	-0.020	-0.010	0.005
	(0.012)	(0.015)	(0.005)
β_2	0.002	0.011***	0.017***
	(0.001)	(0.002)	(0.001)
α_1	-0.010	-0.006	0.100***
	(0.021)	(0.025)	(0.009)
α_2	-0.026***	-0.021**	0.009***
	(0.005)	(0.007)	(0.003)
Observations	20726	16060	99081
R^2	0.003	0.003	0.008

Panel B: 1986-1999 Entry Cohorts

	Large Chain	Small Chain	Independent
	(1)	(2)	(3)
β_1	-0.031	-0.030	0.037***
	(0.016)	(0.017)	(0.009)
β_2	0.006***	0.015***	0.025***
	(0.002)	(0.002)	(0.001)
α_1	-0.032	-0.053	0.020
	(0.028)	(0.030)	(0.015)
α_2	-0.050***	-0.036***	-0.004
	(0.006)	(0.008)	(0.004)
Observations	12816	12876	55239
R^2	0.005	0.005	0.028

Panel C: 2000-2009 Entry Cohorts

	Large Chain	Small Chain	Independent
	(1)	(2)	(3)
β_1	-0.007	0.040	-0.040***
	(0.018)	(0.027)	(0.006)
β_2	-0.002	0.001	0.008***
	(0.002)	(0.003)	(0.001)
α_1	0.014	0.117*	0.151***
	(0.030)	(0.047)	(0.010)
α_2	0.006	0.006	0.018***
	(0.007)	(0.014)	(0.003)
Observations	7910	3184	43842
R^2	0.001	0.003	0.036

Notes: This table summarizes the tests for the relationship between excess failure rate and an entry cohort's distance to census data release separately for large chains (more than 10 outlets), small chains (between 2-10 outlets) and independent establishments. Column (1) reports results on large chains, Column (2) reports results on small chains, and Column (3) reports results on independent establishments. Panel A uses the full sample, Panel B uses the pre-2000 sample and Panel uses the post-2000 sample. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

Table A.13: Chain vs Independent using Alternative Definition of Large Chain (50 Locations)

Panel A: Full Sample

	Large Chain	Small Chain	Independent
	(1)	(2)	(3)
β_1	-0.011	-0.022	0.005
	(0.013)	(0.013)	(0.005)
β_2	0.000	0.011***	0.017***
	(0.001)	(0.001)	(0.001)
α_1	-0.007	-0.017	0.100***
	(0.023)	(0.023)	(0.009)
α_2	-0.017***	-0.033***	0.009***
	(0.005)	(0.006)	(0.003)
Observations	16910	20093	99081
R^2	0.001	0.003	0.008

Panel B: 1986-1999 Entry Cohorts

	Large Chain	Small Chain	Independent
	(1)	(2)	(3)
β_1	-0.021	-0.032*	0.037***
	(0.019)	(0.016)	(0.009)
β_2	0.004*	0.017***	0.025***
	(0.002)	(0.002)	(0.001)
α_1	-0.020	-0.051	0.020
	(0.032)	(0.027)	(0.015)
α_2	-0.041***	-0.051***	-0.004
	(0.007)	(0.007)	(0.004)
Observations	10153	15312	55239
R^2	0.003	0.006	0.028

Panel C: 2000-2009 Entry Cohorts

	Large Chain	Small Chain	Independent
	(1)	(2)	(3)
β_1	-0.001	-0.000	-0.040***
	(0.018)	(0.023)	(0.006)
β_2	-0.003	0.001	0.008***
	(0.002)	(0.002)	(0.001)
α_1	0.008	0.057	0.151***
	(0.031)	(0.039)	(0.010)
α_2	0.012	-0.004	0.018***
	(0.008)	(0.011)	(0.003)
Observations	6757	4781	43842
R^2	0.000	0.004	0.036

Notes: This table summarizes the tests for the relationship between excess failure rate and an entry cohort's distance to census data release separately for large chains (more than 50 outlets), small chains (between 2-50 outlets) and independent establishments. Column (1) reports results on large chains, Column (2) reports results on small chains, and Column (3) reports results on independent establishments. Panel A uses the full sample, Panel B uses the pre-2000 sample and Panel uses the post-2000 sample. Parentheses contain standard errors clustered at the census-tract level. Significance: *** p<0.001, ** p<0.01, * p<0.05.

APPENDIX B

APPENDIX FOR CHAPTER 2

B.1 List of Alternative Data Categories and Keywords

Category (1)	Definition (2)	Keywords (3)	Example from Analyst Report (4)
App Usage	These data track the number of downloads, the number of active users, and the time spent on mobile apps.	active user App Annie* AppData* Jiguang* QuestMobile* Sensor Tower* SimilarWeb* TalkingData*	"The UBS Evidence Lab analyzed App data that provides wait times for the 24 Shanghai Disneyland attractions *that* wait times associated with them. Our analysis covers the thirteen-week period from November 6, 2016 through January 29, 2017." [Issued by UBS on 04/06/17 for WALT DISNEY CO]

Category (1)	**Definition** (2)	**Keywords** (3)	**Example from Analyst Report** (4)
Employee	These data include job postings to evaluate corporate growth and strategy. These data also include management and employee sentiment extracted from statements in earnings conference calls and sites such as Glassdoor, among others.	earnings call transcript indeed.com job posting job trend mining of earnings calls online hiring text analytics transcriptlytics web analytics web mining web scraping	"We do track Apple's overall job postings and have seen a notable increase over the past 4-5 months in the number of engineering positions for Siri and ML, with a total of 205 specific mentions of 'Siri,' 'deep learning,' 'computer vision,' '*natural language processing (NLP)*' or '*machine learning*' in March job postings up from 64 mentions back in November." [Issued by GUGGENHEIM on 04/10/18 for APPLE]

Category (1)	**Definition** (2)	**Keywords** (3)	**Example from Analyst Report** (4)
Geospatial	These data include store-location data to evaluate growth and analyze competition, often overlayed with local income data and other demographic information to assess demand.	branch network model branch rationalization tool demographic analysis Foursquare* geospatial* market quality analysis store overlap within [...] drive within [...] miles	"We utilized the Alpha-Wise Branch Network Model to preview markets where we think JPM will likely invest. Seven factors drove our rankings, including wealth, income and population growth, competitive intensity, and small business opportunities. We calculate this as average deposits per branch. Our view is that areas with more deposits per branch are attractive for two reasons: 1) *it's* indicative of concentrated wealth and 2) it could suggest the area is underserved by low branch count." [Issued by MORGAN STANLEY on 02/21/18 for J.P.MORGAN]

Category (1)	Definition (2)	Keywords (3)	Example from Analyst Report (4)
Point of Sale	These data include merchant-level transaction data, product-level purchase data, and pricing data.	1010Data* airbnb + listing compared online prices discount tracker financial rate monitor First Data SpendTrend* footlocker.com footwear scrapes hotel tracker hotel tracking listing monitor MasterCard Advisors* Nielson* online price survey online pricing study our proprietary datasets Point-Of-Sale*	"CS Proprietary Home Pricing Tracker (median home price trends on an individual store basis across entire store base) shows similar trends in HD/LOW markets." [Issued by CREDIT SUISSE on 02/19/16 for HOME DEPOT]

