# Information Waves and Firm Investment*

Feng Chi†

February 2025

**Abstract**

This paper measures the impact of information quality on the success of firms' investment decisions using the U.S. census as an empirical context. The decennial census provides comprehensive demographic information widely used by businesses, especially for retail establishment entry decisions. However, as local demographics change over the decade, this snapshot becomes outdated, potentially leading to worse investment outcomes. Using establishment-level data, I find that outdated census information increases failure rates. A 10-year gap between censuses raises the failure rate of new establishments by 16 percentage points, which represents a 32% increase over the baseline failure rate. This effect is stronger for geographic areas that experience large changes in demographics, for industries that rely on precise information in small trade areas, and for smaller firms that lack alternative sources of demographic information. Finally, I show that these patterns are not driven by business cycle dynamics and government spending.

*Keywords:* Economics of Data; Firm Investment; Industry Dynamics

*I thank Panle Jia Barwick, Nicholas Bloom, Ryan Chahrour, Allan Collard-Wexler, Steven Davis, Olivier Dessaint (discussant), Erica Groshen, Brett Hollenbeck, Byoung-Hyoun Hwang, Seohee Kim, Ben Leyden, Shanjun Li, Xuelin Li (discussant), Filippo Mezzanotti (discussant), William Mullins, Justin Murfin, David Ng, Robert Seamans, Sophia Xue (discussant), Nathan Yang, and Xiaoyun Yu (discussant) as well as participants at the 3rd Annual Holden Conference in Finance and Real Estate, 2024 SFS Cavalcade North America, the 2023 Stanford SITE Conference on the Macroeconomics of Uncertainty and Volatility, the 33rd Annual Conference on Financial Economics and Accounting, the 5th Future of Financial Information Conference and Cornell University for helpful comments and suggestions.

†Cornell University. Email: fc354@cornell.edu.

# 1 Introduction

Information is a critical input to the firm decision-making process, yet empirically assessing its impact on investment 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.

The U.S. census is a key source of demographic information for various business decisions, such as site selection, product offering, advertising and inventory management.[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 breadth in capturing consumer demographics (Farhi, 1990). Even when firms do not directly obtain census data, they frequently rely on intermediaries such as data analytics firms, market research companies, and commercial real estate agents that integrate census demographic insights into their decision-making tools, often without realizing that the original source of this data is the census.[2] Small businesses, too, engage with census data in ways they may not fully recognize, such as when preparing loan applications that require demographic insights to demonstrate market demand and growth potential.

However, the quality of information derived from census data declines over time, as the U.S. census is only conducted once every 10 years. Despite its initial accuracy, the data gradually becomes outdated, particularly in small geographic areas sensitive to demographic changes from migration. As local demographics shift over time, the census data no longer captures these changes, creating a widening gap between captured and actual market realities, until the next census brings in fresh information.[3]

[1] See report by the National Research Council panel for examples of business uses of census data in a variety of industries, including retail and restaurant, banks and other financial institutions, media and advertising, insurance, utilities, health care, and others. A specific case study of retail expansion involves predicting potential revenue for each market. "Some data were from business sources, but census data provided an essential component for analysis" (National Research Council, 1995).

[2] Armas (2001) and Thau (2014) provide examples of intermediaries, such as market research company Claritas and location intelligence company Esri, that integrate census demographic data into software systems that allows for data manipulation and market analysis, mapping, and the preparation of reports to assist businesses in site selection decisions.

[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."

These periodic fluctuations in the quality of demographic information are what I term *information waves*. For firms that make investment decisions using census data, these shifts create predictable changes in information quality. As census data becomes increasingly outdated, firms continue to rely on stale information, leading to investment decisions that are progressively misaligned with actual market conditions. I hypothesize that this misalignment results in worse investment outcomes over time.[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 locations 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 financial 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. Distinguishing between these groups allows me to isolate the impact of information quality on new entrants from the broader economic conditions affecting all establishments. Moreover, geographic variations provide 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 during the 30 years from 1985 to 2014, I find that excess failure rates across entry-year cohorts follow the wave pattern that I hypothesized. The failure rate increases by a statistically significant rate of 1.6 percentage points each intercensual year during my sample period. A 10-year gap between two decennial census would result in a 16 percentage 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 percentange points, decade-old data could raise a firm's baseline likelihood of failure by 32%.

[4]While positive surprises can help reduce failure rates, as demonstrated in the heterogeneity tests in section 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.

As a placebo test, I create hypothetical census schedules by randomly assigning a year in each decade as the census year. Out of the 10,000 randomly created schedules, only 193 of them produce estimates greater than or equal to the original estimate using the true census schedule. Therefore, the probability of observing by random chance an effect as large as the original estimate is only 1.9%.

Next, I investigate how the effects of outdated census data on establishment failure rates vary across different contexts. The goal is to show that the impact of information quality on investment outcomes is more pronounced in situations where census demographic data is more critical. These cross-sectional tests help verify that the main effect operates through the information channel as hypothesized.

First, I find stronger effects in geographic areas that experience sizable demographic shifts between two censuses. It is in these areas that the disparity between static census data and evolving market conditions becomes most pronounced. In terms of magnitude, the absolute value of the moderating effect ranges from 3 percentage points to 12 percentage points, equivalent to 19% to 75% of the main effect. Conversely, in areas with stable demographics, the effect of outdated census data on failure rates is negligible.

Second, the effects are most pronounced in industries offering highly localized products and services, notably restaurants and grocery stores. These industries tend to operate within small geographic areas due to their need for proximity to their customer base. They rely heavily on up-to-date local demographic data, as small geographic areas can experience rapid demographic changes. In contrast, industries that cater to large trade areas, such as motor vehicle dealers and furniture stores, show diminished sensitivity to variations in census data quality. Adding to this narrative, non-store retailers—those without a physical storefront, such as Electronic Shopping and Mail-Order Houses—appear to be largely unaffected by changing census data quality.

Third, when evaluating the influence of census data among various firm sizes, I find that smaller firms, including independent establishments and small chains, are most affected. In contrast, large chains have been less impacted by the declining reliability of census data, especially in the latter half of the sample period. With the advent of the digitization era, large chains increasingly have access to emerging technologies and alternative data sources, which are often inaccessible to their smaller counterparts.

I conduct additional analyses to rule out alternative explanations. First, I demonstrate that my findings are not driven by business cycles. In the cross-industry analysis, both restaurant and grocery industries exhibit similar responses to outdated census data, despite their differences in cyclicality. Moreover, the impact of outdated census data on failure rate is comparable across local areas that are more/less impacted by the early 1990s and early

2000s recessions. Second, I examine whether the results could be explained by government programs that allocate funding based on census data. For such policies to influence my findings, 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 eligibility criteria of these programs, I show that the results are robust for 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. The analysis finds 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 diminishing the likelihood of entry delays. These findings suggest that the impact of information staleness on exit outcomes is unlikely to be driven by firms' strategic timing of entry based on variations in information quality over time.

This paper makes several contributions to the literature. First, it advances our understanding of how information quality affects firm decision-making. While existing research has explored how firms endogenously acquire and use data (Brynjolfsson and McElheran, 2016; Wu, Hitt and Lou, 2020; Leisten, 2021) or leveraged exogenous changes in data availability as natural experiments (Craft, 1998; Gao and Huang, 2020; Nagaraj, 2022), this study takes a novel approach. By examining the predetermined, recurring release of U.S. census data, it captures the continuous variation in information quality over time. This approach provides a more realistic view of how firms navigate investment decisions under gradually changing conditions, offering nuanced insights that can inform better policy design.

Second, to the best of my knowledge, this is the first study to examine the economic value of the U.S. Census in the context of firm investment. Unlike many other countries, the U.S. Census is unique in its widespread availability and accessibility, creating an ecosystem of census data users (Donnelly, 2019). Although the primary purpose of the census is to determine political representation and allocate federal spending (Serrato and Wingender, 2016), this study highlights its impact on economic efficiency through shaping firms' investment decisions.

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 work emphasizes the role of information quality in shaping these outcomes. The findings align with Collard-Wexler (2013), who demonstrate the significant impact of

uncertainty reduction on firm turnover. The results have broader implications for market structure dynamics, given that census data is open access and non-rival (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; Farboodi et al., 2019).

The rest of the paper is organized as follows. The institutional background about the U.S. decennial census is provided in section 2, and a description of data is in section 3. My empirical strategy is discussed in section 4, followed by the main findings about the effects of outdated census data on establishment failure in section 5. Furthermore, I highlight some variations in failure rate patterns across geographical, industry, and firm dimensions in section 6. Finally, I discuss potential alternative explanations in section 7 and the potential impact of delayed entry in section 8. A concluding summary are provided in section 9.