Category (1)	Definition (2)	Keywords (3)	Example from Analyst Report (4)
		price comparisons	
		price intelligence	
		price monitoring	
		price observations	
		pricing monitor	
		pricing study	
		pricing tracker	
		property listing	
		SG2*	
		spend tracker	
		Standard Media Index*	
		SuperData*	
		vehicle listing	
		web analytics	
		web mining	
		web scraping	
		zillow	

Category (1)	Definition (2)	Keywords (3)	Example from Analyst Report (4)
Satellite Image	Satellite images can be used to track consumer traffic and gauge inventory levels as well as production activities at mines, construction sites, and plants, among other facilities.	Orbital Insight* parking lot fill rate parking lot traffic proprietary satellite data remote sensing Remote Sensing Metrics* RS Metrics* satellite analysis satellite image traffic analysis	"The satellite analysis points to a y/y Q1 parking lot fill rate change of +0.4%; however, the y/y change in fill rate became progressively worse over the quarter. Based on this data and the headwinds faced in Q1 from cash for appliances and weather, we feel comfortable with our Q1E comp of +1%." [Issued by PIPER SANDLER on 05/12/11 for HOME DEPOT]

Category (1)	Definition (2)	Keywords (3)	Example from Analyst Report (4)
Sentiment	These data include social-media feeds and news flow that help gauge consumer sentiment regarding products and services.	brand sentiment CMS Data* consumer sentiment customer rating customer review customer satisfaction rating customer satisfaction trend facebook analysis facebook data facebook like facebook likes facebook post facebook track facebook user guest sentiment instagram data instagram engagement instagram follower Internet World Stats* Investing Analytics* Medicare Plan Finder* Merchant Centric* net sentiment	"In this report, we introduce our proprietary consumer sentiment analysis, using information from Merchant Centric, a company that works with multi-location brands across consumer and service industries to *help* them manage and learn from guests' *online feedback*." ... "For our purposes, Merchant Centric tracks location-specific, user-generated reviews across multiple social media platforms. The reviews are user-generated and tied to a specific location, and then sourced from the following social media sites: Facebook, Google, Yelp, Trip Advisor, Superpages, and CitySearch. Our specific, filtered data set utilizes reviews on a representative sample of some 500 McDonald's locations across the country, as well as reviews for just over 2,400 Bojangles, Burger King, Del Taco, Dunkin' Donuts, Jack in the Box, Sonic, Taco Bell, and Wendy's units located within the same zip code. We have chosen to *exclude independent/local operators, and* focus on a sample of national and regional competitors." [Issued by JEFFERIES on 12/05/17 for MCDONALD'S CORP]

Category	Definition	Keywords	Example from Analyst Report
(1)	(2)	(3)	(4)
		NetBase*	
		online customer review	
		online review	
		Prosper Insights*	
		ratings on tripadvisor	
		review analytics	
		scoring released by cms	
		sentiment analysis	
		sentiment data	
		social media analysis	
		social media engagement	
		social media follower	
		star(s) rating	
		tracking on twitter	
		tripadvisor ratings	
		twitter analysis	
		twitter data	
		twitter purchase intent	
		twitter sentiment	
		web analytics	
		web mining	
		web scraping	
		yelp	

Category (1)	Definition (2)	Keywords (3)	Example from Analyst Report (4)
Web Traffic	These data track what users search for on the Internet and how frequently for how long users visit given websites.	baidu analysis baidu data baidu search data baidu search index baidu search volume ComScore* daily traffic google search analysis google search trend google trend google-searched iphone monitor iphone tracker Scrapehero* search interest search trend search volume smartphone tracker Thinknum* traffic analysis traffic monitor web hit activity web search	"The *AlphaWise* Smartphone Tracker has been developed by Morgan Stanley's *AlphaWise* using multi-country web search analysis using Google Trends. The approach accounts for different search criteria in multiple countries, as well as the differential between search and sales data seasonality, where appropriate. The in-sample period consists of 2008-2011 for Apple and 2010-2012 for Samsung Galaxy." [Issued by MORGAN STANLEY on 09/18/13 for APPLE]

Category (1)	Definition (2)	Keywords (3)	Example from Analyst Report (4)
Other	These include alternative data which do not fit cleanly into any of the above categories (e.g., data on senders and recipients of barrels loaded onto vessels).	BuildFax* climatology ClipperData* Collateral Verifications* Dodge* DrillingInfo* Dun & Bradstreet* Edmunds* Entgroup* EPFR* evidence lab macro Flightglobal* formulary coverage home improvement tracker IFI Claims* Innovata*	"The most effective way to measure the integrated advantage is our proprietary use of *ClipperData*, which can track the shipper ID of barrels loading onto vessels in the Gulf of Mexico. For this analysis, we do not isolate the loadings to export barrels, but look at Jones Act activity as well, as the ability to move its crude production anywhere is the proper reflection of the business model's advantage, in our view." [Issued by WOLFE RESEARCH on 09/28/18 for EXXON MOBIL CORP]

Category	Definition	Keywords	Example from Analyst Report
(1)	(2)	(3)	(4)
		lower end spending	
		m2m	
		macro-to-micro	
		network traffic lab	
		nowcast	
		One Click Retail*	
		Qokla*	
		OpenSignal*	
		Root Metrics*	
		Rystad*	
		STR Data*	
		wait time monitor	
		Wards Automotive*	
		weather monitor	

B.2 Variable Definition

Variables	Definition
Acc	We first calculate the proportional mean absolute forecast error, *PMAFE*, as the difference between the absolute forecast error of an analyst's annual earnings forecast and the average absolute forecast error across all analysts, scaled by the average absolute forecast error. When we compute the average absolute forecast error, we exclude all analysts who (also) report drawing from alternative data in their coverage of the corresponding firm in the respective forecast period. Since negative (positive) values of *PMAFE* indicate above (below) average performance, *Acc* is defined as $PMAFE \times (-1)$.
$I[Alternative\ Data]$	An indicator variable that equals one if the corresponding analyst issues *an earnings* forecast explicitly supported by alternative data and zero otherwise.
Forecast Age	The logarithm of one plus the number of calendar days between the forecast date and the corresponding earnings report date.
Analyst/Firm Experience	The number of years since the corresponding analyst first issued a forecast for the corresponding firm.
Analyst Experience	The number of years since the corresponding analyst first issued a forecast for any firm in the IBES database.
#Firms Covered	The logarithm of one plus the number of firms the corresponding analyst covers in the corresponding year.
Forecast Frequency	The logarithm of one plus the number of forecasts made by the corresponding analyst in the corresponding year.
Broker Size	The number of analysts working at the corresponding analyst's broker in the year of the forecast.
Trading Commissions	For each written report, we consider all trades involving the stock discussed in the report within three months after the report was issued and aggregate the total dollar value of the corresponding commissions paid to the corresponding broker.

Variables	Definition
$I[In-House\ Data\ Science\ Team]$	An indicator variable that equals one if the corresponding analyst works for a broker that has an in-house data-science team.
Σ *Colleagues Alternative Data*	The number of analysts that draw from alternative data and work for the same broker in the same city as the corresponding analyst in the year of the forecast.
Number of 8-Ks	The total number of Form 8-Ks filed during the previous annual forecast period.
Return Volatility	The standard deviation of daily stock returns during the previous annual forecast period.
Earnings Surprise	The most recent earnings surprise in the previous forecast period based on diluted earnings per share, excluding extraordinary items, and applying a seasonal random walk (Livnat and Mendenhall, 2006).
$I[Earnings\ Restatement]$	An indicator variable that equals one if a given firm has issued restatements in the past.
Discretionary Accruals	The most recent discretionary accruals based on the Modified Jones model matched to another from the same industry and year with the closest ROA (Kothari, Leone, and Wasley, 2005).
$I(Lack\ of\ Preferential\ Access\ to\ Management)$	An indicator variable that equals one if the corresponding firm did not participate in a conference hosted by the corresponding analyst's broker over the previous year.
Size	The market capitalization of the corresponding firm at the end of the previous fiscal year in billions.
M/B	The market value of equity divided by the book value of equity at the end of the previous fiscal year.
Momentum	Buy-and-hold return on the corresponding stock over the previous six months.
$I(Category = X)$	An indicator variable that equals one if a given analyst explicitly references the use of alternative data from alternative-data category X.
Σ *Categories*	The number of distinct alternative-data categories an analyst references in her report.

Variables	Definition
$I(Source = Proprietary\ Data)$	An indicator variable that equals one if a given analyst explicitly references the use of proprietary alternative data.
$I(Source = Accessible\ Data)$	An indicator variable that equals one *if* a given analyst explicitly references the use of accessible alternative data.

APPENDIX C

APPENDIX FOR CHAPTER 3

C.1 Tax Measure Construction

Our analysis draws on the Panel Database on Incentives and Taxes (PDIT), which standardizes the fiscal environment facing a hypothetical new facility. The database's objective is to place taxes and incentives on a comparable economic scale by projecting a twenty-year stream of payments and benefits and converting them to present-value terms. This procedure ensures that both recurring tax liabilities and one-time incentive disbursements are expressed in real 2015 dollars and aggregated in a manner consistent with firm-level decision making. The construction proceeds in four steps:

- **Scenario setup.** The simulation begins by positing the establishment of a representative facility (e.g., a retail branch) in a given state and city in a specific year. The financial profile of this facility (profits, property holdings, and employment) is benchmarked to national averages for the relevant industry.
- **Twenty-year projection.** The model applies the state and local tax code and incentive policies prevailing in the start year, assuming these remain fixed over the next twenty years. This yields a stream of annual tax liabilities (corporate, property, and other state taxes) and offsetting incentive payments (such as job-training subsidies or investment credits).
- **Present-value aggregation.** To combine one-time incentives with ongoing obligations, all flows are first expressed in 2015 dollars using the GDP deflator. Ongoing tax liabilities are inflation-adjusted but not discounted, as they represent contemporaneous fiscal obligations that recur annually. Incentive streams, in contrast, are discounted to present value using a real annual rate of 12 percent, consistent with corporate hurdle rates that place greater weight on near-term cash flows.[1] This distinction ensures that the measure

[1] Appendix A provides computational details and confirms that only incentive components are discounted, while tax components are merely deflated to 2015 dollars.

reflects the net present value of fiscal flows as perceived by firms, accounting for both inflation and time preference.

Normalization of fiscal flows. Both tax and incentive streams are discounted at a real annual rate of 12 percent and then divided by the present value of the facility's value-added, which is discounted using the same rate. This parallel treatment ensures that the effective rate tax_{st} is scale-neutral and comparable across sectors and states. Because identical discounting is applied to the numerator and denominator, the reported value can be interpreted as the real, present-value-adjusted share of value-added absorbed by net taxes.

Although the twenty-year simulation formally discounts both taxes and incentives at the same real annual rate of 12 percent, the discounting of tax liabilities cancels out in the computation of the effective rate. Because both the numerator (the present value of tax payments) and the denominator (the present value of value-added) are discounted at the same rate, the resulting ratio is equivalent to the statutory tax burden expressed in real 2015 dollars. Incentive streams, by contrast, are typically one-time or time-limited payments, so discounting materially affects their present value. This clarification ensures that all components are treated symmetrically in the underlying simulation, while the reported tax rates remain directly interpretable as real statutory rates.

An important feature of the PDIT is its distinction between export-base industries (such as manufacturing and technology) and non-export-base industries (such as retail). Retail trade is classified in the latter category, which is critical for interpretation because most states concentrate their largest incentive programs on export-base sectors. Consequently, the average incentive rate for retail facilities is markedly lower: in 2015, non-export-base industries received incentives averaging only 0.16 percent of value-added, compared with 1.42 percent for export-base firms.

To summarize, the tax measure provides a coherent and internally consistent framework for interpreting state tax and incentive policies as a common, economically meaningful rate expressed in real 2015 dollars.

C.2 Retail Sectors in NETS Data

We present a tabulation of the establishments across different retail sectors (i.e., SIC 4-digit classification) in Table C.1. This table shows that each retail sector is well-represented. The retail sectors that have disproportionately more establishments include grocery stores and women's clothing. Among the sectors with the lowest establishment counts are newsstands and luggage stores.

Table C.1: Tabulation of Establishment Counts (2014) Across SIC 4-Digit Classification

SIC		Counts	SIC		Counts
5211	Lumber and Other Building Materials Dealers	51262	5713	Floor Covering Stores	48942
5231	Paint, Glass, and Wallpaper Stores	30364	5714	Drapery, Curtain, and Upholstery Stores	4612
5251	Hardware Stores	26637	5719	Miscellaneous home furnishings Stores	24682
5261	Retail Nurseries, Lawn and Garden Supply Stores	20818	5722	Household Appliance Stores	14207
5271	Mobile Home Dealers	5783	5731	Radio, Television, and Consumer Electronics Stores	22959
5311	Department Stores	27484	5734	Computer and Computer Software Stores	37180
5331	Variety Stores	37335	5735	Record and Prerecorded Tape Stores	14647
5399	Miscellaneous General Merchandise Stores	11771	5736	Musical Instrument Stores	10862
5411	Grocery Stores	200385	5912	Drug Stores and Proprietary Stores	62802
5421	Meat and Fish (Seafood) Markets, Including Freezer Provisioners	12402	5921	Liquor Stores	38473
5431	Fruit and Vegetable Markets	8578	5932	Used Merchandise Stores	62829
5441	Candy, Nut, and Confectionery Stores	9692	5941	Sporting Goods Stores and Bicycle Shops	63001
5451	Dairy Products Stores	4023	5942	Book Stores	19902
5461	Retail Bakeries	52622	5943	Stationery Stores	10023
5499	Miscellaneous Food Stores	40350	5944	Jewelry Stores	48372
5511	Motor Vehicle Dealers (New and Used)	60057	5945	Hobby, Toy, and Game Shops	34934
5521	Motor Vehicle Dealers (Used Only)	48300	5946	Camera and Photographic Supply Stores	2212
5531	Auto and Home Supply Stores	73681	5947	Gift, Novelty, and Souvenir Shops	83989
5541	Gasoline Service Stations	73420	5948	Luggage and Leather Goods Stores	1803
5551	Boat Dealers	9104	5949	Sewing, Needlework, and Piece Goods Stores	16186
5561	Recreational Vehicle Dealers	3836	5961	Catalog and Mail-Order Houses	21569
5571	Motorcycle Dealers	10448	5962	Automatic Merchandising Machine Operators	19079
5599	Automotive Dealers, Not Elsewhere Classified	21407	5963	Direct Selling Establishments	50045
5611	Men's and Boys' Clothing and Accessory Stores	17597	5983	Fuel Oil Dealers	4838
5621	Women's Clothing Stores	91767	5984	Liquefied Petroleum Gas (Bottled Gas) Dealers	6837
5632	Women's Accessory and Specialty Stores	13009	5989	Fuel Dealers, Not Elsewhere Classified	1512
5641	Children's and Infants' Wear Stores	9048	5992	Florists	35803
5651	Family Clothing Stores	23799	5993	Tobacco Stores and Stands	9336
5661	Shoe Stores	29260	5994	News Dealers and Newsstands	1808
5699	Miscellaneous Apparel and Accessory Stores	32398	5995	Optical Goods Stores	23012
5712	Furniture Stores	53926	5999	Miscellaneous Retail Stores, Not Elsewhere Classified	250989