# 2 Institutional Background

## 2.1 The Decennial Census

The decennial censuses from 1970 to 2000 use fairly consistent collection methods and variables throughout this time period. 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 of the population. 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, tabulated data are released on a flow basis after the census was conducted.[5] Basic population counts are published in March the next year, followed by more detailed population, housing and socioeconomic characteristics. Most of the data that appeal to broad business interests, including Summary File 1 and Summary File 3, is made available within 18 months after the initial release.

## 2.2 Decreased Accuracy In-between Censuses

Although highly accurate at the time of collection, the reliability of census data declines over the course of a decade. Small areas, in particular, can experience rapid population and

[5]Since 1930 "Census Day" has been April 1 in the first year of each decade (ending in zero). Census attempts to capture a snapshot of 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).

Table 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

economic changes during intercensual years.

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

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

## 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.

[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).

### 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, often through a network of affiliated organizations such as state executive departments, chambers of commerce, councils of governments, and university research departments. Businesses have been significant users of this data.[8] The South Carolina Census State Data Center estimated that 35 percent of the annual requests received for census data were from businesses (National Research Council, 1995).

The 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. These libraries often have business centers staffed by knowledgeable librarians who assist entrepreneurs in finding crucial demographic data.[9] Such services are important to small businesses because, as part of the loan application process, financial institutions often require a detailed business plan, which includes a comprehensive market research section.[10] This section involves identifying target customers and estimating market size, both of which rely heavily on demographic data.[11]

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

[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 examples 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.

beneficial for small businesses.[12]

### 2.3.2 Indirect Influence through Intermediaries

While direct access to census data became more prevalent with the advent of the internet, many businesses still relied 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 databases, enhancing 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 its 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 A1 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 is often directly obtained from census sources. Furthermore, this demographic data influences market prices for commercial real estate, including rents, thereby impacting the operational costs incurred by businesses for their locations. Mispricing driven by stale data can lead to businesses paying higher rents than justified and facing unexpected competition, ultimately squeezing profit margins and increasing failure rates.

# 3 Data and Summary Statistics

## 3.1 Establishment Data

### 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. Dunn & Bradstreet collects information from various sources, in-

[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.

cluding 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 establishments.

Focusing on the retail and restaurant industries offers two key advantages. Firms in these industries are organized by geographic locations and largely 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 provides 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 section D.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.

### 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

[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 A6 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.

[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).

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).

I measure investment outcomes using the failure rate within the first 5 years of entry, to address potential right censoring for the latter years of my sample. 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 assessment of the market characteristics when the initial entry decision was made.

My sample period does not permit me to identify exits with certainty for establishments that enter after 2009, or time of entry for establishments in the 1985 sample. Therefore, I measure 5-year failure rates for establishments that entered during 1986-2009, leaving me with 24 entry-year cohorts to study.

### 3.1.3 Market Definition

Establishments in the retail and restaurant industries have relatively small trade areas,[17] which is precisely why the location choice is essential to success in this business. I define markets at the census-tract level and map an establishment's geographic coordinates to corresponding census tract using the 2010 TIGER/Line Shapefiles, so that the physical boundary of the markets remains fixed throughout my sample and can be easily linked to census demographic information.[18]

[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).

[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).

### 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 A1 reports the number of establishments across NAICS 4-digit sub-industries during my sample period. Table 2 summarizes key market dynamics, including the number of establishments, new entrants, and failure rates at the market-year level.

Table 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.

## 3.2 Demographic Data

### 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 decennial censuses across geographic areas.

### 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

[19] To verify that my results are not driven by New York City specific dynamics, I report in Table A6 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.

[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.

Consumer Price Index research series using current methods (CPI-U-RS) from St Louis Fed. Table A3 provides details on the construction of each demographic variable.

Table 3 tabulates the distribution of changes in demographic variables between two decennial censuses. Census-tract level demographics exhibit relatively large variation, highlighting that small geographic areas can experience large unexpected shifts in demographics over the course of a decade.

Table 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 A3 provides details on the construction of each demographic variable.

# 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 are faced with varying levels of census data quality when they make their entry decisions. Establishments entering early in the decade benefit from recent census data that accurately reflects current market conditions. In contrast, those entering later in the decade rely on the same data, which has become progressively outdated, putting 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.

While information quality plays an important role, other time-varying factors can also significantly influence the failure rate of an entry cohort. For instance, if a store opens one year before the financial crisis and failed, its failure is more likely due to the macroeconomic condition than a poor location choice.

To account for the influence of time-varying factors, I use the failure rate of all existing establishments from a local market in a calendar year as the baseline failure rate for that year.[21] Assuming time-varying factors impact failure rates similarly across establishments in the same market, irrespective of their entry timing, the baseline failure rate helps eliminate potential confounding effects.[22] This approach allows 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:

$$\Delta f_{imt} = f_{imt} - F_{mt} = \frac{\text{Exit}_{imt}}{B_{imt}} - \frac{\sum_i \text{Exit}_{imt}}{\sum_i B_{imt}}, \tag{1}$$

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:

$$\text{Excess failure rate}_{im} = \sum_{t=i+1}^{i+5} \frac{\Delta f_{imt} * B_{imt}}{B_{im}}. \tag{2}$$

A potential challenge to this approach is whether external shocks systematically impact new entrants differently from incumbents. A potential challenge to this approach is whether external shocks systematically impact new entrants differently from incumbents. In section 7, I examine whether factors such as business cycles or government funding tied to census data could explain the observed pattern of failures. The results do not suggest that these alternative channels drive the differential failure rates.

[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 section 7.1

# 5 Main Results

## 5.1 Failure Rate Pattern

My central hypothesis is that failure rates across entry-year cohorts follows a wave pattern, characterized by worse investment decisions over time as the quality of information derived from census data deteriorates and a reversal when the new census data are released.

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

$$\text{Excess failure rate}_{im} = \sum_{i=1986}^{2009} \beta_i \times I(\text{Entry-Year} = i) + \varepsilon_{im}, \tag{3}$$

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 visualizes the coefficient estimates from Equation 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 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. When fresh data from the 2000 census arrives, the failure rate does not fall right away, indicating that firms now benefit more from detailed socioeconomic data released later. 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

[23] For a detailed schedule, please refer to Table 1.

[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).

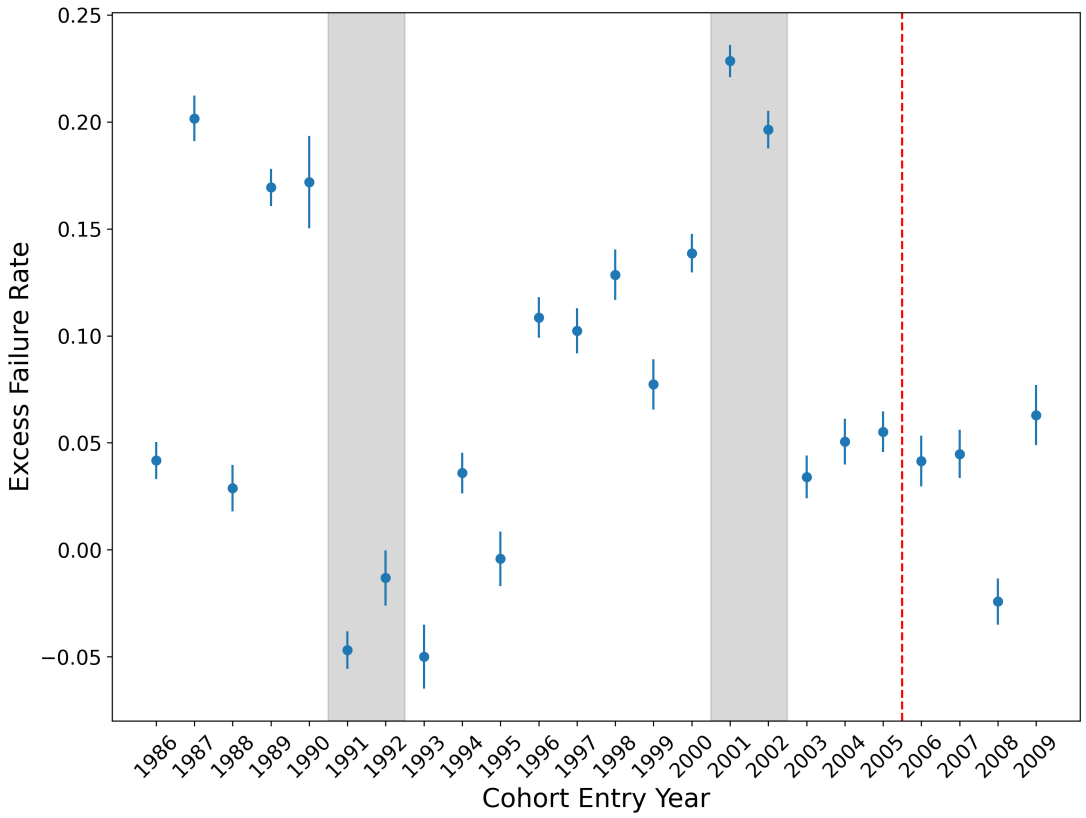



Figure 1: Excess Failure Rate by Entry Cohorts

*Notes*: The figure plots coefficients and associated 95% confidence intervals of the entry-year indicator variables in Equation 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.