C.3 Alternative Population Measures

Additional empirical analysis we consider is focused on the county population, which is an important proxy for market size. Because the decennial census is conducted every ten years, raw population counts can be noisier in intercensal years (Chi, 2024), potentially introducing measurement errors that could attenuate or distort the estimated tax effects. To gauge the sensitivity of our results to population measurement, we performed a robustness check using an alternative construction of population based on US Census intercensal population estimates.[2] In particular, we re-estimate the baseline border-RD specification in (3.1), where the main difference is that $\widetilde{\text{pop}}_{mt}$ is a smoothed measure of county population in market m and year t. All other variables and fixed effects remain unchanged relative to the baseline.

Table C.2 reports estimates using $\widetilde{\text{pop}}_{mt}$. The results are virtually indistinguishable from the baseline: the tax coefficient for all chains remains small and statistically insignificant across all geography controls; the coefficient for big chains on net tax remains negative and statistically significant; and finally, the tax estimates for small chains are close to zero and insignificant. Finally, estimates on population appear to be stable across specifications. Overall, the robustness check reinforces our main conclusions. That is, tax changes primarily affect large-chain entry while leaving small-chain entry largely unaffected.

C.4 Model Extensions

This section extends the theoretical model to examine the robustness and generality of its key insights. We consider alternative frameworks that relax core assumptions, including the functional form of tax policies (Section C.4.1), as well as asymmetries in sunk entry costs (Section C.4.2). These extensions highlight the broader applicability of the model and explore how its predictions adapt under varying market structures and policy environments.

[2] Please see `https://www.nber.org/research/data/us-census-intercensal-population-estim` for more details.

C.4.1 Proportional Tax Policy

Under a proportional tax regime, the total entry costs are scaled to $\kappa\tau$. When $\tau < 1$, this represents a subsidy, while $\tau > 1$ corresponds to a tax. The equilibrium entry probabilities become the following.

$$\tilde{\sigma}_1 = \sigma_1^* - \frac{(\tau-1)\kappa\rho}{1-\rho(1-\rho)}, \qquad \tilde{\sigma}_2 = \sigma_2^* - \frac{(\tau-1)\kappa(1-\rho)}{1-\rho(1-\rho)}.$$

The absolute responsiveness of each firm to proportional taxation is

$$|\tilde{\sigma}_1 - \sigma_1^*| = \frac{|\tau-1|\kappa\rho}{1-\rho(1-\rho)}, \qquad |\tilde{\sigma}_2 - \sigma_2^*| = \frac{|\tau-1|\kappa(1-\rho)}{1-\rho(1-\rho)}.$$

Taking their difference yields

$$|\tilde{\sigma}_1 - \sigma_1^*| - |\tilde{\sigma}_2 - \sigma_2^*| = \frac{|\tau-1|\kappa}{1-\rho(1-\rho)}(2\rho-1).$$

Proposition 3. *Under proportional taxation: (i) when $\rho > 0.5$, firm 1 is more responsive than firm 2; when $\rho < 0.5$, firm 2 is more responsive; and at $\rho = 0.5$ the two firms respond equally. (ii) The magnitude of the difference in responsiveness grows linearly in the entry cost κ, independent of whether $\tau > 1$ (tax) or $\tau < 1$ (subsidy).*

Proof. From the equilibrium shifts above,

$$\left|\tilde{\sigma}_1 - \sigma_1^*\right| = \frac{|\tau-1|\kappa\rho}{1-\rho(1-\rho)}, \qquad \left|\tilde{\sigma}_2 - \sigma_2^*\right| = \frac{|\tau-1|\kappa(1-\rho)}{1-\rho(1-\rho)}.$$

Hence

$$\left|\tilde{\sigma}_1 - \sigma_1^*\right| - \left|\tilde{\sigma}_2 - \sigma_2^*\right| = \frac{|\tau-1|\kappa}{1-\rho(1-\rho)}(2\rho-1),$$

which is positive when $\rho > 0.5$, zero when $\rho = 0.5$, and negative when $\rho < 0.5$, proving part (i). Because κ multiplies the entire expression linearly and appears with $|\tau - 1|$, the magnitude of the difference scales one for one with κ regardless of the sign of $(\tau - 1)$, proving part (ii). □

Proportional taxes change entry probabilities in proportion to the size weights, with ρ for firm 1 and $1-\rho$ for firm 2. Consequently, the firm with the larger profit weight (i.e., the stronger competitor) is always more sensitive in absolute terms to a given percentage change in entry cost. When $\rho > 0.5$, this corresponds to firm 1; when $\rho < 0.5$, it corresponds to firm 2.

C.4.2 Asymmetric Entry Costs

We now generalize the proportional-tax framework to allow for asymmetric entry costs across firms. Suppose firm 1 faces entry cost κ_1 and firm 2 faces κ_2, with $\kappa_1 \neq \kappa_2$ in general. Under a proportional tax policy, the total entry cost for firm i is $\kappa_i \tau$. Solving the two best-response equations yields equilibrium shifts that capture both own- and cross-firm effects:

$$\tilde{\sigma}_1 - \sigma_1^* = \frac{(\tau - 1)\big[(1-\rho)\kappa_2 - \kappa_1\big]}{1 - \rho(1-\rho)}, \qquad \tilde{\sigma}_2 - \sigma_2^* = \frac{(\tau - 1)\big[\rho\kappa_1 - \kappa_2\big]}{1 - \rho(1-\rho)}.$$

The absolute responsiveness of each firm to proportional taxation is therefore

$$\big|\tilde{\sigma}_1 - \sigma_1^*\big| = \frac{|\tau - 1|\big|(1-\rho)\kappa_2 - \kappa_1\big|}{1 - \rho(1-\rho)}, \qquad \big|\tilde{\sigma}_2 - \sigma_2^*\big| = \frac{|\tau - 1|\big|\rho\kappa_1 - \kappa_2\big|}{1 - \rho(1-\rho)}.$$