datasets become available and the failure rate experiences a more gradual increase.[25]

## 5.2 Failure Rate and Distance to Census Data Release

Building on the descriptive pattern in Figure 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

[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. Like the decennial census, ACS data is released in the subsequent year, e.g., 2005 data is released in 2006.

this period, the impact on failure rates is less predictable, as firms may incorporate different batches of census data at different times. The second phase begins after the first two years, once all major census files have been released. From this point forward, as census data becomes increasingly outdated, its effect on firm failure becomes 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. The longer this gap, the more outdated the information firms rely on, increasing their likelihood of failure. Specifically, for my sample period, I define $\mathrm{S}_i$ as:

$$\mathrm{S}_i = \begin{cases} i - 1981, & \text{if } 1981 \leq i < 1991 \\ i - 1991, & \text{if } 1991 \leq i < 2001 \\ i - 2001, & \text{if } 2001 \leq i < 2010 \end{cases} \tag{4}$$

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::

$$\begin{aligned} \text{Excess failure rate}_{im} = {} & \alpha_1 I(\mathrm{S}_i < 2) + \alpha_2 I(\mathrm{S}_i \geq 2) \\ & + \beta_1(\mathrm{S}_i - 2)I(\mathrm{S}_i < 2) + \beta_2(\mathrm{S}_i - 2)I(\mathrm{S}_i \geq 2) + \varepsilon_{im} \end{aligned} \tag{5}$$

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 $\alpha_1$ and $\beta_1$ represent the intercept and slope of the first phase, during the initial two years after a decennial census, when data are being released in stages. 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 $\beta_1$ is highly sensitive to small variations. By contrast, the coefficients $\alpha_2$ and $\beta_2$ 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

[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.

Table 4: Excess Failure Rate and Distance to Census Data Release

| | Full Sample (1) | 1986-1999 Cohorts (2) | 2000-2009 Cohorts (3) |
|---|---|---|---|
| $\beta_1$ | 0.006 | 0.034*** | -0.032*** |
| | (0.005) | (0.008) | (0.006) |
| $\beta_2$ | 0.016*** | 0.025*** | 0.007*** |
| | (0.001) | (0.001) | (0.001) |
| $\alpha_1$ | 0.105*** | 0.021 | 0.164*** |
| | (0.009) | (0.014) | (0.010) |
| $\alpha_2$ | 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 $\alpha_1$ and $\beta_1$ represent the intercept and the slope of the first phase; the coefficients $\alpha_2$ and $\beta_2$ represent the intercept and the slope of the second phase, respectively. The main coefficient of interest $\beta_2$ 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.

census data quality and firm outcomes. Therefore, My primary focus is on interpreting $\beta_2$.

Table 4 reports the coefficient estimates. The $\beta_2$ coefficient estimates indicate that the distance between an establishment's entry time and the previous census data release has a positive and statistically significant impact on the establishment's failure rate. 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 if firms have to rely on the most outdated information. Considering that the average 5-year failure rate in my sample is 50 percentage points, decade-old information could increase a firm's baseline failure rate by 32%.

I further divide the entry cohorts into two groups: those before 2000 (pre-2000) and those after 2000 (post-2000). In both time periods, the coefficient estimates of $\beta_2$ are positive and statistically significant. However, these effects are notably larger in the pre-2000 period. One contributing factor to this difference is the introduction of the American Community

Survey, which provided more timely data to supplement the decennial census. This transition occurred within the broader context of the information age, where new technologies provided firms greater access to alternative sources of demographic data.[27] Overall, the evidence suggests that the impact of census data on firms is influenced by the presence of alternative information sources.

### 5.2.1 Robustness

To assess the robustness of my results, I conduct additional analyses. Full results are reported in the Online Appendix. First, I test whether the results depend on how the two linear segments in my specification are modeled. Instead of allowing separate intercepts, I also estimate a specification where the segments are constrained to connect at the break-point. As shown in Table A7, the results remain similar. Second, I account for potential spatial correlation within the same entry cohort by estimating standard errors using alternative clustering approaches. Table A8 reports that the estimates remain statistically significant under these adjustments. Third, I test whether the benchmark failure rate influences the results. Table A9 presents estimates excluding the entry cohort from the failure rate calculation to ensure it does not mechanically affect the benchmark. Table A10 uses an alternative benchmark based only on establishments that entered within the past five years to addressing the concern that newer firms may be more vulnerable to market conditions than more established ones. Both yield similar estimates.

## 5.3 Placebo Test

To more closely evaluate whether the pattern in Figure 1 occurs by chance, I use a placebo test that shuffles the treatment assignment. To implement this test, I randomly shuffle the census years and re-estimate the main specification in Equation 5. In particular, I randomly assign a year from each decade in the 1970s, 1980s, 1990s, and 2000s as the census year, resulting in 10,000 possible combinations and thus 10,000 alternative census schedules.[28] For each hypothetical schedule, I then recalculate the distance to the most recent census for each entry-year cohorts and then re-estimate Equation 5.

Out of the 10,000 randomly created schedules, 193 produce $\beta_2$ estimates greater than or equal to the original estimate using the true census schedule. The probability of generating

[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.

[28] For example, one hypothetical set of census schedule could be 1975, 1988, 1992, and 2005. In this case, an establishment that enters in 1986 has to rely on data from the 1975 census and its distance to the most recent census release is 10 years. I need to draw a census date from the 1970s in this scenario as the hypothetical 1980s census occur after the start of my sample.

by chance a coefficient estimate as large as the original estimate is thus 1.9%. Since this p-value is not based on standard errors at the geographic market level, it helps verify that spatial correlation is unlikely to drive the main patterns I infer.

# 6 Heterogeneity in Failure Rate Patterns

In this section, I explore potential heterogeneity in the impact of outdated census data on failure rates. I examine whether the main effect documented in the previous section is more pronounced for (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. My findings suggest that the effect on the failure rate is more pronounced in scenarios where census data plays a more crucial role. Additionally, I use placebo tests to show no discernible effect in instances where census data is expected to have no significant impact.

## 6.1 Failure Rate and Shifts in Demographics

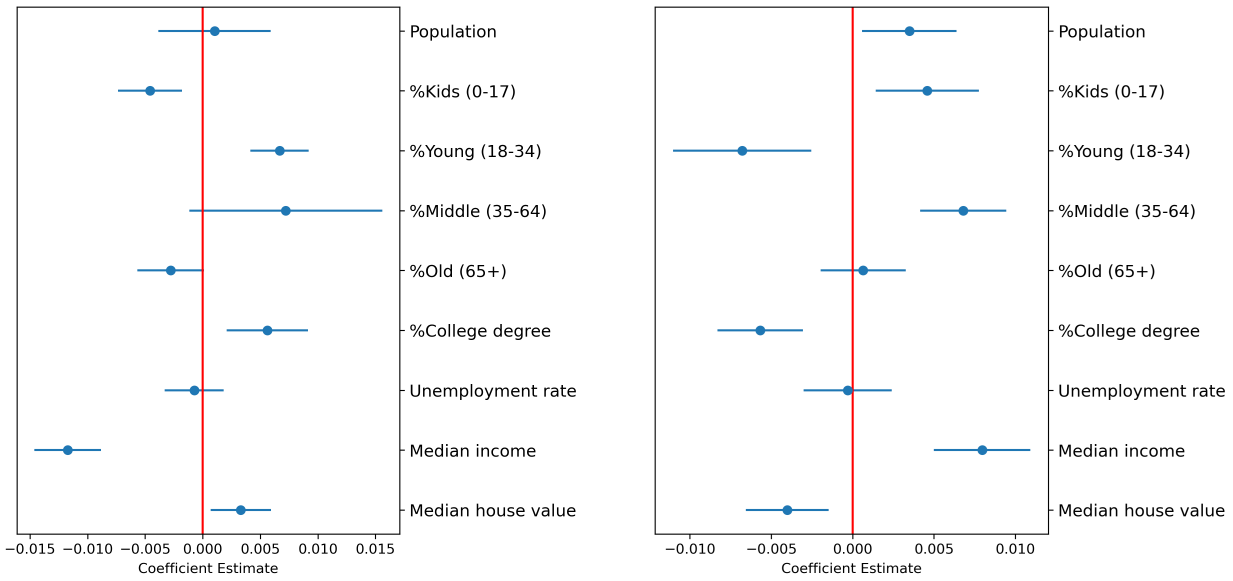
### 6.1.1 Incremental Effect

This section explores how the relevance of census data might differ across various geographic areas. Geographic variations create cross-sectional differences in the relevance of census data beyond timeliness. In areas where demographics remain stable over time, census data collected from the past continues to provide insight into current conditions. However, in areas undergoing rapid demographic shifts, the relevance of census data diminishes quickly. Consequently, I expect a more pronounced impact of outdated census data on failure rates in areas experiencing rapid demographic changes.