Subtracting these expressions gives

$$\big|\tilde{\sigma}_1 - \sigma_1^*\big| - \big|\tilde{\sigma}_2 - \sigma_2^*\big| = \frac{|\tau - 1|}{1 - \rho(1-\rho)}\Big(\big|(1-\rho)\kappa_2 - \kappa_1\big| - \big|\rho\kappa_1 - \kappa_2\big|\Big).$$

Proposition 4. *Under asymmetric entry costs, responsiveness to proportional taxation depends jointly on relative firm strength (ρ) and cost parameters (κ_1, κ_2):*

1. *When $\kappa_1 = \kappa_2$, the symmetric case of Appendix C.4.1 is recovered, with the stronger firm (for $\rho > 0.5$) being more responsive.*
2. *When $\kappa_1 \neq \kappa_2$, the sign of the responsiveness gap depends on whether $\big|(1-\rho)\kappa_2 - \kappa_1\big|$ exceeds $\big|\rho\kappa_1 - \kappa_2\big|$.*

Proof. The expression above directly implies parts (i) and (ii). For $\kappa_1 = \kappa_2$, the terms in absolute value simplify and the sign is governed by $2\rho - 1$, reproducing the symmetric benchmark. When costs differ, the relative magnitudes of κ_1 and κ_2 jointly with ρ determine which firm's entry probability adjusts more to proportional taxation. □

This extended specification preserves the main qualitative insight, whereby under typical conditions, the *stronger* firm remains more tax-responsive. To see this, let $\theta = \kappa_1/\kappa_2$ denote the cost ratio.

For the case $\rho > 0.5$ (firm 1 is stronger), the stronger firm is more responsive whenever

$$\theta \;\geq\; \frac{2-\rho}{1+\rho}.$$

For example, if $\rho = 0.6$, this threshold equals 0.875, meaning that as long as the stronger firm's cost is at least 87.5% of its rival's, it remains more responsive to proportional taxes. Only when the stronger firm's cost advantage is markedly distinct (i.e., κ_1/κ_2 falls below this bound) does the weaker firm exhibit the greater absolute adjustment.

In summary, these comparative statics highlight that the qualitative conclusion of the symmetric model (i.e., greater responsiveness among stronger firms) broadly holds. When both size and cost advantages align (e.g., $\rho > 0.5$ and $\kappa_1 \geq \kappa_2$), proportional taxes magnify disparities in responsiveness; when they oppose each other, the weaker firm's reaction can dominate. Hence, asymmetries in entry costs can either amplify or dampen distortions from uniform proportional taxation, suggesting that differentiated or targeted policies may be warranted when cost heterogeneity is material.

Table C.2: Retail Chain Establishment Counts and State Taxes with Alternative Population Measure

All chains				
Net Tax (percent)	-0.02	-0.02	-0.02	-0.02
	(0.02)	(0.02)	(0.02)	(0.02)
Population (10,000)	0.08***	0.08***	0.08***	0.08***
	(0.00)	(0.00)	(0.00)	(0.00)
Observations	16044	16044	16044	16044
R^2	0.92	0.92	0.92	0.92
Big chains				
Net Tax (percent)	-0.05**	-0.05**	-0.05**	-0.05**
	(0.02)	(0.02)	(0.02)	(0.02)
Population (10,000)	0.08***	0.08***	0.08***	0.08***
	(0.00)	(0.00)	(0.00)	(0.00)
Observations	16044	16044	16044	16044
R^2	0.91	0.91	0.91	0.91
Small chains				
Net Tax (percent)	0.04	0.04	0.04	0.04
	(0.03)	(0.03)	(0.03)	(0.03)
Population (10,000)	0.07***	0.07***	0.07***	0.07***
	(0.00)	(0.00)	(0.00)	(0.00)
Observations	16044	16044	16044	16044
R^2	0.81	0.81	0.81	0.81
Border FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Geography Control	No	Linear	Quadratic	Cubic

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The dependent variable, y_{mt}, represents the count of retail establishments of a certain type (e.g., large or small chains) in market m at time t, scaled by their respective national averages. Each column pertains to different functional form assumptions for the geographic controls. Column (1) has no geographic controls, Column (2) uses linear, Column (3) uses quadratic, and Column (4) cubic functions to control for geography. Standard errors are clustered by county.

BIBLIOGRAPHY

Adair, Bill. 1991. "Census helps firms target consumers." *Tampa Bay Times*.

Addoum, Jawad M, David T Ng, and Ariel Ortiz-Bobea. 2020. "Temperature shocks and establishment sales." *The Review of Financial Studies*, 33(3): 1331–1366.

Adelino, Manuel, Song Ma, and David Robinson. 2017. "Firm age, investment opportunities, and job creation." *The Journal of Finance*, 72(3): 999–1038.

Agrawal, Ajay, Joshua S Gans, and Avi Goldfarb. 2019. "Artificial intelligence: the ambiguous labor market impact of automating prediction." *Journal of Economic Perspectives*, 33(2): 31–50.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

Arcidiacono, Peter, Patrick Bayer, Jason R Blevins, and Paul B Ellickson. 2016. "Estimation of dynamic discrete choice models in continuous time with an application to retail competition." *The Review of Economic Studies*, 83(3): 889–931.

Armas, Genaro C. 2001. "Census Data Big for Businesses." *Associated Press*.

Asker, John, Allan Collard-Wexler, and Jan De Loecker. 2014. "Dynamic inputs and resource (mis) allocation." *Journal of Political Economy*, 122(5): 1013–1063.

Asplund, Marcus, and Volker Nocke. 2006. "Firm turnover in imperfectly competitive markets." *The Review of Economic Studies*, 73(2): 295–327.

Atkinsa, Rachel MB, Pablo Hernández-Lagosb, Cristian Jara-Figueroad, and Robert Seamansc. 2023. "JUE Insight: What is the Impact of Opportunity Zones on Job Postings?" *Journal of Urban Economics*, forthcoming.

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen. 2017. "Concentrating on the Fall of the Labor Share." *American Economic Review*, 107(5): 180–185.

Baker, Scott R, Nicholas Bloom, and Stephen J Terry. 2024. "Using disasters to estimate the impact of uncertainty." *Review of Economic Studies*, 91(2): 720–747.

Baker, Scott R, Nicholas Bloom, and Steven J Davis. 2016. "Measuring economic policy uncertainty." *The Quarterly Journal of Economics*, 131(4): 1593–1636.

Barber, Brad M, Xing Huang, Philippe Jorion, Terrance Odean, and Christoph Schwarz. 2023. "A (sub) penny for your thoughts: Tracking retail investor activity in TAQ." *The Journal of Finance*. Forthcoming.

Barnatchez, Keith, Leland Dod Crane, and Ryan Decker. 2017. "An assessment of the national establishment time series (nets) database." FEDS working paper.

Bartik, Timothy J. 2017. "A new panel database on business incentives for economic development offered by state and local governments in the United States." *Working paper*.

Basker, Emek. 2005. "Job creation or destruction? Labor market effects of Wal-Mart expansion." *Review of Economics and Statistics*, 87(1): 174–183.

Basker, Emek, Chris Vickers, and Nicolas L Ziebarth. 2018. "Competition, productivity, and survival of grocery stores in the Great Depression." *International Journal of Industrial Organization*, 59: 282–315.

Begenau, Juliane, Maryam Farboodi, and Laura Veldkamp. 2018. "Big data in finance and the growth of large firms." *Journal of Monetary Economics*, 97: 71–87.

Ben-David, Itzhak, Justin Birru, and Andrea Rossi. 2019. "Industry familiarity and trading: Evidence from the personal portfolios of industry insiders." *Journal of Financial Economics*, 132(1): 49–75.

Bengtzen, Martin. 2017. "Private investor meetings in public firms: the case for increasing transparency." *Fordham J. Corp. & Fin. L.*, 22(1): 33–132.

Bertrand, Marianne, and Francis Kramarz. 2002. "Does entry regulation hinder job creation? Evidence from the French retail industry." *The Quarterly Journal of Economics*, 117(4): 1369–1413.

Bhattacharya, Nilabh, Hemang Desai, and Kumar Venkataraman. 2013. "Does earnings quality affect information asymmetry? Evidence from trading costs." *Contemporary Accounting Research*, 30(2): 482–516.

Bloom, Nicholas. 2009. "The impact of uncertainty shocks." *Econometrica*, 77(3): 623–685.

Bloom, Nick, Stephen Bond, and John Van Reenen. 2007. "Uncertainty and investment dynamics." *The Review of Economic Studies*, 74(2): 391–415.

Bradley, Daniel. 2023. "Financial Analysts." In *Handbook of Financial Decision Making*. 356–374. Edward Elgar Publishing.

Bradley, Daniel, Sinan Gokkaya, and Xi Liu. 2017. "Before an analyst becomes an analyst: does industry experience matter?" *The Journal of Finance*, 72(2): 751–792.

Bresnahan, Timothy F, and Peter C Reiss. 1991. "Entry and competition in concentrated markets." *Journal of Political Economy*, 99(5): 977–1009.

Brown, Lawrence D, Andrew C Call, Michael B Clement, and Nathan Y Sharp. 2015. "Inside the "black box" of sell-side financial analysts." *Journal of Accounting Research*, 53(1): 1–47.

Busse, Jeffrey A, Lei Tong, Qing Tong, and Zhipeng Zhang. 2019. "Trading regularity and fund performance." *The Review of Financial Studies*, 32(1): 374–422.

Cao, Sean, Wei Jiang, Jun-Lis Wang, and Baolian Yang. 2023. "From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses." National Bureau of Economic Research w31131.

Chi, Feng. 2024. "Information Waves and Firm Investment." *Working paper*.

Chi, Feng, Byoung-Hyoun Hwang, and Yaping Zheng. 2024. "The use and usefulness of big data in finance: Evidence from financial analysts." *Management Science*.

Chodorow-Reich, Gabriel, Owen Zidar, and Eric Zwick. 2024. "Lessons from the Biggest Business Tax Cut in US History." *Journal of Economic Perspectives*, 38(3): 61–88.

Clement, Michael B. 1999. "Analyst forecast accuracy: Do ability, resources, and portfolio complexity matter?" *Journal of Accounting and Economics*, 27(3): 285–303.

Coleman, Brennan, Kenneth Merkley, and Joseph Pacelli. 2022. "Human versus machine: A comparison of robo-analyst and traditional research analyst investment recommendations." *The Accounting Review*, 97(5): 221–244.

Collard-Wexler, Allan. 2013. "Demand fluctuations in the ready-mix concrete industry." *Econometrica*, 81(3): 1003–1037.

Corinth, Kevin, and Naomi Feldman. 2024. "Are Opportunity Zones an Effective Place-Based Policy?" *Journal of Economic Perspectives*, 38(3): 113–136.

Couture, Victor, and Jessie Handbury. 2020. "Urban revival in America." *Journal of Urban Economics*, 119: 103267.

Craft, Erik D. 1998. "The value of weather information services for nineteenth-century Great Lakes shipping." *American Economic Review*, 1059–1076.

Criscuolo, Chiara, Ralf Martin, Henry G Overman, and John Van Reenen. 2019. "Some causal effects of an industrial policy." *American Economic Review*, 109(1): 48–85.

Cropper, Matthew, Jerome N. McKibben, David A. Swanson, and Jeff Tayman. 2012. "Vendor Accuracy Study 2010 Estimates versus Census 2010." Esri.

Currie, Janet, Stefano DellaVigna, Enrico Moretti, and Vikram Pathania. 2010. "The effect of fast food restaurants on obesity and weight gain." *American Economic Journal: Economic Policy*, 2(3): 32–63.

Da Rin, Marco, Marina Di Giacomo, and Alessandro Sembenelli. 2011. "Entrepreneurship, firm entry, and the taxation of corporate income: Evidence from Europe." *Journal of Public Economics*, 95(9-10): 1048–1066.

Davis, Steven J, John Haltiwanger, Ron Jarmin, Javier Miranda, Christopher Foote, and Eva Nagypal. 2006. "Volatility and dispersion in business growth rates: Publicly traded versus privately held firms." *NBER macroeconomics annual*, 21: 107–179.

Decker, Ryan A, John Haltiwanger, Ron S Jarmin, and Javier Miranda. 2016. "Declining business dynamism: What we know and the way forward." *American Economic Review*, 106(5): 203–207.

Dell, Melissa. 2010. "The persistent effects of Peru's mining mita." *Econometrica*, 78(6): 1863–1903.

De Loecker, Jan, and Chad Syverson. 2021. "An industrial organization perspective on productivity." In *Handbook of Industrial Organization*. Vol. 4, 141–223. Elsevier.

Dessaint, Olivier, Thierry Foucault, and Laurent Frésard. 2023. "Does Alternative Data Improve Financial Forecasting? The horizon effect." *The Journal of Finance*. Forthcoming.

Devereux, Michael P, Rachel Griffith, and Helen Simpson. 2007. "Firm location decisions, regional grants and agglomeration externalities." *Journal of Public Economics*, 91(3-4): 413–435.

Dixit, Avinash K, and Robert S Pindyck. 1994. *Investment under uncertainty.* Princeton university press.

Donnelly, Frank. 2019. *Exploring the US Census: Your Guide to America's Data.* SAGE Publications.

Dunne, Timothy, Mark J Roberts, and Larry Samuelson. 1989. "The growth and failure of US manufacturing plants." *The Quarterly Journal of Economics*, 104(4): 671–698.

Ericson, Richard, and Ariel Pakes. 1995. "Markov-perfect industry dynamics: A framework for empirical work." *The Review of Economic Studies*, 62(1): 53–82.

Fajgelbaum, Pablo D, Eduardo Morales, Juan Carlos Suárez Serrato, and Owen Zidar. 2019. "State taxes and spatial misallocation." *The Review of Economic Studies*, 86(1): 333–376.

Fang, Limin, and Nathan Yang. 2024. "Measuring Deterrence Motives in Dynamic Oligopoly Games." *Management Science*, 70(6): 3527–3565.

Farboodi, Maryam, Roxana Mihet, Thomas Philippon, and Laura Veldkamp. 2019. "Big data and firm dynamics." *AEA Papers and Proceedings*, 109: 38–42.

Farhi, Paul. 1990. "For Business, Census is a Marketing Data Motherlode." *The Washington Post*. (accessed April 19, 2022).