To empirically test the incremental effect of outdated information on failure rates in areas with substantial demographic shifts, I interact the slope terms from Equation 5 with an indicator variable that flag these areas and estimate the following regression:

$$\begin{aligned}\text{Excess failure rate}_{im} = {} & \alpha_1 I(\mathrm{S}_i < 2) + \alpha_2 I(\mathrm{S}_i \geq 2) \\ & + \beta_1(\mathrm{S}_i - 2)I(\mathrm{S}_i < 2) + \beta_2(\mathrm{S}_i - 2)I(\mathrm{S}_i \geq 2) + \beta_3 \widetilde{\Delta X_m} \\ & + \gamma_1 \widetilde{\Delta X_m}(\mathrm{S}_i - 2)I(\mathrm{S}_i < 2) + \gamma_2 \widetilde{\Delta X_m}(\mathrm{S}_i - 2)I(\mathrm{S}_i \geq 2) + \varepsilon_{im}\end{aligned} \tag{6}$$

where $\widetilde{\Delta X_{im}}$ is an indicator variable equal to 1 if the value of a demographic variable $X$




(a) Negative Shift

(b) Positive Shift

Figure 2: Coefficient Estimates ($\gamma_2$) by Demographic Variable

*Notes*: This figure reports coefficient estimates $\gamma_2$ and associated 95% confidence intervals from the regression Equation 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.

in market $m$ changes above a threshold between the two decennial censuses surrounding entry-year $i$. Because different demographic variables might have heterogeneous effects on business success, I estimate Equation 6 for each demographic variable $X$, and for positive and negative change separately. Table A3 provides detailed definition of each demographic variable and Table 3 tabulates changes in demographic variables at the census-tract level. Other variables are defined in Equation 5.

In areas with large changes in demographics, the impact of outdated census data on the failure rate of new entry cohorts over time is represented by $\beta_2 + \gamma_2$. The main coefficient of interest $\gamma_2$ captures the incremental effect of stale information. Figure 2 plots the coefficient estimates for $\gamma_2$ 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 evidence of statistically significant incremental impact for age, education, income and housing value. For example, when an area has a markedly reduced proportion of young individuals (*%Young 18-34*) than firms

[29]To assess the sensitivity of this threshold, I also consider specifications using alternative cutoffs at $\pm 15\%$ and $\pm 20\%$ in Table A11. The coefficient estimates are similar in magnitude to results from the 10% cutoff.

had expected based on data from the previous census, the increase in failure rate is further amplified, as indicated by the positive $\gamma_2$ estimate in Figure 2a. Correspondingly, a favorable surprise moderates the increase in failure rate, as indicated by the negative $\gamma_2$ estimate in Figure 2a. 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 decreases in the percentage of kids (*%Kids*) and *%Median income* reduce failure rates.

Businesses have heterogeneous preferences regarding the ideal demographic profile.[30] Given these diverse considerations, the effects outlined above reflect average outcomes across the firms in my sample. 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 $\gamma_2$ ranges from 3 percentage points to 12 percentage points, equivalent to 19% to 75% of the main effect $\beta_2$.

On the other hand, I do not find consistent evidence of incremental effects related to demographic variables such as population and unemployment rate. Unlike the other demographic variables, local population and unemployment statistics are available at annual and even monthly frequencies.[31] Businesses can access this information regularly without waiting for the census data release, which explains why population and unemployment data in the decennial census have a limited impact.

Large shifts in the demographics could also reflect local market volatility. The construct of the excess failure rate alleviates this concern by taking out the average failure rate of existing establishments within the same market and calendar year. Nevertheless, new entrants might be more vulnerable to volatile market conditions than established businesses. 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 because they are available at higher frequencies from other sources further confirms that the incremental effects documented here are through the information channel, rather than market volatility.

Table 5: Census Tracts with Little Change in Demographic Variables

| | (1) |
|---|---|
| $\beta_1$ | 0.010 |
| | (0.089) |
| $\beta_2$ | -0.000 |
| | (0.012) |
| $\alpha_1$ | 0.099 |
| | (0.150) |
| $\alpha_2$ | 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.

### 6.1.2 Placebo Test

I conduct a placebo test using areas that experience little changes in demographics between two censuses. In these areas, timing of entry relative to census data releases is unlikely to be associated with significant variation of census information quality. Specifically, I construct

[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 can be inferred from birth and death records National Vital Statistics System (NVSS) and migration data from the IRS, while monthly unemployment statistics are released by Bureau of Labor Statistics.

a sub-sample of census tracts that have less than 10% change in absolute value for *%Young*, *%College degree*, *Median house value*, *%Kids* and *%Median income*, census demographic variables shown to have a significant impact on firm failure in the previous section.[32] I replicate Table 4 on this "no surprise" sub-sample and confirm that distance to census data release has no effect on failure rate in scenarios where information is plausibly stable over time. As reported in Table 5, the coefficient estimate on $\beta_2$ is practically 0 (-.0000901).

## 6.2 Failure Rate and Industry

To explore heterogeneity in the impact of census data across industries, I estimate Equation 5 for each NAICS 4-digit industry. In this context, the benchmark failure rate is specific to establishments within the same industry from the same neighborhood.

Figure 3 reports the coefficient estimates of $\beta_2$ across 33 NAICS 4-digit industries. Notably, restaurants and grocery stores exhibit the highest sensitivity to such changes. 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 3 hints at a general relationship between the size of an establishment's trade area and its reliance on census data. Businesses that offer highly localized products and services depend heavily on up-to-date demographic data from their immediate surroundings, as small areas can exhibit rapid demographic changes. Conversely, retailers selling specialty goods like motor vehicles and furniture cater to a wider audience. Their consumers are often ready to travel considerable distances to search and compare, providing them with broader trade areas. As a result, localized demographic fluctuations within these expansive regions tend to average out, thereby reducing their dependency on immediate census updates.

To proxy for the size of an establishment's trade area, I categorize retail industries into durable and non-durable goods sectors, in accordance with the Bureau of Economic Analysis' classification of manufacturers of durable and non-durable goods.[33] Durable goods retailers