Fazio, Catherine, Jorge Guzman, and Scott Stern. 2019. "The impact of state-level R&D tax credits on the quantity and quality of entrepreneurship." National Bureau of Economic Research.

Fort, Teresa C, John Haltiwanger, Ron S Jarmin, and Javier Miranda. 2013. "How firms respond to business cycles: The role of firm age and firm size." *IMF Economic Review*, 61(3): 520–559.

Frank, Morgan R, David Autor, James E Bessen, Erik Brynjolfsson, Manuel Cebrian, David J Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, et al. 2019. "Toward understanding the impact of artificial intelligence on labor." *Proceedings of the National Academy of Sciences*, 116(14): 6531–6539.

Froot, Kenneth A, Namho Kang, Gideon Ozik, and Ronnie Sadka. 2017. "What do measures of real-time corporate sales say about earnings surprises and post-announcement returns?" *Journal of Financial Economics*, 125(1): 143–162.

Gao, Meng, and Jiekun Huang. 2020. "Informing the market: The effect of modern information technologies on information production." *The Review of Financial Studies*, 33(4): 1367–1411.

Gerken, William C, and Marcus O Painter. 2023. "The value of differing points of view: Evidence from financial analysts' geographic diversity." *The Review of Financial Studies*, 36(2): 409–449.

Geurts, Karen, and Johannes Van Biesebroeck. 2016. "Firm creation and post-entry dynamics of de novo entrants." *International Journal of Industrial Organization*, 49: 59–104.

Giroud, Xavier, and Joshua Rauh. 2019. "State taxation and the reallocation of business activity: Evidence from establishment-level data." *Journal of Political Economy*, 127(3): 1262–1316.

Green, T Clifton, Russell Jame, Stanimir Markov, and Musa Subasi. 2014. "Access to management and the informativeness of analyst research." *Journal of Financial Economics*, 114(2): 239–255.

Grennan, Jillian, and Roni Michaely. 2020. "Artificial intelligence and high-skilled work: Evidence from analysts." Working paper, Duke University.

Gutiérrez, Germán, and Thomas Philippon. 2017. "Declining Competition and Investment in the US." National Bureau of Economic Research.

Haltiwanger, John, Ron Jarmin, and Cornell John Krizan. 2010. "Mom-and-pop meet big-box: complements or substitutes?" *Journal of Urban Economics*, 67(1): 116–134.

Hanner, Daniel, Daniel Hosken, Luke M Olson, and Loren K Smith. 2015. "Dynamics in a mature industry: Entry, exit, and growth of big-box grocery retailers." *Journal of Economics & Management Strategy*, 24(1): 22–46.

Harford, Jarrad, Feng Jiang, Rong Wang, and Fei Xie. 2019. "Analyst career concerns, effort allocation, and firms' information environment." *The Review of Financial Studies*, 32(6): 2179–2224.

Harju, Jarkko, Aliisa Koivisto, and Tuomas Matikka. 2022. "The effects of corporate taxes on small firms." *Journal of Public Economics*, 212: 104704.

Hoberg, Gerard, and S Katie Moon. 2017. "Offshore activities and financial vs. operational hedging." *Journal of Financial Economics*, 125(2): 217–244.

Hoberg, Gerard, and S Katie Moon. 2019. "The offshoring return premium." *Management Science*, 65(6): 2876–2899.

Hollenbeck, Brett. 2017. "The economic advantages of chain organization." *The RAND Journal of Economics*, 48(4): 1103–1135.

Hollenbeck, Brett, and Renato Zaterka Giroldo. 2022. "Winning big: Scale and success in retail entrepreneurship." *Marketing Science*, 41(2): 271–293.

Holmes, Thomas J. 1998. "The effect of state policies on the location of manufacturing: Evidence from state borders." *Journal of Political Economy*, 106(4): 667–705.

Hopenhayn, Hugo A. 1992. "Entry, exit, and firm dynamics in long run equilibrium." *Econometrica*, 1127–1150.

Hsieh, Chang-Tai, and Esteban Rossi-Hansberg. 2022. "The industrial Revolution in Services." *Journal of Political Economy Macroeconomics*, forthcoming.

Huang, Jian. 2018. "The customer knows best: The investment value of consumer opinions." *Journal of Financial Economics*, 128(1): 164–182.

Hu, Gang, Kye M Jo, Y. Aaron Wang, and Jin Xie. 2018. "Institutional trading and Abel Noser data." *Journal of Corporate Finance*, 52: 143–167.

Hughes-Cromwick, Ellen, and Julia Coronado. 2019. "The value of US government data to US business decisions." *Journal of Economic Perspectives*, 33(1): 131–146.

Hwang, Byoung-Hyoun, Jose M Liberti, and Jason Sturgess. 2019. "Information sharing and spillovers: Evidence from financial analysts." *Management Science*, 65(8): 3624–3636.

Igami, Mitsuru, and Nathan Yang. 2016. "Unobserved heterogeneity in dynamic games: Cannibalization and preemptive entry of hamburger chains in Canada." *Quantitative Economics*, 7(2): 483–521.

Jame, Russell. 2018. "Liquidity provision and the cross section of hedge fund returns." *Management Science*, 64(7): 3288–3312.

Jame, Russell, R Johnston, Stanimir Markov, and Michael C Wolfe. 2016. "The value of crowdsourced earnings forecasts." *Journal of Accounting Research*, 54(4): 1077–1110.

Jarmin, Ronald S, Shawn D Klimek, and Javier Miranda. 2009. "The role of retail chains: National, regional and industry results." In *Producer dynamics: New evidence from micro data*. 237–262. University of Chicago Press.

Javorcik, Beata S, and Yue Li. 2013. "Do the biggest aisles serve a brighter future? Global retail chains and their implications for Romania." *Journal of International Economics*, 90(2): 348–363.

Jones, Charles I, and Christopher Tonetti. 2020. "Nonrivalry and the Economics of Data." *American Economic Review*, 110(9): 2819–58.

Jovanovic, Boyan. 1982. "Selection and the Evolution of Industry." *Econometrica*, 649–670.

Julio, Brandon, and Youngsuk Yook. 2012. "Political uncertainty and corporate investment cycles." *The Journal of Finance*, 67(1): 45–83.

Kang, Jung Koo, L Stice-Lawrence, and Y T F Wong. 2021. "The firm next door: Using satellite images to study local information advantage." *Journal of Accounting Research*, 59(2): 713–750.

Katona, Zsolt, Marcus Painter, Panos N Patatoukas, and Jean Zeng. 2023. "On the capital market consequences of alternative data: Evidence from outer space." *Journal of Financial and Quantitative Analysis*. Forthcoming.

Kellogg, Ryan. 2014. "The effect of uncertainty on investment: Evidence from Texas oil drilling." *American Economic Review*, 104(6): 1698–1734.

Kim, Hyunseob, and Howard Kung. 2017. "The asset redeployability channel: How uncertainty affects corporate investment." *The Review of Financial Studies*, 30(1): 245–280.

Kline, Patrick, and Enrico Moretti. 2014. "Local economic development, agglomeration economies, and the big push: 100 years of evidence from the Tennessee Valley Authority." *The Quarterly journal of economics*, 129(1): 275–331.