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

[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

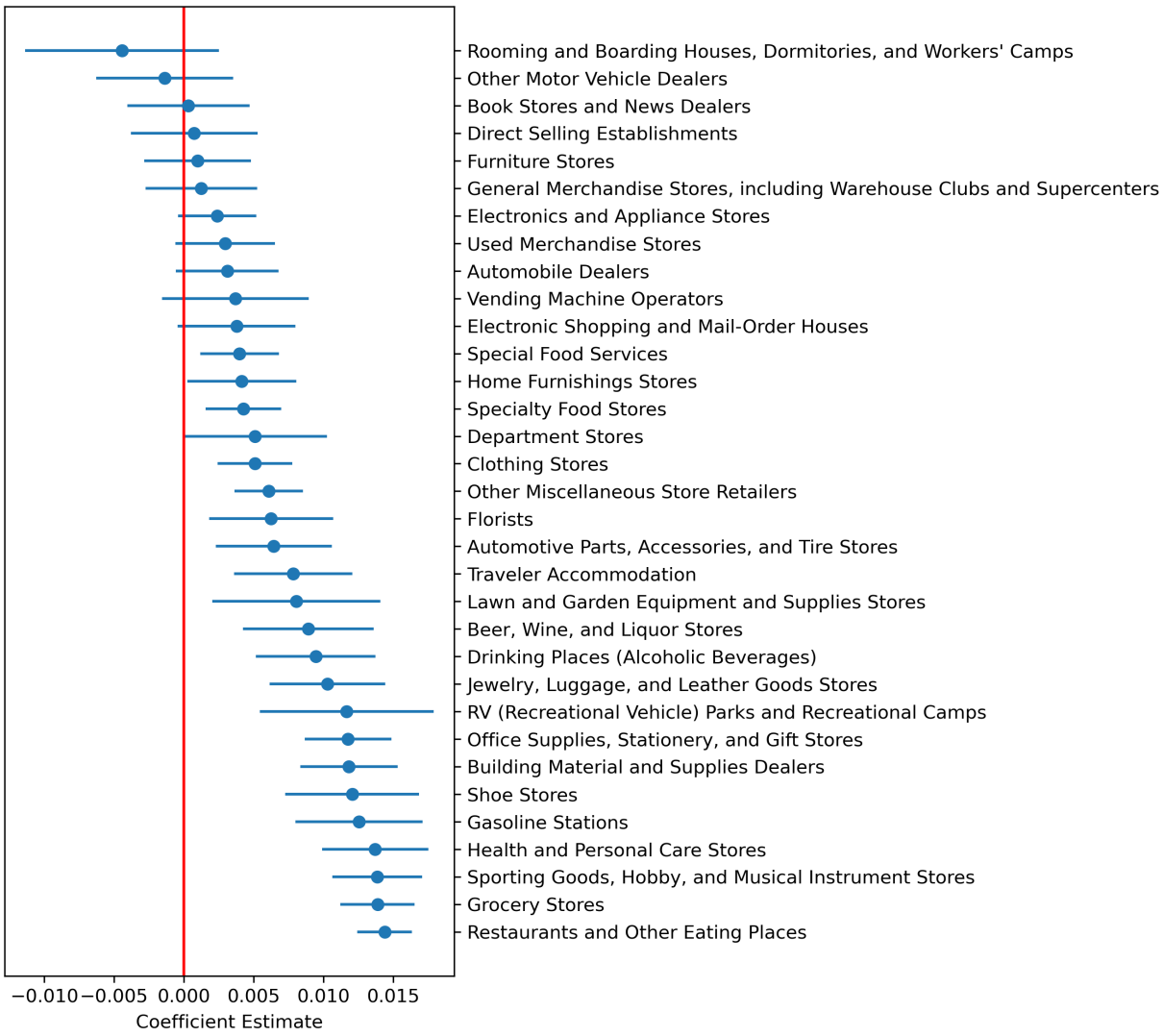



Figure 3: Coefficient Estimates ($\beta_2$) by NAICS 4-digit Industry

*Notes*: This figure plots $\beta_2$ coefficients estimated from Equation 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.

are expected to have larger trade areas, while non-durable goods retailers are likely to have smaller trade areas. As reported in Figure 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 further corroboration, I examine the NAICS 454 Non-store Retailers category, which

Stores), and 4539 (Other Miscellaneous Store Retailers) were omitted due to their ambiguous nature in strictly categorizing as durable or non-durable goods sectors.

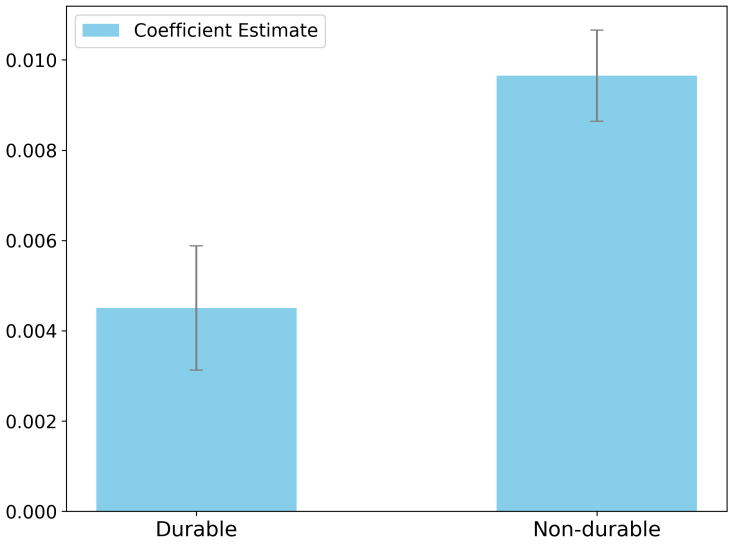



Figure 4: Comparison of Coefficient Estimates ($\beta_2$) between Durable and Non-durable Goods Retailers

*Notes*: This figure plots $\beta_2$ coefficient estimates from Equation 5 and associated 95% confidence intervals across durable and non-durable goods retailers. Standard errors are clustered at the census-tract level.

includes Electronic Shopping and Mail-Order Houses, Vending Machine Operators, and Direct Selling Establishments. Unlike other retailers, non-store retailers do not need a physical store front close to their customers, making them much less reliant on local demographic information. As reported in Figure 3, the impact of census data on failure rate is statistically insignificant for all 3 industries in this category.

## 6.3 Failure Rate and Firm Size

This section examines whether the impact of census data varies across firms of different sizes. One notable difference between large and small firms relates to their ability to obtain information. Large firms can collect information from their existing customers; they also have the resources to conduct market surveys and purchase proprietary data from vendors. Although census data forms the foundation of their market research, they can readily turn to alternative sources when census data 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 is not always up-to-date. For example, when selecting business locations, entrepreneurs often rely on leasing agents, who in turn utilize census-derived metrics to

assess whether the local trade area's demographics are suitable for the proposed business.[34] Both the entrepreneur and the agent might accept this data at face value. Consequently, small firms might be more vulnerable to the effects of outdated census data compared to 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 5 separately for large chains, small chains, and local independents.[35] Large chains are defined as firms with more than 20 locations.[36] This threshold is consistent with existing regulations related to chains,[37] but the patterns are similar using alternative definitions of large chain.[38] The benchmark failure rates are specific to establishments within each size group located in the same census tract, and standard errors are clustered at the census-tract level.

Table 6 reports the full set of coefficient estimates across firm sizes, where the $\beta_2$ 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 $\beta_2$ 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 $\beta_2$ 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, reflecting their greater access to alternative information sources.

The comparison in Panel B reveals a distinct shift in the pre- and post-2000 periods.

[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 A1 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) 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 A2 for details.

[36] Table A1 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 A12 for results setting the threshold at 10 establishments, and Table A13 for results setting the threshold at 50 establishments

## Table 6: Chain vs Independent

Panel A: Full Sample

| | Large Chain (1) | Small Chain (2) | Independent (3) |
|---|---|---|---|
| $\beta_1$ | -0.017 | -0.016 | 0.005 |
| | (0.013) | (0.014) | (0.005) |
| $\beta_2$ | 0.001 | 0.011*** | 0.017*** |
| | (0.001) | (0.002) | (0.001) |
| Adjusted $\beta_2$ | 0.005 | 0.037 | 0.033 |
| $\alpha_1$ | -0.008 | -0.014 | 0.100*** |
| | (0.022) | (0.024) | (0.009) |
| $\alpha_2$ | -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 (1) | Small Chain (2) | Independent (3) | Large Chain (4) | Small Chain (5) | Independent (6) |
| $\beta_1$ | -0.023 | -0.036* | 0.037*** | -0.012 | 0.034 | -0.040*** |
| | (0.018) | (0.016) | (0.009) | (0.018) | (0.025) | (0.006) |
| $\beta_2$ | 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 $\beta_2$ | 0.026 | 0.053 | 0.049 | -0.015 | 0.007 | 0.016 |
| $\alpha_1$ | -0.020 | -0.061* | 0.020 | 0.005 | 0.105* | 0.151*** |
| | (0.031) | (0.028) | (0.015) | (0.031) | (0.044) | (0.010) |
| $\alpha_2$ | -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 $\beta_2$ 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.

In the earlier pre-2000 period, the effect of census information on failure rate is positive and significant across all firm types, including large chains. Moving into the 2000s, the proliferation of digital data collection methods likely provided more timely demographic insights, mitigating the reliance on decennial census updates.[39] The decline in the magnitude of the effect for all types of firms, and the persistence of the effect for independents in the 2000s, may be indicative of the growing availability of alternative data sources that complement or substitute for census information, particularly benefiting larger firms with more resources.

# 7 Alternative Explanations

## 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 different phases of the business cycle.[40] In particular, the early 90s and early 00s recessions overlap with outdated census data in terms of timing, which may drive the pattern. However, the industry breakdown in Figure 3 shows that my results are strongest for restaurants and groceries, despite the marked differences in cyclicality between these two sectors.

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-

[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.

[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.

Table 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) |
| $\beta_1$ | 0.023** | -0.018* | 0.008 | 0.004 |
| | (0.007) | (0.008) | (0.008) | (0.008) |
| $\beta_2$ | 0.016*** | 0.016*** | 0.015*** | 0.019*** |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| $\alpha_1$ | 0.119*** | 0.083*** | 0.103*** | 0.107*** |
| | (0.012) | (0.013) | (0.012) | (0.013) |
| $\alpha_2$ | -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.

level unemployment rate before and after the recession.[41] As reported in Table 7, I find similar effects on establishment failure rate ($\beta_2$) across sub-samples of local areas that are more/less impacted by the recessions. Therefore, it seems unlikely that these recessions are driving the empirical patterns I document.

## 7.2 Government Policies

This section explores how government funding programs tied to census data affect excess failures. US Government Accountability Office (2009) reviews the ten largest federal assistance programs that rely at least in part on the decennial census data. These programs 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

[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).

[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

consumer-directed programs, while state-level allocations are based on census data, within-state benefits are distributed according to individual eligibility criteria, making it unlikely that misallocation drives 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. Unlike broadly applied programs that impact both new and existing firms, 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). A prominent policy in the New York State relevant to retail and restaurant businesses 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]

Using these institutional details, I create sub-samples based on a census tract's eligibility for the Economic Development Zones program and estimate Equation 5 separately for these two groups. Table 8 reports the results. The coefficient estimate of $\beta_2$ for census tracts that are not eligible for Empire Zones funding is very similar to the main result. Note that the estimated effect is actually 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 offer an alternative explanation for the increase in establishment failure as census data become more outdated.

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 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.

Table 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) |
| $\beta_1$ | -0.001 | 0.045*** | -0.006 | 0.052*** |
| | (0.006) | (0.013) | (0.006) | (0.011) |
| $\beta_2$ | 0.017*** | 0.014*** | 0.017*** | 0.013*** |
| | (0.001) | (0.002) | (0.001) | (0.001) |
| $\alpha_1$ | 0.098*** | 0.140*** | 0.089*** | 0.163*** |
| | (0.009) | (0.021) | (0.010) | (0.018) |
| $\alpha_2$ | 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.

# 8 Entry Patterns

The release of census data significantly reduces uncertainty about market conditions, which may influence firms' entry decisions. The literature on investment under uncertainty (Dixit and Pindyck, 1994) suggests that firms facing high levels of uncertainty are likely to delay irreversible investment decisions 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 releases, to better understand how the availability of updated information influences firm behavior under uncertainty, particularly in the context of retail and restaurant sectors.

The investment under uncertainty framework is well-established in industries characterized by high irreversibility and long-term commitment, such as manufacturing, mining, and energy. However, retail and restaurant sectors differ from these industries in two key ways. 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

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

high reversibility allows assets to be repurposed or sold off with minimal loss of value, which reduces the value of delaying investment. Second, competition in these industries exhibit a lot of geographic specificity due to their potentially small trade areas, which limits the number of firms that can enter and thrive. 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]

Table 9: Entry Patterns Surrounding Census Data Release

| | (1) All firms | (2) Large chain | (3) Small chain | (4) 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 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.

period.

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

$$\begin{aligned}\text{Entry}_{mt} = \beta_0 &+ \beta_1 \text{Pre-census}_t + \beta_2 \text{Census}_t + \beta_3 \text{Post-census}_t \\ &+ \beta_4 \text{GDP growth}_{t-1} + \beta_5 \text{Expected GDP growth}_{t-1} + \beta_6 \text{VXO}_{t-1} \\ &+ \beta_7 \text{EPU}_{t-1} + \beta_8 \text{Forecast dispersion}_{t-1} + \alpha_m + \epsilon_{mt}\end{aligned} \tag{7}$$

To examine the entry pattern surrounding the census data releases, I regress market-level entry each year on various controls for economic conditions and uncertainty, as specified in Equation 7. The indicator variables *Pre-census*, *Census*, and *Post-census* refer 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 can serve as a benchmark for entry cohorts, the entry regressions lack a comparable control group. Including time controls in the entry analysis would absorb the variations of interest. Consequently, for the entry regressions, I rely on proxies for macroeconomic conditions and uncertainty to control for potential confounding factors that affect firms' entry decisions.

To control 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 account for economic uncertainty, I employ three measures: the VXO index from the Chicago Board Options Exchange, which measures the implied volatility of the stock market; the composite Economic Policy Uncertainty (EPU) index for New York state, measuring policy-related economic uncertainty (Baker, Bloom and Davis, 2016); and forecast dispersion, calculated as the standard deviation of GDP forecasts from the Livingston Survey, reflecting disagreement among professional forecasters.

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

The results presented in Table 9 do not provide evidence 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 not statistically significant 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 for the lack of observed entry delays around census releases is that firms might simply be unaware that the data they are using is stale. As detailed in subsubsection 2.3.1, businesses frequently rely on intermediaries such as data analytics companies, market research firms, and commercial real estate agents for demographic insights.

These intermediaries incorporate census data into their products and services, but the end users may not be aware of the underlying data's update frequency. Even when firms look up the statistics directly, they might not fully grasp the implications of using potentially outdated data. As a result, firms might make entry decisions based on what they perceive as current market information, thus showing no apparent pattern of delaying their entry to coincide with or follow census data releases.

# 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 document a 1.6 percentage point per year increase in the 5-year failure rate. Over the decade-long gap between decennial censuses, this would accumulate to a 16 percentage point increase, equivalent to a 32% rise over the baseline failure rate.

These findings underscore the value of timely and accurate census data, with implications for the 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 (Marshall, 2015). In the U.S., new confidentiality measures in the 2020 census compromised the reliability of detailed block-group-level data, limiting its usefulness for local analysis (Wines, 2022; Hotz et al., 2022). While census data is primarily designed for political representation and public planning, this paper highlights a channel through which lack of data can also distort market decisions, particularly for small establishments that depend on public data sources. Future studies 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 for critical business decisions. For example, manufacturing firms use census data to assess labor markets and plan production locations, logistics companies analyze population density patterns to optimize distribution networks, and advertising agencies leverage demographic trends for consumer segmentation 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 industries where investment decisions are less directly tied to geographic entry and exit data.

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 pro-

ductivity (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.

# References


**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.

**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.

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

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

**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.

**Brynjolfsson, Erik, and Kristina McElheran.** 2016. "The rapid adoption of data-driven decision-making." *American Economic Review*, 106(5): 133–139.

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

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

**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.

**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.

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

**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.

**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).

**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.

**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.

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

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

**Hotz, V Joseph, Christopher R Bollinger, Tatiana Komarova, Charles F Manski, Robert A Moffitt, Denis Nekipelov, Aaron Sojourner, and Bruce D Spencer.** 2022. "Balancing data privacy and usability in the federal statistical system." *Proceedings of the National Academy of Sciences*, 119(31): e2104906119.

**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.

**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.

**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.

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

**Leisten, Matthew.** 2021. "Information, managerial incentives, and scale: Evidence from hotel pricing."

**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.

**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.

**Marshall, Aarian.** 2015. "The Tragedy of Canada's Census." *Bloomberg*. (accessed August 15, 2024).

**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.

**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.

**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.

**Sedláček, Petr, and Vincent Sterk.** 2017. "The growth potential of startups over the business cycle." *American Economic Review*, 107(10): 3182–3210.

**Serrato, Juan Carlos Suárez, and Philippe Wingender.** 2016. "Estimating local fiscal multipliers." National Bureau of Economic Research.

**Slattery, Cailin, and Owen Zidar.** 2020. "Evaluating state and local business incentives." *Journal of Economic Perspectives*, 34(2): 90–118.

**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*.

**Wines, Michael.** 2022. "The 2020 Census Suggests That People Live Underwater. There's a Reason." *The New York Times*. (accessed April 22, 2022).

**Wu, Lynn, Lorin Hitt, and Bowen Lou.** 2020. "Data analytics, innovation, and firm productivity." *Management Science*, 66(5): 2017–2039.