Kolko, Jed. 2012. "Broadband and local growth." *Journal of Urban Economics*, 71(1): 100–113.

Kothari, Suraj P, Andrew J Leone, and Charles E Wasley. 2005. "Performance matched discretionary accrual measures." *Journal of Accounting and Economics*, 39(1): 163–197.

Levine, David I, Michael W Toffel, and Matthew S Johnson. 2012. "Randomized government safety inspections reduce worker injuries with no detectable job loss." *Science*, 336(6083): 907–911.

Lin, Mengwei. 2024. "Local Policies and Firm Location: Measuring the "Unmeasurable"." *Working Paper*.

Livnat, Joshua, and Richard R Mendenhall. 2006. "Comparing the post–earnings announcement drift for surprises calculated from analyst and time series forecasts." *Journal of Accounting Research*, 44(1): 177–205.

Logan, John R, Zengwang Xu, and Brian J Stults. 2014. "Interpolating US decennial census tract data from as early as 1970 to 2010: A longitudinal tract database." *The Professional Geographer*, 66(3): 412–420.

Maican, Florin G, and Matilda Orth. 2018. "Entry regulations, welfare, and determinants of market structure." *International Economic Review*, 59(2): 727–756.

McMahon, Dinny, and Kathy Chu. 2012. "Clampdown in China on corporate sleuthing." *ht tps://www.wsj.com*.

Merkley, Kenneth, Roni Michaely, and Joseph Pacelli. 2017. "Does the scope of the sell-side analyst industry matter? An examination of bias, accuracy, and information content of analyst reports." *The Journal of Finance*, 72(3): 1285–1334.

Mian, Atif, and Amir Sufi. 2014. "What explains the 2007–2009 drop in employment?" *Econometrica*, 82(6): 2197–2223.

Mukherjee, Abhiroop, George Panayotov, and Janghoon Shon. 2021. "Eye in the sky: Private satellites and government macro data." *Journal of Financial Economics*, 141(1): 234–254.

Nagaraj, Abhishek. 2022. "The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry." *Management Science*, 68(1): 564–582.

National Research Council. 1995. *Modernizing the U.S. Census.* The National Academies Press.

National Research Council. 2015. *Realizing the potential of the American Community Survey: Challenges, tradeoffs, and opportunities.* National Academies Press.

Nelson, Phillip. 1970. "Information and consumer behavior." *Journal of Political Economy*, 78(2): 311–329.

Neumark, David, Brandon Wall, and Junfu Zhang. 2011. "Do small businesses create more jobs? New evidence for the United States from the National Establishment Time Series." *The Review of Economics and Statistics*, 93(1): 16–29.

New York State Office of the State Comptroller. 2004. "Assessing the Empire Zones Program Reforms Needed to Improve Program Evaluation and Effectiveness." *https://web.osc.state.ny.us/osdc/empirezone3-2005.pdf*.

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

Puckett, Andy, and Xuemin Yan. 2011. "The interim trading skills of institutional investors." *The Journal of Finance*, 66(2): 601–633.

Pugsley, Benjamin Wild, and Ayşegül Şahin. 2019. "Grown-up business cycles." *The Review of Financial Studies*, 32(3): 1102–1147.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter. 2021. "Diverging trends in national and local concentration." *NBER Macroeconomics Annual*, 35(1): 115–150.

Schuetz, Jenny, Jed Kolko, and Rachel Meltzer. 2012. "Are poor neighborhoods "retail deserts"?" *Regional Science and Urban Economics*, 42(1-2): 269–285.

Seasholes, Mark S, and Ning Zhu. 2010. "Individual investors and local bias." *The Journal of Finance*, 65(5): 1987–2010.

Sedláček, Petr, and Vincent Sterk. 2017. "The growth potential of startups over the business cycle." *American Economic Review*, 107(10): 3182–3210.

Sedlácek, Petr, and Vincent Sterk. 2019. "Reviving american entrepreneurship? tax reform and business dynamism." *Journal of Monetary Economics*, 105: 94–108.

Serrato, Juan Carlos Suárez, and Philippe Wingender. 2016. "Estimating local fiscal multipliers." National Bureau of Economic Research.

Shambaugh, Jay, Ryan Nunn, Audrey Breitwieser, and Patrick Liu. 2018. "The state of competition and dynamism: Facts about concentration, start-ups, and related policies." *Hamilton Project. Washington, DC: Brookings Institution*.

Shoag, Daniel, and Stan Veuger. 2018. "Shops and the city: Evidence on local externalities and local government policy from big-box bankruptcies." *Review of Economics and Statistics*, 100(3): 440–453.

Slattery, Cailin. 2022. "Bidding for Firms: Subsidy Competition in the U.S." *Working paper*.

Slattery, Cailin, and Owen Zidar. 2020. "Evaluating state and local business incentives." *Journal of Economic Perspectives*, 34(2): 90–118.

Soltes, Eugene. 2014. "Private interaction between firm management and sell-side analysts." *Journal of Accounting Research*, 52(1): 245–272.

Stulz, René M. 2007. "Hedge funds: Past, present, and future." *Journal of Economic Perspectives*, 21(2): 175–194.

Suárez Serrato, Juan Carlos, and Owen Zidar. 2016. "Who benefits from state corporate tax cuts? A local labor markets approach with heterogeneous firms." *American Economic Review*, 106(9): 2582–2624.

Suzuki, Junichi. 2013. "Land use regulation as a barrier to entry: evidence from the Texas lodging industry." *International Economic Review*, 54(2): 495–523.

Thau, Barbara. 2014. "How Big Data Helps Chains Like Starbucks Pick Store Locations – An (Unsung) Key To Retail Success." *Forbes*. (accessed April 19, 2022).

The Council of Economic Advisers. 2000. "The Uses of Census Data: An Analytical Review." The Council of Economic Advisers.

Thomadsen, Raphael. 2005. "The effect of ownership structure on prices in geographically differentiated industries." *RAND Journal of Economics*, 908–929.

Tsui, Jennifer, Jana A Hirsch, Felicia J Bayer, James W Quinn, Jesse Cahill, David Siscovick, and Gina S Lovasi. 2020. "Patterns in geographic access to health care facilities across neighborhoods in the United States based on data from the national establishment time-series between 2000 and 2014." *JAMA Network Open*, 3(5): e205105–e205105.

US Government Accountability Office. 2009. "Formula Grants: Funding for the Largest Federal Assistance Programs is Based on Census-Related Data and Other Factors. (GAO Publication No. 10-263)." *Washington, D.C.: U.S. Government Printing Office*.

Wigglesworth, Robin. 2016. "Investors mine big data for cutting-edge strategies." *https://www.ft.com*.

Wilson, W Mark. 2008. "An empirical analysis of the decline in the information content of earnings following restatements." *The Accounting Review*, 83(2): 519–548.

Zhang, Jonathan Z, Chun-Wei Chang, and Scott A Neslin. 2022. "How physical stores enhance customer value: The importance of product inspection depth." *Journal of Marketing*, 86(2): 166–185.

Zhu, Chen. 2019. "Big data as a governance mechanism." *The Review of Financial Studies*, 32(5): 2021–2061.