# Online Appendix

## Table of Contents


**A Example of Census Data Usage** **2**

**B Additional Descriptive Statistics** **3**

**C Demographic Variable Definition** **5**

**D Robustness and Sensitivity Analyses** **6**

D.1 Validation of NETS Data Using CBP data . . . . . 6

D.2 Census Response Rate . . . . . 9

D.3 Excluding New York City . . . . . 11

D.4 Alternative Specification . . . . . 12

D.5 Alternative Clustering . . . . . 13

D.6 Alternative Benchmark Excluding the Entry Cohort . . . . . 14

D.7 Alternative Benchmark Based on Establishments that Entered in the Recent 5 Years . . . . . 15

D.8 Alternative Thresholds for Demographic Changes . . . . . 16

D.9 Alternative Thresholds for Chain Size . . . . . 16

# A Example of Census Data Usage

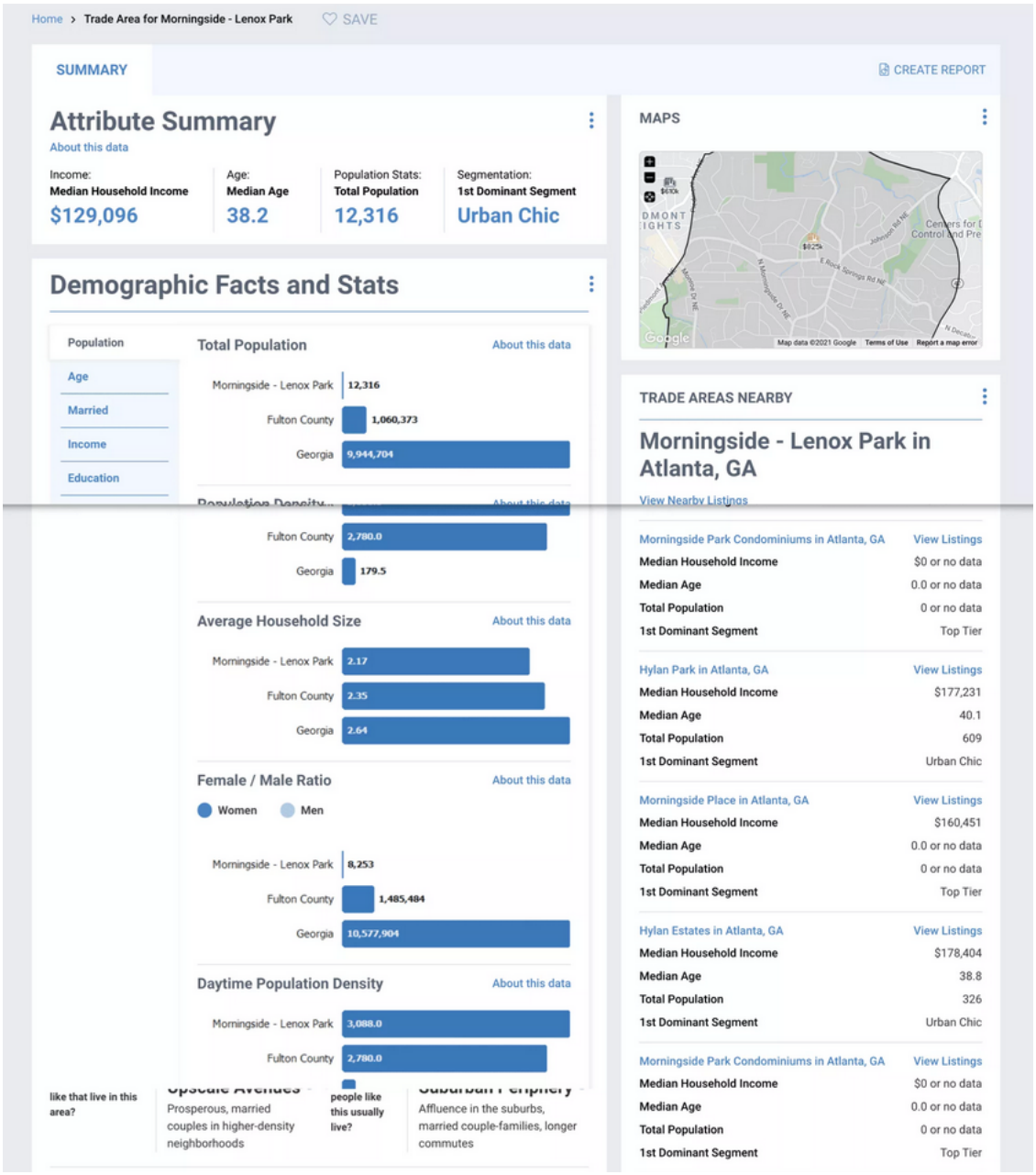



Figure A1: Example Trade Area Analysis Report Tool For Commercial Real Estate Brokers

*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/.

# B Additional Descriptive Statistics

Table A1: 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 A2: 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.

# C Demographic Variable Definition

Table A3: 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 |

# D Robustness and Sensitivity Analyses

## D.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,[48] 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.

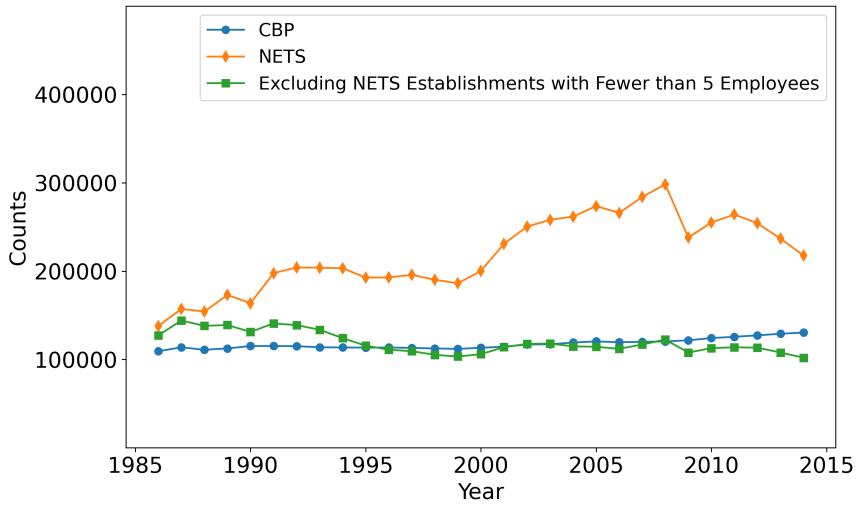
Previous research has examined the alignment between NETS and CBP data, addressing key discrepancies and identifying conditions under which the datasets show consistency. Barnatchez, Crane and Decker (2017) note that NETS includes many nonemployer establishments, which are not covered by CBP and tend 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. Figure A2a 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. Figure A2b reports the cross-sectional correlation across counties in establishment counts between NETS and CBP data.

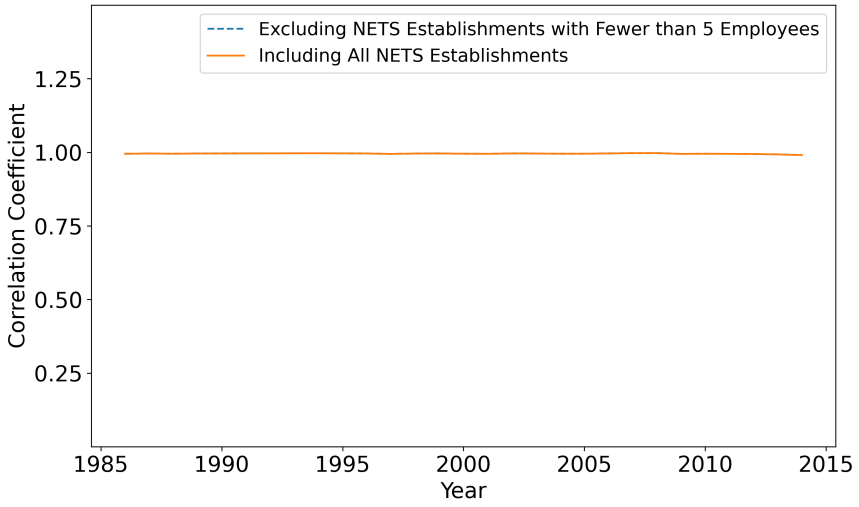
[48] The Longitudinal Business Database (LBD) is constructed using administrative records such as payroll tax filings, and survey data from the Economic Census. While these sources provide a comprehensive foundation for tracking establishments, certain factors can lead to timing discrepancies in the 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.

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 A4 replicates Table 6 using only establishments with at least 5 employees. The results are consistent with the main findings: $\beta_2$ is still positive and statistically significant for small chains and independent establishments but remains insignificant for large chains.




(a) Establishment Count




(b) County-level Correlation

Figure A2: NETS and CBP Data Comparison

Table A4: Excluding Establishments with Fewer than 5 Employees

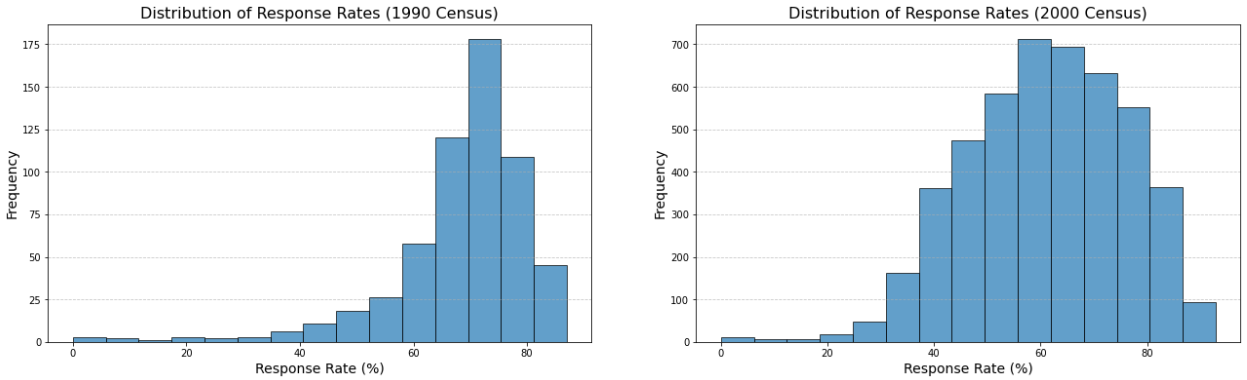
| | Large Chain (1) | Small Chain (2) | Independent (3) |
|---|---|---|---|
| $\beta_1$ | -0.022 | -0.033* | 0.015 |
| | (0.013) | (0.017) | (0.008) |
| $\beta_2$ | -0.002 | 0.008*** | 0.014*** |
| | (0.001) | (0.002) | (0.001) |
| Adjusted $\beta_2$ | -0.012 | 0.029 | 0.028 |
| $\alpha_1$ | -0.022 | -0.048 | 0.215*** |
| | (0.022) | (0.029) | (0.013) |
| $\alpha_2$ | -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 6 using establishments with at least 5 employees. The adjusted $\beta_2$ 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.

## D.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 A3 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.




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

Figure A3: 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 A5 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 $\beta_2$ estimates across response rate groups are consistent with the original results from Table 4, particularly for areas above the lowest response threshold.

Table A5: Subsample Analysis Based on Census Response Rates

| | Low Response Rate (< 43%) (1) | Medium Response Rate (43%–73%) (2) | High Response Rate (> 73%) (3) |
|---|---|---|---|
| $\beta_1$ | 0.058*** | 0.015* | -0.031** |
| | (0.018) | (0.007) | (0.012) |
| $\beta_2$ | 0.011*** | 0.016*** | 0.018*** |
| | (0.002) | (0.001) | (0.001) |
| $\alpha_1$ | 0.164*** | 0.121*** | 0.055** |
| | (0.029) | (0.011) | (0.020) |
| $\alpha_2$ | 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 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.

## D.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 A6 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 A6: Excess Failure Rate and Distance to Census Data Release: NYC vs Others

| | (1)<br>NYC Census Tracts | (2)<br>Non-NYC Census Tracts |
|---|---|---|
| $\beta_1$ | 0.041*** | -0.020*** |
| | (0.007) | (0.007) |
| $\beta_2$ | 0.017*** | 0.016*** |
| | (0.001) | (0.001) |
| $\alpha_1$ | 0.142*** | 0.076*** |
| | (0.012) | (0.012) |
| $\alpha_2$ | -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 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.

## D.4 Alternative Specification

Table A7: Excess Failure Rate and Distance to Census Data Release (Connected Segments)

| | Full Sample (1) | 1986-1999 Cohorts (2) | 2000-2009 Cohorts (3) |
|---|---|---|---|
| $\beta_1$ | -0.046*** | 0.016*** | -0.108*** |
| | (0.002) | (0.003) | (0.003) |
| $\beta_2$ | 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 $\beta_1$ and $\beta_2$ 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.

## D.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 A8 reports results using these alternative clustering approaches and the results remain statistically significant.

Table A8: Alternative Clustering Approaches

| | (1) | (2) | (3) |
|---|---|---|---|
| $\beta_1$ | 0.006 | 0.006 | 0.006 |
| | (0.010) | (0.032) | (0.125) |
| $\beta_2$ | 0.016*** | 0.016*** | 0.016*** |
| | (0.001) | (0.002) | (0.004) |
| Wild bootstrap p-value | | | 0.013 |
| $\alpha_1$ | 0.095*** | 0.095** | 0.095 |
| | (0.019) | (0.044) | (0.181) |
| $\alpha_2$ | 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 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 $\beta_2$. Significance: *** p<0.001, ** p<0.01, * p<0.05.

## D.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 A9, the results are qualitatively and quantitatively very similar to the main result from the original baseline specification.

Table A9: Alternative Benchmark Excluding the Entry Cohort

| | Full Sample (1) | 1986-1999 Cohorts (2) | 2000-2009 Cohorts (3) |
|---|---|---|---|
| $\beta_1$ | 0.007 | 0.053*** | -0.051*** |
| | (0.006) | (0.010) | (0.007) |
| $\beta_2$ | 0.019*** | 0.028*** | 0.008*** |
| | (0.001) | (0.001) | (0.001) |
| $\alpha_1$ | 0.118*** | 0.040** | 0.169*** |
| | (0.010) | (0.016) | (0.011) |
| $\alpha_2$ | 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.

## D.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 A10, 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 A10: Alternative Benchmark Based on Establishments that Entered in the Recent 5 Years

| | Full Sample (1) | 1986-1999 Cohorts (2) | 2000-2009 Cohorts (3) |
|---|---|---|---|
| $\beta_1$ | 0.015** | 0.083*** | -0.060*** |
| | (0.007) | (0.011) | (0.009) |
| $\beta_2$ | 0.021*** | 0.034*** | 0.006*** |
| | (0.001) | (0.001) | (0.001) |
| $\alpha_1$ | 0.061*** | 0.046** | 0.059*** |
| | (0.011) | (0.019) | (0.013) |
| $\alpha_2$ | -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.

## D.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 A11 shows that the coefficient estimates are similar in magnitude across these alternative cutoffs.

Table A11: 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 $\gamma_2$ coefficients and associated standard errors are estimated from the regression Equation 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.

## D.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 A12: Chain vs Independent using Alternative Definition of Large Chain (10 Locations)

Panel A: Full Sample

| | Large Chain (1) | Small Chain (2) | Independent (3) |
|---|---|---|---|
| $\beta_1$ | -0.020 | -0.010 | 0.005 |
| | (0.012) | (0.015) | (0.005) |
| $\beta_2$ | 0.002 | 0.011*** | 0.017*** |
| | (0.001) | (0.002) | (0.001) |
| $\alpha_1$ | -0.010 | -0.006 | 0.100*** |
| | (0.021) | (0.025) | (0.009) |
| $\alpha_2$ | -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 (1) | Small Chain (2) | Independent (3) |
|---|---|---|---|
| $\beta_1$ | -0.031 | -0.030 | 0.037*** |
| | (0.016) | (0.017) | (0.009) |
| $\beta_2$ | 0.006*** | 0.015*** | 0.025*** |
| | (0.002) | (0.002) | (0.001) |
| $\alpha_1$ | -0.032 | -0.053 | 0.020 |
| | (0.028) | (0.030) | (0.015) |
| $\alpha_2$ | -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 (1) | Small Chain (2) | Independent (3) |
|---|---|---|---|
| $\beta_1$ | -0.007 | 0.040 | -0.040*** |
| | (0.018) | (0.027) | (0.006) |
| $\beta_2$ | -0.002 | 0.001 | 0.008*** |
| | (0.002) | (0.003) | (0.001) |
| $\alpha_1$ | 0.014 | 0.117* | 0.151*** |
| | (0.030) | (0.047) | (0.010) |
| $\alpha_2$ | 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 A13: Chain vs Independent using Alternative Definition of Large Chain (50 Locations)

Panel A: Full Sample

| | Large Chain (1) | Small Chain (2) | Independent (3) |
|---|---|---|---|
| $\beta_1$ | -0.011 | -0.022 | 0.005 |
| | (0.013) | (0.013) | (0.005) |
| $\beta_2$ | 0.000 | 0.011*** | 0.017*** |
| | (0.001) | (0.001) | (0.001) |
| $\alpha_1$ | -0.007 | -0.017 | 0.100*** |
| | (0.023) | (0.023) | (0.009) |
| $\alpha_2$ | -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 (1) | Small Chain (2) | Independent (3) |
|---|---|---|---|
| $\beta_1$ | -0.021 | -0.032* | 0.037*** |
| | (0.019) | (0.016) | (0.009) |
| $\beta_2$ | 0.004* | 0.017*** | 0.025*** |
| | (0.002) | (0.002) | (0.001) |
| $\alpha_1$ | -0.020 | -0.051 | 0.020 |
| | (0.032) | (0.027) | (0.015) |
| $\alpha_2$ | -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 (1) | Small Chain (2) | Independent (3) |
|---|---|---|---|
| $\beta_1$ | -0.001 | -0.000 | -0.040*** |
| | (0.018) | (0.023) | (0.006) |
| $\beta_2$ | -0.003 | 0.001 | 0.008*** |
| | (0.002) | (0.002) | (0.001) |
| $\alpha_1$ | 0.008 | 0.057 | 0.151*** |
| | (0.031) | (0.039) | (0.010) |
| $\alpha_2$ | 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.